Circular code motifs in protein-coding genes and ribosomal RNAs

Gopal Krishna Dila

▶ **To cite this version:**

Gopal Krishna Dila. Circular code motifs in protein-coding genes and ribosomal RNAs. Quantitative Methods [q-bio.QM]. Université de Strasbourg, 2020. English. NNT : 2020STRAD027 . tel-03855884

HAL Id: tel-03855884
https://theses.hal.science/tel-03855884

Submitted on 16 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE MATHEMATIQUES, SCIENCES DE L'INFORMATIQUE ET DE L'INGENIEUR

CSTB, ICube, UMR-7357

THÈSE présentée par :

Gopal Krishna DILA

soutenue le : **12 Novembre 2020**

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

Discipline/ Spécialité : Informatique

Motifs de codes circulaires dans les gènes codant les protéines et les ARN ribosomaux

THÈSE dirigée par :

M. MICHEL Christian	Professeur, Université de Strasbourg

RAPPORTEURS :

M. BERNOT Gilles	Professeur, Université de Nice Sophia Antipolis
Mme GASPIN Christine	Directeur de recherche, INRAE, Toulouse

AUTRES MEMBRES DU JURY :

Mme THOMPSON Julie	Directeur de recherche, Université de Strasbourg
M. POCH Olivier	Directeur de recherche, Université de Strasbourg
M. TIMSIT Youri	Directeur de recherche, Institut Méditerranéen d'Océanologie

UNIVERSITY OF STRASBOURG

DOCTORAL SCHOOL MATHEMATICS, INFORMATION SCIENCES AND ENGINEERING

CSTB, ICube, UMR-7357

THESIS presented by :

Gopal Krishna DILA

defended on : **12 November 2020**

to obtain the grade of: **Doctor of the University of Strasbourg**

Discipline/ Speciality : Computer Science

Circular code motifs in protein-coding genes and ribosomal RNAs

THESIS supervised by :

M. MICHEL Christian	Professor, University of Strasbourg

REFEREES :

M. BERNOT Gilles	Professor, University of Nice Sophia Antipolis
Mme GASPIN Christine	Director of Research, INRAE, Toulouse

OTHER MEMBERS OF THE JURY :

Mme THOMPSON Julie	Director of Research, University of Strasbourg
M. POCH Olivier	Director of Research, University of Strasbourg
M. TIMSIT Youri	Director of Research, Mediterranean Institute of Oceanography

Acknowledgements

First of all, I would like to thank Gilles Bernot. Christine Gaspin, Julie Thompson, Olivier Poch and Youri Timsit, for accepting to be a part of the jury and to examine my work. I am truly honoured to have you in my jury.

This PhD endeavour has been a genuinely life-changing experience for me, and I can say that it was one of the best decisions of my educational career. As Tony Robbins put it, *"The only difficult path is the one you never start"*. During these three years there have been numerous ups and downs, but this journey would not have been possible without the love, support and encouragement I received from many people.

I would like to express my sincere gratitude to my thesis advisor Christian Michel, who gave me the opportunity to work in this interesting subject. Your hard work and contribution to the field of circular codes is truly amazing. You have always shown me the right path from the very beginning when I was in India. Thank you, Christian, for your patience, constant motivation and expert suggestions that have always guided me during this PhD and writing of this thesis. Thank you very much for giving me the opportunity to work with a great team, I could not have imagined having a better group to work with.

I would like to express my warmest gratitude to the entire CSTB family, to all the current and former members. I will always be grateful to Olivier Poch and Pierre Collet, for integrating me in their team and providing me with the best working environment. Both scientifically and culturally, it was altogether a different experience for me. The work atmosphere is incredible with the right mixture of fun and seriousness. I discovered the French culture of pot, gallette des rois, Noël celebrations, barbeque outings, and many more. Thank you, Olivier, for your teachings, being the father-figure of the team, illuminating the mood with music, your chess lessons and particularly your humorous, ironic remarks. Your optimistic attitude and passion for serious science is felt by all those who have come in touch with you.

I would like to express my heartfelt gratitude to Julie Thompson, for her faith, support and countless advice she has given me over the last three years. I still remember our talk on skype before I even began my thesis. Your calm composure and insightful deductions have always been helpful. From correcting my manuscripts to translating discussions during meetings, you have helped me a lot. You are the perfect role model for many of us. These few lines cannot explain the impact you have had on shaping my PhD career.

I would like to express my sincere gratitude to Claudine, Laetitia, Odile, Luc, Jean-Sébastien, Anne J., Pierre P., Rabih, Catherine, who have always been helpful to me, and for their scientific and cultural teachings. I am really grateful to Anne N., who has always been

very kind and has helped me out with administration things from the very beginning. I would like to express my warmest gratitude to Raymond, Arnaud, Yannis, Audrey and Kirsley, my office mates. You have made room 555 a great place to work together, I am really going to miss sharing an office with such wonderful people. A special thank you to Raymond for all your assistance with my coding and teaching me how to use the bioinformatics toolbox from Gscope. Thank you, Arnaud, for helping me out a number of times, and for the awesome meme culture. Thank you, Yannis, Audrey, Kirsley, Julio, Nicolas, Thomas, Anna, Pierre W., Romain, François, for the stimulating discussions, helpful suggestions, and for all the fun we have had in the last three years. I know I have missed quite a few of the *sacré chapati* meals together, but I enjoyed each one with the team. I would also like to thank all the interns who have worked with us, especially Camille, Christelle, Célia, Corentin, Sarah, Oussama and Clémentine.

I would like to express my gratitude to the University of Strasbourg for funding my thesis and for offering the best research facilities. The courses offered by the university and the doctoral school were very beneficial for a growing researcher. I would also like to express my gratitude to all the university staff who helped me during these three years.

As for people outside my workplace, thank you all, for your love, friendship and motivation. I apologize for not mentioning all of you. First of all, I am very grateful to Jayash and Anurag for helping me settle in when I first arrived in Strasbourg, and for all the thrilling journeys we have had together. Thank you, for being patient and listening to my questions, both technical and personal. Thank you, Prashant, Sandeep, Jiten, Uma, Soumya, Ashish, Sai, Sarita and Himani for your unfailing support and, in particular, for tolerating me for more than 10 years. Thank you, Sumeet, for being such a good friend since our school days, and staying in touch whenever you can. Thank you, Harish, Chloé, Gégé, Danai, Sibylle, Lara, Teresa, Tommy, and Dimitris, for the countless memories and for encouraging me when I was not my best. I would like to express my gratitude to Strasbourg Strollers and the players associated with it, for giving a wonderful opportunity to play cricket in France and in neighbouring countries. Last but not least, I am so grateful to Spyridoula, the person who has always been with me in my happy and rough times, for her unconditional love, care, patience, inspiration, and understanding. Thank you for everything, and particularly for the several meals you made while I was busy writing my thesis.

Finally, I would like to express my warmest thanks to my parents, my aunt and my cousins, for their love and affection, for always believing in me, and for encouraging me to follow my dreams. And to help in any way they could through this difficult time. Thank you for all the sacrifices you have made, so that I can succeed and have a better life. I will be forever indebted to my parents, without whom I would not be here.

I dedicate this thesis to my parents, for their unconditional love and support.
I miss you and love you dearly.

Table of Contents

List of Figures

List of Figures

List of Tables

List of abbreviations

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
mRNA	Messenger ribonucleic acid
rRNA	Ribosomal ribonucleic acid
tRNA	Transfer ribonucleic acid
miRNA	Micro ribonucleic acid
lncRNA	Long non-coding ribonucleic acid
RNA pol	Ribonucleic acid polymerase
pre-mRNA	Precursor messenger ribonucleic acid
poly-A	polyadenylation
SGC	Standard genetic code
LSU	Large subunit
SSU	Small subunit
A-site	Aminoacyl site
P-site	Peptidyl site
E-site	Exit site
A-tRNA	Aminoacyl transfer ribonucleic acid
P-tRNA	Peptidyl transfer ribonucleic acid
E-tRNA	Exit transfer ribonucleic acid
LUCA	Last universal common ancestor
BLAST	Basic Local Alignment Search Tool
CDS	Coding sequence
SQL	Structured Query Language
PA	Protein abundance
AA	Amino acids
PTC	Peptidyl transferase center
CPK	Central pseudoknot
AES	Ancestral expansion segments
GAC	GTP (guanosine triphosphate) Associated Center
ORFs	Open reading frames

Résumé de thèse

La théorie du code circulaire X

Les travaux présentés dans cette thèse s'intéressent aux motifs de codes circulaires identifiés dans les gènes codant les protéines et les ARN ribosomaux. Plus précisément, il s'agit de motifs construits à partir du code circulaire X. Le code circulaire X est un ensemble de 20 trinucléotides (codons, mots de 3 lettres sur l'alphabet génétique à 4 lettres $\{A, C, G, T\}$) qui a été découvert en 1996 avec une analyse statistique des gènes des procaryotes, des eucaryotes, des plasmides et des virus (Michel, 2017; Arquès et Michel, 1996). Il est formé des trinucléotides ayant la plus forte occurrence moyenne dans la phase de lecture (phase 0) par rapport aux deux phases décalées (phases 1 et 2); la phase de lecture étant la phase utilisée pour décoder l'information génétique codée dans l'ARN messager (ARNm) pour synthétiser les protéines. Le code circulaire X est un code correcteur d'erreurs qui a la capacité de retrouver, maintenir et synchroniser la phase de lecture dans les gènes. Par conséquent, les motifs construits à partir du code circulaire X, appelés motifs X, ont la propriété de synchroniser la phase de lecture. Les codes circulaires ont la propriété mathématique de circularité, c'est-à-dire que tout mot (séquence de nucléotides, gènes) écrit sur un cercle (la dernière lettre étant la première lettre sur le cercle) a une décomposition unique en trinucléotides du code circulaire. Ainsi tout motif X peut retrouver la phase de lecture; aucun autre signal de phase (codons d'initiation ou stop) n'est nécessaire pour identifier la phase de lecture. Le code circulaire X possède de fortes propriétés mathématiques; en particulier, il est auto-complémentaire, c'est-à-dire que 10 trinucléotides de X sont complémentaires aux 10 autres trinucléotides de X. Il est maximal, c'est-à-dire que X ne peut pas être contenu dans des codes circulaires de cardinalité supérieure. Un code circulaire maximal contient au maximum 20 trinucléotides. De plus, les permutations circulaires $+1$ et $+2$ du code circulaire X sont également des codes circulaires maximaux (propriété C^3). Le code circulaire X appartient à la classe des 216 codes circulaires C^3 auto-complémentaires maximaux. Les codes “comma-free” (“sans virgule”) forment une variante plus restrictive des codes circulaires dans laquelle la phase de lecture est retrouvée immédiatement (3 nucléotides consécutifs, un nucléotide étant une lettre). Dans le cas du code circulaire X, la phase de lecture est retrouvée après un maximum de 13 nucléotides consécutifs.

Suite à l'identification du code circulaire X en 1996, un nombre important de recherches a été mené pour étudier ses propriétés, et selon différentes approches: statistique, traitement du signal, combinatoire et théorie des graphes. Dans une étude statistique à grande échelle portant sur 138 génomes d'eucaryotes complets, il a été montré que les motifs X apparaissent

préférentiellement dans les gènes (El Soufi et Michel, 2016): la proportion de motifs *X* trouvés dans les gènes par rapport aux régions non codantes était significativement élevée (un ratio proche de 8). Dans une autre analyse statistique, cette fois-ci au niveau du génome, un enrichissement significatif des motifs *X* est à nouveau observé dans les gènes du génome complet de l'eucaryote *Saccharomyces cerevisiae* (Michel et al., 2017). Il a été proposé que les motifs *X* trouvés dans les gènes pouvaient être la conséquence de traces évolutives d'un code primitif qui aurait été utilisé pour la traduction primitive des protéines. Il a également été fait l'hypothèse que ces motifs *X* pouvaient être encore fonctionnels dans les gènes actuels pour la synthèse des protéines et le mécanisme de correction des erreurs de traduction.

Contributions

Nos travaux de thèse poursuivent cet axe de recherche. Pour la première fois, nous analysons les motifs *X* en utilisant des alignements multiples de gènes issus d'organismes des trois domaines de la vie, les archées, les bactéries et les eucaryotes. Les résultats obtenus dans cette thèse renforcent la théorie du code circulaire et apportent des solutions à des problèmes ouverts depuis son identification en 1996.

I. Conservation évolutive des motifs de codes circulaires dans les gènes codant les protéines

Une première analyse s'intéresse aux gènes provenant de deux ensembles d'organismes différents: quatre mammifères et neuf levures (Dila et al., 2019a). Ces organismes représentent une large distribution phylogénétique et une grande variété de structures géniques, allant de simples gènes (absence d'intron), par exemple dans *S. cerevisiae*, jusqu'à la structure intron/exon très complexe des gènes humains. De plus, les mammifères représentent une évolution d'espèces étroitement apparentée (partageant un ancêtre commun il y a environ 300 millions d'années), alors que les levures ont subi une évolution plus divergente (partageant un ancêtre commun il y a environ un milliard d'années). Nous avons ainsi construit des alignements multiples de gènes à la fois pour les mammifères et les levures. Une longueur minimale de 12 nucléotides est choisie pour identifier les motifs *X* dans les alignements multiples de gènes afin que chaque motif *X* puisse retrouver la phase de lecture avec une probabilité de 100%. Le premier résultat obtenu montre un fort enrichissement des motifs *X* (aussi bien en nombre qu'en longueur) dans les alignements multiples de gènes des mammifères et des levures, confirmant ainsi les études antérieures sur l'enrichissement des motifs *X* dans les gènes. De plus, la définition de divers paramètres de conservation évolutive montre que les motifs *X* sont mieux conservés par rapport au reste des séquences de gènes avec un ratio (dN/dS) plus faible de substitutions non synonymes (mutations ne conservant pas l'acide aminé codé) par rapport à des

substitutions synonymes (mutations conservant l'acide aminé codé). Cette propriété évolutive est associée à une sélection purificatrice des motifs X. Nous avons également effectué une étude approfondie des substitutions synonymes dans les motifs X. Les résultats obtenus suggèrent deux types de pression de sélection: une première sélection qui permet de préserver les acides aminés des protéines codés par les gènes et une deuxième sélection qui ne s'applique qu'aux motifs X. Nous mettons ainsi en évidence une nouvelle propriété de conservation des motifs X par acide aminé. Enfin, nous montrons une forte corrélation entre les niveaux d'expression des protéines et l'enrichissement des motifs X dans les gènes. Dans le futur, ce résultat pourrait constituer une nouvelle stratégie expérimentale pour optimiser les gènes.

II. Motifs de codes circulaires universels dans les ARN ribosomaux: un processus d'évolution de la traduction ?

Une analyse innovatrice a concerné l'étude des motifs X dans les ARN ribosomiques (ARNr). Nous avons choisi un ensemble de 133 organismes représentatif des trois domaines de la vie (32 eucaryotes, 65 bactéries et 36 archées) (Dila et al., 2019b). Les séquences d'ARNr étant plus courtes et plus conservées, une longueur minimale de 8 nucléotides est choisie pour identifier les motifs X dans les alignements multiples de séquences des ARNr. De plus, nous avons introduit un paramètre "d'universalité" : un "motif X universel" (motif uX) est défini comme une suite de nucléotides dans l'alignement multiple de séquences ayant au moins 6 positions consécutives appartenant à un motif X, chaque position ayant une universalité supérieure à 90% (c'est-à-dire observé avec un minimum de 119 espèces sur 133). Des alignements multiples de séquences sont construits pour la petite sous-unité (SSU ARNr) et la grande sous-unité (LSU ARNr) du ribosome. Nous avons identifié 32 motifs uX (13 dans la SSU et 19 dans la LSU), dont la plupart sont dans des régions impliquant des fonctions importantes du ribosome, notamment le centre de la peptidyl transférase (PTC) et le centre de décodage qui forment le "proto-ribosome" primordial. Nous constatons également que ces motifs uX dans les ARNr ne sont pas nécessairement conservés en termes de nucléotides mais en fonction de trinucléotides. Ce résultat montre une information génétique dans les ribosomes basée sur des trinucléotides, comme dans les gènes, observation qui n'avait jamais été publiée auparavant. Des analyses structurelles (2D et 3D) ont été également effectuées. Elles révèlent que la plupart des motifs uX sont en interaction avec différentes molécules, notamment l'ARN messager, l'ARN de transfert et les protéines ribosomiques. Notamment, 11 des 32 motifs uX sont en contact avec les ARN de transfert des sites A, P et E; 11 des 13 motifs uX dans le SSU et 16 des 19 motifs uX dans le LSU sont en contact avec des protéines ribosomiques. Les diverses interactions des motifs uX sont ainsi associées à des fonctions ribosomales majeures localisées en particulier au niveau du tunnel et du cliquet. Nous nous sommes également

intéressés à l'évolution du ribosome. En s'appuyant sur les modèles d'accrétion existants, nous proposons que les codes circulaires ont représenté une étape importante dans l'émergence du code génétique standard. Ainsi, les codes circulaires auraient permis simultanément le codage des acides aminés et la synchronisation de la phase de lecture dans des systèmes primitifs de traduction.

III. Le code circulaire *X* et les erreurs de phase

Diverses études statistiques et biochimiques ont montré que le code génétique (standard) est optimisé pour réduire l'impact des erreurs de traduction. Les erreurs de décalage de phase du ribosome peuvent conduire à la synthèse de protéines tronquées ou mal repliées, entraînant la perte de la fonction protéique. En se basant sur plusieurs propriétés physico-chimiques des acides aminés, nous avons comparé l'optimalité du décalage de phase du code circulaire X avec le code génétique et avec les 216 codes circulaires C^3 auto-complémentaires maximaux (Dila et al., 2020). Un ensemble de 13 propriétés biochimiques fondamentales des acides aminés a été sélectionné: la charge, l'hydrophobicité, le point isoélectrique, le point de fusion, le poids moléculaire, la rotation optique, la polarité, la polarisabilité, la taille, l'effet stérique, le volume, l'hélice alpha et la conformation en feuille bêta. Nous mesurons les différences des propriétés biochimiques des acides aminés codés par un code, sans décalage et après un décalage de phase. Les déphasages vers l'avant $(+1)$ et l'arrière (-1) sont considérés séparément puisque leurs mécanismes biologiques sont différents. Cette analyse a montré que le code circulaire X minimise les effets des erreurs de traduction ribosomique par déphasage $+1$, qui sont les plus fréquents. De plus, le code circulaire X a la meilleure optimalité de déphasage dans sa classe combinatoire des 216 codes circulaires, aussi bien en déphasage $+1$ que -1. Cette nouvelle propriété fournit une réponse à une question ouverte depuis 1996 concernant la sélection du code circulaire X parmi les 216 codes circulaires.

Bibliographie

Arquès D.G., Michel C.J. (1996). A complementary circular code in the protein coding genes. *Journal of Theoretical Biology* 182, 45-58. https://doi.org/10.1006/jtbi.1996.0142.

El Soufi K., Michel C.J. (2016). Circular code motifs in genomes of eukaryotes. *Journal of Theoretical Biology* 408, 198-212. https://doi.org/10.1016/j.jtbi.2016.07.022.

Michel C.J. 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. *Life* 7, 20, 1-16. https://doi.org/10.3390/life7020020.

Michel C.J., Ngoune V.N., Poch O., Ripp R., Thompson J.D. (2017). Enrichment of Circular Code Motifs in the Genes of the Yeast Saccharomyces cerevisiae. *Life 7*, 52,1-20. https://doi.org/10.3390/life7040052.

Chapter 1

1. Introduction

Here we give a general introduction to this thesis. The work presented in this thesis focuses on circular code motifs in protein-coding genes and ribosomal RNAs; more precisely motifs from the circular code X discovered in 1996 (Arquès & Michel, 1996). In this thesis, we provide a new dimension in the study of circular code motifs in genes. Before going into details about the circular code X and its significance in biology, we would like to introduce the biological context of the work.

1.1. Biological context

The origin of life is one of the most fundamental yet controversial topics in the study of evolution. The reconstruction of the "Tree of Life" led to the discovery of the three domains of life as we know them, viz. archaea, bacteria and eukaryotes. When we talk about how life emerged nearly 4.2 billion years ago, the important thing to consider is how the first cells came into being from the primordial chemical soup. Cells are known as the "building blocks of life", since they are the basic structural and functional unit of all living species. Cells can be divided into two main categories: prokaryotic (archaea and bacteria) and eukaryotic. Eukaryotic cells contain a nucleus that separates the genetic material (the genome) from other membrane-bound cellular components. In eukaryotes, the genetic material is located mainly within the cell nucleus, but a small amount can also be present within the mitochondria (powerhouse of the cell) and chloroplasts of plants. In contrast, prokaryotic cells do not have a nucleus; but the genetic material is located in a specialized region known as the nucleoid. However, both eukaryotic and prokaryotic cell types have a cell membrane and the water-based environment within the cell membrane is known as cytoplasm. The cytoplasm is filled with intracellular biomolecules including nucleic acids, proteins, lipids, and so on; life on earth is dependent on these biomolecules and their diverse interactions. In addition to the cell membrane (plasma membrane), prokaryotic cells have an outer cell wall that acts as a protective layer allowing them to survive in harsh environments. Archaea and bacteria are unicellular organisms, whereas the eukaryotes contain both unicellular and multicellular organisms.

In his Origin of Species, Darwin described evolution as a process that started from "simple beginnings" and gave rise to more "complex forms" (Darwin, 1859). He also provided a mechanism to explain the complex and characteristic adaptations of living beings: Natural

Selection. In molecular biology, one of the most fascinating questions is how the basic structures of life evolved and what were the evolutionary pressures acting on them? An interesting thing to consider in the evolution of life is that at some point in the past, any two living organisms, no matter how different, shared a common ancestor. We, as humans, share an ancestor with the chimpanzee that lived about 5 million years ago and with any existing bacteria about 3 billion years ago. Consequently, both prokaryotes and eukaryotes share many similar biological mechanisms needed for life, which originated from the common primordial ancestor. Before going into further details on the biological mechanisms needed for life, we will start with the introduction of genetic information to understand how information is stored in genes and passed from one generation to the next.

1.1.1. Genetic information and its storage

A gene is considered the basic functional unit of heredity. Genes containing information for the production of proteins are called protein-coding genes. Other genes do not code for proteins, but non-coding RNA (known as non-coding genes) that control gene activity and have other regulatory functions. Genes are made up of DNA, the biomolecule responsible for storing genetic information in cells. Genetic information is extremely important for the survival and development of an organism, and is passed on from one generation to next. This information is encoded in the DNA as a code composed of four nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order of the bases in the DNA determines the encoded information, similar to the way alphabets are ordered to form words or sentences. Our understanding of how genetic material is transferred from one generation to the next comes from the research on pea plants by Gregor Mendel, which took him 8 years to discover the fundamental laws of inheritance (Mendel, 1865). Because of his insights into how traits are passed from parents to offspring, he is regarded as the "Father of genetics". A few years later, DNA was first observed by Frederich Miescher in 1869 (Dahm, 2005). However, its significance was not understood until 1944, when it was shown to be the carrier of genetic information (Avery et al., 1944). Nevertheless, following the scientific breakthrough that unveiled the structure of DNA as a double-helix polymer (Watson & Crick, 1953), we have come a long way to understanding the intricate mechanisms of how genetic information is stored and transmitted. In the next section, we will introduce nucleic acids (DNA and RNA) and their role in transmitting genetic information.

1.1.2. Nucleic acids and amino acids

The term "nucleic acid" refers to both DNA (deoxyribonucleic acid) and RNA (ribonucleic acid); macromolecules deemed to be the basis of life. These nucleic acids are

polymers (long chains) that are made up of repeating units known as nucleotides or nucleotide bases. RNA molecules are made up of a single chain of these nucleotides, whereas DNA molecules are made up of two chains of nucleotides wrapped around each other to form a double helix structure (Figure 1.1). Nucleotides are small molecules composed of a nitrogenous base (either adenine (A), cytosine (C), guanine (G), and thymine (T) in DNA, or uracil (U) instead of T in RNA), a five carbon sugar (ribose or deoxyribose) and a phosphate group. The nucleotides are connected by covalent bonds between one nucleotide's phosphate and the next nucleotide's sugar, creating a sugar-phosphate backbone. As shown in the Figure 1.1, the two strands of DNA are attached together by hydrogen bonds between the bases, i.e. complementary base pairing: G pairs with C, and A pairs with T.

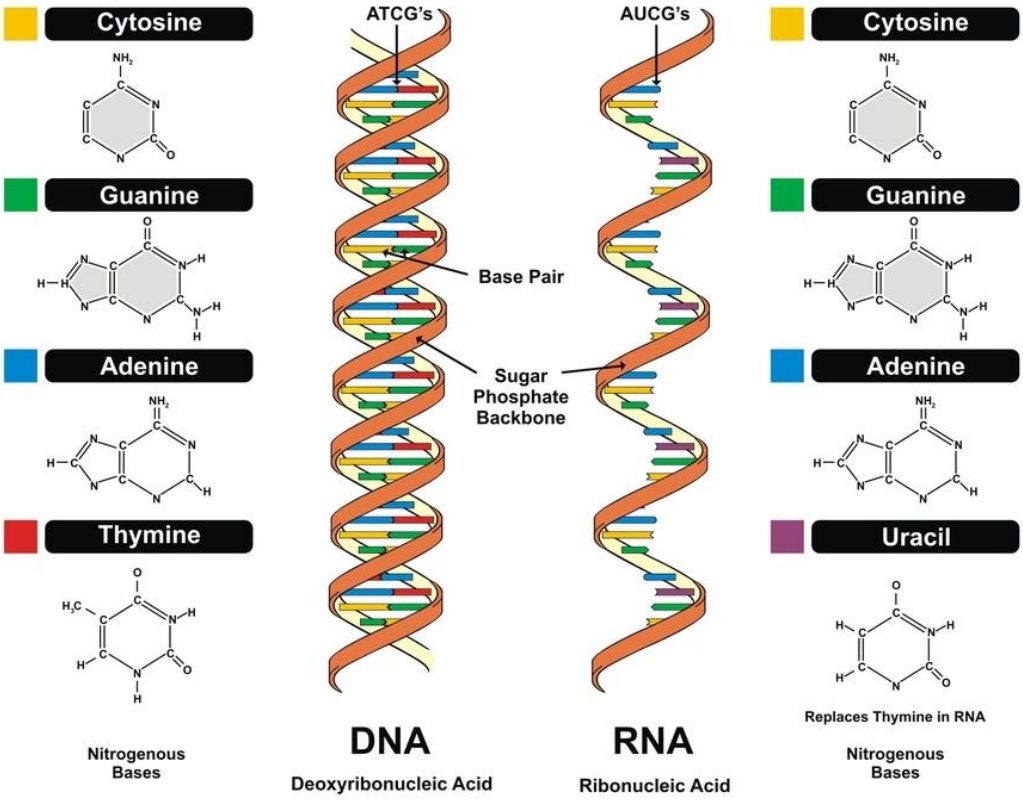


Figure 1.1. Structure of nucleic acids: DNA and RNA. Both are composed of repeating units of nucleotides. Structure of the nitrogenous bases: adenine, guanine, cytosine, thymine, and uracil. (Picture taken from https://www.shutterstock.com/image-illustration/dna-rna-108604994)

Nucleic acid strands have directionality due to their structure and their sequences are usually written in the 5′ to 3′ direction. An important property of the DNA is replication; each strand in the double helix can serve as a pattern for replication. As can be seen in the Figure 1.2, the two strands of DNA run in opposite directions, which means that the 5′-end of one strand is paired up with the 3′-end of its matching strand. This property of DNA is termed antiparallel orientation, and is crucial in DNA replication. For example, if we take the DNA sequence of

one strand as $5' - ATGCTGGGCTAG - 3'$, then the complementary strand can be written as $3' - TACGACCCGATC - 5'$.

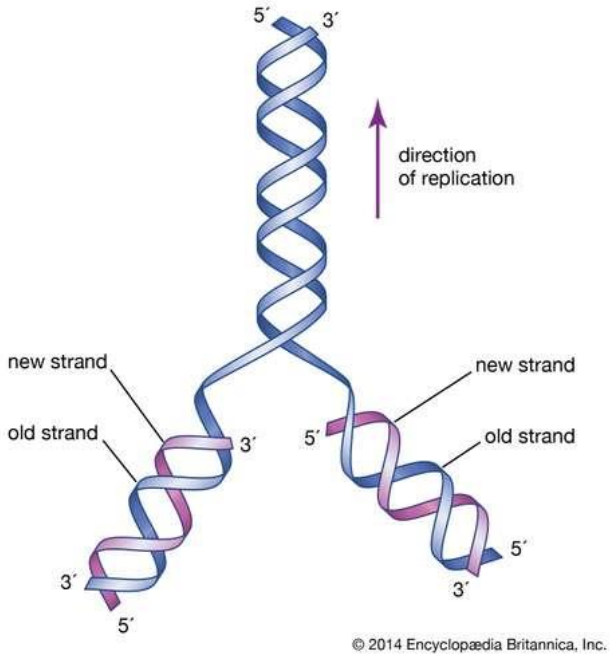


Figure 1.2. Initial proposed structure of DNA and means of replication by Watson and Crick. (Picture taken from https://www.britannica.com/science/DNA#/media/1/167063/90501, Encyclopædia Britannica)

Although DNA has the predominant role to store genetic information in the cell, the role of RNAs are far more complex. One hypothesis is that RNAs had the role of storing and passing of genetic information in primordial systems, before the advent of DNA; we will address this later in the thesis. In extant organisms, RNAs play a pivotal role in protein synthesis. Proteins are essential to life, as they perform a wide variety of tasks needed for the functioning of cells, from synthesizing new cellular components to making copies of DNA during cell division. Next, we will discuss on the composition of proteins and explain how they are synthesized inside a cell.

Proteins are made up of chains of molecules known as amino acids. As the name suggests, amino acids contain an amino group (NH_2), a carboxyl group (COOH) which is acidic and a variable side chain R that differentiates one amino acid from the other (Figure 1.3). There are over 500 amino acids found in nature. However, there are only 20 distinct types of amino acids coded by the genetic code, shown in Figure 1.4 with their chemical structure and some important physicochemical properties. These 20 amino acids combine (in chains) in a specific order to make a protein (amino acid polymers) and the sequence of amino acids defines a protein's distinctive 3D structure and its specific function.

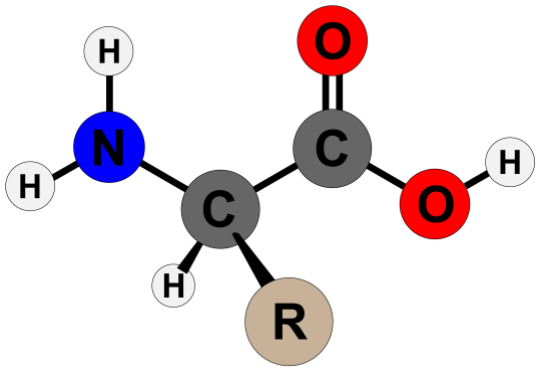


Figure 1.3. Structure of an amino acid with an amino group (NH2), a carboxyl group (COOH) which is acidic and a variable side chain R (different R for different amino acids). (Original image by Techguy78 [Public Domain], via Wikimedia Commons, shared under a Creative Commons license).

Proteins are synthesized using precise instructions encoded in the DNA to assemble amino acids according to the genetic code, which is a complex process. In order to understand the process of protein synthesis, let us come back to RNAs and discuss their diverse roles. Different RNA molecules are found in cells, including messenger RNAs (mRNAs), ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), micro RNAs (miRNAs), and long non-coding RNAs (lncRNAs). The mRNA is a single-stranded copy of the protein-coding DNA, which encodes the genetic information needed for protein synthesis. The rRNAs are complex structures that form the functional component of the ribosome, the machinery that decodes the information encoded in the mRNA sequence, whereas the tRNAs act as the carrier of amino acids during protein synthesis. The miRNAs are short non-coding RNAs that can block the process of protein synthesis, and lncRNAs often have gene regulatory functions.

Returning to the process of protein synthesis, it involves two major steps: transcription and translation. Transcription is the process of making an RNA copy of a gene sequence, the mRNA. During the process of transcription, the DNA sequence of a protein-coding gene serves as a template for the mRNA, where *T* is replaced by *U* for the mRNA molecule. As we can see in Figure 1.5, information encoded in the double-stranded DNA is copied into a single strand of mRNA by a protein complex called RNA polymerase. The transcription process for all species starts with the binding of the RNA polymerase to the protein-coding DNA sequence. However, the number and composition of RNA polymerases is different for prokaryotes and eukaryotes. The prokaryotes have one type of RNA polymerase, but the eukaryotes have three types: RNA polymerase I, II, and III. In eukaryotes, the RNA pol I transcribes rRNAs, RNA pol II transcribes mRNAs and other regulatory RNAs, and RNA pol III transcribes tRNAs.

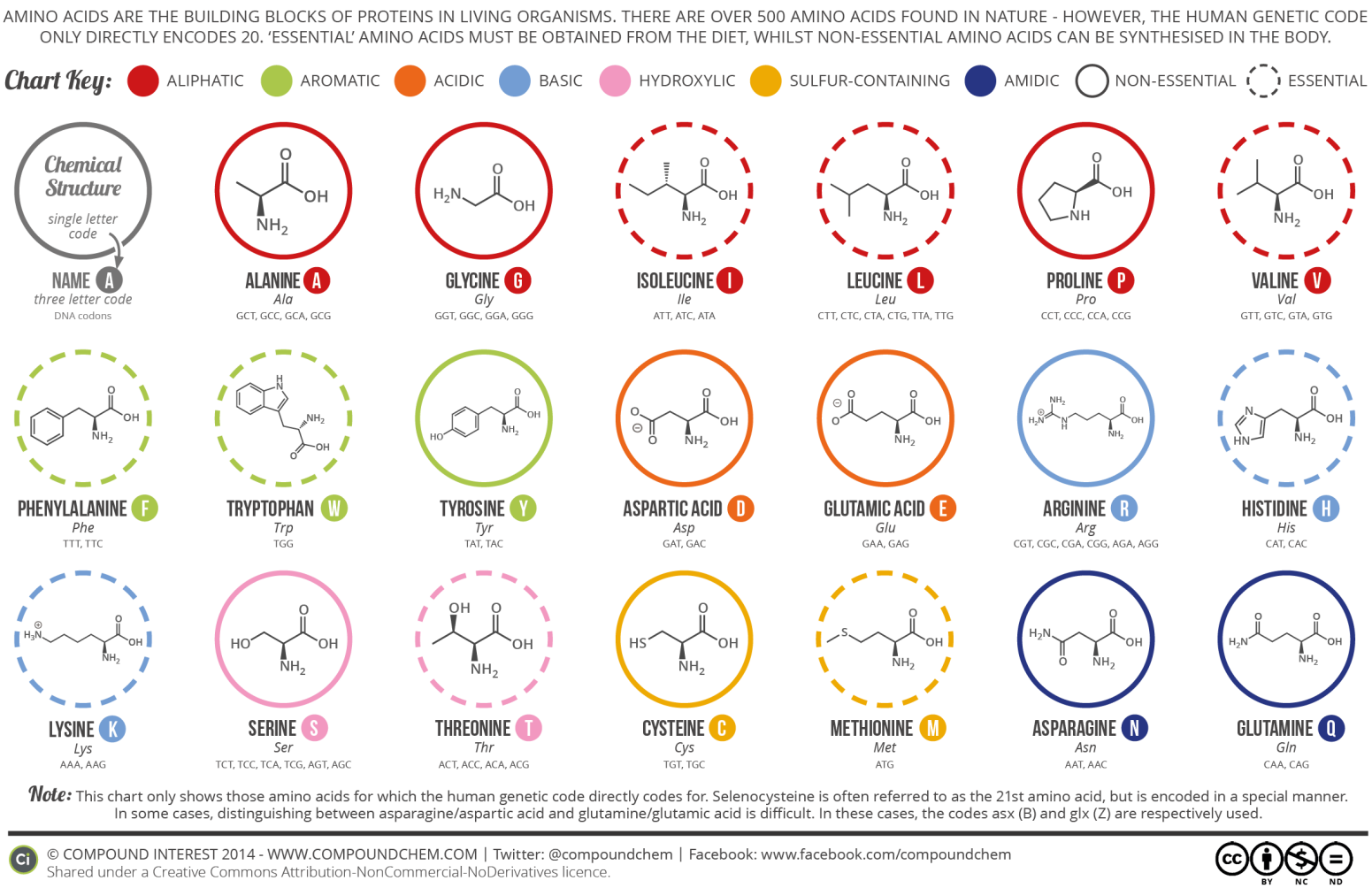


Figure 1.4. Chart showing the structure and basic properties of the 20 amino acids encoded by the standard genetic code. (Picture taken from www.compoundchem.com, shared under a Creative Commons license)

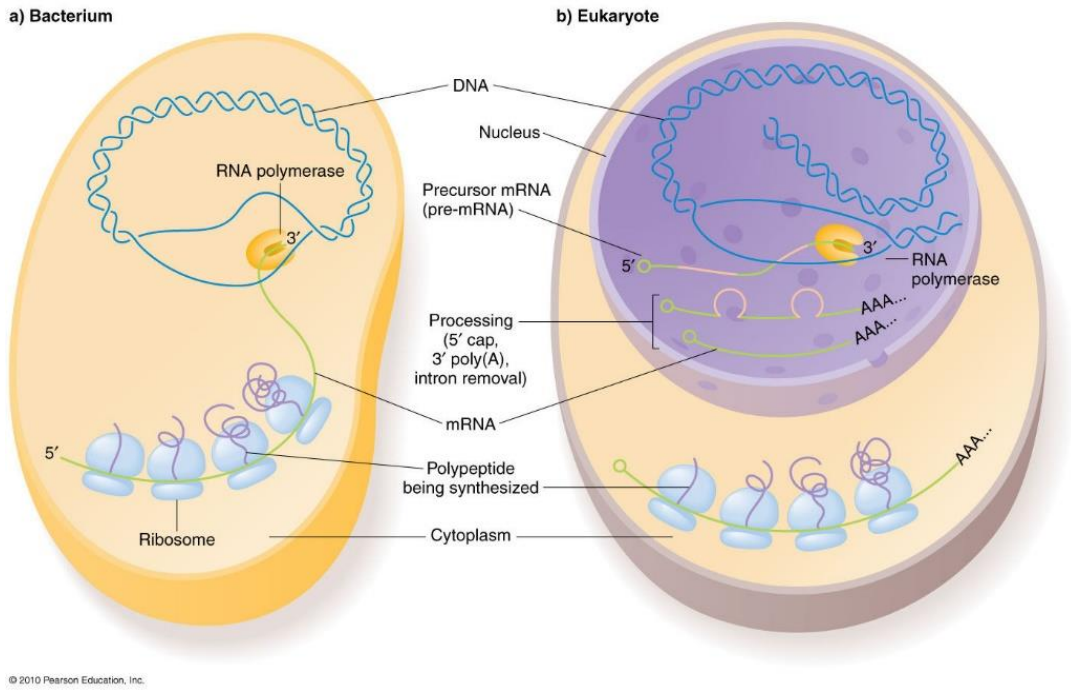


Figure 1.5. The process of protein synthesis in (a) a prokaryote (bacterium), and (b) a eukaryote. RNA polymerase copies the protein-coding DNA into an mRNA during transcription. The transcribed mRNA undergoes several modifications in the case of eukaryotes (Figure 1.6), before it is released to the cytoplasm for translation. Whereas, for the prokaryotes the transcribed mRNA is ready to be read by the ribosome to synthesize the coded protein (polypeptide chain). (Picture taken from https://www.mun.ca/biology/scarr/iGen3_05-09.html, Figure copyright 2010 PJ Russell, iGenetics 3rd ed.)

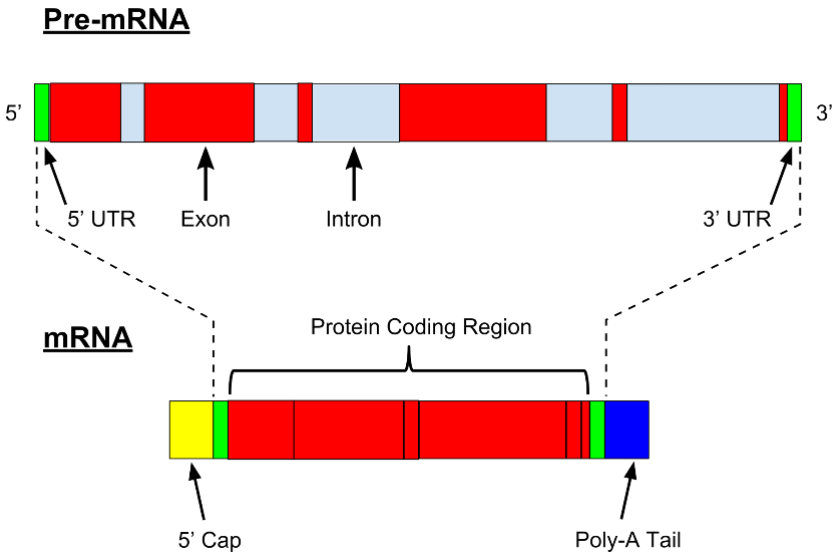


Figure 1.6. Processing of an eukaryotic pre-mRNA into a mature RNA: 5′ mRNA capping, 3′-polyadenylation (poly-A tail), and alternative RNA splicing. UTR is the untranslated region of the mRNA which regulates translation. (Picture by Nastypatty [Public domain], via Wikimedia Commons, shared under a Creative Commons license)

In Figure 1.5, we can clearly see the difference in the process of protein synthesis for prokaryotes and eukaryotes. In the case of prokaryotes, the mRNA sequence obtained after transcription is ready to be translated by the ribosome. Whereas in eukaryotes, the product obtained is a precursor mRNA (pre-mRNA), which contains both exons (information for protein synthesis) and introns (non-coding). This pre-mRNA undergoes several modifications such as: 5′ mRNA capping, 3′-polyadenylation (poly-A tail), and alternative RNA splicing (removing introns and joining the exons) (Figure 1.5, Figure 1.6). After the modifications, the final product is a mature mRNA or simply mRNA, which leaves the nucleus and is ready to be read by the ribosome to produce the encoded protein molecule (amino acid chain) during protein translation.

Before we discuss in detail the translation of an mRNA sequence into the coded protein, let us discuss how a sequence made up of only four nucleotides (A, C, G and U) encodes information according to the genetic code.

1.1.3. Breaking the genetic code

To recall, proteins are made from 20 different amino acids, each protein having a specific order of amino acids joined together in a long polypeptide chain. When we read an mRNA sequence from one end, any of the four nucleotides: adenine (A), cytosine (C), guanine (G) or uracil (U), can be found at a particular position. Thus, we have a combinatorial problem of how these four nucleotides can code for 20 amino acids. If we consider that each of the four nucleotides can code for an amino acid, then only four amino acids will be possible. Similarly, if we consider a dinucleotide code ($AA, AC, \dots, UU$), i.e. two nucleotides coding for an amino acid, it is still not possible for the 16 dinucleotides ($4^2 = 16$ possibilities) to code for 20 amino acids. But, if we consider a trinucleotide code ($AAA, \dots, UUU$), i.e. three nucleotides coding for an amino acid, we have 64 trinucleotides ($4^3 = 64$ possibilities) which would allow coding for 20 amino acids. Clearly, we have now more available options to choose from, thereby introducing redundancy.

Francis Crick and co-workers conducted experiments in 1961 revealing that the genetic code is effectively based on a code of trinucleotides or codons (Crick et al., 1961). Soon after, Marshall Nirenberg and Heinrich Matthaei made one of the most remarkable breakthroughs through a series of experiments using synthetic RNA, where they demonstrated that the UUU trinucleotide codes for the amino acid phenylalanine (Nirenberg & Matthaei, 1961). Within a span of just 5 years, due to the combined efforts of Marshall Nirenberg and Har Govind Khorana, all the 64 possible triplets (codons) were deciphered (Figure 1.7), coding for the 20 amino acids and three special codons called stop codons.

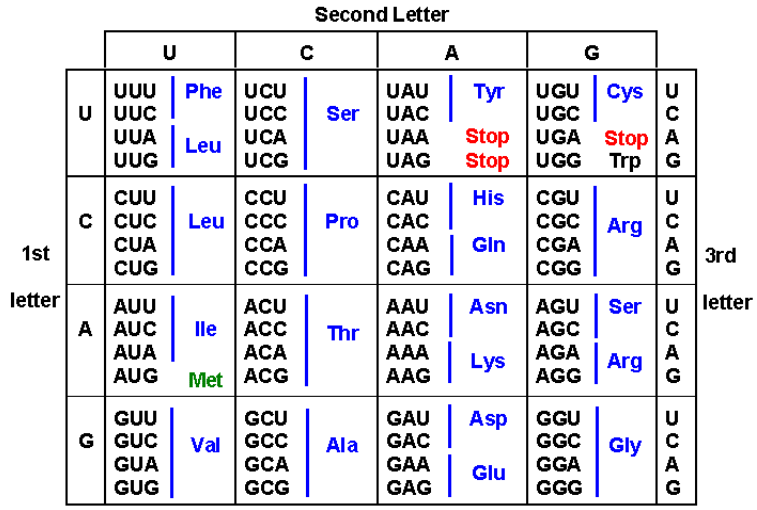


Figure 1.7. The standard genetic code specifies the 64 codons coding for the 20 amino acids. Multiple codons can code for the same amino acid. AUG is the initiation (start) codon shown in green; UAA, UAG and UGA are termination (stop) codons. (Picture taken from http://biology.kenyon.edu/courses/biol114/Chap05/Chapter05.html)

The deciphering of the "standard genetic code" is considered one of most remarkable genetic breakthroughs in the last 60 years. The standard genetic code (SGC) is a set of rules that direct the translation of 64 codons into 20 amino acids along with the start and stop signals. During protein synthesis, the translation process begins with a START codon (generally AUG) and ends with a STOP codon (UAA, UAG or UGA). Some of the amino acids are coded by only one codon, but most of them are coded by multiple codons. Because of this redundancy, the genetic code is said to be degenerate. Usually, the first two positions of a codon are crucial for determining the coded amino acid, and the third position, known as the wobble position, is less critical. In certain cases, if the third base of a codon is modified, it still codes for the same amino acid. For example, the amino acid valine (Val) is coded by 4 codons: GUA, GUC, GUG, and GUU, therefore we can say that valine has degeneracy 4. We can see that the first two bases (GU) of the codons coding for valine are the same; therefore, even if the wobble position changes (either A, C, G, or U), the coded amino acid remains the same. This is one of the ways used by the SGC to minimize translation errors by preventing the incorporation of a wrong amino acid in case of changes in the wobble base. We will talk about the error-minimization properties of the SGC later in the thesis. Next, we will briefly discuss the degeneracy of the SGC.

From a mathematical point of view, the genetic code can be described as a surjective mapping (Definition 2.6), meaning that each amino acid is coded by at least one codon. Only two amino acids (Met and Trp) are coded by a single codon, where the codon AUG coding for Met is also the START codon for protein translation. There are 9 amino acids ($Asn, Asp, Cys, Glu, Gln, His, Lys, Phe,$ and Tyr) with degeneracy 2, which means they are

coded by two codons; one amino acid (Ile) with degeneracy 3; 5 amino acids ($Ala, Gly, Pro, Thr,$ and Val) with degeneracy 4; and 3 amino acids ($Arg, Leu,$ and Ser) with degeneracy 6. This constitutes 61 out of the 64 codons. The remaining three codons do not code for an amino acid, but are used as the STOP signal for protein translation. It is evident that each of the 64 codons has some meaning according to the genetic code. As some amino acids are coded by multiple codons, the genetic code of different organisms is often biased to use one or other of the codons encoding the same amino acid. This also introduces a "codon usage bias" in organisms. The choice of preferred codons is most commonly seen in highly expressed genes. In simpler terms, gene expression refers to the synthesis of the corresponding protein and to do so more efficiently some codons are preferred than others. This brings us back to protein synthesis.

1.1.4. Protein translation

Here we will discuss in detail the second step in protein synthesis, the process of protein translation. The translation machinery that translates mRNA into proteins is called the ribosome. To synthesize a protein, the genetic information encoded in the DNA is first transferred to an mRNA through the process of transcription; the transcribed mRNA is then translated into a protein complex by the ribosome. These processes occur simultaneously in the case of prokaryotes, as shown in Figure 1.5. Whereas in eukaryotes, the transcribed mRNA molecule leaves the nucleus and enters the cytoplasm, where it is translated into a protein. A cell's cytoplasm is packed with ribosomes, RNA polymerases, tRNAs, mRNAs, and enzymes, all performing their respective functions independently. Before going into further details, we will discuss the role of tRNAs in protein translation.

The tRNAs play an important role during protein translation, since they are adaptor molecules that act as carriers for the 20 different amino acids during the translation of an mRNA sequence by the ribosome. tRNAs have a unique L-shaped 3D structure (Figure 1.8); one end of the tRNA attaches (binds) itself to the mRNA sequence by complementary base-pairing (anticodon of tRNA with the codon of mRNA) and the other end (3'-end or CCA-tail) attaches to a specific amino acid. The correct amino acid is added to the 3'-end of a tRNA by enzymes known as aminoacyl tRNA synthetases, during tRNA charging. These carrier molecules ensure that the appropriate amino acid is inserted into the growing polypeptide chain (protein complex). tRNAs have a distinctive 2D-structure with three hairpin loops (D loop, T loop, and anticodon loop) forming a three-leafed clover. In Figure 1.8, we can see the common ways of illustrating tRNAs (3D and 2D structure).

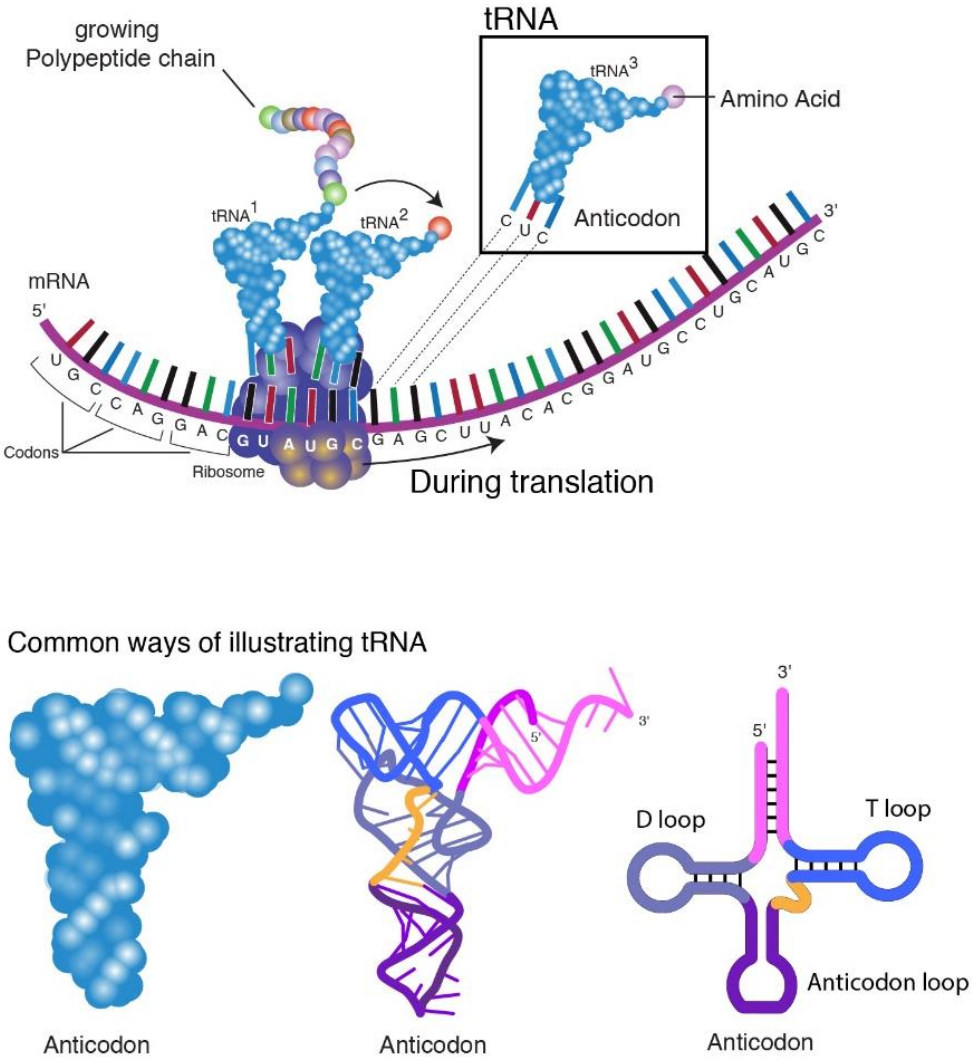


Figure 1.8. The structure of tRNAs and their role is protein translation (Figure 1.10). tRNAs are an important link between the transcribed mRNA and the amino acid they carry. tRNAs attach to the mRNA via anticodon-codon interaction, carrying the amino acid coded by the codon in the mRNA. (Picture taken from https://www.genome.gov/genetics-glossary/Transfer-RNA, Courtesy: National Human Genome Research Institute)

Next, we will discuss the structure of the ribosome and explain how the mRNA sequence is translated into a protein complex.

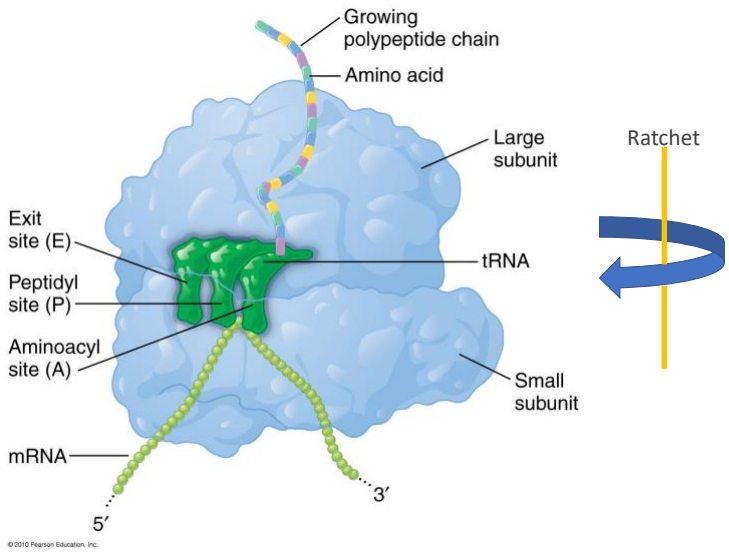


Figure 1.9. Structure of the modern ribosome with three tRNAs, mRNA and polypeptide chain. The ribosome is composed of two subunits: large subunit (LSU) and small subunit (SSU). These subunits are nucleoprotein complexes which come together during the initiation stage of the protein synthesis. (Original picture taken from https://www.mun.ca/biology/scarr/iGen3_06-14.html, Figure copyright 2010 PJ Russell, iGenetics 3rd ed)

Ribosomes are composed of ribosomal proteins and ribosomal RNAs (rRNAs). In Figure 1.9, we show the structure of the ribosome with tRNAs and the mRNA during translation. A ribosome has two main components, a large subunit (LSU) and a small subunit (SSU). Both LSU and SSU are composed of rRNAs and a variety of proteins. The small subunit of the ribosome is responsible for the initiation of the translation process: the small subunit moves along the mRNA strand (made up of codons) in the 5′ to 3′ direction until a start codon AUG is found. The met-tRNA containing the anticodon UAC, then pairs (by complementary base pairing) with the start codon AUG of the mRNA forming the initiation complex. Once the start codon is identified, the large subunit (LSU) attaches itself onto the small subunit and the process of elongation commences.

The ribosome contains three tRNA-binding sites: A-site (aminoacyl), P-site (peptidyl) and E-site (exit) (Figure 1.9). The tRNAs at these three binding sites are called aminoacyl-tRNA (A-tRNA), peptidyl-tRNA (P-tRNA) and exit-tRNA (E-tRNA) respectively. During the process of elongation, an aminoacyl tRNA with the corresponding amino acid enters the A-site in the ribosome decoding center (including the universally conserved nucleotides G530, A1492 and A1493) and cognate/near-cognate tRNAs are identified by codon-anticodon interactions with the mRNA codon (proofreading step). Then, a peptide bond is formed between the methionine of the met-tRNA at the P-site and the amino acid of the aminoacyl-tRNA at the A-site resulting in the deacylation (amino acid released from one end of the tRNA) of the met-tRNA. At this

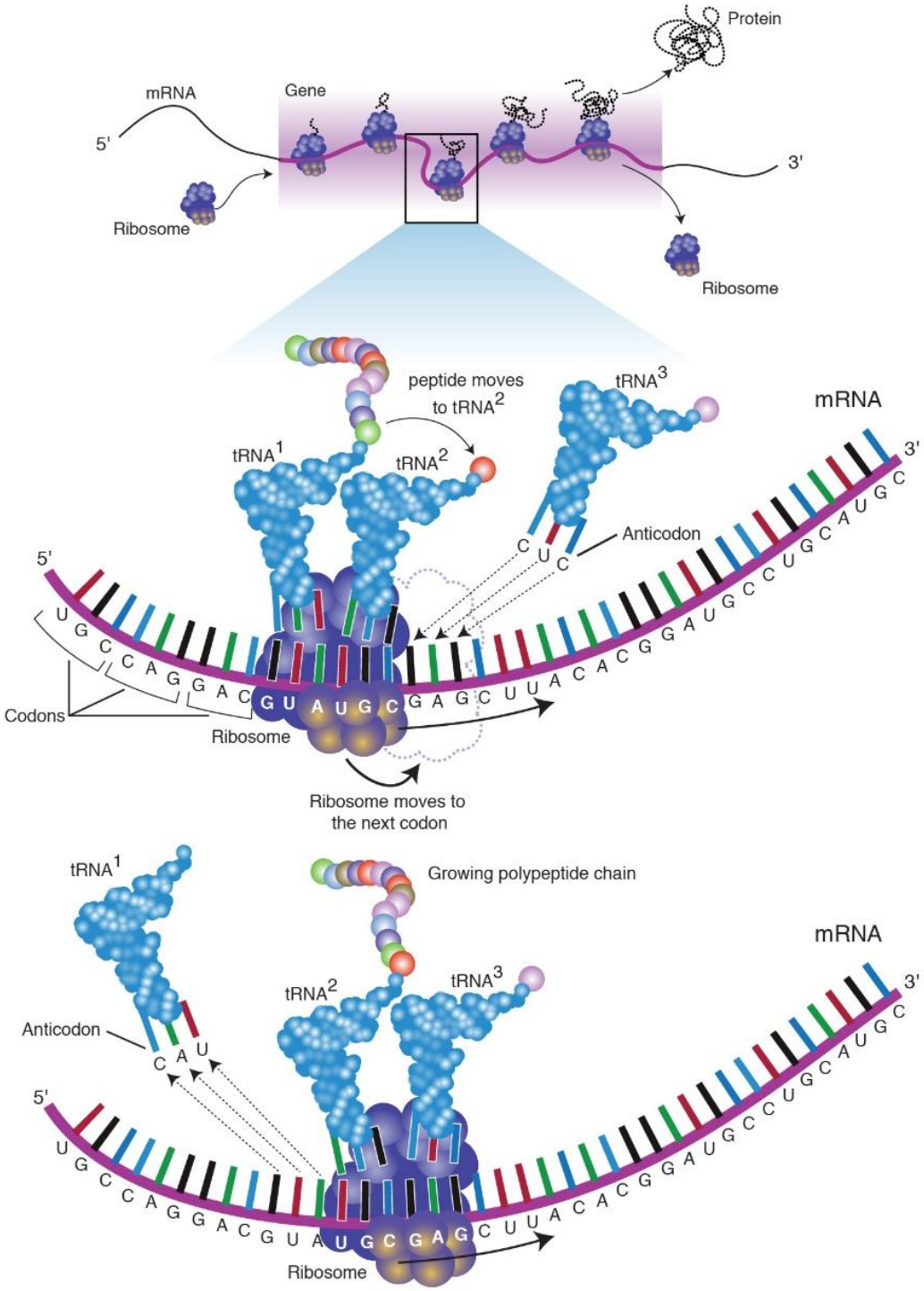


Figure 1.10. Translation of an mRNA sequence into a protein complex by the ribosome. tRNAs carrying amino acids bind (complementary base pairing) with the mRNA at specific sites, thereby playing the role of carriers of amino acids during protein translation. (Picture taken from https://www.genome.gov/genetics-glossary/Gene-Expression, Courtesy: National Human Genome Research Institute)

point after the peptidyl transferase reaction, there is a deacylated-tRNA at the P-site and a dipeptidyl-tRNA at the E-site. Then the ribosome advances a distance of one codon through a ratchet movement and the deacylated tRNA is shifted to the E-site, where it gets detached from the ribosome (Figure 1.10). Another amino acid carrying tRNA then enters the A-site, leading to another peptide-bond formation. The repetition of this process (elongation) results in the formation of a polypeptide chain, until a stop codon is decoded at the A-site, thereby terminating the translation process. The polypeptide chain is released by the ribosome through the exit tunnel. The two subunits of the ribosome separate once the protein synthesis is finished (ribosome recycling). After ribosome recycling, the SSU can initiate translation on a new mRNA.

Through these complex processes, the flow of information from DNA to RNA and eventually to protein molecules is considered as the central dogma of genetics. Apart from a few differences, both prokaryotes and eukaryotes follow similar biological mechanisms to synthesize proteins. It is now clear, mechanisms involved in protein synthesis are extremely vital and the standard genetic code directs the decoding of genetic information into proteins. Next, we will move on to discuss how the standard genetic code (SGC) may have evolved from primitive life to extant organisms.

1.2. Evolution of the genetic code

All known living organisms with only a few exceptions share the same standard genetic code (SGC), which supports the idea that the code evolved in a group of primitive structures preceding the first cells, collectively known as the common ancestor of all life.

The Earth is estimated to be 4.5 billion years old and the events shaping the genetic code took place 3.7–4.1 billion years ago (Nutman et al., 2016). All known modern life forms as of today, trace back to a *last universal common ancestor* (LUCA) that had a cell membrane and a primitive translation machinery. Since LUCA, the same standard genetic code has been used by (almost) all organisms to translate information encoded within the genetic material into proteins or amino acids. The ribosome, described in the previous section, is responsible for translating this specific genetic information transferred by mRNA. LUCA was an intermediate system between the origin of life and the life on earth today. It is generally believed that this intermediate system (LUCA) was far too complex to emerge spontaneously; rather it must have evolved from simpler systems that were capable of performing self-replication with some error-tolerance.

Due to the universality of the genetic code, it is difficult to trace back how it may have evolved as no organisms exist containing a primitive or intermediate genetic code for comparison. Researchers have continuously debated various theories that attempt to explain

how the genetic code may have emerged from the primordial chemical soup (Koonin & Novozhilov, 2009). The stereochemical theory is based on the hypothesis that some kind of stereochemical relationship existed between codons (or anticodons) and assigned amino acids (Pelc & Welton, 1966; Yarus, 2017). The coevolution theory postulates the origin of the genetic code through biosynthetic pathways (Wong, 1975); at first a few amino acids (precursors) were coded and the rest of the amino acids developed biosynthetically from these precursor amino acids. The adaptive theory suggests the code evolved in order to minimize mutation errors and to become maximally robust (Freeland & Hurst, 1998); it also implies similar amino acids being coded by similar codons. The frozen accident theory proposes a random origin of the codon assignments to the amino acids with successive evolution due to different evolutionary pressures which has stayed frozen ever since. The general textbook concept of an early "RNA world" (Gilbert, 1986) has dominated for quite long in the past; suggesting that the RNA was the first molecule that facilitated the origin of life by storing information and catalysing chemical reactions, and the DNA and proteins evolved later. Recent studies (Bowman et al., 2015; van der Gulik & Speijer, 2015; Kunnev & Gospodinov, 2018; Carter & Wills, 2018; Chatterjee & Yadav, 2019; Piette & Heddle, 2020) have suggested that RNA alone may not have been able to perform the functions needed for the origin of life; instead short peptides interacted with the RNA to produce enzymes needed for the origin of life, thereby suggesting an early "peptide–RNA world".

The ribosomal RNAs are considered some of the most conserved biomolecules in evolution, which suggests they must have emerged very early. Primordial rRNAs could produce proteins (short peptides from a few randomly joined amino acids) necessary for their diverse functions and development. Hence, the primordial protein synthesis machinery may have originated from the interaction between ribosomal RNAs and short peptides. Biochemists Stanley Miller and Harold Urey conducted an experiment in one of the most remarkable discoveries that demonstrated the synthesis of several amino acids by simulating the early atmospheric conditions of the Earth (Miller, 1953). In this experiment, they used a mixture of methane (CH_4), ammonia (NH_3), water and hydrogen (H_2), treated with electrical discharge to successfully synthesize 5 amino acids: glycine (Gly), aspartic acid (Asp), α- and β-alanine, and α-aminobutyric acid. More recently, in an adaptation of the Miller-Urey experiment (Ferus et al., 2017), the authors demonstrated the synthesis of all RNA nucleobases: A, C, G and U, along with the simplest amino acid glycine (Gly) by using a mixture of ammonia (NH_3), carbon monoxide (CO) and water treated with electrical discharge and a laser shockwave. These experiments demonstrate how the simpler biomolecules needed for the origin of life may have emerged from the primordial chemical soup.

Complete recreation of the complex processes that facilitated the evolution of early translation systems and the genetic code is quite unlikely. Still, researchers around the globe are

trying to solve the mystery of the evolution of the genetic code. Next, we will discuss how genomes have acquired the ability to encode overlapping signals, in addition to the genetic code.

1.3. Multiple genome codes

It is generally believed that genomes encode multiple "overlapping signals" or "auxiliary genetic information" in the protein-coding regions (Weatheritt & Babu, 2013; Maraia & Iben, 2014; Bergman & Tuller, 2020), which is possible due to the redundancy of the genetic code. In fact, the codon usage in some proteins is highly biased, indicating some additional constraints. By statistical analysis of short sequences (*k*-mers) in the protein-coding regions of nearly 700 different species, it has been shown that from bacteria to eukaryotes, all organisms encode overlapping information (Itzkovitz et al., 2010). Here we discuss the possibilities leading to this multiple encoding of genetic information in addition to the amino acid sequences.

As described above, the standard genetic code (which is a mapping between the 64 codons and the 20 amino acids with the start and stop signals) is degenerate, which allows extra space for other "overlapping codes" found in the genome. After the discovery of the genetic code by Crick (1961), many researchers discovered a wide variety of new codes, such as the operational RNA code (Schimmel et al., 1993), the protein folding code (Rackovsky, 1993), the *X* circular code (Arquès & Michel, 1996), the adhesive code (Redies & Takeichi, 1996), the sequence codes (Trifonov, 1999), nucleosome positioning code (Segal et al., 2006), codon usage code (Yu et al., 2015), the splicing code (Baralle et al., 2019; Baralle & Baralle, 2018), the histone code (Jenuwein & Allis, 2001), the sugar code (Gabius & Roth, 2017), the tubulin code (Verhey & Gaertig, 2007), the ubiquitin code (Komander & Rape, 2012) and many more.

We have given a general introduction to this thesis, explaining the biological context of the work. Before we go any further, we will discuss error-correcting codes; the comma-free codes and the circular codes, to provide an introduction to the results presented in this thesis.

Chapter 2

2. Error-correcting codes

2.1. Introduction

The genetic code is said to have error-mitigating properties (Freeland & Hurst, 1998) and may have evolved explicitly to reduce the effects caused by translation errors (Woese, 1965). According to Warnecke and Hurst, "several features of gene structure and genome design could be adaptations to error-prone gene expression" (Warnecke & Hurst, 2011). Different error-correction hypotheses have been proposed. One such hypothesis, called the ambush hypothesis (Seligmann & Pollock, 2004), proposes the presence of off-frame stop codons in the coding sequences that terminate frameshifted translation (Farabaugh & Björk, 1999; Parker, 1989), thereby reducing energy and resource waste on non-functional proteins. In addition, codons which are more likely to form hidden stops or off-frame stops have a higher usage frequency and bias in their favour among the synonymous codons. Biased codon usage can be a useful signature of additional CDS (coding region of a gene) functionality. In the human genome, out of frame stop codons (also called ambush codons) are significantly more abundant than those codons lacking the ability to transform into a stop codon in a shifted frame (Warnecke & Hurst, 2011). We will address the problem of reading frame maintenance and frameshifts (causing out of frame codons), along with possible mechanisms in place to deal with these problems to reduce errors or reduce the effect of errors already occurred.

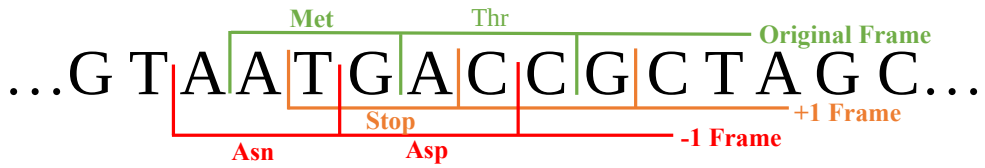


Figure 2.1. Original reading frame in comparison to the two shifted frames $+1$ and -1 results in different read out of amino acids.

An important source of translation errors is ribosomal frameshifting, which occurs with an error rate of around 10^{-5} (Drummond & Wilke, 2009). Since the genetic code has a non-overlapping structure, the codons in a DNA sequence must be decoded in the correct reading frame in order to produce the correct amino acid sequence. A shift of one or two bases into the $+1$ or -1 frames respectively, can have severe effects, including termination of translation if a

stop codon is encountered out-of-frame, or production of a non-functional protein sequence otherwise (Figure 2.1).

2.2. Reading frame maintenance

Similar to any digital transmission of a message, an accurate protein synthesis requires effective means to ensure that the decoding process is synchronized with the correct reading frame of codons. Without the correct reading frame (in case of a frameshift), the protein synthesis process can terminate beforehand or even the protein produced can be non-functional. As every trinucleotide in the genetic code represents one amino acid, loss of frame maintenance results in a completely erroneous translation and frameshift reading errors usually result in different amino acids. The problem of reading frame location and maintenance in mRNA translation has been one of the most important topics of molecular biology. Next, we will introduce "comma-free codes" and "circular codes", in the context of this problem.

2.3. Preliminary definitions

Before going into detail about the error-correcting codes, we will introduce some mathematical definitions here, which are necessary to understand the notations used in this thesis.

Notation 1. Let us denote the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine. The trinucleotide set over B is denoted by $B^3 = \{AAA, \dots, TTT\}$. The set of non-empty words (words, respectively) over B is denoted by B^+ (B^*, respectively).

Figure 2.2. The three frames: frame 0 (original reading frame), frame 1 and frame 2 while reading a sequence of trinucleotides.

Notation 2. Genes or motifs in reading frame have three frames f. By convention here, the reading frame $f = 0$ is set up by a start trinucleotide, classically ATG, and the frames $f = 1$ and $f = 2$ are the reading frame $f = 0$ shifted by one and two nucleotides in the $5' - 3'$ direction (to the right), respectively (Figure 2.2).

Definition 2.1. The *trinucleotide circular permutation map* $\mathcal{P}: B^3 \to B^3$ is defined by $\mathcal{P}(l_0 l_1 l_2) = l_1 l_2 l_0$ for all $l_0, l_1, l_2 \in B$. The 2nd iterate of $\mathcal{P}$ is $\mathcal{P}^2(l_0 l_1 l_2) = l_2 l_0 l_1$, e.g. $\mathcal{P}(ACG) = CGA$ and $\mathcal{P}^2(ACG) = \mathcal{P}\big(\mathcal{P}(ACG)\big) = \mathcal{P}(CGA) = GAC$. 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 $\mathcal{P}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{P}(u)\}$, e.g. $\mathcal{P}(\{ACG, GTA\}) = \{TAG, CGA\}$ and $\mathcal{P}^2(\{ACG, GTA\}) = \mathcal{P}\{TAG, CGA\} = \{GAC, AGT\}$.

Definition 2.2. According to the complementary property of the DNA double helix, the *nucleotide complementarity map* $\mathcal{C}: B \to B$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, $\mathcal{C}(T) = A$. According to the complementary and antiparallel properties of the DNA double helix, the *trinucleotide complementarity map* $\mathcal{C}: B^3 \to B^3$ is defined by $\mathcal{C}(l_0 l_1 l_2) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in B$. By extension to a trinucleotide set S, the set complementarity map $\mathcal{C}: \mathbb{P}(B^3) \to \mathbb{P}(B^3)$, $\mathbb{P}$ being the set of all subsets of B^3, is defined by $\mathcal{C}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{C}(u)\}$, e.g. $\mathcal{C}(\{CGA, GAT\}) = \{ATC, TCG\}$.

Definition 2.3. A set $S \subseteq B^+$ of words is a code if, for each $x_1, \dots, x_n, y_1, \dots, y_m \in S$, $n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \dots, n$, e.g. $B^3 = \{AAA, \dots, TTT\}$ is a code, but the set $Y = \{A, CG, ACG\}$ is not a code as there are two decompositions $A \cdot CG = ACG$.

Definition 2.4. Any non-empty subset of the code B^3 is a code and called *trinucleotide code*.

Definition 2.5. A trinucleotide code $X \subseteq B^3$ is self-complementary if, for each $t \in X$, $\mathcal{C}(t) \in X$, i.e. $X = \mathcal{C}(X)$.

Definition 2.6. The standard genetic code (SGC) is a self-complementary trinucleotide code. It defines a surjective map $g: \tilde{B}^3 \longrightarrow P$ where $\tilde{B}^3 = B^3 \backslash \{TAA, TAG, TGA\}$ and P is the set of the 20 peptide components (amino acids):

$$P = \{Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu,$$
$$Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val\},$$

e.g. $g(GGA) = Gly$, $g^{-1}(Gly) = \{GGA, GGC, GGG, GGT\}$.

2.4. Comma-free codes

Before the genetic code was discovered, Crick for the first time suggested a class of trinucleotide codes, called comma-free codes, that allows a sequence of trinucleotides to code for the 20 amino acids and to retrieve the correct reading frame at the same time (Crick et al., 1957). A code which can be decoded without separation symbols is called "comma-free". The following two points were extremely important considering the construction of the comma-free codes:

(a) There are 64 different trinucleotides possible with the four nucleotides A, C, G and T, but they code for only 20 amino acids rather than 64.

(b) How is the original reading frame detected, or how are the groups of trinucleotides chosen to read the correct amino acids?

There are a few constraints to be considered while constructing a comma-free code. Out of the 64 possible trinucleotides, the four periodic trinucleotides AAA, CCC, GGG and TTT must be excluded; e.g., a sequence containing $AAAAAA$ can be misinterpreted as $\ldots AAA, AAA, \ldots$ or $\ldots A, AAA, AA \ldots$ or $\ldots AA, AAA, A \ldots$, which does not allow the detection of the reading frame and creates ambiguity while choosing the correct frame. The remaining 60 trinucleotides were grouped into 20 sets of three trinucleotides, each set of three being cyclic permutations of one another; e.g. ACG and its cyclic permutations CGA and GAC. After this restricted grouping, at most one triplet from each cyclic set was chosen to code for the 20 amino acids without ambiguity. Indeed, the concatenation of ACG with itself, e.g. $ACGACG$ can be misinterpreted as $\ldots ACG, ACG \ldots$ or $\ldots A, CGA, CG \ldots$ or $\ldots AC, GAC, G \ldots$, if one of the two permuted trinucleotides is included in the code, which does not allow the detection of the reading frame. In other words, trinucleotides in the reading frame make sense, whereas trinucleotides in the shifted frames make nonsense.

In coding theory, such a comma-free code is known as a self-synchronizing code as no external synchronization is necessary. These codes also allow error detection in the coding sequences by rejecting non-valid codons. A comma-free code can notably retrieve the correct reading frame within a window of three consecutive nucleotides. As shown in the Figure 2.3, codons in green in the first row belong to the comma-free code $\{GTA, GTC, GTG, GTT\}$, whereas codons in grey in the two shifted frames shown in the second and third row do not belong to the comma-free code. Therefore, a comma-free code detects the reading frame immediately.

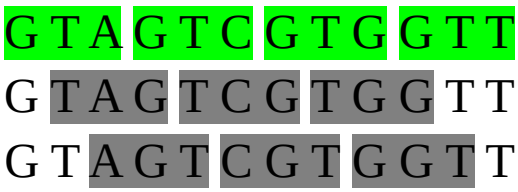


Figure 2.3. A comma-free code $\{GTA, GTC, GTG, GTT\}$ detects the reading frame immediately. All codons in gray in the two shifted frames: frame 1 codons shown in second row $\{TAG, TCG, TGG\}$ and frame 2 codons shown in third row $\{AGT, CGT, GGT\}$ do not belong to the comma-free code.

However, no trinucleotide comma-free codes have been identified statistically in genes. It was later proved that the genetic code could not be a comma-free code when it was discovered that TTT codes for the amino acid phenylalanine (Phe). Other comma-free codes that may have represented primitive codes are mentioned here:

a) The RRY code ($R = \{A, G\}, Y = \{C, T\}$) which contains eight trinucleotides and codes for four amino acids (Crick et al., 1976).

b) The RNY code ($N =$ any nucleotide) which contains 16 trinucleotides and codes for eight amino acids (Eigen & Schuster, 1978; Shepherd, 1981).

c) The GNC code which contains four trinucleotides and codes for four amino acids (Ikehara, 2002).

2.5. Circular codes

A less restrictive variant of the comma-free codes, the circular codes also possess the ability to retrieve, maintain and synchronize the reading frame. These codes possess the circular property, i.e. any word written on a circle (the last letter becoming the first in the circle) has a unique decomposition into trinucleotides of the circular code (Figure 2.4). A circular code naturally excludes the periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$. It also excludes trinucleotides related by circular permutation (Michel, 2008). By excluding the periodic trinucleotides and dividing the 60 remaining trinucleotides into three disjoint classes, a circular code of trinucleotides has at most 20 trinucleotides (called a maximal circular code). There exist 12,964,440 maximal circular codes, although it has been shown that there is no maximal circular code that can code 20 or 19 amino acids and only 10 can code for 18 amino acids (Michel & Pirillo, 2013).

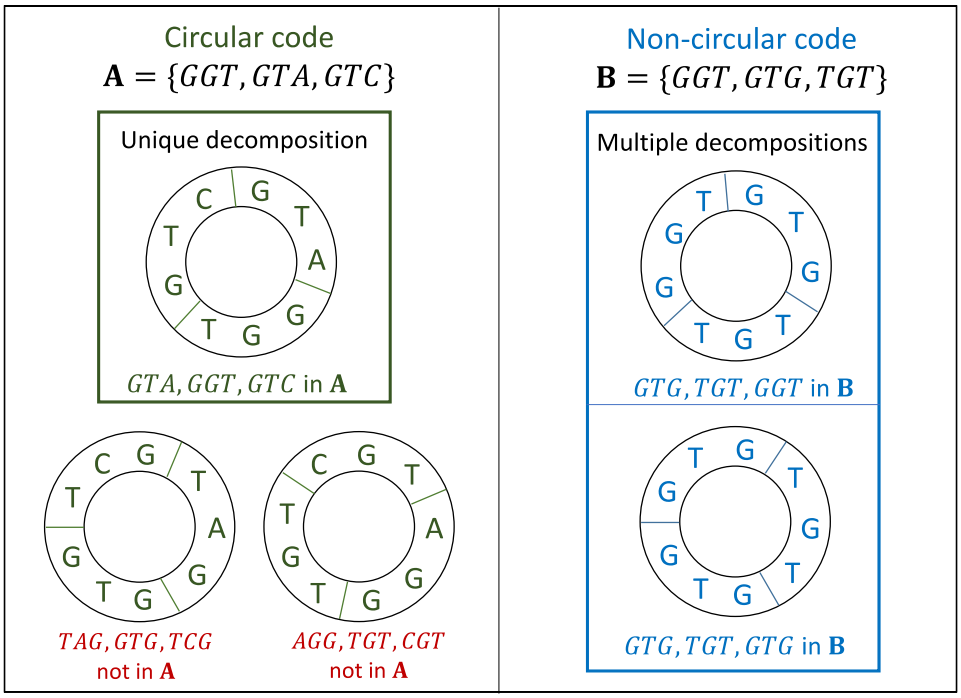


Figure 2.4. For a circular code, e.g. $\{GGT, GTA, GTC\}$, any word when written on a circle has a unique decomposition into the trinucleotides of the circular code, thereby retrieving the original reading frame. But, for a non-circular code, e.g. $\{GGT, GTG, TGT\}$, there can be multiple decompositions, therefore ambiguous.

In 1996, Arquès and Michel performed a statistical analysis of genes of prokaryotes and eukaryotes and identified a circular code X of 20 trinucleotides (codons) that are preferentially present in the reading frame. The circular code X contains the 20 following trinucleotides:

$$\begin{aligned} X = \{&AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\ &GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\} \end{aligned} \quad (1)$$

which code for the 12 following amino acids (three and one letter notation):

$$\begin{aligned} \mathcal{X} &= \{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val\} \\ &= \{A, N, D, Q, E, G, I, L, F, T, Y, V\} \end{aligned} \quad (2)$$

Other circular codes, and notably variations of the circular code X, are hypothesized to exist in different organisms (Frey & Michel, 2003, 2006; Ahmed et al., 2010; Michel, 2017). Motifs obtained from a circular code have the ability to retrieve, maintain and synchronize the original reading frame, not immediately in contrast to the comma-free codes, but after a maximum of 13 consecutive nucleotides (Figure 2.5).

Figure 2.5. The circular code X does not detect the reading frame immediately, but after a maximum of 13 nucleotides. All codons in gray in the two shifted frames 1 and 2 shown in second and third row respectively, do not belong to the code.

The circular code X is denoted as X_0, as the codons of the code are preferentially present in the reading frame (frame 0). The circular code X has strong mathematical properties; in particular, it is self-complementary (Figure 2.6), meaning that if a trinucleotide belongs to X_0, then its complementary trinucleotide also belongs to X_0. Moreover, the $+1$ and $+2/-1$ circular permutations of X_0, denoted X_1 and X_2 respectively (Figure 2.7), are also maximal circular codes and are complementary to each other (Figure 2.8).

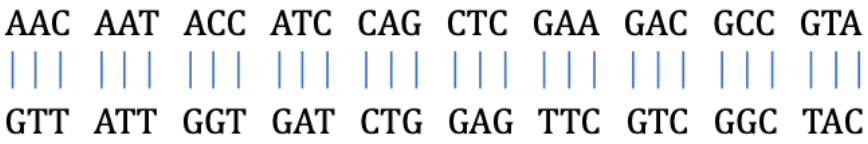


Figure 2.6. The self-complementary property of circular code X, where 10 of its trinucleotides are complementary to the other 10 trinucleotides.

$$X_0 = n_0n_1n_2 \qquad X_1 = n_1n_2n_0 \qquad X_2 = n_2n_0n_1$$

$$X_0 = \{AAC,\ AAT,\ ACC,\ ATC,\ ATT,\ CAG,\ CTC,\ CTG,\ GAA,\ GAC,\ GAG,\ GAT,\ GCC,\ GGC,\ GGT,\ GTA,\ GTC,\ GTT,\ TAC,\ TTC\}$$

$$X_1 = \{ACA,\ ATA,\ CCA,\ TCA,\ TTA,\ AGC,\ TCC,\ TGC,\ AAG,\ ACG,\ AGG,\ ATG,\ CCG,\ GCG,\ GTG,\ TAG,\ TCG,\ TTG,\ ACT,\ TCT\}$$

$$X_2 = \{CAA,\ TAA,\ CAC,\ CAT,\ TAT,\ GCA,\ CCT,\ GCT,\ AGA,\ CGA,\ GGA,\ TGA,\ CGC,\ CGG,\ TGG,\ AGT,\ CGT,\ TGT,\ CTA,\ CTT\}$$

Figure 2.7. The $+1$ and $+2$ circular permutations of X_0, denoted as X_1 and X_2 respectively, are also maximal circular codes and complementary to each other (Figure 2.8).

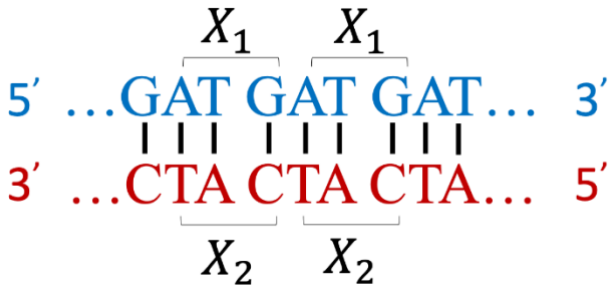


Figure 2.8. Complementary property of the two permuted codes of circular code X (X_0). A codon of X_0 $(n_0n_1n_2)$ in the shifted frame 1 belongs to the permuted code X_1 $(n_1n_2n_0)$ of the strand $5' - 3'$ and is complementary to the codon $(n_2n_0n_1)$ in the shifted frame 2 which belongs to the permuted code X_2 of the strand $3' - 5'$, and vice versa.

2.6. Mathematical definitions and properties of circular code

We recall the mathematical definition of a circular code and a few mathematical properties.

Definition 2.7. A trinucleotide code $X \subseteq B^3$ is circular 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 $i = 1, \ldots, n$.

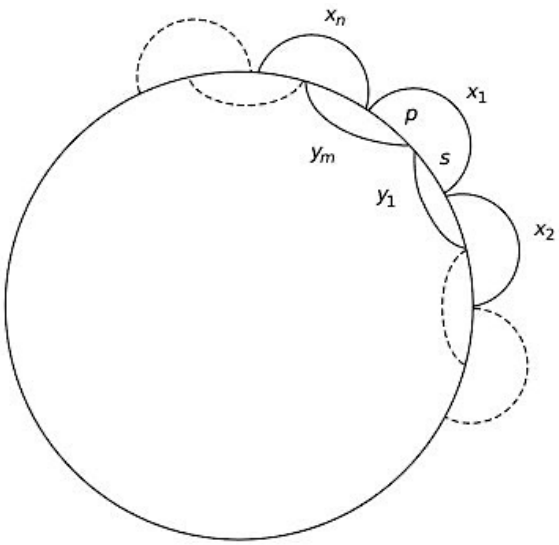


Figure 2.9. Graphical representation of a circular code (Definition 2.7).

Definition 2.8. A trinucleotide circular code $X \subseteq B^3$ is maximal if for all trinucleotide circular codes $Y \subseteq B^3$, we have $|Y| \leq |X|$.

The 60 trinucleotides of $B^3 \backslash \{AAA, CCC, GGG, TTT\}$ when arranged in 20 classes invariant by circular permutations, a trinucleotide circular code has at most one trinucleotide from each class. Therefore, a trinucleotide circular code $X \subseteq B^3$ has at most 20 trinucleotides and the

maximality is 20 trinucleotides. In other words, a maximal circular code cannot be included in another circular code.

Definition 2.9. A trinucleotide circular code $X \subseteq B^3$ is C^3 *self-complementary* if X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes such that $X = \mathcal{C}(X)$ (self-complementary) (Figure 2.6), $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary to each other) (Figure 2.8).

The trinucleotide set X (1) coding the reading frame ($f = 0$) in genes is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996), where the maximal circular code $X_1 = \mathcal{P}(X)$ coding the frame $f = 1$ contains the 20 following trinucleotides:

$$X_1 = \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \quad (3)$$

and the maximal circular code $X_2 = \mathcal{P}^2(X)$ coding the frame $f = 2$ contains the 20 following trinucleotides:

$$X_2 = \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}. \quad (4)$$

The trinucleotide circular codes X_1 and X_2 are related by the permutation map, i.e. $X_2 = \mathcal{P}(X_1)$ and $X_1 = \mathcal{P}^2(X_2)$ (Figure 2.7), and by the complementary map, i.e. $X_1 = \mathcal{C}(X_2)$ and $X_2 = \mathcal{C}(X_1)$ (Figure 2.8) (Bussoli et al., 2012).

There exists 216 maximal C^3 self-complementary trinucleotide circular codes including the X circular code observed in genes. Motifs from the circular code X having the reading frame maintenance property are called X motifs.

2.7. Classes of motifs

Here we define the three different classes of motifs that we have used in the work presented during this thesis.

Definition 2.10. An *X motif* is a motif constructed from the circular code X (1), with cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides, which has the ability to retrieve, maintain and synchronize the reading frame.

Here, the cardinality c of a motif refers to the number of unique codons that belong to the code. Any maximal circular code can contain a maximum of 20 trinucleotides, therefore we have chosen the cardinality $c \leq 20$ trinucleotides for the circular code X. We have excluded the class of X motifs with cardinality $c < 4$ trinucleotides that are mostly associated with the "pure" trinucleotide repeats often found in non-coding regions (Michel et al., 2017). As we have

mentioned previously, an X motif is able to retrieve the reading frame after a maximum of 13 nucleotides. We have chosen the minimal length to be 4 trinucleotides, thereby choosing very strict constraints so that each X motif has the ability to retrieve the reading frame.

Definition 2.11. A non-X *motif* ($m(\bar{X})$) is a word constructed from the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$, excluding the X motifs ($m(X)$) of Definition 2.10. For comparison purposes, we only consider the non-X motifs found in the reading frame.

If we exclude all the X motifs found in the reading frame, the rest of the sequence can be considered as non-X motifs.

Remark 1. The non-X motifs can be of any cardinality and length.

In order to evaluate the statistical significance of X motifs in the protein-coding genes of organisms, we have generated random codes (R).

Definition 2.12. A random motif ($m(R)$) is a motif constructed from a random code R, with cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides.

Any random code R generated for this analysis has similar properties to the circular code X, except its circularity property. Therefore, random motifs do not possess the frame-retrieval property of the circular codes. We generate random codes taking into account the following properties:

(a) R has a cardinality equal to 20 trinucleotides.
(b) In any random code R, the total number of each nucleotide A, C, G and T is equal to 15. The circular code X has this property (Note: $20 \times 3 = 15 \times 4$).
(c) R does not contain any stop trinucleotides $\{TAA, TAG, TGA\}$ and periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$.
(d) R is not a circular code, thereby do not possess the frame-retrieval property.

For comparison purposes, we consider the R random motifs found only in the reading frame of genes. We have generated 100 different random codes R (Appendix Table V).

2.8. Summary

In this chapter we introduced the error-correcting codes, in particular the X circular code and the importance of reading frame retrieval during protein synthesis. We also defined the different classes of motifs that have been used for comparison. After the discovery of the X circular code in 1996, a significant amount of research has been conducted in studying the properties of the X circular code, based on statistical analysis, combinatorics and graph theory.

In a large scale statistical study involving 138 complete eukaryotic genomes, it was shown that X motifs occur preferentially in genes (El Soufi & Michel, 2016). This study involved several statistical analyses of the X motifs in both protein-coding and non-coding regions searching for large X motifs with lengths of at least 15 consecutive trinucleotides and cardinality 10. It was shown that the proportion of the X motifs found in gene regions to that found in non-gene regions is nearly equal to 8. More recently, in a statistical analysis of the complete genome of the existing eukaryote *Saccharomyces cerevisiae*, for the first time the circular code theory was tested analysing the X motifs (Michel et al., 2017). Various properties of X motifs were identified by simple frequency-level statistics, which were tested by comparison with random motifs (30 random codes generated). It was demonstrated that the X motifs were significantly enriched in the protein-coding genes compared to the non-coding regions of the genome. Also, the distribution of X motifs in the three frames of the complete genome suggests the occurrence of X motifs preferentially in the reading frame regardless of length or cardinality. It was the first evidence of significant enrichment of X circular code motifs in the genes of an extant organism. Hence, two hypotheses were proposed: either the X motifs represent evolutionary remnants of a primitive code that was used for early translation systems or they still represent functional elements involved in the process of protein synthesis in extant organisms. During the elaboration of this thesis, the focus of our work was on investigating these questions.

2.9. Thesis outline

As mentioned above, we focus solely on the X circular code which is an error-correcting code that has the ability to retrieve, maintain and synchronize the reading frame in genes. The results/contributions are divided into different chapters. In chapter 3, we will demonstrate the evolutionary conservation of X circular code motifs in the protein-coding genes of eukaryotes by using multiple gene alignments and the possible functional role played by them. In chapter 4, we will discuss in detail the identification of X circular code motifs in the ribosomal RNA (rRNAs) of organisms from prokaryotes to eukaryotes; and the possible role played by circular codes in the evolution of the genetic code. In chapter 5, we discuss the optimality of circular codes (particularly the X circular code) to minimize the effects after frameshift errors in comparison to the optimality of the standard genetic code and among the combinatorial class of 216 maximal C^3 self-complementary trinucleotide circular codes. In chapter 6, we will present the conclusion and perspectives.

Chapter 3

3. Circular code motifs in eukaryotic genomes

3.1. Introduction

In evolutionary biology, a common assumption is that genetic elements that are conserved during evolution are a sign of natural selection, and that such elements are functional in some way. In the famous words of Theodosius Dobzhansky: “Nothing makes sense in biology, except in the light of evolution”.

To investigate whether motifs of the X circular code represent functional elements in genes, we therefore carried out an extensive study of the evolutionary conservation of X motifs in two independent sets of complete genomes. In this detailed analysis of X motifs we selected various extant organisms from two sets; the first set is made up of four highly evolved mammalian genomes, which are very closely related to one another, sharing a common ancestor nearly 300 million years ago. The second set is made up of nine yeast genomes, representing more divergent genome sequences of the simplest eukaryotes sharing a common ancestor nearly 1 billion years ago. Each set includes a well-studied and annotated ‘reference’ genome: the *Homo sapiens* genome (human) for the first set and the *Saccharomyces cerevisiae* genome (baker’s yeast) for the second set. For both sets of organisms we constructed multiple gene alignments of all the protein-coding genes in each reference genome. Multiple gene alignments are an important source for conducting comparative genomics, helping to reconstruct ancestral genomes or finding specific patterns of sequence conservation at the evolutionary level. This chapter is organized as follows. We start with the data used for constructing multiple gene alignments, followed by a detailed explanation and mathematical formulation of the methods. Then we move on to the results section where we highlight specific evolutionary pressures acting on the X motifs and identify important new properties of X motif conservation at the level of encoded amino acids. The results presented here are based on basic frequency statistics and their biological significance is clear. In order to evaluate the statistical significance of the different results presented in this study, we chose an approach that involved comparing the results obtained for the X motifs (Definition 2.10) with those obtained for R random motifs (Definition 2.12) generated by 100 (different) random codes R. This approach avoids the problems associated with defining statistical hypotheses about the nucleotide composition, the length and the random model of the different regions of the genome.

3.2. Gene Alignment Data

In this analysis, separate alignments for the two different sets of organisms are constructed to make the results highly relevant. As mentioned above, the first set of alignments is built from four highly evolved yet closely related mammal genomes. The second set of alignments is built from nine of the most simple yet divergent yeast genomes. We have used two different methods to construct the multiple alignment of gene sequences, thereby avoiding any bias towards a specific sequence alignment algorithm or evolutionary model. We obtained the high quality mammal gene alignments from a previous independent analysis of genome annotation methods (Sharma & Hiller, 2017). The multiple alignments of the yeast genes were constructed using the classical protein alignment method implemented in the ClustalW software (Thompson et al., 1994). Furthermore, well characterized, well annotated genomes (*H. sapiens* and *S. cerevisiae*) were chosen as reference genomes in the study to ensure high quality gene models. We provide the complete details of both constructed alignments.

3.2.1. Mammalian genome alignment

For the set of mammals, we extracted four well annotated genomes from an already available high quality multiple alignment for 144 mammals available in the UCSC site (https://bds.mpi-cbg.de/hillerlab/144VertebrateAlignment_CESAR/, Sharma & Hiller, 2017). In Table 3.1, we provide the scientific names of the genomes with the number of protein-coding genes used for the alignments and the complete length of the genes in nucleotides. For this set of organisms, *Homo sapiens* (*hg38*, $\mathbb{H}$) is taken as the reference genome. The number of genes present in the reference genome $\mathbb{H}$ is equal to 22,352 with a total length of 36,808,167 nucleotides. Each multiple alignment corresponds to one human gene sequence, aligned with the corresponding genes from the other three species (*tupBel1*, *mm10* and *canFam3*) when present. If a gene is absent in one of the organisms, the sequence is replaced by gaps in the multiple alignment.

Table 3.1. Details of the four mammal genomes used to construct the multiple gene alignments.

Genome name	Identification	Number of genes	Length of genes in nucleotides
Homo sapiens	*hg38* ($\mathbb{H}$)	22,352	36,808,167
Canis lupus familiaris	*canFam3*	21,137	34,379,490
Mus musculus	*mm10* ($\mathbb{M}$)	20,178	33,519,381
Tupaia belangeri	*tupBel1*	18,485	23,387,559

3.2.2. Yeast genome alignment

For the set of yeasts, the protein coding sequences of the nine different yeasts and the localisation of the corresponding nucleic acid sequence on the chromosomes were obtained from the NCBI Genbank (https://www.ncbi.nlm.nih.gov/genbank/). We then constructed the multiple gene alignments using classical methods. We provide the scientific names of the genomes with the number of protein-coding genes used for the alignments and the complete length of the genes in nucleotides in Table 3.2.

Table 3.2. Details of the nine yeast genomes used to construct the multiple gene alignments.

Genome name	Identification	Number of genes	Length of genes in nucleotides
Saccharomyces cerevisiae	*Sc* (ℂ)	6008	8,246,529
Kluyveromyces lactis	*Kl* (𝕃)	5085	7,729,998
Kuraishia capsulata	*Kc*	5989	6,911,424
Lodderomyces elongisporus	*Le*	5799	7,110,237
Meyerozyma guilliermondii	*Mg*	5920	6,633,972
Debaryomyces hansenii	*Dh*	6288	7,506,066
Scheffersomyces stipitis	*Ss*	5818	6,991,422
Schizosaccharomyces pombe	*Sp*	4980	5,614,506
Yarrowia lipolytica	*Yl*	6472	6,762,072

For this set of organisms, *Saccharomyces cerevisiae* (*Sc*, ℂ) is taken as the reference genome. The number of genes present in the reference genome ℂ is equal to 6008 with a total length of 8,246,529 nucleotides. To construct the multiple gene alignments, a BLAST database (Altschul et al., 1997) was created containing all the protein sequences of the nine organisms. For each of the protein sequences of the reference genome ℂ, a BLAST search in this database was performed. Then, the protein alignments containing 2 to 9 sequences were obtained using the ClustalW software (Thompson et al., 1994). By localization of each amino acid on the genome, the corresponding nucleic sequence alignments were created. Finally, for each nucleic sequence in the multiple alignments, we localized (localization was done without gaps) the different types of motifs in order to perform various statistical analyses. The BLAST searches, alignments and computations of the data were done using the in-house integrative software platform Gscope (R. Ripp, unpublished, details in Software development).

3.3. Software development

We identified all instances of the different classes of motifs used in the statistical analyses. For each nucleic sequence in the multiple alignments described above, the X motifs and R motifs (100 random codes; Appendix Table V) were localized in the genes using a program developed in the Java language (El Soufi & Michel, 2017). The program takes optional parameters that define the minimum cardinality c (in trinucleotides) and the minimum length l

(in trinucleotides) of the X and R motifs searched. For this study, we used cardinality $c \geq 4$ trinucleotides and length $l \geq c \geq 4$ trinucleotides for the X and R motifs. These localizations of X and R motifs are without gaps; but gaps may occur in the alignments which are inserted during the alignment process.

Gscope is an integrated platform allowing the analysis of all kinds of genomic data. It is written in Tcl/Tk and runs on all platforms. It is specially designed to perform high throughput analysis. Gscope is mainly composed of tools necessary to create the basic data, analysis tools, visualization interfaces. It allows also the creation and feeding of SQL relational databases and the querying and display of the available information through a web based interface (Wscope). To verify the localization of X and R motifs and to perform other computations, we also developed programs in Tcl/Tk using Gscope. We performed various statistical analyses on the motifs found after localization which we will explain in detail in the forthcoming sections.

3.4. Multiple gene alignments of mammal and yeast genomes

Here we introduce the multiple gene alignments of the four mammal genomes and nine yeast genomes used for this analysis. For both sets of organisms, we selected a reference genome, which is well known and annotated. We briefly recall the mathematical notations used.

Definition 3.1. By convention here, the *reference sequence* is the first sequence in the multiple gene alignment. The reference gene sequence $s_1 = \mathcal{R}$ is aligned with its $n-1$ orthologous genes denoted by $s_2, \ldots, s_n$, where $s_2, \ldots, s_n \in B^+$.

The length of the gene sequences $s_1, s_2 \ldots, s_n$ is denoted by $|s_1|, |s_2|, \ldots, |s_n|$ respectively.

These orthologous set of genes have originated from a common ancestor in the past and have diverged since then. The mammals share a common ancestor nearly 300 million years ago, while the yeasts share a common ancestor nearly 1 billion years ago. We have utilised the classical methods of multiple alignments; Definition 3.2 gives the mathematical formalism.

Definition 3.2. We define a multiple gene alignment $s_1, \ldots, s_n, n \geq 2$ as a mapping z on the alphabet $(B \cup \{\varepsilon\})^n \setminus (\{\varepsilon\})^n$ whose projection on the first component is s_1, on the second component is s_2, up to the projection on the nth component is s_n. A multiple gene alignment z of length l is denoted as:

$$z = \begin{pmatrix} \bar{M}_{11} & \cdots & \bar{M}_{l1} \\ \bar{M}_{12} & \cdots & \bar{M}_{l2} \\ \vdots & \vdots & \vdots \\ \bar{M}_{1n} & \cdots & \bar{M}_{ln} \end{pmatrix}$$

Where the reference genome sequence $\mathcal{R} = s_1 = \bar{M}_{11}, \ldots, \bar{M}_{l1}$, the second sequence of the alignment $s_2 = \bar{M}_{12}, \ldots, \bar{M}_{l2}$ up to the nth sequence $s_n = \bar{M}_{1n}, \ldots, \bar{M}_{ln}$, such that $\bar{M}_{ji} \in B \cup \{\varepsilon\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, l$, where ε being classically associated with the gap symbol "–" or ".". The following conditions are true for each multiple gene alignment:

a) An aligned tuple $(\bar{M}_{j1}, \ldots, \bar{M}_{ji}, \ldots \bar{M}_{jn})$ such that $\bar{M}_{j1}, \bar{M}_{ji} \in B$ with $\bar{M}_{j1} \neq \bar{M}_{ji}$ and $i \geq 2$ denotes the substitution of the jth nucleotide $\bar{M}_{ji}$ of $\mathcal{R}$ by the jth nucleotide $\bar{M}_{ji}$ of s_i.

b) An aligned tuple $(\bar{M}_{j1}, \ldots, \bar{M}_{ji}, \ldots \bar{M}_{jn})$ such that $\bar{M}_{j1} \in B$ and $\bar{M}_{ji} \in \{\varepsilon\}$ with $i \geq 2$ denotes the deletion of the jth nucleotide $\bar{M}_{j1}$ of $\mathcal{R}$.

c) An aligned tuple $(\bar{M}_{j1}, \ldots, \bar{M}_{ji}, \ldots \bar{M}_{jn})$ such that $\bar{M}_{j1} \in \{\varepsilon\}$ and $\bar{M}_{ji} \in B$ with $i \geq 2$ denotes the insertion of the jth nucleotide $\bar{M}_{ji}$ of s_i.

The X motifs $m(X)$, non-X motifs $m(\bar{X})$ and the random motifs $m(R)$ defined above may contain gaps in the multiple gene alignment, such that $m(X), m(\bar{X}), m(R) \in B \cup \{\varepsilon\}$.

The three classes of motifs located in the reference gene sequence $\mathcal{R}$ are denoted as $m(X, \mathcal{R}), m(\bar{X}, \mathcal{R}), m(R, \mathcal{R})$. Similarly, motifs located in any of the gene sequence s_i, where $i = 1, \ldots, n$ are denoted as $m(X, s_i), m(\bar{X}, s_i), m(R, s_i)$.

Next, we provide an example of a multiple gene alignment from our analysis, for better understanding.

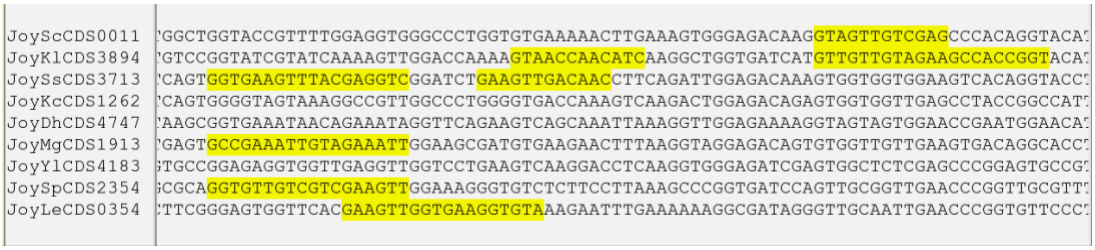


Figure 3.1. A part of the yeast multiple gene alignments. *Saccharomyces cerevisiae* (*Sc*, $\mathbb{C}$) is taken as the reference genome.

In Figure 3.1, we can observe that the X motifs (cardinality $4 \leq c \leq 20$ trinucleotides and length $l \geq c \geq 4$ trinucleotides) identified in the multiple gene alignment of yeasts are coloured in yellow. We have chosen strict conditions for the length and cardinality of the X motifs so that each motif is able to detect and maintain the reading frame of genes. The first sequence "JoyScCDS0011" ($s_1 = \mathcal{R} =$ *Saccharomyces Cerevisiae, Sc*), which is the reference genome in this multiple gene alignment contains one X motif $m(X, \mathcal{R})$ in the reading frame and the sequence without colour excluding the X motif are the non-X motifs $m(\bar{X}, \mathcal{R})$. Similarly, X motifs $m(X, s_i)$ found in the other yeasts genes are also shown coloured in yellow.

3.5. Codon substitution matrix of *X* motifs and random motifs

Here we will provide the mathematical formalism of the codon substitution matrices (for X motifs and R motifs), explaining their construction from the multiple gene alignments of mammal and yeast.

Definition 3.3. We define the codon substitution matrix $\mathbf{A}(m)$, $m = m(\mathfrak{I},\mathcal{R})$, where $\mathfrak{I} \in \{X, R\}$ denotes X motifs and R random motifs in the genes of a reference genome $R = s_1$ in the multiple gene alignment $s_1, s_2 \ldots, s_n$ of mammals and yeasts. The motifs m are based on 20 trinucleotides and each trinucleotide of m can be substituted into the 64 trinucleotides B^3. The codon substitution matrix $\mathbf{A}\big(m(\mathfrak{I},\mathcal{R})\big) = \big[a_{ij}\big]_{1\leq i\leq 64, 1\leq j\leq 20}$ has a size 64×20 (rectangular matrix) such that the 64 rows are associated with the 64 trinucleotides B^3 and the 20 columns are associated with the 20 trinucleotides of X or R (random codes). The matrix $\mathbf{A}(m(\mathfrak{I},\mathcal{R})) = \big[a_{ij}\big]_{1\leq i\leq 64, 1\leq j\leq 20}$ with element $a_{ij} = N(j \to i)$ in row i and column j refers to the number of substitutions of codon j of the motifs m (in the reference genome R) into the aligned codon i (codon $i \in B^3$) of the $n-1$ genomes $s_2 \ldots, s_n$.

Definition 3.4. We define the normalized matrix $\mathbf{B}(m(\mathfrak{I},\mathcal{R})) = \big[b_{ij}\big]_{1\leq i\leq 64, 1\leq j\leq 20}$ with element $b_{ij} = a_{ij}/\sum_{k=1}^{64} a_{kj}$ for $1 \leq i \leq 64$ and $1 \leq j \leq 20$ such that $\sum_{k=1}^{64} a_{kj} \neq 0$. Instead of a complete matrix normalization, the normalization per column allows the codons to be compared whatever may be the codon usage.

Remark 2. The diagonal elements a_{ii} of **A** and b_{ii} of **B** can be different from 0.

Remark 3. The 20 codon columns of **A** and **B** vary with each random code *R* (100 random codes); which is different from the 20 codon columns of the *X* circular code.

Next, we will explain the construction of these matrices with the help of examples. First we will construct these matrices by taking an example of a multiple gene alignment, and then with a part of the multiple gene alignment of mammals.

Example 1. We will provide an example of a gene alignment containing four genomes in Table 3.3. We will demonstrate the construction of the matrices $\mathbf{A}(m)$ and $\mathbf{B}(m)$ from the gene alignment in Table 3.3.

Table 3.3. Example of a gene alignment in a multiple global alignment containing four genomes, where *R* is the reference genome.

R	*GAG*	*GAC*	*ATC*	*CTG*	*GAC*	*CTG*	*AAC*	*CAG*
s_2	*GAC*	*GAC*	*ATC*	*CCA*	*GGC*	*CTG*	*AGT*	*CAG*
s_3	*GAA*	*GAC*	*ATC*	*CCG*	*GGC*	*CCA*	*CAT*	*CAC*
s_4	*GAG*	*GAC*	*ATC*	*CGG*	*GGC*	*CTG*	*AGC*	*CCG*

Example 2. Here we explain the construction of the codon matrices from the gene alignment in Table 3.3. The first trinucleotide column leads to the submatrix $\mathbf{A}(m)$ given in Table 3.4. The procedure is iterated for each trinucleotide column and leads to the submatrix $\mathbf{A}(m)$ given in Table 3.5.

Table 3.4. Codon substitution submatrix **A** of the first trinucleotide column from the example of gene alignment in Table 3.3.

A	*GAG*	
GAA	1	$GAG \longrightarrow GAA$
GAC	1	$GAG \longrightarrow GAC$
GAG	1	$GAG \longrightarrow GAG$

Table 3.5. Codon substitution matrix $\mathbf{A}(m)$ from the example of gene alignment in Table 3.3. The codon rows and column which are equal to 0 are not shown.

A	*AAC*	*ATC*	*CAG*	*CTG*	*GAC*	*GAG*	⋯
AGC	1						
AGT	1						
ATC		3					
CAC			1				
CAG			1				
CAT	1						
CCA				2			
CCG			1	1			
CGG				1			
CTG				2			
GAA						1	
GAC					3	1	
GAG						1	
GGC					3		
⋮							

Example 3. The codon substitution matrix $\mathbf{A}(m)$ given in Table 3.5 is normalized per column to obtain the normalized matrix $\mathbf{B}(m)$, which is given in Table 3.6.

Table 3.6. Normalized matrix $\mathbf{B}(\boldsymbol{m})$ from the example of gene alignment in Table 3.3. The codon rows and column which are equal to 0 are not shown.

B	*AAC*	*ATC*	*CAG*	*CTG*	*GAC*	*GAG*	⋯
AGC	1/3						
AGT	1/3						
ATC		1					
CAC			1/3				
CAG			1/3				
CAT	1/3						
CCA				1/3			
CCG			1/3	1/6			
CGG				1/6			
CTG				1/3			
GAA						1/3	
GAC					1/2	1/3	
GAG						1/3	
GGC					1/2		
⋮							
Sum	1	1	1	1	1	1	

Next, we will explain the construction of codon substitution matrices for *X* motifs identified in a part of the multiple gene alignment of mammals (Figure 3.2) with reference genome *hg38*.

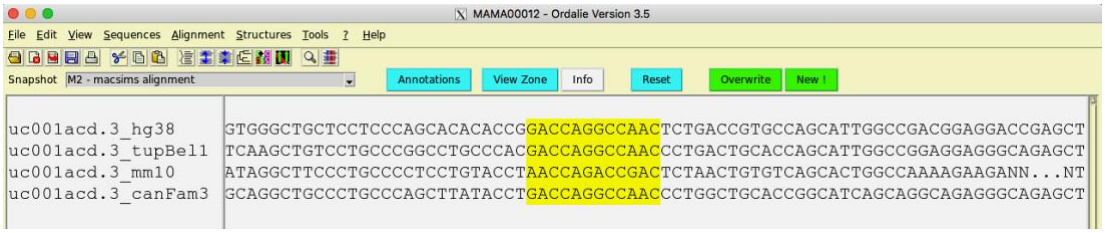


Figure 3.2. A part of the multiple gene alignment of four mammals, where the reference genome $\mathcal{R} = hg38$ and *X* motifs highlighted in yellow are perfectly aligned in all the four genomes.

Example 4. Here we demonstrate the construction of the matrices $\mathbf{A}(m(X, \mathcal{R} = hg38))$ and $\mathbf{B}(m(X, \mathcal{R} = hg38))$ for the *X* motifs in Figure 3.2. The gene alignment for the *X* motifs corresponding to the *X* motif in the human reference genome *hg38* is given in Table 3.7.The first trinucleotide column leads to the submatrix $\mathbf{A}(m)$ given in Table 3.8. The procedure is iterated for each trinucleotide column and leads to the codon substitution matrix $\mathbf{A}(m)$ given in Table 3.9. After normalization per column, the normalized matrix $\mathbf{B}(m)$ is given in Table 3.10.

Table 3.7. Gene alignment for the *X* motifs in Figure 3.2, where the reference genome $\mathcal{R} = hg38$ and *X* motifs are highlighted in yellow.

R	*GAC*	*CAG*	*GCC*	*AAC*
s_2	*GAC*	*CAG*	*GCC*	*AAC*
s_3	*AAC*	*CAG*	*ACC*	*GAC*
s_4	*GAC*	*CAG*	*GCC*	*AAC*

Table 3.8. Codon substitution submatrix **A** of the first trinucleotide column from the gene alignment in Table 3.7.

A	*GAC*	
GAC	1	$GAC \longrightarrow GAC$
AAC	1	$GAC \longrightarrow AAC$
GAC	1	$GAC \longrightarrow GAC$

Table 3.9. Codon substitution matrix **A** from the gene alignment in Table 3.7. The codon rows and column which are equal to 0 are not shown.

A	*AAC*	*CAG*	*GAC*	*GCC*	⋯
AAC	2		1		
ACC				1	
CAG		3			
GAC	1		2		
GCC				2	
⋮					

Table 3.10. Normalized matrix **B** from the gene alignment in Table 3.7. The codon rows and column which are equal to 0 are not shown.

B	*AAC*	*CAG*	*GAC*	*GCC*	⋯
AAC	2/3		1/3		
ACC				1/3	
CAG		1			
GAC	1/3		2/3		
GCC				2/3	
⋮					
Sum	1	1	1	1	

The normalized matrix $\mathbf{B}(m)$ for the R random motifs $m(R, \mathcal{R} = hg38)$ is constructed similarly for each of the 100 R random codes. The codon substitution matrix **A** and the normalized matrix **B** are given in the appendix (Table I and Table II).

We have followed the same procedure for all the X motifs identified in the multiple gene alignments of both mammals and yeasts separately. Similarly, codon matrices are calculated for the 100 random codes (Appendix Table V) for comparison.

3.6. Evolutionary conservation of *X* motifs in mammal and yeast genes

In this section, we discuss the various evolutionary constraints acting on the X motifs in the genes of mammals and yeasts. We performed various statistical analyses to verify the biological significance of X motifs compared to the R random motifs and the non-X motifs in the protein coding genes of mammals and yeasts. We first show the clear enrichment of X motifs

compared to the R random motifs. Then, we define various conservation parameters in the multiple gene alignments for both set of organisms.

3.6.1. Enrichment of X motifs in mammal and yeast genes

After the construction of multiple gene alignments for both sets of organisms, we calculated the enrichment of the X motifs in the reading frames of all genes and compared it with the enrichment for R random motifs. In Figure 3.3 for mammal genes and Figure 3.4 for yeast genes, we show a very high enrichment of X motifs compared to the R random motifs from the 100 random codes R.

The number of X motifs in the mammal genes is equal to 173390, whereas the mean number of R motifs is equal to 60330. The comparison of this mean number 60330 to 173390 leads to a Student one sided value $p \approx 10^{-82}$. The number of X motifs in yeast genes is equal to 35833, whereas the mean number of R motifs is equal to 15853. The comparison of this mean number 15853 to 35833 leads to a Student one sided value $p \approx 10^{-75}$.

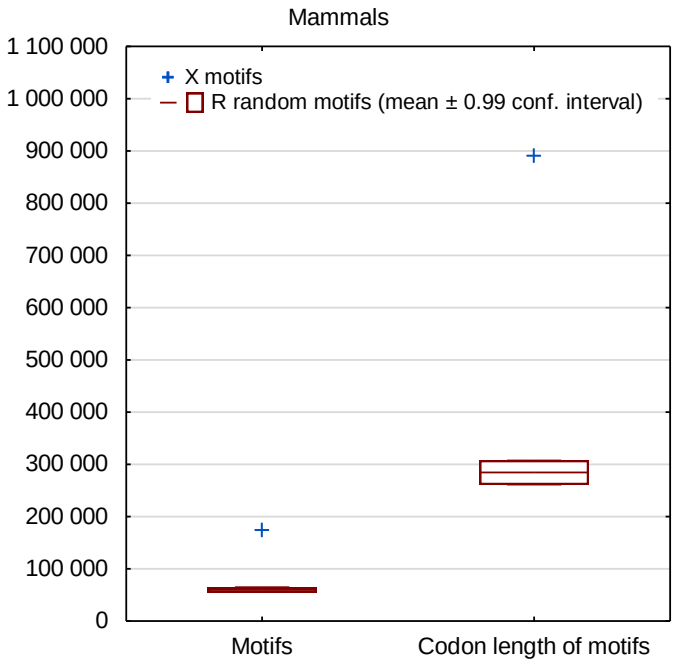


Figure 3.3. Comparison of the number of X and R random motifs and their codon length in the mammal genes. The number of X motifs is represented with a blue cross. The distribution of the R random motifs from the 100 random codes R is indicated by boxplots representing the mean and ±0.99 confidence interval.

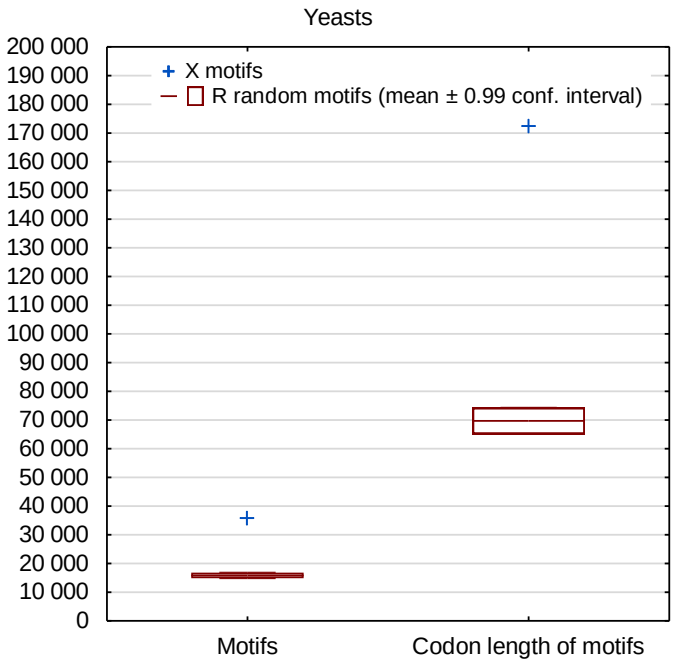


Figure 3.4. Comparison of the number of *X* and *R* random motifs and their codon length in the yeast genes. The number of *X* motifs is represented with a blue cross. The distribution of the *R* random motifs from the 100 random codes *R* is indicated by boxplots representing the mean and ±0.99 confidence interval.

This result is an additional and clear confirmation of the enrichment of *X* motifs in the genes which was previously identified in the yeast *Saccharomyces cerevisiae* (Michel et al., 2017).

3.6.2. Positional conservation

Next, we considered whether the position of the *X* motifs is preserved within the genes from different organisms. To check for the positional conservation of *X* motifs, we considered each column of a multiple gene alignment as one position (column with at least one sequence present). For each of these positions in the multiple gene alignments, we calculated the number of organisms with an *X* motif.

In Figure 3.5 and Figure 3.6, we show the calculation of the number of organisms with an *X* motif at a particular position (column). For the mammal gene alignment in Figure 3.5, the number of organisms having an *X* motif ranges from 0 to 4, as we have only four species in this set of alignments. For the yeast gene alignment in Figure 3.6, the number of organisms having an *X* motif ranges from 0 to 9, as we have nine species in this set of alignment. This number of *X* motifs present at a particular position in the alignment is then normalized by the number of species having at least one nucleotide at that position in the multiple gene alignment and not a gap.

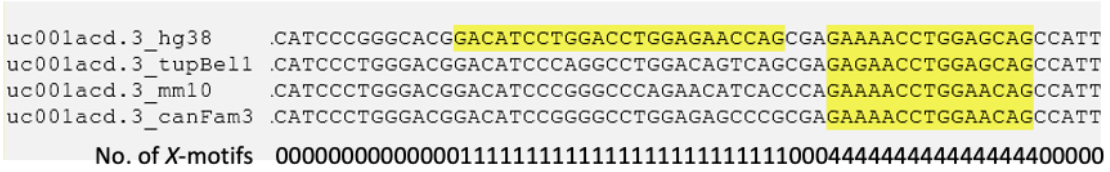


Figure 3.5. A part of the mammal gene alignment with X motifs highlighted in yellow. The number of X motifs is used in the calculation of the positional conservation parameter.

Figure 3.6. A part of the yeast gene alignment with X motifs highlighted in yellow. The number of X motifs is used in the calculation of the positional conservation parameter.

3.6.2.1. Positional conservation scores of X motifs and R random motifs

Here, we define a simple statistical parameter for analysing the positional conservation of all motifs $m(\mathfrak{I},\mathcal{R})$, $\mathfrak{I} \in \{X,R\}$ in the genes of a reference genome $\mathcal{R}$ in the multiple alignments of mammals and yeast. We give a mathematical formalism of the positional conservation score.

Definition 3.5. The positional conservation score $S_{pc}(m)$ of all motifs $m = m(\mathfrak{I},\mathcal{R})$, where $\mathfrak{I} \in \{X,R\}$ denotes X motifs and R random motifs, of letter lengths $|m|, m$ on the alphabet $B \cup \{\varepsilon\}$ (with gaps), in the genes of a reference genome $s_1 = \mathcal{R}$ in the multiple alignments of n genes $s_1, s_2 \ldots, s_n$ is equal to

$$S_{pc}(m) = S_{pc}(m(\mathfrak{I},\mathcal{R})) = \frac{1}{\sum_{m\in\mathcal{R}}|m|}\sum_{m\in\mathcal{R}}\sum_{j=1}^{|m|}\frac{1}{N_j}\sum_{i=1}^{n}\delta_{i,j}$$

where

$$\delta_{i,j} = \begin{cases} 1 \text{ if } L_{ji} \in B \text{ and } L_{ji} \in m(\mathfrak{I},s_i) \\ 0 \quad \text{otherwise} \end{cases},$$

N_j is the number of nucleotides without gaps at position j in the multiple gene alignments of n genes $s_1, s_2 \ldots, s_n$, $2 \le N_j \le n$ for $j = 1, \ldots, |m|$.

The condition used, $L_{ji} \in B$ and $L_{ji} \in m(\mathfrak{I},s_i)$ denotes that at the jth position of the gene sequence s_i, the letter L_{ji} is a nucleotide and not a gap, and belongs to a motif m.

Remark 4. The positional conservation score $S_{pc}(m) = S_{pc}(m(\mathfrak{I},\mathcal{R})) \in]0,1]$.

Remark 5. $S_{pc}(m) \approx 0$, when the motif m in the reference genome $\mathcal{R}$ is aligned with zero motifs in the other genomes of the multiple gene alignments. The positional conservation S_{pc} of the motif m is lowest in this case.

Remark 6. $S_{pc}(m) = 1$, when the motif m in the reference genome $\mathcal{R}$ is aligned with a motif in all of the other genomes of the multiple gene alignments. The positional conservation S_{pc} of the motif m is the highest in this case where all the genes in the multiple gene alignments have motifs in the same position as the reference genome but not a gap.

Remark 7. For a multiple alignment of n genes, the score $S_{pc}(m(\mathfrak{I},\mathcal{R}))$ takes $\left|(S_{pc}(m(\mathfrak{I},\mathcal{R}))\right| = \sum_{i=1}^{n} i - 1 = (n-1)(n+2)/2$ values.

Example 5. For a mammal alignment of $n = 4$ genomes, $\left|(S_{pc}(m(\mathfrak{I},\mathcal{R}))\right| = 9$ and for a yeast alignment of $n = 9$ genomes $\left|(S_{pc}(m(\mathfrak{I},\mathcal{R}))\right| = 44$.

For each set of multiple gene alignments (mammals and yeasts), Table 3.11 and Table 3.12 show the possible positional conservation scores at a particular position in the alignment depending on the number of organisms having a nucleotide and not a gap, and the number of organisms having an *X* motif at that particular position.

Table 3.11. Possible positional conservation scores at a particular position in the multiple gene alignments for mammals.

Number of nucleotides at position *p*	Number of *X* motifs at position *p*	Positional conservation score S_{pc}
4	4	1.00
3	3	1.00
2	2	1.00
4	3	0.75
3	2	0.67
4	2	0.50
2	1	0.50
3	1	0.33
4	1	0.25

Table 3.12. Possible positional conservation scores at a particular position in the multiple gene alignments for yeasts.

Number of nucleotides at position *p*	Number of *X* motifs at position *p*	Positional conservation score S_{pc}
9	9	1.00
8	8	1.00
7	7	1.00
6	6	1.00
5	5	1.00
4	4	1.00
3	3	1.00
2	2	1.00
9	8	0.89
8	7	0.88
7	6	0.86
6	5	0.83
5	4	0.80
9	7	0.78
8	6	0.75
4	3	0.75
7	5	0.71
9	6	0.67
6	4	0.67
3	2	0.67
8	5	0.63
5	3	0.60
7	4	0.57
9	5	0.56
8	4	0.50
6	3	0.50
4	2	0.50
2	1	0.50
9	4	0.44
7	3	0.43
5	2	0.40
8	3	0.38
9	3	0.33
6	2	0.33
3	1	0.33
7	2	0.29
8	2	0.25
4	1	0.25
9	2	0.22
5	1	0.20
6	1	0.17
7	1	0.14
8	1	0.13
9	1	0.11

Next, we will explain the calculation of the positional conservation score with the help of an example.

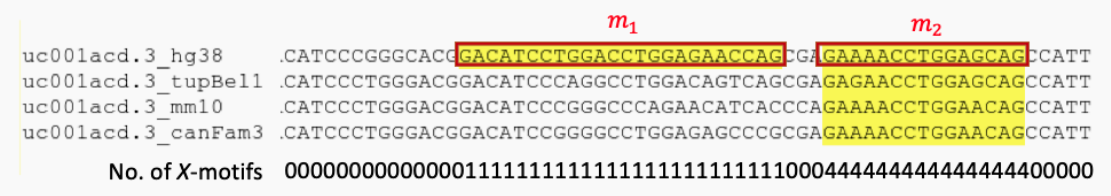


Figure 3.7. Calculation of positional conservation score (S_{pc}) for the X motifs m_1 and m_2 in Figure 3.5 showing a part of the multiple gene alignment for mammals.

Example 6. For the alignment shown in Figure 3.7, we will calculate the positional conservation score $S_{pc}(m(X,\mathcal{R}))$ (Definition 3.5) for the X motifs highlighted in yellow. We have two X motifs in the alignment. Let us denote the motifs $m_1 = GACATCCTGGACCTGGAGAACCAG$ and $m_2 = GAAAACCTGGAGCAG$. We observe that the motif m_1 in the reference genome *hg38* is aligned with zero motifs in the other genomes of the multiple gene alignment, whereas the motif m_2 in the reference genome *hg38* is aligned with motifs in all the other genomes of the multiple gene alignment.

As shown in the Figure 3.7, for each position (column) in the multiple gene alignment, we calculate the number of X motifs found in the other genomes that are aligned with the X motifs (here m_1 and m_2) identified in the reference genome *hg38*. We will calculate the positional conservation scores (Definition 3.5) for each motif m_1 and m_2:

$$S_{pc}\big(m_1(X,hg38)\big) = S_{pc}(GACATCCTGGACCTGGAGAACCAG)$$

$$= \frac{1}{\sum_{m_1 \in hg38}|m_1|}\sum_{m_1 \in hg38}\sum_{j=1}^{|m_1|}\frac{1}{N_j}\sum_{i=1}^{n}\delta_{i,j} = \frac{1}{24}\sum_{m_1 \in hg38}\sum_{j=1}^{24}\frac{1}{N_j}\sum_{i=1}^{4}\delta_{i,j}$$

$$= \frac{1}{24}\sum_{m_1 \in hg38}\sum_{j=1}^{24}\frac{1}{4}\big(\delta_{1,j}+\delta_{2,j}+\delta_{3,j}+\delta_{4,j}\big)$$

$$= \frac{1}{24}\Big\{\frac{1}{4}\big(\delta_{1,1}+\delta_{2,1}+\delta_{3,1}+\delta_{4,1}\big)+\frac{1}{4}\big(\delta_{1,2}+\delta_{2,2}+\delta_{3,2}+\delta_{4,2}\big)+\cdots$$

$$+\frac{1}{4}\big(\delta_{1,24}+\delta_{2,24}+\delta_{3,24}+\delta_{4,24}\big)\Big\}$$

$$= \frac{1}{24}\Big\{\frac{24}{4}(1+0+0+0)\Big\} = \frac{1}{4} = 0.25.$$

$$S_{pc}\big(m_2(X,hg38)\big) = S_{pc}(GAAAACCTGGAGCAG)$$

$$= \frac{1}{\sum_{m_2 \in hg38}|m_2|}\sum_{m_2 \in hg38}\sum_{j=1}^{|m_2|}\frac{1}{N_j}\sum_{i=1}^{n}\delta_{i,j} = \frac{1}{15}\sum_{m_2 \in hg38}\sum_{j=1}^{15}\frac{1}{N_j}\sum_{i=1}^{4}\delta_{i,j}$$

$$= \frac{1}{15}\sum_{m_2 \in hg38}\sum_{j=1}^{15}\frac{1}{4}\big(\delta_{1,j}+\delta_{2,j}+\delta_{3,j}+\delta_{4,j}\big)$$

$$= \frac{1}{15}\left\{\frac{1}{4}\left(\delta_{1,1} + \delta_{2,1} + \delta_{3,1} + \delta_{4,1}\right) + \frac{1}{4}\left(\delta_{1,2} + \delta_{2,2} + \delta_{3,2} + \delta_{4,2}\right) + \cdots \right.$$
$$\left. + \frac{1}{4}\left(\delta_{1,15} + \delta_{2,15} + \delta_{3,15} + \delta_{4,15}\right)\right\}$$
$$= \frac{1}{15}\left\{\frac{15}{4}(1 + 1 + 1 + 1)\right\} = 1.$$

The positional conservation score ($S_{pc} = 0.25$) of the motif m_1 is lowest in this multiple gene alignment, as it is not aligned with any other motifs from other genomes. Whereas in the case of motif m_2, the positional conservation score ($S_{pc} = 1$) is the highest in this multiple gene alignment as it is aligned with a motif in each of the other genomes for all the positions.

This parameter allows us to evaluate the presence of X motifs at a particular position throughout the multiple gene alignments. We calculated the positional conservation scores for each position of the multiple alignments of genes (mammals and yeasts separately). Next, we will evaluate the conservation of X motifs on the basis of positional conservation scores.

3.6.2.2. Positional conservation of *X* motifs in mammal and yeast genes

We calculated the positional conservation score S_{pc} (Definition 3.5) for all the motifs $m(\mathfrak{I},\mathcal{R}), \mathfrak{I} \in \{X, R\}$ in the multiple gene alignments, corresponding to the 22,352 mammal genes and the 6008 yeast genes. In order to evaluate the biological significance of the positional conservation of the X motifs found in the reference genome we compared the results with those obtained from R random motifs (100 R random codes generated) for both sets of alignments.

The positional conservation scores with their respective probabilities are shown in Figure 3.8 and Figure 3.9 for mammals and yeasts, respectively. For both mammals and yeasts, the number of X motifs with the highest positional conservation score ($S_{pc} = 1$) was higher when compared to the number of R motifs. In contrast, the number of X motifs with the lowest positional conservation score ($S_{pc} < 0.25$) was much lower when compared to the number of R motifs. A one sample Wilcoxon signed rank test indicated that the X motifs and the R motifs have significantly different medians with two-sided values $p = 0.031$ for the mammals and $p = 0.016$ for the yeasts. We observe that, the positional conservation probabilities (hence number) are greater in the case of X motifs than R motifs, for higher positional conservation scores.

The results obtained here show that the enrichment of X motifs in protein-coding genes is not at all random. Rather, X motifs are more likely to be conserved in the same position in orthologous genes. We can also observe that, the positional conservation probabilities are higher in the case of mammals than the yeasts. This is due to the diversity of the yeast genomes in the

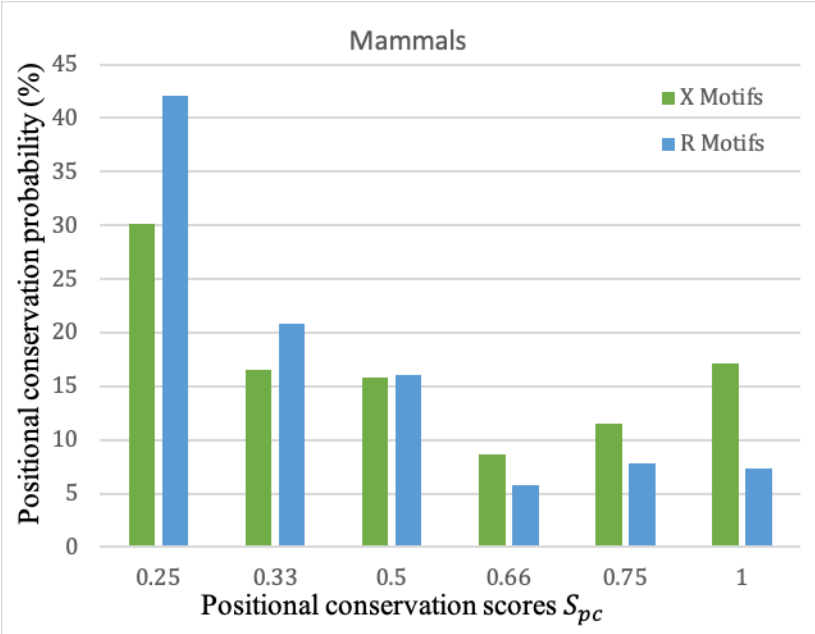


Figure 3.8. Positional conservation probability (%) of *X* motifs and *R* random motifs in the mammal multiple gene alignments with respect to the human reference genome (*hg38*) varying from 0 (no conservation in the alignment) to 1 (highest conservation in the alignment).

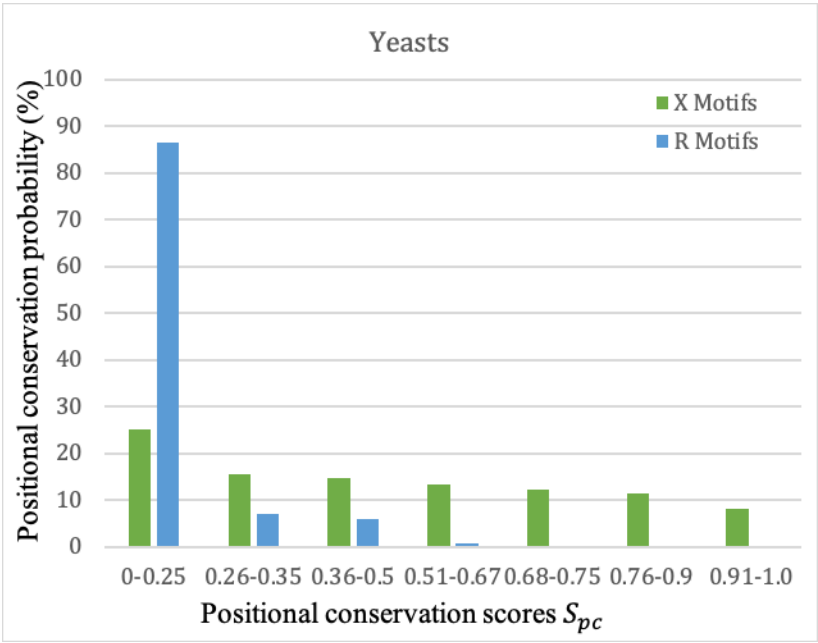


Figure 3.9. Positional conservation probability (%) of *X* motifs and *R* random motifs in the yeast multiple gene alignments with respect to the *Saccharomyces cerevisiae* reference genome varying from 0 (no conservation in the alignment) to 1 (highest conservation in the alignment).

multiple gene alignment, since yeasts diverged much earlier than mammals. To recall, yeasts shared a common ancestor nearly 1 billion years ago, whereas mammals shared a common ancestor nearly 300 million years ago. Therefore, we observe more evolutionary diversity in the yeast genomic sequences than the mammals.

3.6.3. Sequence conservation

In the previous section, we showed that the positions of the X motifs in genes from mammals and yeasts are significantly more conserved than would be expected by chance. To investigate whether this positional conservation was due to a high conservation of the nucleotide sequences, we investigated the level of sequence conservation of X motifs from pairwise gene alignments. We computed various pairwise alignment parameters (defined later in this section). We also computed a particularly useful statistic known as dN/dS ratio (Spielman & Wilke, 2015), which is the ratio of non-synonymous to synonymous substitutions (non-synonymous substitutions are nucleotide changes that alter the amino acid sequence, whereas synonymous substitutions do not). This ratio is used to infer purifying selection ($dN/dS < 1$, deleterious in nature), neutral selection ($dN/dS \approx 1$) or diversifying selection ($dN/dS > 1$, advantageous in nature). The genetic code indicates that almost all substitutions at the second nucleotide position of codons result in amino acid replacement whereas a fraction of the nucleotide changes at the first and third positions are synonymous. Therefore, we were motivated to check for selection pressures (if any) for the X motifs identified in the multiple gene alignments. Before going into the results, we will explain various parameters used in the analysis with the help of examples.

3.6.3.1. Pairwise alignment of *X* motifs and non-*X* motifs

A pairwise alignment is a multiple gene alignment according to Definition 3.2 with $n = 2$ sequences of letter length l such that the nucleotides belonging to the two genome sequences are with gaps, i.e. $N \in B \cup \{\varepsilon\}$. We will describe some classical parameters for pairwise alignments that are used to estimate the conservation of the motifs.

Definition 3.6. A pairwise alignment is a multiple gene alignment z of two sequences $s_1 = N_1 N_2 \cdots N_l$ and $s_2 = N'_1 N'_2 \cdots N'_l$, where $N, N' \in B \cup \{\varepsilon\}$ (with gaps), of nucleotide letter length l.

Definition 3.7. The percentage identity $Pid(m)$ of identical nucleotides of all motifs $m = m(\mathfrak{I}, \mathcal{R})$, where $\mathfrak{I} \in \{X, \bar{X}\}$ denotes X motifs and non-X motifs, of letter lengths $|m|, m$ on the alphabet $B \cup \{\varepsilon\}$ (with gaps), in the genes of a reference genome $s_1 = \mathcal{R}$ in all the pairwise gene alignments of s_1 and s_2 is equal to

$$Pid(m) = Pid\big(m(\mathfrak{I}, \mathcal{R})\big) = \frac{1}{\sum_{m \in \mathcal{R}} |m|} \sum_{m \in \mathcal{R}} \sum_{i=1}^{|m|} \delta_i$$

where $1 \leq i \leq |m|$ and operator δ_i associated with a pair of nucleotide letters N is defined by

$$\delta_i = \begin{cases} 1 \text{ if } N_{i1} \in B \text{ and } N_{i1} = N_{i2} \\ 0 \text{ otherwise} \end{cases}.$$

Definition 3.8. Let $f_i(c)$ and $g_i(c)$ be the fraction of synonymous and non-synonymous substitutions respectively, at the ith site of a given codon $c = N_1N_2N_3$, with $i = 1, 2, 3$.
The number $Ns(c)$ of synonymous sites and the number $Nns(c)$ of non-synonymous sites for a given codon c (Nei & Gojobori, 1986), are defined as:

$$Ns(c) = \sum_{i=1}^{3} f_i(c), \text{ and}$$
$$Nns(c) = \sum_{i=1}^{3} g_i(c) = \sum_{i=1}^{3} \big(1 - f_i(c)\big) = 3 - Ns(c).$$

The definitions of $Ns(c)$ and $Nns(c)$ for a given codon are naturally extended to a series of codons or a motif m.

Definition 3.9. The expected numbers $Ns(m)$ of synonymous substitutions and $Nns(m)$ of non-synonymous substitutions for a motif $m = m(\mathfrak{I}, \mathcal{R})$, where $\mathfrak{I} \in \{X, \bar{X}\}$ denotes X motifs and non-X motifs, of letter lengths $|m|, m$ on the alphabet B (without gaps), are equal to:

$$Ns(m) = \sum_{c \in m} Ns(c) \text{ and } Nns(m) = |m| - Ns(m),$$

where $Ns(c)$ is defined in Definition 3.8. $Ns(m)$ and $Nns(m)$, the expected number of synonymous and non-synonymous substitutions respectively for all motifs m in the reference genome sequence ($s_1 = \mathcal{R}$) is computed for the pairwise alignments of s_1 and s_2.

Definition 3.10. Let $Os(m)$ and $Ons(m)$ be the observed numbers of synonymous and non-synonymous substitutions respectively of a motif $m = m(\mathfrak{I}, \mathcal{R})$ in the reference genome $s_1 = \mathcal{R}$, where $\mathfrak{I} \in \{X, \bar{X}\}$ denotes X motifs and non-X motifs, in the motif m' in the gene sequence s_2 of letter lengths $|m| = |m'|, m, m'$ on the alphabet B (without gaps) in all the gene pairwise alignments s_1 and s_2.

Remark 8. $Os(m) + Ons(m) = |m| - \sum_{i=1}^{|m|} \delta_i$, where δ_i is defined in Definition 3.7.

Definition 3.11. The ratio $Rs(m)$ of synonymous substitutions and $Rns(m)$ of non-synonymous substitutions of a motif $m = m(\mathfrak{I}, \mathcal{R})$ in the reference genome $s_1 = \mathcal{R}$, where $\mathfrak{I} \in \{X, \bar{X}\}$ denotes X motifs and non-X motifs, in the motif m' in the gene sequence s_2 of letter lengths $|m| = |m'|, m, m'$ on the alphabet B (without gaps) in all the gene pairwise alignments s_1 and s_2 is defined as

$$Rs(m) = Os(m)/Ns(m) \text{ and } Rns(m) = Ons(m)/Nns(m),$$

where $Os(m)$ and $Ons(m)$ are defined in Definition 3.10, and $Ns(m)$ and $Nns(m)$ in Definition 3.9.

We will now explain the parameters defined above with the help of examples.

In Figure 3.10, the X motifs (Definition 2.10) located in the reading frame of genes are highlighted in yellow, the rest of the sequence is considered to be non-X motifs (Definition 2.11). The reference gene sequence $s_1 = \mathcal{R}$ ($\mathbb{H}$, *Homo sapiens*, $hg38$) contains two X motifs $m(X, \mathbb{H})$ (Table 3.13) and three non-X motifs $m(\bar{X}, \mathbb{H})$ (Table 3.14). The 2nd (*Tupaia belangeri*, *tupBel1*), 3rd ($\mathbb{M}$, *Mus musculus*, *mm10*) and 4th (*Canis lupus familiaris*, *canFam3*) gene sequences each contain an X motif $m(X)$ and two non-X motifs $m(\bar{X})$ in the reading frame. We used s_1 ($\mathbb{H}$, *Homo sapiens*, $hg38$) and s_3 ($\mathbb{M}$, *Mus musculus*, *mm10*) for the pairwise gene alignment.

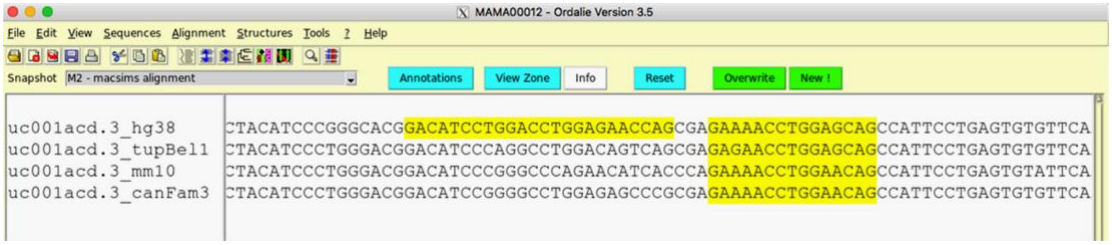


Figure 3.10. A part of multiple gene alignment for mammals with X motifs highlighted in yellow.

Table 3.13. The pairwise gene alignment of the two X motifs $m(X, \mathbb{H})$ of total length l=39 nucleotides in the reference genome $\mathbb{H}$ (*hg38*) and the genome $\mathbb{M}$ (*mm10*) from Figure 3.10.

Reference gene $\mathbb{H}$	1st X motif $m(X, \mathbb{H})$								2nd X motif $m(X, \mathbb{H})$				
Protein $g(s_1)$ of $\mathbb{H}$	*D*	*I*	*L*	*D*	*L*	*E*	*N*	*Q*	*E*	*N*	*L*	*E*	*Q*
Gene s_1 of $\mathbb{H}$	*GAC*	*ATC*	*CTG*	*GAC*	*CTG*	*GAG*	*AAC*	*CAG*	*GAA*	*AAC*	*CTG*	*GAG*	*CAG*
Gene s_2 of $\mathbb{M}$	*GAC*	*ATC*	*CCG*	*GGC*	*CCA*	*GAA*	*CAT*	*CAC*	*GAA*	*AAC*	*CTG*	*GAA*	*CAG*
Protein $g(s_2)$ of $\mathbb{M}$	*D*	*I*	*P*	*G*	*P*	*E*	*H*	*H*	*E*	*N*	*L*	*E*	*Q*

In the pairwise gene alignment of the two X motifs located in the reference genome $\mathbb{H}$ (*hg38*) and the genome $\mathbb{M}$ (*mm10*) given in Table 3.13, there are 30 identical pairs of nucleotides, and 9 different pairs of nucleotides (underlined). The protein alignment associated with the pairwise gene alignment is given by applying the universal genetic code map g (Definition 2.6) to each trinucleotide/codon in the reading frame of the gene.

Table 3.14. The pairwise gene alignment of the three non-X motifs $m(\bar{X})$ of total length l=36 nucleotides in the reference genome $\mathbb{H}$ (*hg38*) and the genome $\mathbb{M}$ (*mm10*) from Figure 3.10.

Reference gene $\mathbb{H}$	1st non-X motif $m(\bar{X})$					2nd $m(\bar{X})$	3rd non-X motif $m(\bar{X})$					
Protein $g(s_1)$ of $\mathbb{H}$	*Y*	*I*	*P*	*G*	*T*	*R*	*P*	*F*	*L*	*S*	*V*	*F*
Gene s_1 of $\mathbb{H}$	*TAC*	*ATC*	*CCG*	*GGC*	*ACG*	*CGA*	*CCA*	*TTC*	*CTG*	*AGT*	*GTG*	*TTC*
Gene s_2 of $\mathbb{M}$	*TAC*	*ATC*	*CC<u>T</u>*	*GG<u>G</u>*	*ACG*	*C<u>C</u>A*	*CCA*	*TTC*	*CTG*	*AGT*	*GT<u>A</u>*	*TTC*
Protein $g(s_2)$ of $\mathbb{M}$	*Y*	*I*	*P*	*G*	*T*	*P*	*P*	*F*	*L*	*S*	*V*	*F*

In the pairwise gene alignment of the three non-X motifs $m(\bar{X})$ in the reference genome $\mathbb{H}$ (*hg38*) and the genome $\mathbb{M}$ (*mm10*) given in Table 3.14, there are 32 identical pairs of nucleotides and 4 different pairs of nucleotides (underlined). The protein alignment associated with the pairwise gene alignment is given by applying the universal genetic code map g (Definition 2.6) to each trinucleotide/codon in the reading frame of the gene.

Example 7. The percentage identity (Definition 3.7) from the pairwise alignment of X motifs in Table 3.13, $Pid(m(X,\mathbb{H})) = 30/39 = 76.92\ \%$.

Example 8. We will calculate the number of synonymous ($Ns(c)$) and non-synonymous ($Nns(c)$) sites for the codon $c = CTG$. The codon CTG codes for the amino acid $g(c) = Leu$ according to the standard genetic code.

Therefore, $f_1(CTG) = \frac{1}{3}$ as only the 1st site substitution $CTG \longrightarrow TTG$ is synonymous out of *ATG*, *TTG* and *GTG*, $f_2(CTG) = 0$ as there is no 2nd site synonymous substitution out of *CAG*, *CCG* and *CGG*, and $f_3(CTG) = \frac{3}{3} = 1$ as all the 3rd site substitutions are synonymous out of *CTA*, *CTC* and *CTT*. So, $Ns(CTG) = \frac{1}{3} + 0 + 1 = \frac{4}{3}$ and $Nns(CTG) = \left(3 - \frac{4}{3}\right) = \frac{5}{3}$.

We then calculate $Ns(c)$ and $Nns(c)$ for each of the codon from the pairwise alignment of X motifs in Table 3.13.

Example 9. We will calculate the expected numbers $Ns(m)$ and $Nns(m)$ of synonymous and non-synonymous sites respectively for the two X motifs from the pairwise alignment in Table 3.13.

The total length of the motifs $m(X,\mathbb{H})$ in the reference genome $\mathbb{H}$ is l=39 nucleotides or 13 codons. So, the expected numbers $Ns(m)$ and $Nns(m)$ for the two X motifs $m(X,\mathbb{H})$ are:

$$\begin{aligned} Ns(m) = {} & Ns(GAC) + Ns(ATC) + Ns(CTG) + Ns(GAC) + Ns(CTG) + Ns(GAG) \\ & + Ns(AAC) + Ns(CAG) + Ns(GAA) + Ns(AAC) + Ns(CTG) + Ns(GAG) \\ & + Ns(CAG) \\ = {} & \frac{1}{3} + \frac{2}{3} + \frac{4}{3} + \frac{1}{3} + \frac{4}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{1}{3} + \frac{4}{3} + \frac{1}{3} + \frac{1}{3} = \frac{23}{3} \approx 7.67 \end{aligned}$$

and $Nns(m) = \left(39 - \frac{23}{3}\right) = \frac{94}{3} \approx 31.33$.

Example 10. We will calculate $Os(m)$ and $Ons(m)$, the observed numbers of synonymous and non-synonymous substitutions respectively of the two X motifs $m(X, \mathbb{H})$ from the pairwise alignment in Table 3.13. We have four synonymous substitutions:

- $CTG\ (L) \longrightarrow CCA\ (P)$ on the 3rd site
- $GAG\ (E) \longrightarrow GAA\ (E)$
- $AAC\ (N) \longrightarrow CAT\ (H)$ on the 3rd site
- $GAG\ (E) \longrightarrow GAA\ (E)$

So, $Os(m(X, \mathbb{H})) = 4$ and we have five non-synonymous substitutions:

- $CTG\ (L) \longrightarrow CCG\ (P)$
- $GAC\ (D) \longrightarrow GGC\ (G)$
- $CTG\ (L) \longrightarrow CCA\ (P)$ on the 2nd site
- $AAC\ (N) \longrightarrow CAT\ (H)$ on the 1st site
- $CAG\ (Q) \longrightarrow CAC\ (H)$

So, $Ons(m(X, \mathbb{H})) = 5$.

Example 11. We will then calculate the ratio $Rs(m)$ of synonymous substitutions and $Rns(m)$ of non-synonymous substitutions for the two X motifs $m(X, \mathbb{H})$ from the pairwise alignment given in Table 3.13. So, $Rs(m(X, \mathbb{H})) = \frac{4}{\frac{23}{3}} = \frac{12}{23} \approx 0.52$ and $Rns(m(X, \mathbb{H})) = \frac{5}{\frac{94}{3}} = \frac{15}{94} \approx 0.16$.

Example 12. Similarly, we calculate the various parameters for the three non-X motifs $m(\bar{X})$ of total length l=36 nucleotides from the pairwise alignment given in Table 3.14.

- $Pid(m(\bar{X}, \mathbb{H})) = 32/36 = 88.89\ \%$
- $Ns(m(\bar{X}, \mathbb{H})) = \frac{1}{3} + \frac{2}{3} + \frac{3}{3} + \frac{3}{3} + \frac{3}{3} + \frac{4}{3} + \frac{3}{3} + \frac{1}{3} + \frac{4}{3} + \frac{1}{3} + \frac{3}{3} + \frac{1}{3} = \frac{29}{3} \approx 9.67$,
  $Nns(m(\bar{X}, \mathbb{H})) = \left(36 - \frac{29}{3}\right) = \frac{79}{3} \approx 26.33$
- $Os(m(\bar{X}, \mathbb{H})) = 3$ (three synonymous substitutions: $CCG\ (P) \longrightarrow CCT\ (P)$, $GGC\ (G) \longrightarrow GGG\ (G)$ and $GTG\ (V) \longrightarrow GTA\ (V)$)
  $Ons(m(\bar{X}, \mathbb{H})) = 1$ (one non-synonymous substitution: $CGA\ (R) \longrightarrow CCA\ (P)$),
- $Rs(m(\bar{X}, \mathbb{H})) = \frac{3}{\frac{29}{3}} = \frac{9}{29} \approx 0.31$ and $Rns(m(\bar{X}, \mathbb{H})) = \frac{1}{\frac{79}{3}} = \frac{3}{79} \approx 0.04$.

Next, we will discuss the results obtained from the pairwise alignments of mammals and yeasts.

3.6.3.2. Sequence conservation of *X* motifs in mammal and yeast genes

We defined various pairwise alignment parameters above. For the pairwise alignments, we selected two organisms from each of the mammal and yeast gene sets. For the set of mammals, 14,681 pairwise alignments of human ($\mathbb{H}$) and mouse ($\mathbb{M}$) genes were used, whereas for the yeasts, 1088 pairwise alignments of *S. cerevisiae* ($\mathbb{C}$) and *K. lactis* ($\mathbb{L}$) genes were used. The Pid observed in X motifs was 87.44% for $\mathbb{H}$ - $\mathbb{M}$ alignments, and 59.88% for $\mathbb{C}$ - $\mathbb{L}$ alignments. In comparison, The Pid observed in non-X motifs was 77.56% for $\mathbb{H}$ - $\mathbb{M}$ alignments, and 53.94% for $\mathbb{C}$ - $\mathbb{L}$ alignments. For the 14,681 $\mathbb{H}$ - $\mathbb{M}$ alignments, a χ^2 test showed a highly significant difference between the Pid observed in X motifs and non-X motifs with one-sided value $p \approx 10^{-110}$; whereas for the 1088 $\mathbb{C}$ - $\mathbb{L}$ alignments, a χ^2 test showed a significant difference between the Pid observed in X motifs and non-X motifs with one-sided value $p \approx 0.005$. Therefore after examining the results from the pairwise alignments of protein-coding genes, we can say that X motif sequences are generally more conserved in terms of nucleic acids than non-X motif sequences.

The increased sequence conservation of X motifs suggests that they are preserved in the process of natural selection, and may represent variations in the strengths of positive selection or purifying selection. To better understand the relative contributions of these selection modes, we calculated the ratio of non-synonymous to synonymous substitutions (dN/dS). This ratio is commonly used to infer purifying ($dN/dS < 1$) or positive selection ($dN/dS > 1$) in the protein-coding genes. It is important to note that a non-synonymous substitution implies a change in the amino acid in the translated protein sequence, whereas a synonymous substitution only changes the codon: another codon coding for the same amino acid replaces the original codon. In this work, the ratio of non-synonymous to synonymous substitutions (dN/dS) is defined as

$$\frac{dN}{dS} = \frac{Rns(m)}{Rs(m)},$$

where $Rns(m)$ and $Rs(m)$ are defined in Definition 3.11, $m = m(\mathfrak{I}, \mathcal{R})$, where $\mathfrak{I} \in \{X, \bar{X}\}$ denotes X motifs and non-X motifs.

Table 3.15. Comparison of non-synonymous and synonymous substitutions for X motifs and non-X motifs in pairs of aligned genes for Human and Mouse ($\mathbb{H}$ and $\mathbb{M}$).

$\mathbb{H}$-$\mathbb{M}$ alignment	Nns	Ns	Ons	Os	Rns	Rs	dN/dS
X motifs	1,611,224	480,358	99,670	184,643	0.06	0.38	0.16
$non-X$ motifs	19,772,931	8,225,136	1,524,889	2,572,797	0.08	0.31	0.25

Table 3.16. Comparison of non-synonymous and synonymous substitutions for X motifs and non-X motifs in pairs of aligned genes for *S. cerevisiae* and *K. lactis* ($\mathbb{C}$ and $\mathbb{L}$).

$\mathbb{C}$ - $\mathbb{L}$ alignment	Nns	Ns	Ons	Os	Rns	Rs	dN/dS
X motifs	369,426	93,981	103,766	80,081	0.28	0.85	0.33
$non-X$ motifs	5,310,908	1,973,266	1,580,781	1,362,399	0.30	0.69	0.43

We compared the results obtained for the two sets of pairwise gene alignments $\mathbb{H}$ and $\mathbb{M}$ for the mammals, and $\mathbb{C}$ and $\mathbb{L}$ for the yeasts. From Table 3.15 and Table 3.16, we observe that the rates of non-synonymous and synonymous substitutions (Rns and Rs respectively) for $\mathbb{H}$ - $\mathbb{M}$ pairwise alignments is less than that obtained from $\mathbb{C}$ - $\mathbb{L}$ pairwise alignments. Due to the smaller phylogenetic distance between $\mathbb{H}$ and $\mathbb{M}$ than $\mathbb{C}$ and $\mathbb{L}$, the values of Rns and Rs are lower for both X motifs and non-X motifs. In other words, as $\mathbb{H}$ and $\mathbb{M}$ are more closely related than $\mathbb{C}$ and $\mathbb{L}$ phylogenetically, we would expect less substitutions, both synonymous and non-synonymous. Also, the values of Rns are lower than Rs for X motifs and non-X motifs in both sets of genes; as non-synonymous substitutions have a greater impact on the translated protein and occur less frequently than synonymous substitutions.

Importantly, we observed significantly lower ratios of non-synonymous to synonymous substitutions (dN/dS) in X motifs than in non-X motifs. Notably, X motifs show more synonymous substitutions than non-synonymous substitutions compared to non-X motifs, indicating more evolutionary constraints on X motifs. When both classes of substitutions are neutral, i.e. are not affected by selection, then a dN/dS ratio of 1 is expected. In contrast, if mutations that replace the amino acid are selected against and only rarely are allowed to become fixed, while synonymous substitutions are effectively neutral, then a lower dN/dS ratio is expected. This analysis on synonymous and non-synonymous substitutions motivated us to carry out an analysis to evaluate the amino acid conservation in the motifs identified in the multiple gene alignments, which we will discuss next.

3.6.4. Amino acid conservation

In the previous sections, we checked for positional and sequence conservation of X motifs. Here we will check for the amino acid conservation of X motifs and R random motifs. We will first define mathematically the statistical parameter used for evaluating the amino acid conservation of motifs. Then, we provide the results for the conservation of X codons per amino acid p (peptide) coded by all motifs $m = (m(\mathfrak{I}, \mathcal{R}))$, where $\mathfrak{I} \in \{X, R\}$ denotes X motifs and R random motifs. In a multiple alignment consisting of orthologous gene sequences, our goal is to evaluate whether the codons in X motifs are conserved to allow synonymous substitutions and if they evolved under the selective pressure on the amino acids they encode.

3.6.4.1. Amino acid conservation parameter of *X* motifs and random motifs

We define a simple statistical parameter for evaluating the conservation of X motifs and R random motifs ($m(\mathfrak{I},\mathcal{R}), \mathfrak{I} \in \{X,R\}$) for the 12 amino acids $\mathcal{X}$ (2) coded by the X circular code, in the genes of a reference genome $s_1 = \mathcal{R}$ in the multiple gene alignment $s_1, s_2 \ldots, s_n$ of mammals and yeasts.

Definition 3.12. The percentage $Paac(m(\mathfrak{I},\mathcal{R}),p)$ of conservation of X codons per amino acid p (peptide) coded by all motifs $m = (m(\mathfrak{I},\mathcal{R}))$, where $\mathfrak{I} \in \{X,R\}$ denotes X motifs and R random motifs, in the genes of a reference genome $s_1 = \mathcal{R}$ in all the multiple gene alignments $s_1, s_2 \ldots, s_n$, is equal to

$$Paac(m(\mathfrak{I},\mathcal{R}),p) = \frac{1}{Card(\mathcal{g}^{-1}(p) \cap \mathfrak{I})} \sum_{\substack{i,j \in \mathcal{g}^{-1}(p) \\ i,j \in \mathfrak{I}}} b_{ij}\big(m(\mathfrak{I},\mathcal{R})\big)$$

where $p \in \mathcal{X} = \{A,N,D,Q,E,G,I,L,F,T,Y,V\}$ (2), $b_{ij}\big(m(\mathfrak{I},\mathcal{R})\big)$ is the element of the normalized matrix $\boldsymbol{B}$ defined in Definition 3.4 and the inverse map $\mathcal{g}^{-1}$ in Definition 2.6.

Definition 3.13. The mean percentage $\bar{P}aac(m(\mathfrak{I},\mathcal{R}),\mathcal{X})$ of conservation of X codons in the 12 amino acids $\mathcal{X}$ (2) coded by all the motifs $m(\mathfrak{I},\mathcal{R})$, where $\mathfrak{I} \in \{X,R\}$ denotes X motifs and R random motifs, in the genes of a reference genome $s_1 = \mathcal{R}$ in all the multiple gene alignments $s_1, s_2 \ldots, s_n$, is equal to

$$\bar{P}aac(m(\mathfrak{I},\mathcal{R}),\mathcal{X}) = \frac{1}{Card(\mathcal{X})} \sum_{p \in \mathcal{X}} Paac(m(\mathfrak{I},\mathcal{R}),p)$$

where $p \in \mathcal{X} = \{A,N,D,Q,E,G,I,L,F,T,Y,V\}$ (2) and $Paac(m(\mathfrak{I},\mathcal{R}),p)$ is defined in Definition 3.12.

Definition 3.14. To attain a strong statistical significance, we use the data from the 100 random codes R ($R_1, R_2 \ldots, R_{100}$), the percentage $Paac(\bar{m}(R,\mathcal{R}),p)$ of conservation of X codons per amino acids p coded by the R mean random motifs $\bar{m}(R,\mathcal{R})$ in the genes of a reference genome $s_1 = \mathcal{R}$ in all the multiple gene alignments $s_1, s_2 \ldots, s_n$, is equal to

$$Paac(\bar{m}(R,\mathcal{R}),p) = \frac{1}{\sum_{k=1}^{100} \delta_k} \sum_{k=1}^{100} Paac(m(R_k,\mathcal{R}),p)$$

where $p \in \mathcal{X} = \{A,N,D,Q,E,G,I,L,F,T,Y,V\}$ (2) and $Paac(m(R_k,\mathcal{R}),p)$ is defined in Definition 3.12 and $\delta_k = 1$, if $\mathcal{g}^{-1}(p) \cap R_k \neq \emptyset$ (i.e. a random code R_k can code for the amino acid p) and $\delta_k = 0$ otherwise.

Definition 3.15. The mean percentage $\bar{P}aac(\bar{m}(R,\mathcal{R}),\mathcal{X})$ of conservation of X codons in the 12 amino acids $\mathcal{X}$ (2) coded by the R mean random motifs $\bar{m}(R,\mathcal{R})$ in the genes of a reference genome $s_1 = \mathcal{R}$ in all the multiple gene alignments $s_1, s_2 \ldots, s_n$, is equal to

$$\bar{P}aac(\bar{m}(R,\mathcal{R}),\mathcal{X}) = \frac{1}{Card(\mathcal{X})}\sum_{p\in\mathcal{X}} Paac(\bar{m}(R,\mathcal{R}),p)$$

where $p \in \mathcal{X} = \{A,N,D,Q,E,G,I,L,F,T,Y,V\}$ (2) and $Paac(\bar{m}(R,\mathcal{R}),p)$ is defined in Definition 3.14.

Remark 9. The mean percentage $\bar{P}aac$ gives the same statistical weight for each amino acid.

Remark 10. For the R mean random motifs $\bar{m}(R,\mathcal{R})$, we only analyse in the mean matrix $\bar{\boldsymbol{B}}$ the trinucleotides coding the 12 amino acids $\mathcal{X}$ coded by the X circular code.

3.6.4.2. Synonymous substitutions of trinucleotides in *X* motifs

We chose to consider only those positions in the multiple gene alignment with a preserved amino acid, i.e. involving synonymous substitutions, to evaluate whether trinucleotides are preserved in X motifs beyond what might be predicted by chance if they evolved under the selective pressure on the amino acids they encode.

For the X motifs in all the mammal and yeast multiple gene alignments, we constructed the codon substitution matrices $\boldsymbol{A}(m)$ (defined in Definition 3.3). These matrices were normalized to produce the normalized codon substitution matrices $\boldsymbol{B}(m)$ (defined in Definition 3.4). We then extracted the rows and columns from the normalized codon substitution matrices that correspond to the synonymous substitutions of the X trinucleotides (Appendix Table I and Table II). We also constructed the equivalent matrices $\boldsymbol{B}(m)$ for the R random motifs and extracted the rows and columns that correspond to the synonymous substitutions of the X trinucleotides. Finally, we calculated the percentages $Paac(m(X,\mathcal{R}),p)$ (defined in Definition 3.12) of conservation of X trinucleotides per amino acid $p \in \mathcal{X}$ for the two mammals and yeasts multiple gene alignment (Appendix Table III and Table IV). We provide the summary of these results in Table 3.17 and Table 3.18. We also include the mean percentages $\bar{P}aac(m(X,\mathcal{R}),\mathcal{X})$ (defined in Definition 3.13) of conservation of X trinucleotides in the 12 amino acids $\mathcal{X}$ (2) for both sets of multiple gene alignments. For comparison, we provide the values of $Paac(\bar{m}(R,\mathcal{R}),p)$ (defined in Definition 3.14) and $\bar{P}aac(\bar{m}(R,\mathcal{R}),\mathcal{X})$ (defined in Definition 3.15) for the R mean random motifs from the 100 random codes R. The distribution of $Paac$ and $\bar{P}aac$ values for the R random motifs with a $\pm$ 0.99 confidence interval are shown in Figure 3.11 and Figure 3.12. From these results, we identified a new and strong property of the X motifs.

Table 3.17. Amino acid conservation parameters in the mammal multiple gene alignments. Comparison of mean percentage $\bar{P}aac(m(X,\mathbb{H}),\mathcal{X})$ (Definition 3.13) and percentages $Paac(m(X,\mathbb{H}),p)$ (Definition 3.12) given in the first row, with $\bar{P}aac(\overline{m}(R,\mathbb{H}),\mathcal{X})$ (Definition 3.15) and $Paac(\overline{m}(R,\mathbb{H}),p)$ (Definition 3.14) given in the second row, for the mammal multiple gene alignments with human reference genes $s_1 = \mathbb{H}$.

	Mean	*A*	*D*	*E*	*F*	*G*	*I*	*L*	*N*	*Q*	*T*	*V*	*Y*
$m(X,\mathbb{H})$	78.1	66.1	89.4	90.3	78.9	77.3	84.9	78.3	85.6	80.7	63.5	65.7	76.1
$\overline{m}(R,\mathbb{H})$	67.6	60.2	73.0	76.8	77.7	68.1	68.0	67.1	70.5	71.3	58.1	62.6	74.6

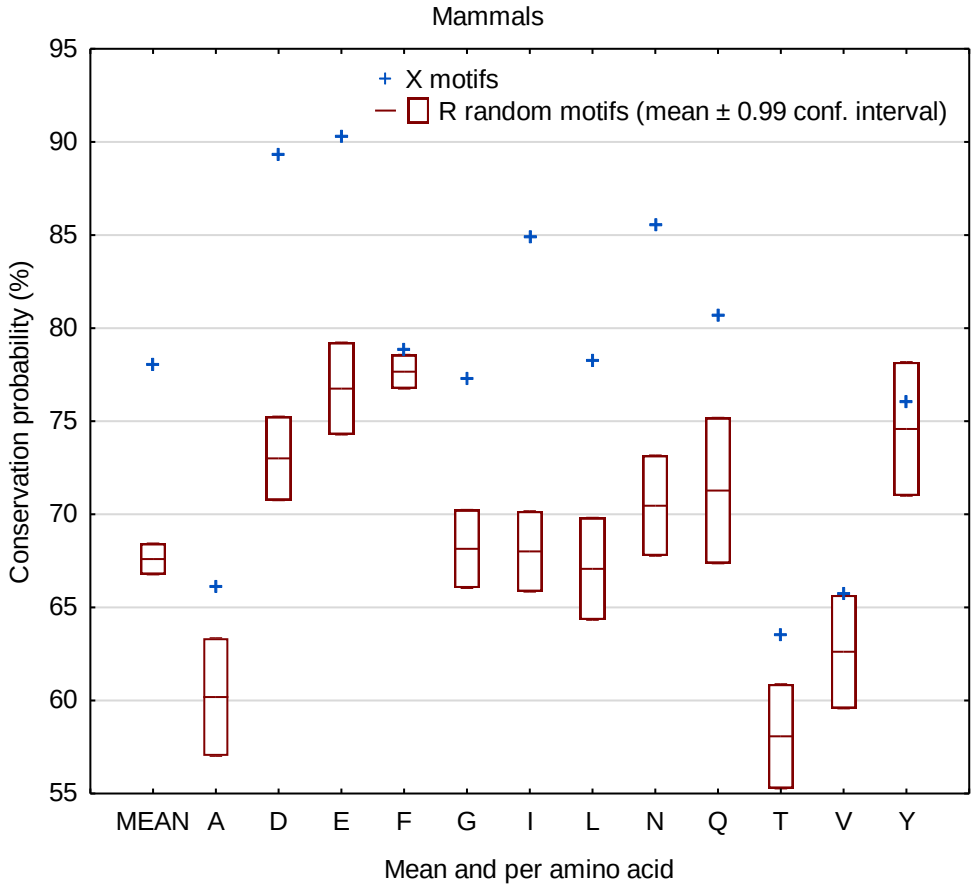


Figure 3.11. Graphical representation of Table 3.17. Comparison of mean percentage $\bar{P}aac(m(X,\mathbb{H}),\mathcal{X})$ (Definition 3.13) and percentages $Paac(m(X,\mathbb{H}),p)$ (Definition 3.12) shown by blue cross for the X motifs, with $\bar{P}aac(\overline{m}(R,\mathbb{H}),\mathcal{X})$ (Definition 3.15) and $Paac(\overline{m}(R,\mathbb{H}),p)$ (Definition 3.14) shown as boxplots for the R random motifs, for the mammal multiple gene alignment with human reference genes $s_1 = \mathbb{H}$.

The average percentage ($\bar{P}aac$) conservation of X codons is significantly higher in X motifs than in the R random motifs in the mammal gene alignments (one-sided Student's *t*-test with $p \approx 10^{-55}$) (Table 3.17 and Figure 3.11). Furthermore, this is true for 11 out of 12 amino acids (percentage $Paac$) coded by the X circular code. For the amino acid *Y*, the conservation of X codons in X motifs is higher than in R motifs although the difference is not significant at 0.99. We can formalize this new property in a simple way by the following inequalities:

$$\begin{cases} \bar{P}aac(m(X,\mathbb{H}),\mathcal{X}) > \bar{P}aac(\bar{m}(R,\mathbb{H}),\mathcal{X}) \\ Paac(m(X,\mathbb{H}),p) > Paac(\bar{m}(R,\mathbb{H}),p) \quad \forall p \in \mathcal{X} \end{cases}$$

Table 3.18. Amino acid conservation parameters in the yeast multiple gene alignments. Comparison of mean percentage $\bar{P}aac(m(X,\mathbb{C}),\mathcal{X})$ (Definition 3.13) and percentages $Paac(m(X,\mathbb{C}),p)$ (Definition 3.12) given in the first row, with $\bar{P}aac(\bar{m}(R,\mathbb{C}),\mathcal{X})$ (Definition 3.15) and $Paac(\bar{m}(R,\mathbb{C}),p)$ (Definition 3.14) given in the second row, for the yeast multiple gene alignment with *S. cerevisiae* reference genes $s_1 = \mathbb{C}$.

	Mean	*A*	*D*	*E*	*F*	*G*	*I*	*L*	*N*	*Q*	*T*	*V*	*Y*
$m(X,\mathbb{C})$	29.3	16.1	45.3	41.8	29.9	40.4	39.3	10.6	33.5	13.6	14.9	33.6	33.0
$\bar{m}(R,\mathbb{C})$	22.3	19.1	26.9	25.2	29.8	27.1	20.3	21.8	22.3	20.9	15.6	18.2	33.0

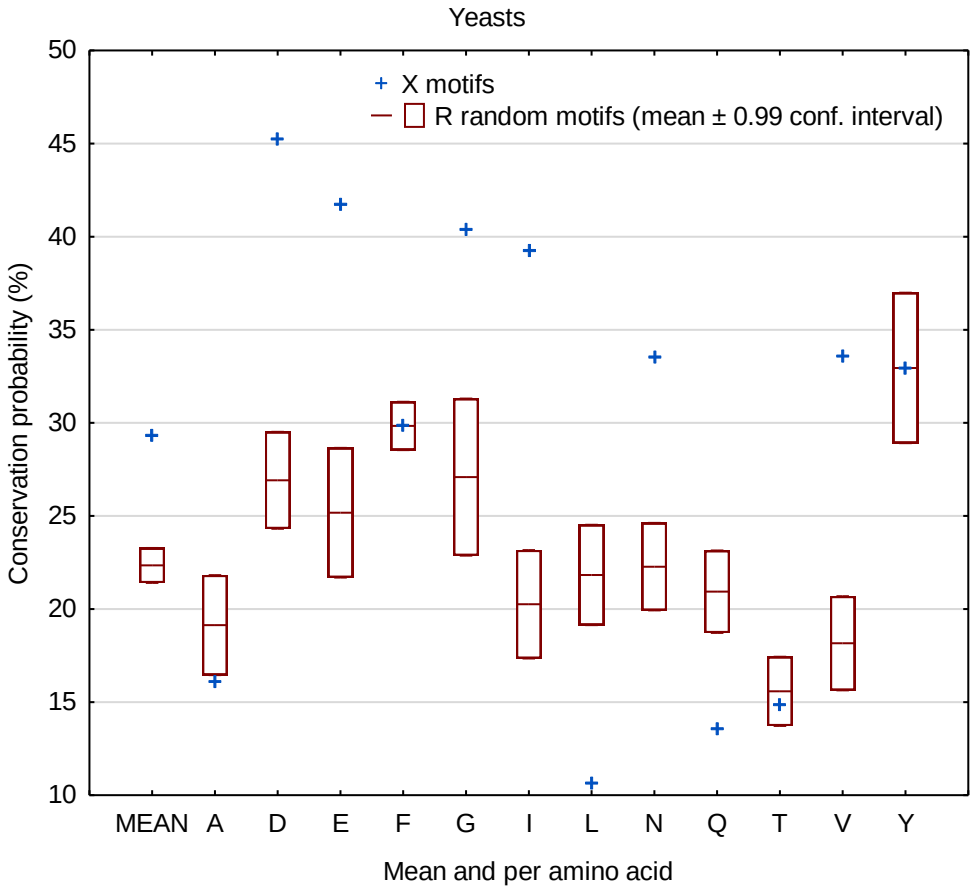


Figure 3.12. Graphical representation of Table 3.18. Comparison of mean percentage $\bar{P}aac(m(X,\mathbb{C}),\mathcal{X})$ (Definition 3.13) and percentages $Paac(m(X,\mathbb{C}),p)$ (Definition 3.12) shown by blue cross for the X motifs, with $\bar{P}aac(\bar{m}(R,\mathbb{C}),\mathcal{X})$ (Definition 3.15) and $Paac(\bar{m}(R,\mathbb{C}),p)$ (Definition 3.14) shown as boxplots for the R random motifs, for the yeast multiple gene alignment with *S. cerevisiae* reference genes $s_1 = \mathbb{C}$.

The average percentage ($\bar{P}aac$) conservation of X codons is significantly higher in X motifs than in the R mean random motifs in the yeast gene alignments (one-sided Student's *t*-test with $p \approx 10^{-35}$) (Table 3.18 and Figure 3.12). For 6 out of 12 amino acids, the conservation (percentage $Paac$) of X codons in X motifs is higher than in R random motifs. In contrast, for

the amino acids *A*, *L*, *Q* and *T*, the conservation is lower than in the *R* motifs. For the amino acids *F* and *Y*, the conservation is similar to the *R* motifs. This property is formalized simply by the following inequalities:

$$\begin{cases} \bar{P}aac(m(X,\mathbb{C}),\mathcal{X}) > \bar{P}aac(\overline{m}(R,\mathbb{C}),\mathcal{X}) \\ Paac(m(X,\mathbb{C}),p) > Paac(\overline{m}(R,\mathbb{C}),p) \quad \forall p \in \mathcal{X}\backslash\{A,L,Q,T\} \end{cases}.$$

From these results, we can summarize that the conservation values of the motifs observed in the yeast multiple gene alignments is lower than that observed in the human multiple gene alignments. This is expected since it is well known that the yeasts diverged much earlier (more synonymous and non-synonymous substitutions) than the mammals. This evolutionary diversity in the yeasts used in this study may also clarify the exception with the four amino acids observed with this simple statistical parameter $Paac$. We emphasize the fact that the identified conservation property of X codons in X motifs with respect to the encoded amino acids is independent of the codon usage, the GC content, the nucleotide composition, the length of genes, etc.

3.6.5. Union of circular codes associated with each amino acid

We carried out various statistical analyses, the results of which indicate that the 20 trinucleotides of the X circular code (above) are strongly related to the amino acids they encode, resulting in the 20 trinucleotides of X being divided into 12 trinucleotide classes, each of which is associated with an amino acid $p \in \mathcal{X}$ (above). This property allows us to suggest that the extant genetic code may have resulted from a union of circular codes: the sub-codes of the circular code X associated with each amino acid (Table 3.19 and Figure 3.13). Notably, a sub-code of a circular code is also circular. In section 2.4, we introduced the comma-free codes and their reading frame maintenance ability. There exists a stronger class of the comma-free codes, known as strong comma-free codes that has the ability to retrieve the reading frame after a maximum of 2 nucleotides. Furthermore using an approach developed earlier (Michel, 2014), we determine the circular class of each trinucleotide code involved in Table 3.19.

Table 3.19. Classes of codes (non-circular NC, circular C, comma-free CF, strong comma-free SCF) of the 12 amino acids $\mathcal{X}$ (above) with respect to the circular code X (above) and the universal genetic code (SGC).

AA	Circular code X	Class	Union	Class	Genetic code	Class
Asn	$N_X = \{AAC, AAT\}$	*SCF*			$N = \{AAC, AAT\}$	*SCF*
Asp	$D_X = \{GAC, GAT\}$	*SCF*			$D = \{GAC, GAT\}$	*SCF*
Gln	$Q_X = \{CAG\}$	*SCF*	$\{CAA\}$	*SCF*	$Q = \{CAA, CAG\}$	*SCF*
Glu	$E_X = \{GAA, GAG\}$	*CF*			$E = \{GAA, GAG\}$	*CF*
Phe	$F_X = \{TTC\}$	*SCF*	$\{TTT\}$	*NC*	$F = \{TTC, TTT\}$	*NC*
Tyr	$Y_X = \{TAC\}$	*SCF*	$\{TAT\}$	*CF*	$Y = \{TAC, TAT\}$	*CF*
Ile	$I_X = \{ATC, ATT\}$	*SCF*	$\{ATA\}$	*CF*	$I = \{ATA, ATC, ATT\}$	*CF*
Ala	$A_X = \{GCC\}$	*SCF*	$\{GCA, GCG, GCT\}$	*CF*	$A = \{GCA, GCC, GCG, GCT\}$	*CF*
Gly	$G_X = \{GGC, GGT\}$	*SCF*	$\{GGA, GGG\}$	*NC*	$G = \{GGA, GGC, GGG, GGT\}$	*NC*
Thr	$T_X = \{ACC\}$	*SCF*	$\{ACA, ACG, ACT\}$	*CF*	$T = \{ACA, ACC, ACG, ACT\}$	*CF*
Val	$V_X = \{GTA, GTC, GTT\}$	*SCF*	$\{GTG\}$	*CF*	$V = \{GTA, GTC, GTG, GTT\}$	*CF*
Leu	$L_X = \{CTC, CTG\}$	*CF*	$\{CTA, CTT, TTA, TTG\}$	*CF*	$L = \{CTA, CTC, CTG, CTT, TTA, TTG\}$	*C*

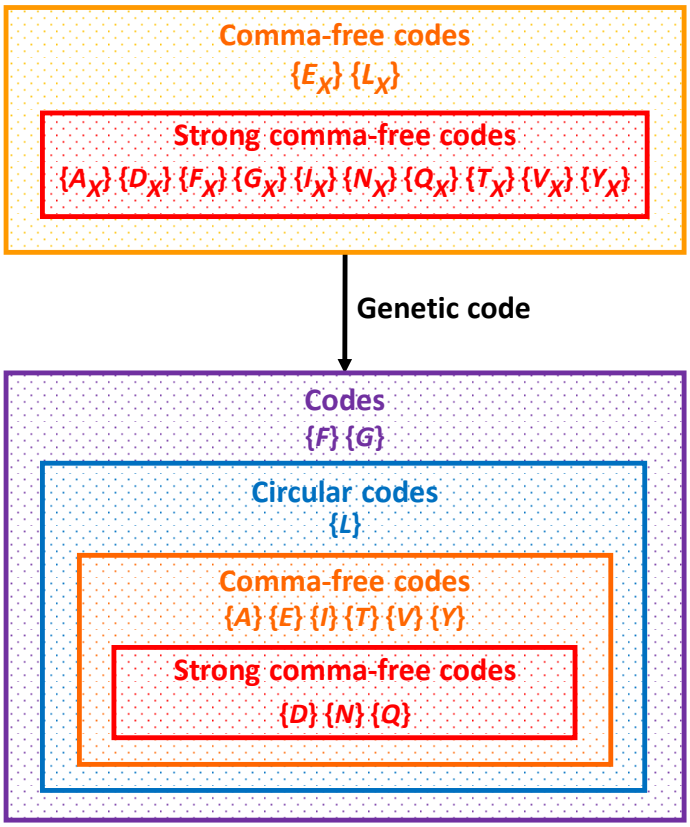


Figure 3.13. (associated with Table 3.19). Evolution of the genetic code by union of circular codes associated with each amino acid from the circular code X (above).

Here we put forward a new hypothesis of evolution of the genetic code. The evolution of the standard genetic code may have started from the circular codes with the most stringent constraints; motifs from the strong comma-free codes and the comma-free codes retrieving the reading frame after the reading of 2 and 3 nucleotides, i.e. a nucleotide length of a codon or anticodon in the primitive systems. It is tempting to suggest that such circular codes in the primordial chemical soup may have originated independently. These highly constrained coding systems, however, may not have been feasible in the long term. By relaxing the constraints, they may have evolved into circular codes with flexible motifs for retrieving the reading frame after reading a maximum of 13 nucleotides, and into non-circular codes without the ability to retrieve the reading frame. Among the 12 amino acids $\mathcal{X}$ (above) coded by the X circular code (above), 10 amino acids are coded by strong comma-free codes and 2 amino acids E_X and L_X of X, by comma-free codes (Table 3.19 and Figure 3.13). In the extant genetic code, only 3 amino acids *D*, *N* and *Q* are still coded by strong comma-free codes, 6 amino acids *A*, *E*, *I*, *T*, *V* and *Y*, by comma-free codes, 1 amino acid *L*, by a circular code, and 2 amino acids *F* and *G*, by simple codes (not circular). The union of circular codes allows us to extend the amino acid coding. For example, the union of the strong comma-free code $Q_X = \{CAG\}$ of X and the strong comma-free code $\{CAA\}$ leads to the strong comma-free code $Q = \{CAA, CAG\}$ of the genetic code, etc.

Obviously, the union of 2 comma-free codes does not imply that the resulting code is comma-free, see for example the case of the amino acid *L* (Table 3.19). The 8 remaining amino acids could have been generated by mutations in circular codes. From a mathematical point of view, the standard genetic code is a code (Definition 2.3), however it is not circular, i.e. it does not have the ability to retrieve the reading frame in genes.

3.7. Functionality of *X* motifs in mammal and yeast genes

We identified specific evolutionary constraints on the *X* motifs in the previous section of this chapter. Moreover, the nucleotides in the *X* motifs display a significant excess of synonymous substitutions compared to the non-*X* motifs. These results suggest that the *X* motifs located in the genomes used in this study have evolved under purifying selection. In addition, the average conservation of codons in *X* motifs is significantly higher than predicted in the case where the substitution process was random. These findings indicate a potential functional role of *X* motifs, possibly as elements of the complex genome decoding system. In order to examine the potential role of *X* motifs on the translation of protein-coding genes, we compared the frequency of *X* motifs in the genes of the four mammalian and nine yeast species with existing experimental data on protein expression and protein production. We will demonstrate that the experimental data can generally be explained by circular code motifs, i.e. motifs having the reading frame retrieval property.

3.7.1. Dicodons associated with reduced protein synthesis are absent in *X* motifs

Recently, experimental studies were performed in the genes of *S. cerevisiae* (Gamble et al., 2016) to investigate the effects of different codons on the efficiency of translation elongation. The authors measured the expression levels of over 35,000 synthetic protein variants in which three adjacent codons of the coding sequence were randomized, and found that the translation efficiency is modulated by adjacent codon pairs. No individual codons had consistent effects on gene expression. However, 17 pairs of adjacent codons (called in the following dicodons) were identified associated with reduced expression level of genes (when they were present in-frame in the coding sequence). They proposed that “an interplay between tRNAs at adjacent ribosome sites modulates the translation performance”. We provide the list of the 17 dicodons that are associated with reduced expression level of genes in Table 3.20. In this list, we identified those codons belonging to the *X* circular code (symbol *X*) and those which are not (symbol *N*). Surprisingly, none of these 17 dicodons are made up of two *X* codons which means they cannot be located in an *X* motif.

Table 3.20. List of the 17 dicodons associated with reduced expression level of the genes (Gamble et al., 2016) (1st and 3rd columns). Class of the dicodons according to its codons belonging to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd and 4th columns).

Dicodon	Class	Dicodon	Class
AGGCGA	*NN*	*CGAGCG*	*NN*
AGGCGG	*NN*	*CTCCCG*	*XN*
ATACGA	*NN*	*CTGATA*	*XN*
ATACGG	*NN*	*CTGCCG*	*XN*
CGAATA	*NN*	*CTGCGA*	*XN*
CGACCG	*NN*	*GTACCG*	*XN*
CGACGA	*NN*	*GTACGA*	*XN*
CGACGG	*NN*	*GTGCGA*	*NN*
CGACTG	*NX*		

3.7.2. Dicodons associated with protein production in correlation with *X* motifs

In a similar study of dicodons (Diambra, 2017), Diambra performed a comparative study of dicodon usage frequencies over two sets of proteins: a low protein abundance (PA) set and a high PA set, consisting of nine diverse organisms including three prokaryotes, one plant, one yeast (*S. cerevisiae*), two multicellular eukaryotes and two mammals. In terms of translation efficiency, the research hypothesis was that sequences encoding rich abundance proteins are optimized. He found a significant difference in the use of dicodons depending on the PA and calculated which dicodons were associated statistically with low or high abundance proteins. The usage frequency of single codons did not justify the preferences observed in the study. The statistical analysis of coding sequences of the chosen nine organisms revealed that in many cases dicodon preferences are commonly shared between related organisms. We identified those codons belonging to the *X* circular code to determine the possible functionality of *X* motifs in the case of protein production.

We list the 16 dicodons associated with low protein abundance along with the class they belong in Table 3.21. As none of these dicodons have two consecutive codons that belong to the *X* circular code, these 16 dicodons cannot be located in *X* motifs. Combining these 16 dicodons with the 17 dicodons previously identified in the genes of *S. cerevisiae* associated with reduced expression levels (Table 3.20), we have identified 33 low abundance dicodons that support the circular code theory. In addition, 40 dicodons were identified in the study, which were shared among various organisms and used preferably with high protein abundances (Table 3.22). Of these, 27 (67.5%) dicodons are made up of two *X* codons and can thus be present in *X* motifs.

Table 3.21. List of the 16 dicodons associated with low protein abundance (Diambra, 2017) (1st and 3rd columns). Class of the dicodons according to its codons belong to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd and 4th columns).

Dicodon	Class	Dicodon	Class
AAAATA	*NN*	*CAGAAA*	*XN*
AATGCA	*XN*	*GAAAGT*	*XN*
AATTGG	*XN*	*GAACTA*	*XN*
AGTAAG	*NN*	*GCATTT*	*NN*
AGTGTG	*NN*	*TATAAA*	*NN*
ATAGGT	*NX*	*TATCCG*	*NN*
ATTAAA	*XN*	*TTTCAG*	*NX*
CAAAGT	*NN*	*TTTTTT*	*NN*

Table 3.22. List of the 40 dicodons associated with high protein abundance (Diambra, 2017) (1st, 3rd, 5th and 7th columns). Class of the dicodons according to its codons belong to the circular code *X* (symbol *X*) or not (symbol *N*) (2nd, 4th, 6th and 8th columns).

Dicodon	Class	Dicodon	Class	Dicodon	Class	Dicodon	Class
AACAAC	*XX*	*ACCTTC*	*XX*	*GACACC*	*XX*	*GTCACC*	*XX*
AACAAG	*XN*	*ATCAAC*	*XX*	*GACTAC*	*XX*	*GTCATC*	*XX*
AACACC	*XX*	*ATCAAG*	*XN*	*GATGCT*	*XN*	*GTTGCC*	*XX*
AAGTCC	*NN*	*ATCACC*	*XX*	*GCCAAC*	*XX*	*TACAAC*	*XX*
ACCAAC	*XX*	*ATCATC*	*XX*	*GCCAAG*	*XN*	*TACAAG*	*XN*
ACCAAG	*XN*	*ATTGCC*	*XX*	*GCCACC*	*XX*	*TCCACC*	*NX*
ACCACC	*XX*	*CCACCA*	*NN*	*GCCATC*	*XX*	*TTCAAC*	*XX*
ACCATC	*XX*	*CGTCGT*	*NN*	*GCCGCC*	*XX*	*TTCAAG*	*XN*
ACCATT	*XX*	*GACAAC*	*XX*	*GGTGTC*	*XX*	*TTCACC*	*XX*
ACCGCC	*XX*	*GACAAG*	*XN*	*GTCAAG*	*XN*	*TTCATC*	*XX*

3.7.3. Classification of genes as low or high abundance

We performed statistical tests to classify genes as low abundance or high abundance according to the circular code theory. We identified 33 low abundance dicodons and 40 high abundance dicodons, from previous experimental (Gamble et al., 2016; Table 3.20 and Table 3.21) and statistical (Diambra, 2017; Table 3.22) studies and summarized in the Table 3.23. Here we identify an important and new factor for the classification of genes.

Table 3.23. Classification of low/high abundance protein related to the presence/absence of dicodons *XX* (deduced from Table 3.20, Table 3.21 and Table 3.22).

	XX	{*NN*,*NX*,*XN*}	Total
Low abundance protein	0	33	33
High abundance protein	27	13	40
Total	27	46	73

We performed a $\mathcal{X}^2$ test to determine the relation between presence/absence of dicodons that belong to the *X* circular code and low/high protein abundance. The test shows a strongly significant relation between the presence/absence of dicodons *XX* and protein

abundancy with a one sided value $p \approx 10^{-9}$ from Table 3.23. We also calculated the probabilities of low and high abundance protein related to the presence of dicodons XX from Table 3.23:

$$P(\text{Low abundance protein} \mid XX) = 0/33 = 0\% \text{ and}$$

$$P(\text{High abundance protein} \mid XX) = 27/40 = 67.5\%.$$

Based on these findings, the presence-absence of XX dicodons in a gene can be associated with low or high protein abundance, possibly a new gene-classifying factor. We did further analysis to evaluate if the enrichment of X motifs in a gene is associated with gene expression levels.

3.7.4. Correlation of *X* motifs with gene expression level

Here we analyse "wild type" genes and "optimized" genes (SGDB database) for the presence of X motifs and compare the results obtained with those obtained from R random motifs (100 random codes). The SGDB database (Wu et al., 2007) contains gene expression data for genes that are experimentally re-engineered to enhance gene expression. In order to enhance the expression of genes, codons in the wild type gene are replaced with optimal codons thereby enhancing the expression system. Generally, it is achieved by replacing rare codons with the most frequently used codons in the organism, as most amino acids are coded by 2 or more synonymous codons. In this analysis, we only considered the re-engineered genes that did not involve non-synonymous changes (without altering the amino acid). Thus, we analysed 42 re-engineered genes that had increased gene expression and 3 re-engineered genes that had no significant increase in gene expression. We searched for X motifs and R random motifs in the wild type genes and the genes optimized for gene expression. We also calculated the mean number and the mean nucleotide length of X and R motifs per sequence for comparison (Table 3.24).

Table 3.24. Mean number and mean nucleotide length of X and R random motifs (100 random codes) per wild type gene and per optimized gene taken from the SGDB database (Wu et al., 2007).

		Mean number of X motifs	Mean number of R motifs	Mean length of X motifs	Mean length of R motifs
42 genes with increased expression	Wild type	5.4	3.6	86.1	53.7
	Optimized gene	11.2	3.7	188.6	58.2
3 genes with no increased expression	Wild type	5.3	2.6	80.0	35.8
	Optimized gene	5.0	3.8	80.0	55.6

For the optimized genes where there is no significant increase of gene expression, we observed a non-significant difference in the mean number of X motifs ($5.0 - 5.3 = -0.3$, one tailed Wilcoxon test with value $p = 0.50$) and no difference in the mean length ($80.0 - 80.0 = 0$) between the optimized genes and the wild type genes. These differences are also not significant for R random motifs (mean number : $3.8 - 2.6 = 1.2$ and mean length : $55.6 - 35.8 = 19.8$). Unfortunately, since this test is based on only three cases, the results remain to be confirmed.

In comparison, for the optimized genes where there is a significant increase of gene expression, we observed a strong enrichment in the mean number of X motifs ($11.2 - 5.4 = 5.8$, one tailed Wilcoxon test with value $p \approx 10^{-6}$). In addition, in the optimized genes the X motifs covered a larger proportion of the genes (mean length : $188.6 - 86.1 = 102.5$, one tailed Wilcoxon test with value $p \approx 10^{-6}$) compared to the wild type genes (Figure 3.14). In Figure 3.14, we ordered the genes according to the coverage of wild type genes by X motifs in ascending order. These differences are not observed with the R random motifs (one tailed Wilcoxon test, p values equal to 0.24 and 0.12, respectively).

The results shown here indicate that most of the optimized genes contain codons (consecutive codons form motifs) from the X circular code, which is not the case for the wild type genes. We obtained a significant increase in coverage by X motifs after optimization of genes, which may suggest a new strategy of efficient gene optimization by using codons from the X circular code.

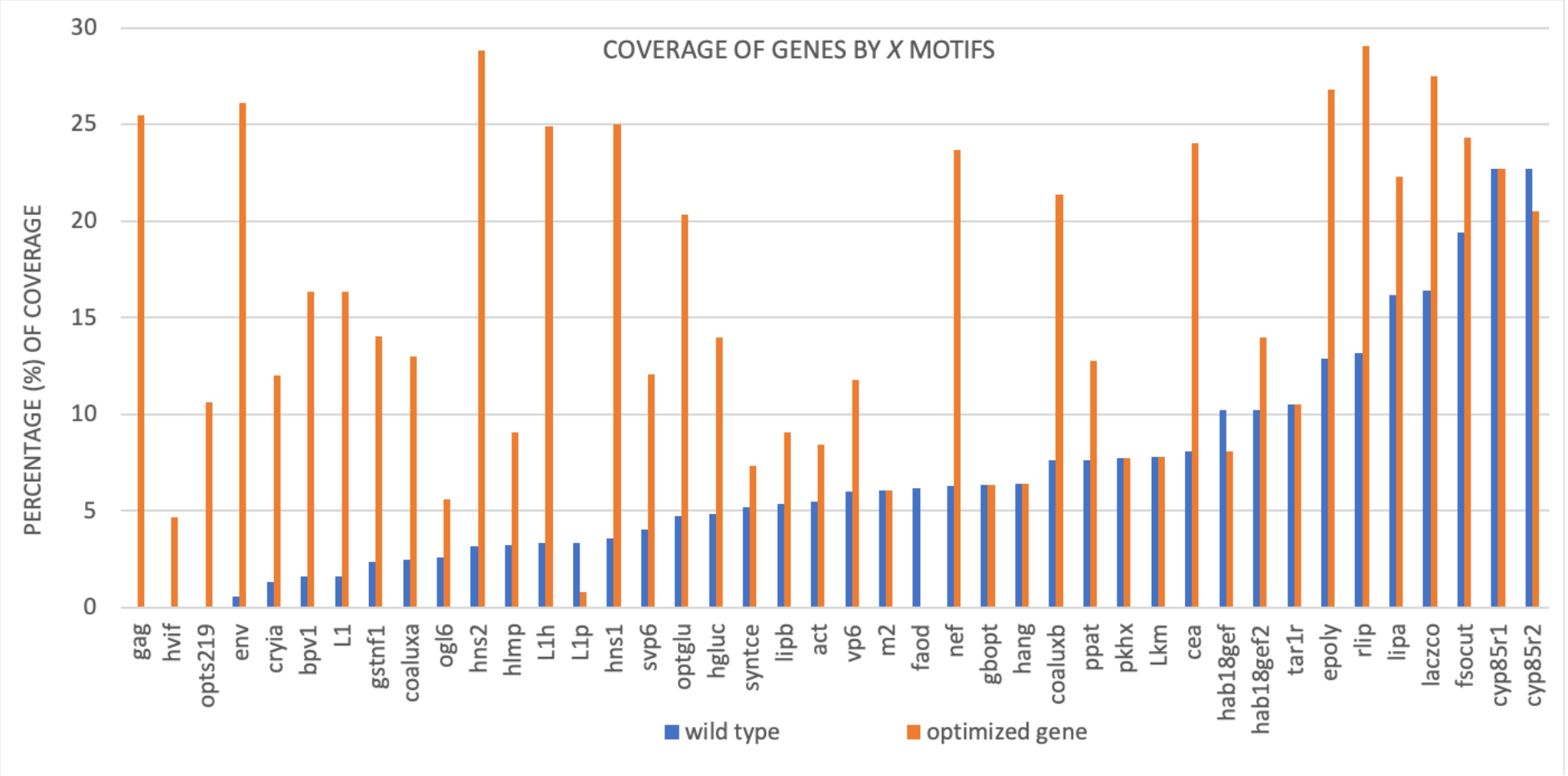


Figure 3.14. Percentage coverage (total length of *X* motifs divided by the total length of genes) of 42 wild type and optimized genes by *X* motifs.

3.8. Summary

The work presented in this chapter addressed two questions: are X motifs conserved during evolution? and do they continue to play a functional role in the processes of genome decoding and protein synthesis?

The organisms chosen in this extensive study of two sets of organisms represent a large phylogenetic distribution, and a wide variety of gene structures, ranging from the simple, single exon genes of *S. cerevisiae* to the highly complex intron/exon structure of human genes. We identified a strong enrichment of X motifs (number and length) in both mammal and yeast multiple gene alignments, thus confirming the previous studies on enrichment of X motifs in protein-coding genes. With the help of various parameters of evolutionary conservation, we showed that the X motifs are more conserved compared to the rest of the gene sequences, with a lower dN/dS ratio of non-synonymous to synonymous substitutions, indicating purifying selection. We also carried out in-depth investigation of synonymous substitutions in X motifs. The sequence conservation of X motifs suggests two types of selection pressure : first to preserve the amino acids of the respective proteins encoded by the genes and second that applies only to X motifs, thereby highlighting a new conservation property of X motifs per amino acid. This led us to propose a new hypothesis for the evolution of the genetic code as a union of circular codes associated with each amino acid.

The increased conservation of X motifs and the specific evolutionary constraints suggest that X motifs may represent an additional, overlapping function within the protein-coding regions of genomes. In section 1.3 we discussed the degeneracy of the genetic code and its ability to encode overlapping information. It is well known that different synonymous codons encoding the same amino acid are not used by different organisms with the same frequency. The genetic code also contains information affecting the rate and efficiency of translation. However, these codon-mediated regulatory mechanisms are still not clear (Brule & Grayhack, 2017). Many recent studies have been performed to try to explain the different codon usages observed and their effects on translation. Analysis involving increased expression level of genes suggest that synonymous codons may be selected for some specific translational properties (Brar, 2016). Also, not only specific codons are selected to carry out some specific translational properties, but moreover the selection of optimal codons for specific properties is also co-ordinated with the modification of their respective tRNA anticodons, which suggests complex biological modifications are co-ordinated with the present of specific optimal codons to improve translational efficiency. In particular, it has been shown that the efficiency of translating a particular codon is influenced by the nature of the immediately adjacent flanking codons (Chevance & Hughes, 2017; Diambra, 2017; Gamble et al., 2016). There have been significant

discussions about the idea that codon pairs have a different effect on the translation efficiency than individual codons. To explain these findings, we explored whether these findings could be explained in terms of the circular code theory. In two related studies on effects of dicodons (Diambra, 2017; Gamble et al., 2016), a total of 33 dicodons were found to be associated with low protein abundance, and 40 dicodons associated with high abundance proteins. We showed a strong correlation between the level of protein abundance and the dicodons that belong to the X circular code. To further examine this link between the presence of X motifs in a gene and the expression level, we compared a set of re-engineered genes with the original wild type genes. Again, we observed a positive correlation between the number and length of X motifs and protein expression levels; which may suggest a new strategy for efficient gene optimization.

If our hypothesis is correct that the X circular code plays a role in gene expression, the next question is how do X motifs in mRNA sequences influence translation? During the introduction of this thesis, we briefly described the molecular machinery responsible for translation, i.e. the ribosome, and stressed the importance of ribosomal RNAs (rRNAs) in decoding. This motivated us to carry out an extensive study of X motifs in the rRNAs of representative organisms covering all the three domains of life. In the next chapter, we will discuss the X circular code motifs identified in the ribosome.

Chapter 4

4. Circular code motifs in the ribosome

4.1. Introduction

In the previous chapter we presented the results from an extensive study of multiple alignment of genes from two different sets of organisms. The results obtained strengthen the results on the enrichment of X motifs in the protein-coding regions of the genes of most organisms, from bacteria to eukaryotes. We observed specific evolutionary pressures acting on the X motifs compared to the other regions of the gene, and we examined the possibility of a functional role being played by the X motifs in the translation process in extant organisms.

Here we discuss the idea that circular codes played an important role in the evolution of the genetic code in primitive systems. We investigate the possibility that the X circular code was used in primordial translation systems to encode fewer amino acids along with the ability to maintain the reading frame before the evolution of the modern genetic code. In order to examine this, we searched for X circular code motifs in the rRNAs, which are considered to be the most conserved over the evolutionary period. The ribosome is universal in all extant organisms (Melnikov et al., 2012), which suggests that its formation can be traced back to the time of the *last universal common ancestor* (LUCA). Although there are some differences in size and structure of the ribosomes from different domains of life, the translation mechanisms are generally similar with some variations. For the first time, we show the evolutionary conservation of trinucleotide (codon) motifs in the rRNAs. We show that traces of the X circular code still exist in modern rRNAs and in the evolution of the translation machinery, which suggests that the X circular code might be a predecessor of the standard genetic code. We would like to recall some previous discussions, to explain why the identification of X motifs in the ribosome is crucial to understand the evolution of the genetic code and translation machinery in primitive organisms.

In section 1.2 we discussed the origin of the genetic code, which is still a mystery. Various theories have been suggested to understand how the genetic code may have originated from the primordial chemical soup. According to the general textbook concept about the history of life on Earth, an initial “RNA world” (Gilbert, 1986) where RNA polymers acted both as a carrier of genetic information and as a catalyst for translation preceded the advent of protein synthesis (proteins and DNA). RNA alone carried out the function of carrying genetic information (carried out by DNA later) and catalyzing chemical reactions (carried out by

proteins). However, the RNA world fails to explain the evolution of the modern genetic code. Recent studies have suggested the possibility of a "peptide-RNA world", where the primitive RNA polymers and peptide molecules interacted and co-existed in the prebiotic environment (Bowman et al., 2015; Carter & Wills, 2018; Kunnev & Gospodinov, 2018; van der Gulik & Speijer, 2015; Wills & Carter, 2018; Chatterjee & Yadav, 2019; Gospodinov & Kunnev, 2020; Piette & Heddle, 2020). According to the "peptide-RNA world" theory, the origin of life was facilitated by interactions between short peptides and RNA (this peptide-RNA complex acted as the carrier of genetic information), that produced enzymes for catalytic activities from two types of prebiotic amino acids (instead of 20).

Researchers have debated that the RNA was not versatile enough to carry out alone the functions needed for the origin of life. It is widely accepted that increased chemical complexity in the prebiotic chemical soup led to the formation of RNA-like oligomers. These RNA-like oligomers interacted with short oligopeptides (prebiotic amino acids) present in the prebiotic chemical soup which provided important catalytic functions thereby stabilizing the early systems (Szathmáry, 1999; Plankensteiner et al., 2005; van der Gulik & Speijer, 2015). Various mechanisms have been proposed to try to explain the interaction between anticodons/codons and their cognate amino acids (prebiotic) which likely reflects a 'proto-translation system' (Yarus et al., 2009; Ma, 2010; Noller, 2012). Thus, a prebiotic nucleoprotein complex (early ribosome) may have consisted of rRNAs stabilized by a few small peptides containing glycine, alanine, aspartic acid and/or valine, essential for its structure (Fournier et al., 2010; Maier et al., 2013). According to this theory, peptide molecules and RNA co-evolved interactively, which gave this early translation system the ability to translate genetic information (Kunnev & Gospodinov, 2018) and the ability to self-replicate (Banwell et al., 2018). Consequently, the early translation systems would have been RNA-based and the genetic code preceded the emergence of DNA (Chatterjee & Yadav, 2019; Root-Bernstein & Root-Bernstein, 2019). However, there is growing evidence supporting that the idea that the evolution of the genetic code may have co-evolved progressively with the ribosome (Hartman & Smith, 2014; Johnson & Wang, 2010).

In the past, various primitive codes were proposed as the ancestors of the standard genetic code, including comma-free codes such as the RRY, RNY or GNC codes, and the X circular code (above), an error-correcting code which has the ability to identify and maintain the reading frame of genes. Previous studies on X motifs (El Soufi & Michel, 2014, 2015; Michel, 2012, 2013) showed that conserved X motifs (of short lengths) were found in tRNAs and rRNAs; specifically, the universally conserved nucleotides G530, A1492 and A1493 in the ribosome decoding center are located in short X motifs. Due to the self-complementary property of the X circular code (Definition 2.9) there is a possibility of some kind of interaction between X motifs

in the mRNA of protein-coding genes and the rRNA *X* motifs in order to maintain the correct reading frame during translation.

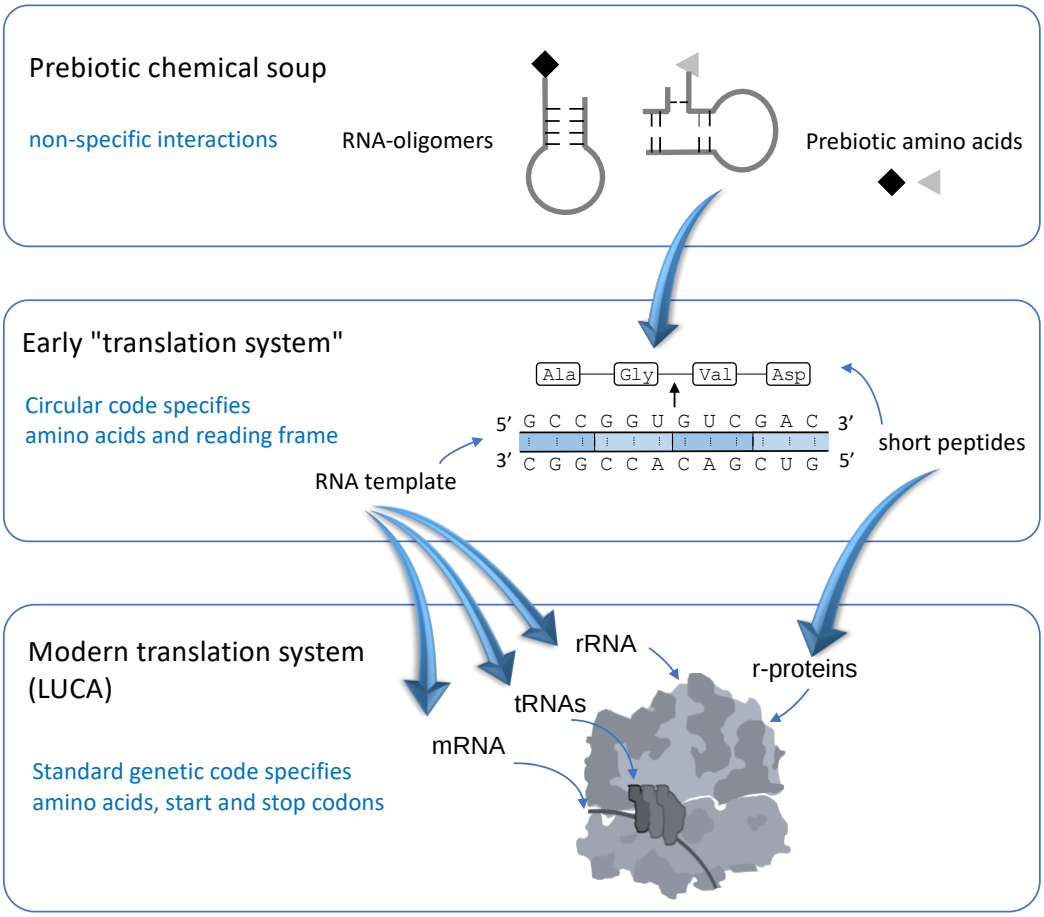


Figure 4.1. Hypothesis that the circular codes represent an intermediate coding system. Prebiotic chemical soup contained RNA-oligomers that interacted non-specifically with the prebiotic amino acids. This gave rise to the early "translation system" based on RNA template and more specific mapping between trinucleotides and early amino acids. This early translation system then evolved to form the RNA building blocks of the modern ribosome.

Here we test our hypothesis that the *X* circular code represents an intermediate coding system between the primordial, non-specific RNA-peptide interactions and the modern ribosome-based translation machinery (Figure 4.1). We performed a large-scale study of extant rRNA sequences from 133 organisms (covering the three domains of life: archaea, bacteria and eukaryote), in order to identify universally conserved *X* motifs. Universally conserved *X* motifs are motifs that have been conserved throughout evolution and in all three domains of life. In a detailed study of ribosome structural data, we showed that these universally conserved *X* motifs (denoted by uX motifs) are located in important functional sites including the decoding center and the peptidyl transferase center (PTC) in the ribosome. In fact, these functional sites are widely recognized as the 'essential building blocks' of the primordial 'proto-ribosome', a

primitive translation system that existed in the LUCA (Smith et al., 2008; Bokov & Steinberg, 2009; Hsiao et al., 2009, 2013; Petrov et al., 2015; Agmon, 2017, 2018; Bowman et al., 2020).

4.2. Ribosomal RNA data

Here we provide information about the rRNA data used in this analysis. The prokaryotic (archaea and bacteria) ribosome is composed of LSU rRNAs (23S and 5S) and SSU rRNAs (16S); whereas the eukaryotic ribosome is composed of LSU rRNAs (28S, 5.8S and 5S) and SSU rRNAs (18S). We considered two multiple alignments corresponding to the LSU rRNAs and SSU rRNAs containing sequences from the three domains of life. We obtained the multiple sequence alignments for LSU rRNAs (23S/28S and 5S) and SSU rRNAs (16S/18S) from the Center for Ribosomal Origins and Evolution's RiboVision web server at http://apollo.chemistry.gatech.edu/RibosomeGallery/index.html (Bernier et al., 2014).

The multiple sequence alignments contain complete sequences for rRNAs from 133 distinct organisms (32 eukaryotes, 65 bacteria, and 36 archaea), representing a broad sampling of the phylogenetic tree of life. The complete sequences for the 133 organisms were originally extracted from the SILVA database at https://www.arb-silva.de (Quast et al., 2013). We provide the list of the organisms present in the multiple sequence alignments in the appendix (Table VI). We recall the classical definition of multiple gene alignment given in Definition 3.2.

Definition 4.1. We define a ribosomal RNA (rRNA) multiple sequence alignment $s_1, \dots, s_n$ of $n = 133$ organisms as a mapping z on the alphabet $(B \cup \{\varepsilon\})^n \setminus (\{\varepsilon\})^n$ whose projection on the first component is s_1, on the second component is s_2, up to the projection on the nth component is s_n. The rRNA multiple sequence alignment z of length l is denoted as :

$$z = \begin{pmatrix} \bar{M}_{11} & \cdots & \bar{M}_{l1} \\ \bar{M}_{12} & \cdots & \bar{M}_{l2} \\ \vdots & \vdots & \vdots \\ \bar{M}_{1n} & \cdots & \bar{M}_{ln} \end{pmatrix}$$

Where the first ribosomal RNA sequence $s_1 = \bar{M}_{11}, \dots, \bar{M}_{l1}$, the second sequence $s_2 = \bar{M}_{12}, \dots, \bar{M}_{l2}$ up to the nth sequence $s_n = \bar{M}_{1n}, \dots, \bar{M}_{ln}$, such that $\bar{M}_{ji} \in B \cup \{\varepsilon\}$ for $i = 1, \dots, n$ and $j = 1, \dots, l$, where ε being classically associated with the gap symbol "–" or ".". For both SSU and LSU rRNA multiple sequence alignments we have chosen the numbering of rRNA sequences as per *Escherichia coli* numbering, which is well-annotated and widely used.

The X motifs $m(X)$ and the random motifs $m(R)$ located in the multiple sequence alignments may contain gaps such that $m(X), m(\bar{X}), m(R) \in B \cup \{\varepsilon\}$.

4.3. Secondary structures of rRNAs

The secondary structures of LSU and SSU rRNAs for *E. coli* were downloaded from http://apollo.chemistry.gatech.edu/RibosomeGallery/index.html.

Mapping of information on to secondary structures was performed with RiboVision (http://apollo.chemistry.gatech.edu/RiboVision) (Bernier et al., 2014). Positions of the expansion segments for LSU and SSU rRNAs and phases in the accretion model were obtained from (Petrov et al., 2015).

4.4. Three-dimensional structures of rRNAs

Coordinates of the high-resolution crystal structure of the *T. thermophilus* ribosome were obtained from the PDB database (https://www.rcsb.org/) (Berman et al., 2000). The PDB entry 4W2F was chosen because it contains mRNA nucleotides, an antibiotic (amicoumacin A) and three deacylated tRNAs in the A, P and E sites. Numbering of the *T. thermophilus* SSU rRNA is the same as for *E. coli*. For the LSU rRNA, *E. coli* numbering is used.

Visualization and analysis of the three-dimensional structures, as well as image preparation were performed with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC).

4.5. Universal X motifs (uX motifs) in rRNA multiple alignments

We previously defined X motifs belonging to the maximal C^3 self-complementary circular code X, which has the ability to retrieve and synchronize the reading frame of the sequence. We already discussed in detail the mathematical properties of the X circular code. For this study of rRNA multiple sequence alignments, we identified X motifs in all the sequences with the minimum length $l \geq 8$ nucleotides, i.e. a minimum of two trinucleotides with either prefixes or suffixes of trinucleotides belonging to the code. Demonstrated previously (Michel, 2012), X motifs with $l \geq 8$ nucleotides are able to retrieve the reading frame with a probability of 99.6% (reading frame retrieval with a probability of 90% with $l \geq 6$ nucleotides and 100% with $l \geq 12$ nucleotides). In the LSU rRNA (23S/28S and 5S) and SSU rRNA (16S/18S) multiple sequence alignments, for each position we also calculated the 'universality' of the X motifs, representing their evolutionary conservation.

Definition 4.2. We define 'X universality' by the number of sequences (denoted as n_X) having an X motif ($m(X)$) at a particular position in the multiple sequence alignment (Definition 4.1) of rRNAs from 133 organisms.

Definition 4.3. A 'universal X motif', denoted as 'uX motif' is defined as a region in the multiple sequence alignment (Definition 4.1) such that at least 6 consecutive positions belong to an X motif with each position having $\geq 90\%$ X universality ($n_X \geq 119$ sequences in the alignment).

To evaluate the statistical significance of both the occurrence number and the nucleotide length of the uX motifs identified in the rRNA alignments, we also defined universal random motifs (Definition 2.12) from 100 random codes (Appendix Table V). Motifs from each of these random codes were identified in the rRNA multiple sequence alignments and their universality was calculated with similar constraints as for the universal X motifs.

Definition 4.4. A 'universal R random motif', denoted as 'uR motif' is defined as a region in the multiple sequence alignment (Definition 4.1) with at least 6 consecutive positions belonging to a R motif (for each 100 R random code) and having $\geq 90\%$ R universality.

It is important to note that, in the case of the rRNA, because the notion of "reading frame" is not relevant, we searched for X motifs starting at any position in the sequences. Thus, the trinucleotides of the X motifs in the different organisms are not necessarily in the same "frames", e.g. one of the uX motifs in the SSU covers the sequences AG, GTA, ACC in *E. coli* and A, GGT, TTC, G in *Homo sapiens*.

4.6. Mapping *uX* motifs to the "rRNA common core"

Primitive translation systems were much simpler than the modern ribosome; nevertheless these primitive systems have evolved gradually to what we know today as the modern ribosome. In section 1.1.4 we discussed the structure of the modern ribosome and its function. The modern ribosome is a highly sophisticated translational machinery consisting of two subunits that translates the mRNA sequences into proteins. Each of its subunits is a large nucleoprotein complex that come together during the initiation process of protein synthesis and eventually split again in tandem with the release of the synthesized protein. In bacteria and archaea, the LSU contains a 23S rRNA and a 5S rRNA, whereas the SSU contains the 16S rRNA. In eukaryotes, the LSU contains a 28S rRNA, a 5S rRNA and a 5.8S rRNA, whereas the SSU contains the 18S rRNA. The rRNA sequences which are considered to be the most conserved over evolution, contain information about the evolution of the translation machinery. In fact, a "common core" of rRNA (Figure 4.2) was identified by the comparison of 3D ribosome structures from different organisms (Hsiao et al., 2009; Petrov et al., 2015; Opron & Burton, 2018).

In Petrov et al., 2015, the authors proposed the evolution of the ribosome in a chronological manner, based on 3D comparative study of the rRNAs. The "accretion model" presented in this analysis explains how the translation system expanded in terms of insertion of rRNA segments from proto-translation systems in the prebiotic environment to form a rRNA common core based on the acquisition of important ribosomal functions. This rRNA common core was found to be conserved over the entire phylogenetic tree covering all the three domains of life, especially in terms of secondary/tertiary structures. Even after the formation of the common core of rRNA which is conserved over all domains, expansion segments have been added to the eukaryotic rRNAs without altering the structure of the common core. Most importantly, this expansion of rRNA segments in eukaryotes is excluded from the important functional regions of the ribosome such as the peptidyl transferase center (PTC), decoding center, tRNA binding sites among other important regions. Here we show the 'universal X motifs' in the common core of rRNA which are preserved in all the three domains of life, most importantly in the regions identified for carrying out important ribosomal functions.

In order to investigate the presence of X motifs in the rRNAs, we identified universal X motifs (denoted uX motifs) in multiple sequence alignments of the LSU rRNAs (23S/28S and 5S) and SSU rRNAs (16S/18S) for 133 representative organisms covering all three domains of life. In Figure 4.4, we provide a detailed view of the distribution of X motifs in the rRNA multiple sequence alignments, including the coverage of X motifs in bacteria, archaea and eukaryote individually, in bacteria and archaea taken together, and 'universal X motifs' preserved in all the three domains of life. To recall, we defined X motifs as universal X motifs if they are present in at least 90% of the aligned sequences and have a length of at least 6 consecutive nucleotides. It is important to note that uX motifs are not necessarily conserved in terms of the nucleotide sequence, e.g. the SSU trinucleotide 1505-1507 is conserved in bacteria and archaea as GUA and conserved in eukaryotes as GUU, thus affecting the sequence conservation but not the universality of the X trinucleotide as both GUA and GUU belong to the X circular code. For the nucleotide and trinucleotide composition of universal X motifs, refer to Nucleotide and trinucleotide composition of uX motifs.

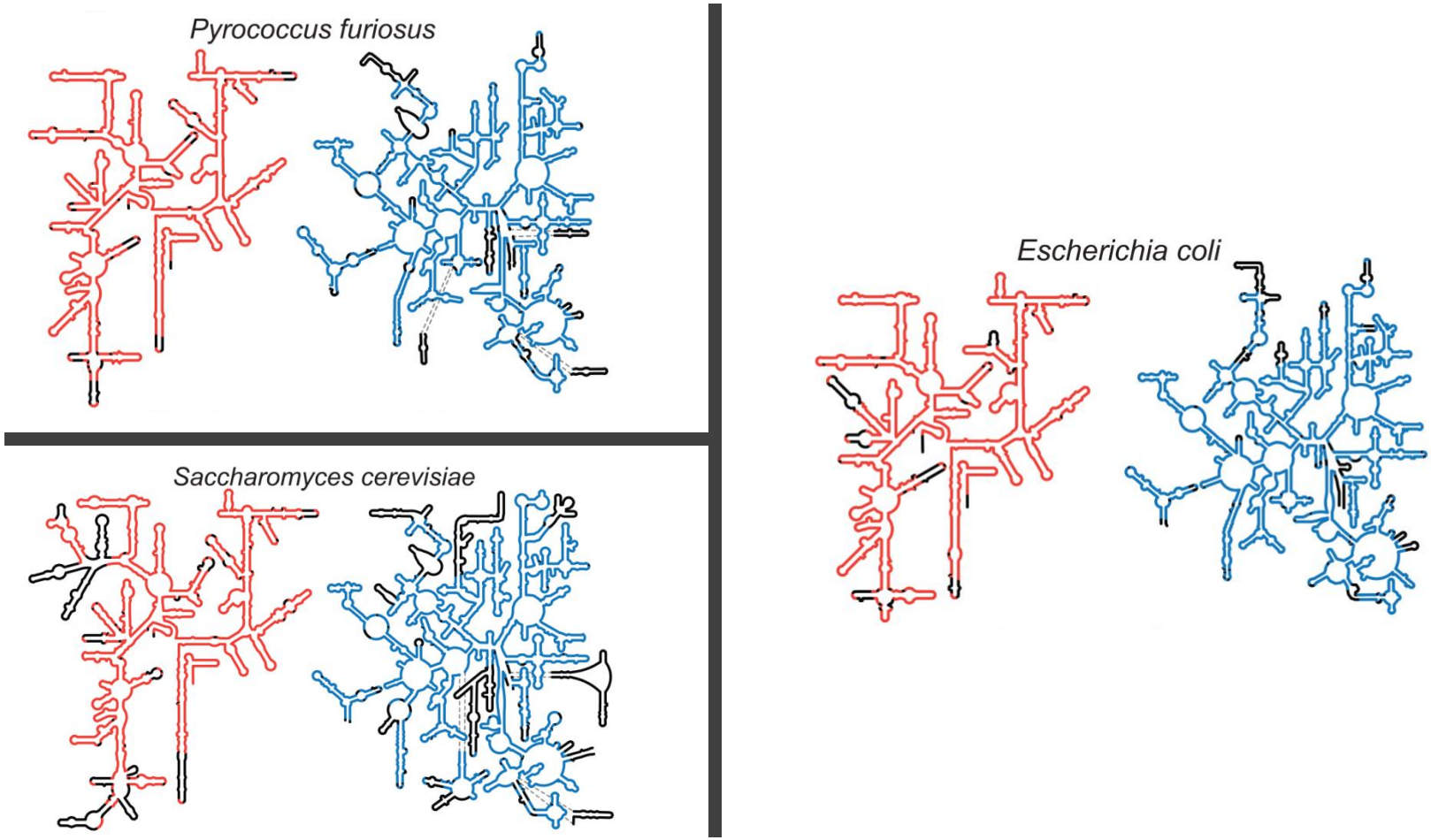


Figure 4.2. Common core of rRNA shown in the rRNA secondary structures of *Pyrococcus furiosus* (archaea), *Saccharomyces cerevisiae* (eukaryota) and *Escherichia coli* (bacteria). Red segments for the SSU and blue segments for the LSU depict the rRNA common core. (Pictures were taken from http://apollo.chemistry.gatech.edu/RibosomeGallery/index.html.)

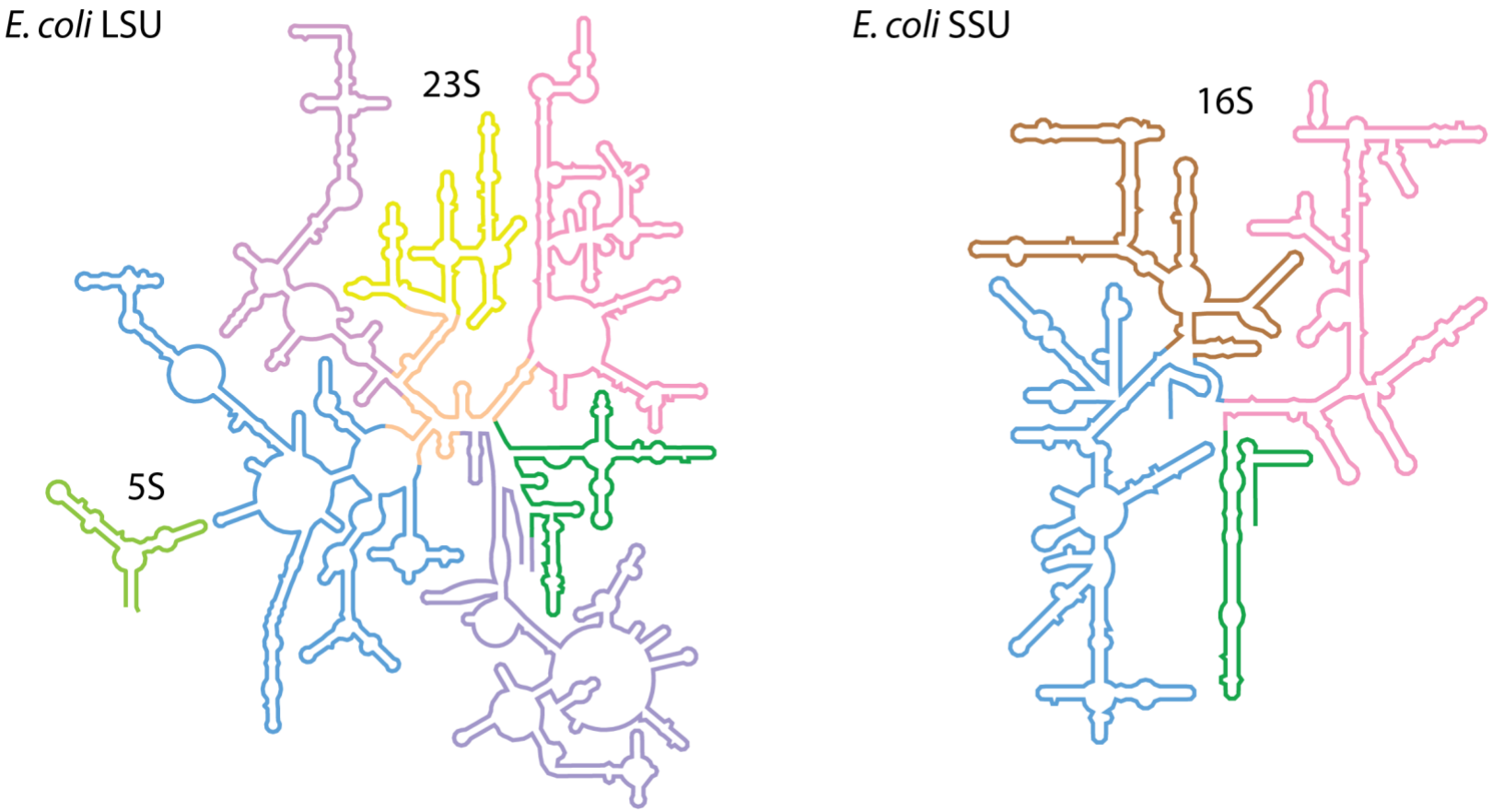


Figure 4.3. rRNA secondary structure of *Escherichia coli*. Colored regions depict the various structural domains in the SSU: light blue for domain 5', olive for the central domain, pink for 3'M and green for 3'm domains and in the LSU: magenta for domain I, blue for domain II, violet for domain III, light orange for domain 0, yellow for domain IV, pink for domain V, green for domain VI. (Picture taken from http://apollo.chemistry.gatech.edu/RibosomeGallery/index.html.)

Figure 4.4. Distribution of the number and total nucleotide lengths of *X* motifs in the SSU (16S/18S) and LSU (23S/28S) rRNA multiple sequence alignments conserved in the 3 domains of life, in bacteria and archaea only, and 'universal *X* motifs' conserved in all 3 domains. *X* motifs are considered to be universal if they are present in at least 90% of the aligned sequences and have a length of at least 6 consecutive nucleotides.

We identified 13 uX motifs in the SSU rRNA alignment (Table 4.1 and Figure 4.5) and 19 uX motifs in the LSU rRNA alignment (Table 4.2 and Figure 4.6), while no uX motifs were found in the 5S rRNA alignment. In Table 4.1 and Table 4.2, the uX motifs are labeled according to the "accretion model" (Petrov et al., 2015), and using capital letters for LSU motifs and small letters for SSU motifs. We provide the location of these uX motifs according to structural domains (Figure 4.3) and helices. We used the *Escherichia coli* numbering (16S and 23S rRNA) for the nucleotide positions in both the LSU and the SSU rRNA multiple sequence alignments. The commas in the given rRNA sequences represent the decomposition of the uX motifs into trinucleotides of the X circular code along with suffixes and prefixes of trinucleotides belonging to the X circular code. The underlined nucleotides in the uX motifs are present in more than 90% of the sequences in the multiple sequence alignments.

Table 4.1. Location of the 13 uX motifs in the SSU rRNA alignment (prokaryotic 16S and eukaryotic 18S), according to structural domains and helices (*E. coli* numbering). uX motifs are labeled according to the accretion model.

uX motif	Start	End	Sequence (*E. coli*)	Domain	Helix
a	1396	1404	*AC,ACC,GCC,C*	3'm	h44
b	1492	1501	*G,GGT,GAA,GTC,GTA,AC*	3'm	h44
c	1503	1514	*AG,GTA,ACC,GTA,GG*	3'm	h45
d	918	926	*A,ATT,GAC,GG*	3'M	h28
e	789	797	*TA,GAT,ACC,CTG,GTA,GTC,CA*	C	h24
f	1368	1377	*AC,GGT,GAA,TAC,GTT,C*	3'M	h43
g	520	525	*GC,CAG,CAG,C*	5'	h18
h	527	536	*GC,GGT,AAT,AC*	5'	h18
i	1186	1197	*G,GAT,GAC,GTC,AA*	3'M	h34
j	1333	1338	*AT,GAA,GTC,GG*	3'M	h42
k	249	257	*TA,GTA,GGT,GG*	5'	h11
l	1064	1073	*GT,CAG,CTC,GT*	3'M	h34, h35
m	1099	1107	*GC,AAC,GAG,C*	3'M	h35

Table 4.2. Location of the 19 *uX* motifs in the LSU rRNA alignment (prokaryotic 23S and eukaryotic 25S/28S), according to structural domains and helices (*E. coli* numbering). *uX* motifs are labeled according to the accretion model.

uX motif	Start	End	Sequence (*E. coli*)	Domain	Helix
A	2479	2484	*AT,ATC,GAC,GGC,GGT,GTT,T*	V	H89
B	2497	2511	*AC,CTC,GAT,GTC,GGC,T*	V	H89, H90
C	2516	2525	*AC,ATC,CTG,GG*	V	H91
D	2574	2586	*GC,GAG,CTG,GGT,TT*	V	H90, H93
E	2587	2596	*AG,AAC,GTC,GT*	V	H90, H93
F	2550	2561	*G,CTG,TTC,GCC,ATT,TA*	V	H92
G	2010	2015	*GT,GAA,ATT,GAA,CTC,GC*	0	H26a
H	513	519	*T,GAA,ACC,GT*	I	H2
I	724	732	*AA,CTG,GAG,GAC,C*	II	H34
J	699	708	*G,CAG,GTT,GAA,GGT,T*	II	H34
K	1975	1983	*GT,AAT,GAT,GGC,CAG,GC*	IV	H65, H67
L	804	812	*AG,CTG,GTT,CTC,C*	II	H32
M	1896	1905	*G,GTA,AAC,GGC,GGC,C*	IV	H68
N	1848	1853	*G,GAA,GGT,TA*	IV	H68
O	2654	2662	*AG,TAC,GAG,A*	V1	H95
P	1124	1131	*G,GAA,GAT,GTA,AC*	II	H41, H42
Q	1057	1062	*GC,CAG,GAT,GTT,GGC,TT*	II	H43
R	47	55	*A,GGC,GAT,GAA,GG*	I	H5
S	1388	1398	*AA,CAG,GTT,AAT,ATT,C*	III	H53

In Figure 4.5 and Figure 4.6, we show the mean percentage conservation of the universal *X* motifs identified in the rRNA multiple sequence alignments. Coloured regions indicate rRNA domains; positions in Table 4.1 for the SSU domains and positions in Table 4.2 for the LSU domains. To recall, we used the *Escherichia coli* numbering (16S and 23S rRNA) for the nucleotide position in both the LSU and the SSU rRNA multiple sequence alignments.

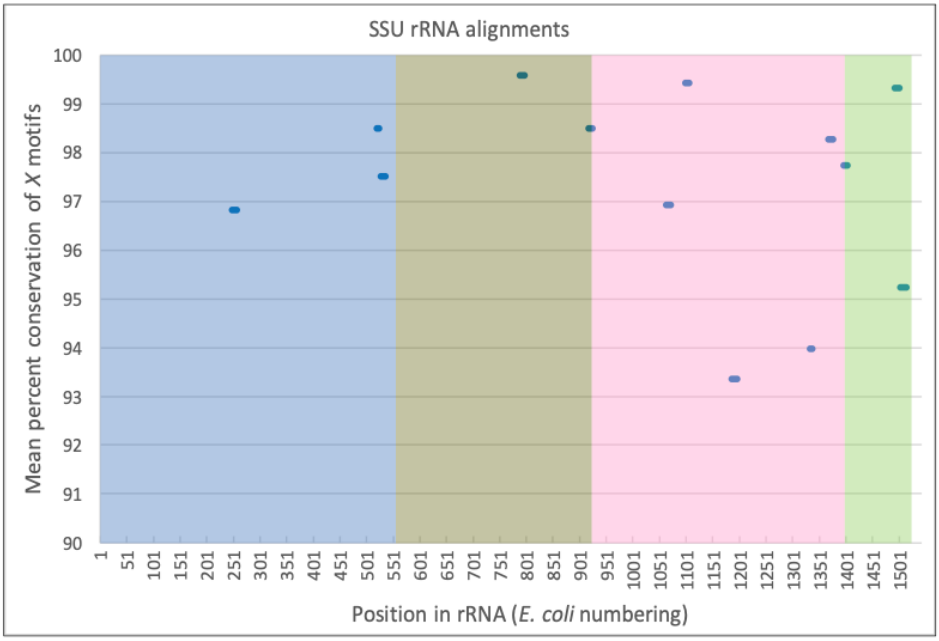


Figure 4.5. Location of the 13 *uX* motifs in the SSU rRNA multiple sequence alignment (prokaryotic 16S and eukaryotic 18S). The abscissa gives the nucleotide position referenced according to the *E. coli* 16S rRNA and the ordinate indicates the level of sequence conservation observed in the *uX* motifs. Colored boxes indicate rRNA domains (positions in Table 4.1) for the SSU: light blue for domain 5', olive for the central domain, pink for 3'M and green for 3'm domains.

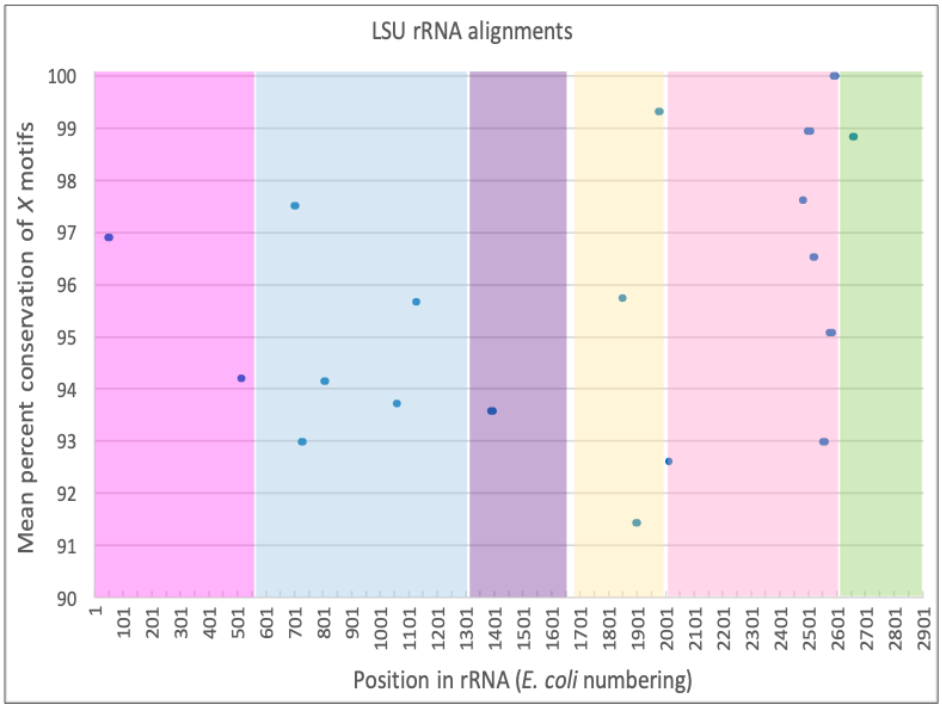


Figure 4.6. Location of the 19 *uX* motifs in the LSU rRNA alignments (prokaryotic 23S and eukaryotic 25S/28S). The abscissa gives the nucleotide position referenced according to the *E. coli* 23S rRNA and the ordinate indicates the level of sequence conservation observed in the *uX* motifs. Colored boxes indicate rRNA domains (positions in Table 4.2) for the LSU: magenta for domain I, blue for domain II, violet for domain III, white for domain 0, yellow for domain IV, pink for domain V, green for domain VI.

We calculated the mean sequence conservation of the SSU and LSU rRNA multiple sequence alignments as well as the mean sequence conservation of the universal X motifs identified in the rRNA multiple sequence alignments. The mean sequence conservation across the full length of the SSU and LSU is 65% and 62% respectively, whereas for the uX motifs mean sequence conservation is 81%.

In Figure 4.7, we provide a detailed comparison of the nucleotide sequence conservation and the universality of the uX motifs in the rRNA multiple sequence alignments containing 133 species. In order to do so, we show the position of the nucleotide conservation (90% identity), universally conserved X codons (conserved in >90% of the sequences) and the universal X motifs (uX motifs) in the 16S/18S rRNA, 23S/28S rRNA and the 5S rRNA multiple sequence alignments. We would like to emphasize the fact that, the universality of the X motifs does not necessarily depend on the sequence conservation, knowing that the rRNA sequences are highly conserved.

Moreover, in Table 4.3 we show the number of conserved positions (nucleotides) and the number of uX motif positions (universally conserved) in each of the 3 rRNA multiple sequence alignments. As shown in columns 2 and 3, we calculated the respective percentage of conservation of nucleotides and the uX motifs in terms of total sequence length for each of the 3 rRNA alignments (*E. coli* numbering). In fact, nearly >28% of the 16S/18S and 23S/28S rRNA multiple sequence alignments covered by uX motifs are not conserved in terms of sequence conservation (column 4). As mentioned earlier, in the 5S alignment we did not identify any universally conserved uX motifs. Also, we calculated the number of conserved positions (among the universally conserved positions) that are not in the uX motifs (column 5).

Table 4.3. Comparison of universally conserved positions and uX motif positions in each of the three rRNA multiple sequence alignments containing 133 organisms (*E. coli* numbering).

rRNA multiple sequence alignments	Total sequence length (*E. coli*)	No. of conserved positions and percentage of total length	No. of uX motif positions and percentage of total length	No. (and %) of uX motif positions that are not conserved	No. (and %) of conserved positions that are not uX motifs
16S/18S	1542	490 (31.8%)	121 (7.8%)	34 (28.1%)	403 (82.2%)
23S/28S	2904	870 (30.0%)	175 (6.0%)	50 (28.6%)	745 (85.6%)
5S	120	20(16.7%)	0 (0%)	0 (0%)	20 (100%)

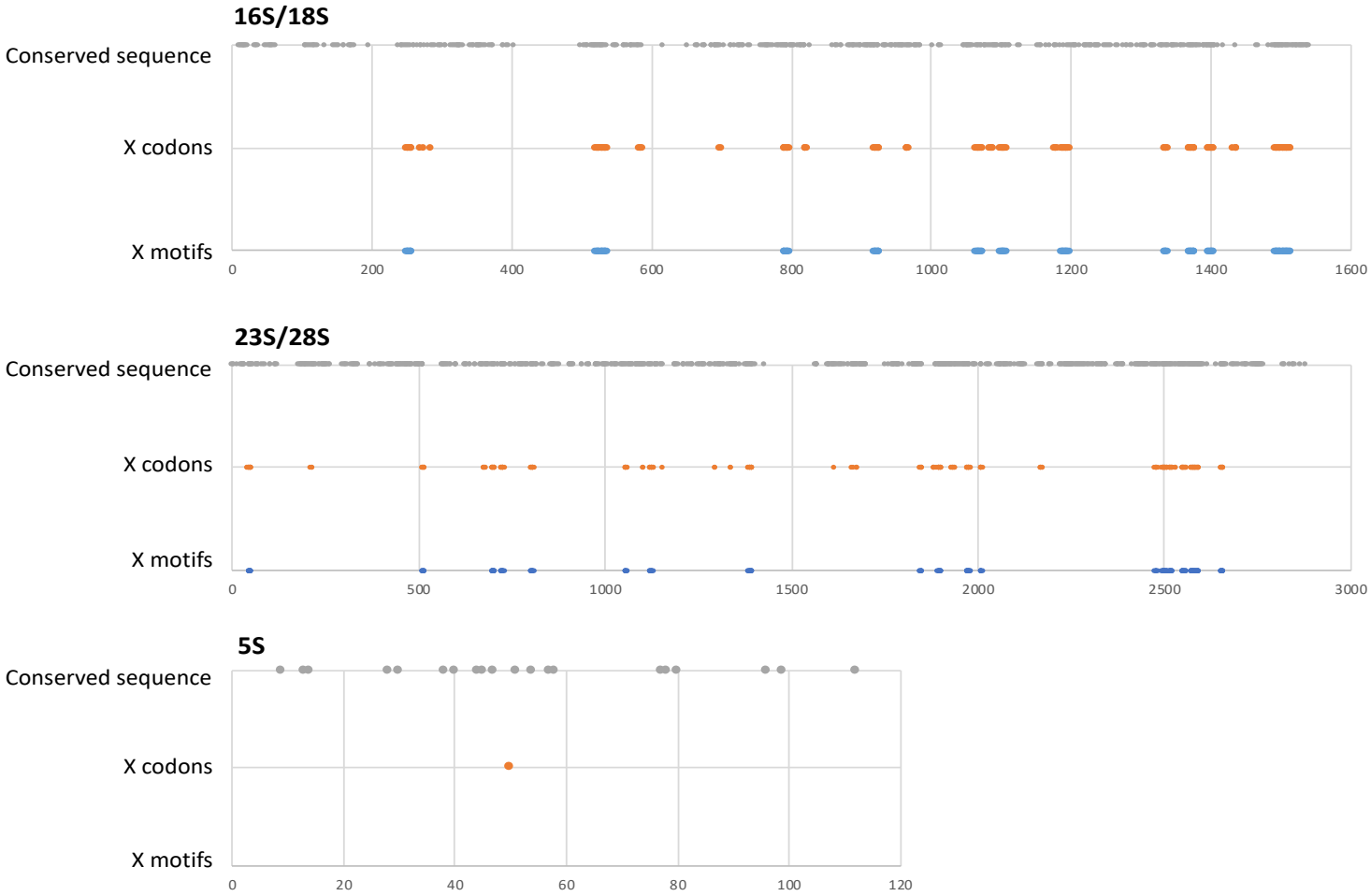


Figure 4.7. Comparison of sequence conservation (90% identity), universal *X* codons and the universal *X* motifs (*uX* motifs) in the rRNA multiple sequence alignments containing 133 species. The x-axis shows the position in each of the three multiple sequence alignments corresponding to the 16S/18S rRNA, the 23S/28S rRNA and the 5S rRNA. Positions with *X* codons in >90% of the sequences are identified as universal *X* codons.

In order to further test for any correlation between the sequence conservation and the universality of X motifs, we performed statistical correlation tests. We compared the sequence conservation with the X motif universality for each position in the universally identified 13 uX motifs in the SSU (Table 4.4) and 19 uX motifs in the LSU (Table 4.5). We observed no significant correlation between universality of X motifs and sequence conservation. For the 13 uX motifs in the SSU rRNA multiple sequence alignments, the following results were obtained from statistical correlation tests: Pearson correlation coefficient $r = 0.37$ ($p < 10^{-4}$), Spearman correlation coefficient $\rho = 0.25$ ($p = 0.006$) and Kendall coefficient $\tau = 0.19$ ($p = 0.007$); which suggests no significant positive correlation. Furthermore, for the 13 uX motifs in the SSU, a two-tailed matched sample signed ranks Wilcoxon test showed that the two distributions are significantly different ($p < 10^{-3}$). We observed similar results from statistical correlation tests for the 19 uX motifs in the LSU rRNA multiple sequence alignments: Pearson correlation coefficient $r = 0.04$, Spearman correlation coefficient $\rho = 0.07$ and Kendall coefficient $\tau = 0.05$; suggesting no significant correlation between sequence conservation and universality of X motifs. Furthermore, for the 19 uX motifs in the LSU, a two-tailed matched sample signed ranks Wilcoxon test showed that the 2 distributions are significantly different ($p < 10^{-4}$). These results demonstrate that the X circular code property exists in certain regions of the ribosome, in addition to sequence level constraints. In the following sections, we will discuss in detail how these regions are essential for various ribosome functions.

Next we will highlight the universal X motifs identified in the rRNA multiple sequence alignments (SSU and LSU) in terms of their coverage in the various RNA domains, which are based on the RNA secondary structures and helices.

Table 4.4. Comparison of *X* motif universality (number of species having an *X* motif) and sequence conservation (percent sequence identity) for each position in the 13 universal *X* motifs (*uX* motifs) in the SSU rRNA multiple sequence alignments (Table 4.1). Each row represents one *uX* motif and each column represents one position within the *uX* motif.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	*X* universality	97.7	97.7	97.7	97.7	97.7	97.7	97.7	97.7	97.7						
	Sequence conservation	97	100	100	100	95.6	100	100	100	98.5						
b	*X* universality	100	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3					
	Sequence conservation	100	100	97	100	100	98.5	98.5	98.5	100	100					
c	*X* universality	94.7	94.7	94.7	94.7	95.5	95.5	95.5	98.5	95.5	95.5	94	94			
	Sequence conservation	100	100	100	100	68.7	56.1	98.5	84.4	95.6	100	100	57.1			
d	*X* universality	98.5	98.5	98.5	98.5	98.5	98.5	98.5	98.5	98.5						
	Sequence conservation	100	100	100	97	100	59.5	98.5	100	100						
e	*X* universality	100	100	99.3	99.3	99.3	99.3	99.3	100	100						
	Sequence conservation	72.2	100	100	100	94.1	100	100	98.5	63.9						
f	*X* universality	97.7	98.5	98.5	98.5	99.3	99.3	99.3	99.3	99.3	93.2					
	Sequence conservation	70.9	92.7	92.6	61.6	89.9	100	100	66.3	97	83.1					
g	*X* universality	98.5	98.5	98.5	98.5	98.5	98.5									
	Sequence conservation	100	100	100	61.7	97	100									
h	*X* universality	97.7	97.7	97	97	97	97.7	97.7	97.7	97.7	97.7	91	91.7	93.2	93.2	
	Sequence conservation	100	100	100	100	98.5	97	100	75.9	64.8	98.5	46.8	33.4	48.9	48.2	
i	*X* universality	91	90.2	92.5	92.5	92.5	92.5	92.5	94	94	94	100	94.7	90.2	90.2	90.2
	Sequence conservation	80.9	79.7	48.3	50.3	44.5	100	64.8	64.8	37.3	74.3	97	49.7	79.8	100	97
j	*X* universality	94	94	94	94	94	94									
	Sequence conservation	98.5	54.6	33.4	33.8	100	100									

k	*X* universality	92.5	93.2	97.7	97.7	97.7	97.7	97.7	98.5	98.5	
	Sequence conservation	92.7	45.3	100	57.8	72	100	87.3	67.7	55.6	
l	*X* universality	97.7	97.7	97.7	97.7	97	97	97	96.2	95.5	95.5
	Sequence conservation	63.6	82.1	64.8	95.6	100	57	95.6	67.1	100	100
m	*X* universality	99.3	99.3	99.3	99.3	99.3	99.3	99.3	100	100	
	Sequence conservation	58.1	55	100	100	98.5	97	94.1	65.6	100	

Table 4.5. Comparison of *X* motif universality (number of species having an *X* motif) and sequence conservation (percent sequence identity) for each position in the 19 universal *X* motifs (*uX* motifs) in the LSU rRNA multiple sequence alignments (Table 4.2). Each row represents one *uX* motif and each column represents one position within the *uX* motif.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	*X* universality	97	97.7	97.7	97.7	97.7	97.7									
	Sequence conservation	63.2	98.5	100	95.6	98.5	57									
B	*X* universality	99.3	99.3	99.3	100	99.3	99.3	100	98.5	98.5	98.5	98.5	98.5	98.5	98.5	98.5
	Sequence conservation	63.6	100	55.3	95.6	94.1	98.5	97	95.6	100	100	98.5	100	88.4	74.3	100
C	*X* universality	100	100	94	94	94	97	97.7	97.7	97.7	93.2					
	Sequence conservation	34	62.4	100	81.7	100	65.3	100	62	86	40.7					
D	*X* universality	94	94	94	94	94	94	94	94	97	97	97	97	96.2		
	Sequence conservation	95.5	60.2	100	100	100	88.4	100	100	100	98.5	100	100	54.4		
E	*X* universality	100	100	100	100	100	100	100	100	100	100					
	Sequence conservation	100	83.2	95.6	49.6	100	98.5	98.5	100	100	84.7					
F	*X* universality	91.7	91.7	93.2	93.2	93.2	93.2	93.2	93.2	93.2	94	94	91.7			
	Sequence conservation	64.8	50.7	100	100	86	87.3	100	62.4	100	100	39.1	60.1			
G	*X* universality	90.2	92.5	93.2	93.2	93.2	93.2									
	Sequence conservation	89.8	49.7	100	100	98.5	51									
H	*X* universality	94.7	94.7	94	94	93.2	94	94.7								
	Sequence conservation	100	98.5	94.1	69.2	77.4	62	37.3								
I	*X* universality	91	90.2	90.2	90.2	91	96.2	95.5	96.2	96.2						
	Sequence conservation	97	100	98.5	100	87.3	100	29.5	79.8	88.6						
J	*X* universality	97.7	98.5	98.5	98.5	97	97	97.7	97	97	96.2					
	Sequence conservation	97	88.6	94.1	31.2	75.1	95.5	100	100	92.7	40.3					

K	*X* universality	94.7	94.7	100	100	100	100	100				
	Sequence conservation	100	39.5	46.5	54.9	100	100	67.4				
L	*X* universality	97.7	92.5	92.5	94	94.7	94	94	94	94		
	Sequence conservation	94.1	95.6	100	97	83	100	100	94.1	98.5		
M	*X* universality	90.2	91	94	94	91	91	91	91	91	90.2	
	Sequence conservation	42.6	84.6	97	61.8	100	100	83.2	94.1	86	100	
N	*X* universality	94.7	94.7	96.2	96.2	96.2	96.2					
	Sequence conservation	89.8	39.2	71.5	100	39	100					
O	*X* universality	99.3	98.5	98.5	98.5	98.5	98.5	99.3	99.3	99.3		
	Sequence conservation	97	100	98.5	100	98.5	100	100	97	100		
P	*X* universality	96.2	95.5	95.5	95.5	95.5	94.7	97	95.5			
	Sequence conservation	37.9	82.9	95.5	79.2	57.8	100	97	67.1			
Q	*X* universality	94	94	94	94	93.2	93.2					
	Sequence conservation	62.4	58.4	98.5	100	46.4	70					
R	*X* universality	97.7	97.7	97.7	97	97	97	97	97	94		
	Sequence conservation	83.3	92.7	95.5	49.1	100	100	95.6	94.1	66.5		
S	*X* universality	93.2	93.2	93.2	93.2	94	94	93.2	94	94	94	93.2
	Sequence conservation	87	95.5	98.5	75.5	56.7	94.1	64.2	98.5	94.1	86	94.1

Secondary structures of RNA are important in understanding 3D structures, RNA folding and how the ribosome operates. Previously shown as coloured regions in Figure 4.3, the SSU rRNA is divided into four structural domains: SSU 5', SSU central, SSU 3'M and SSU 3'm; whereas the LSU rRNA is divided into seven structural domains: LSU 0, LSU I, LSU II, LSU III, LSU IV, LSU V and LSU VI.

Table 4.6. Coverage of rRNA structural domains by *uX* motifs (SSU and LSU). For each rRNA domain, domain length corresponds to nucleotide length with start and end position, *uX* motif length is the total length of *X* motifs located in nucleotides and % coverage is the percentage of nucleotides covered by the universal *X* motifs.

rRNA domain	Domain start	Domain end	Domain length	*uX* motif length	% coverage
SSU 5'	1	559	559	25	4.5
SSU central	560	920	361	12	3.3
SSU 3'M	921	1398	478	56	11.7
SSU 3'm	1399	1542	144	28	19.4
Total SSU	1	1542	1542	121	7.8
LSU 0	disjoint	disjoint	159	6	3.8
LSU I	1	561	561	16	2.9
LSU II	587	1250	664	42	6.3
LSU III	1271	1647	377	11	2.9
LSU IV	1679	1989	311	25	8.0
LSU V	2058	2610	553	66	11.9
LSU VI	2626	2895	270	9	3.3
Total LSU	1	2895	2895	175	6

In Table 4.6, we provide the details of each of the RNA structural domains with start and end positions, length in nucleotides, length of *uX* motifs identified in nucleotides and their coverage (%). The overall coverage of the SSU and LSU rRNAs by *uX* motifs (in terms of nucleotides) is quite similar with 7.8% and 6.0% coverage respectively. However, we see that coverage by *uX* motifs across the different RNA structural domains of both subunits is not homogenous (Table 4.6). Notably, with 19.4% of nucleotides in *uX* motifs, the SSU 3'm domain comprising the central pseudoknot (CPK) and the decoding center has the highest coverage. Furthermore, in comparison to the SSU's central domain and the LSU's 0, I, III and VI domains with around 3% coverage, both the SSU 3'M domain corresponding to the 'head' region and the LSU V domain comprising the PTC (peptidyl transferase center) are enriched with nearly 12% coverage. It is important to note that the RNA structural domains enriched with *uX* motifs cover the central pseudoknot (CPK), the peptidyl transferase center (PTC) and the decoding center which are important functional regions of the ribosome and are found to be conserved in all organisms. The central pseudoknot is very important to the structure of the SSU as it links all four domains. The peptidyl transferase center is the ribosomal site where the peptide bond formation takes place during the elongation stage (growing polypeptide chain) of

protein synthesis. These functional regions are considered to be the oldest parts of the ribosome, which played a significant role in the development and functioning of primitive translation systems.

4.7. Comparison of universal *X* motifs (uX) with universal random motifs (uR)

In order to determine the significance of the observed coverage by universal *X* motifs in the rRNA multiple sequence alignments (both occurrence number and nucleotide length), we chose to compare the coverage obtained from universal random motifs (uR motifs from 100 *R* random codes) (Definition 4.4). To recall, these *R* random codes have similar properties to the *X* circular code, except they are not circular. In Figure 4.8 and Figure 4.9, we show the comparison in occurrence number and nucleotide length respectively, between uX motifs and uR motifs identified in the SSU and LSU rRNA multiple sequence alignments. The number and total length of uX motifs are represented by a blue cross, whereas the distribution of the uR random motifs obtained from 100 random codes *R* is indicated by boxplots representing the mean and ± 0.99 confidence interval. In Figure 4.8, we show that the numbers of uX motifs (SSU = 13 uX motifs; LSU = 19 uX motifs) are higher than those obtained for uR motifs: mean number of 10.2 (in SSU) and 12.8 (in LSU). The number of uX motifs is significantly higher than the mean number of uR motifs according to a one-sided Student's t-test: $p \approx 10^{-21}$ for SSU and $p \approx 10^{-33}$ for LSU.

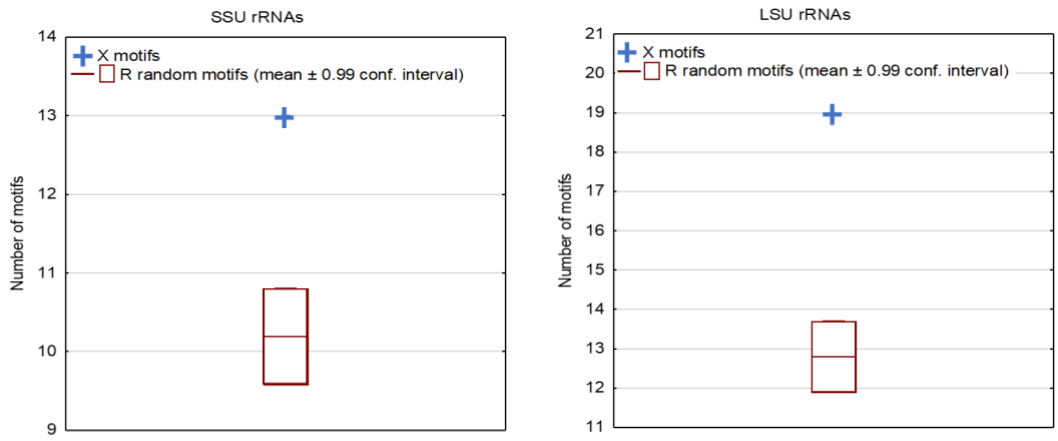


Figure 4.8. Comparison between the number of uX motifs and uR random motifs identified in the SSU and LSU rRNA multiple sequence alignments.

In Figure 4.9 we show that the nucleotide lengths of the uX motifs (SSU = 121 nucleotides; LSU = 175 nucleotides) are higher than the mean nucleotide lengths of uR motifs (SSU = 100.5 nucleotides; LSU = 120.1 nucleotides). The nucleotide length of the uX motifs is also significantly higher than the mean nucleotide length of uR motifs according to a one-sided Student's t-test: $p \approx 10^{-14}$ for SSU and $p \approx 10^{-32}$ for LSU.

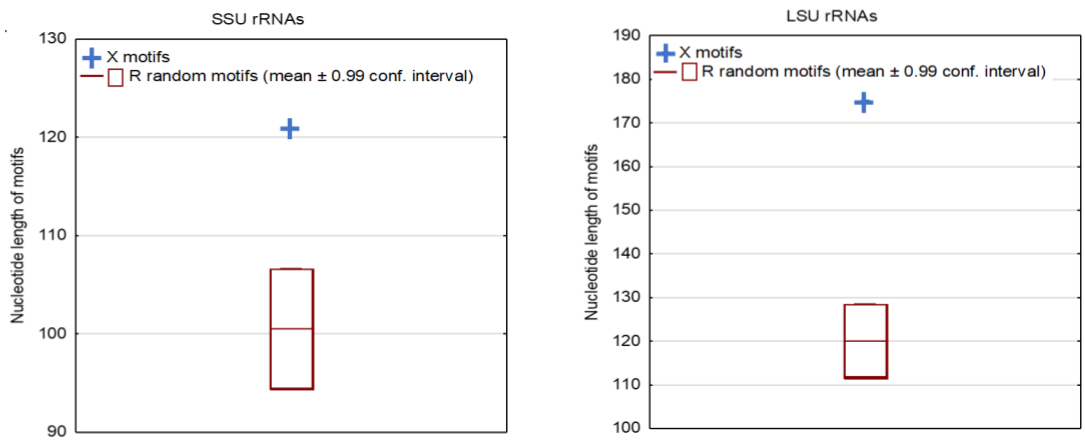


Figure 4.9. Comparison between the nucleotide lengths of uX motifs and uR random motifs identified in the SSU and LSU rRNA multiple sequence alignments.

We also determined how many of the uR motifs display the same level of occurrence and coverage as the uX motifs. In Figure 4.10 we show that none of the R codes had a larger number of motifs than for observed uX motifs (number = 32; 13 in the SSU and 19 in the LSU), whereas 2% of the R codes had the same number of motifs. In Figure 4.11 we show that only 3% of the R codes had a longer total length than the uX motifs (length = 296; 121 in the SSU and 175 in the LSU).

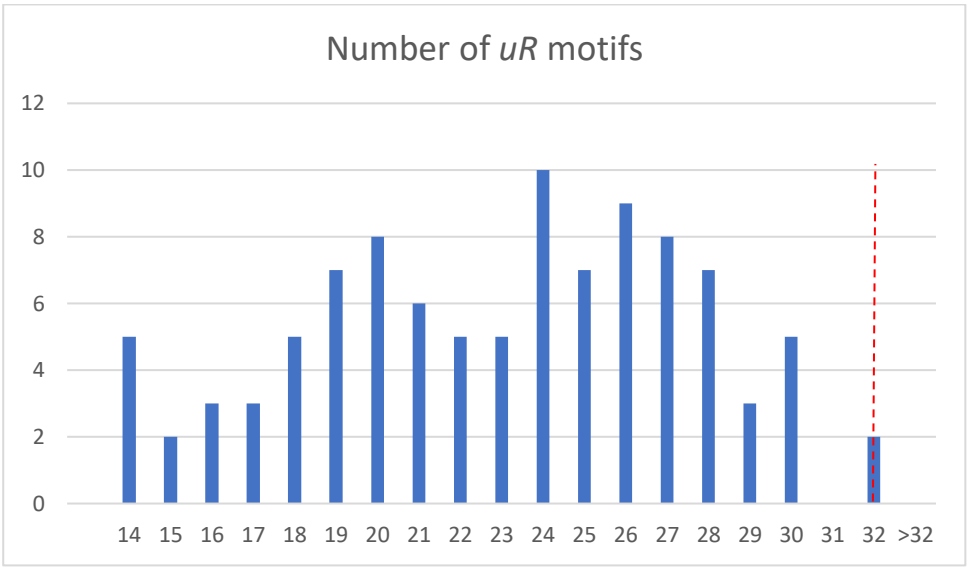


Figure 4.10. Distribution of the total number of the uR random motifs in the SSU and LSU rRNA multiple alignments. The corresponding value for the uX motifs is indicated by a vertical red line. Two percent of the random codes have the same number of universal motifs compared to uX motifs (number = 32; 13 in the SSU and 19 in the LSU).

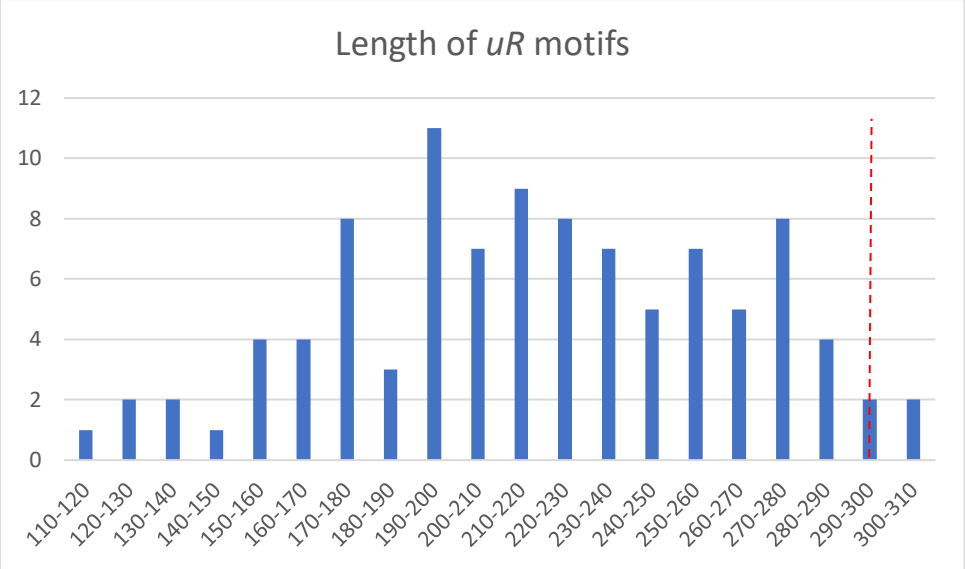


Figure 4.11. Distribution of the total nucleotide lengths of the uR random motifs in the SSU and LSU rRNA multiple alignments. The corresponding value for the uX motifs is indicated by a vertical red line. Three percent of the random codes have the same or larger total length of universal motifs compared to uX motifs (length = 296).

These results reveal an important enrichment of uX motifs in the SSU and LSU rRNAs conserved in the three domains of life, viz. archaea, bacteria and eukaryotes. Next, we will discuss the uX motifs in terms of nucleotide and trinucleotide compositions.

4.8. Nucleotide and trinucleotide composition of uX motifs

We also searched for any compositional bias of the rRNA sequences which might explain the enrichment of X motifs in the rRNA multiple sequence alignments. In order to do so, we calculated the nucleotide composition (Table 4.7) and the trinucleotide composition (Table 4.8) of the rRNA sequences used in the multiple sequence alignments. We observed some bias in the rRNA sequences in terms of nucleotide composition; where G is the most frequent (31.1%) and T is the least frequent (20.5%). This bias can be explained by the GC-content (54.8%) in the rRNA sequences. Whereas, for the X circular code there is no bias in terms of nucleotide composition, with equal frequencies of the four bases A, C, G and T. Concerning the trinucleotide composition of the rRNA sequences, we used the Mann-Whitney U Test between X trinucleotides and non-X trinucleotides to find any kind of bias in terms of trinucleotides. We observed no significant enrichment of X trinucleotides in the rRNA sequences. The mean frequency of X trinucleotides (1.58) is not significantly greater than the mean frequency of non-X trinucleotides (1.55), according to Mann-Whitney U test (z-score $= -0.51419$; $p = 0.61$). We conclude that the enrichment concerns X trinucleotides located within motifs specifically. Therefore, we calculated the nucleotide composition and trinucleotide composition for each of

the 13 uX motifs in the SSU (Table 4.9 and Table 4.11) and the 19 uX motifs in the LSU (Table 4.10 and Table 4.12) rRNA multiple sequence alignments.

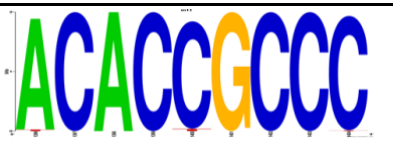
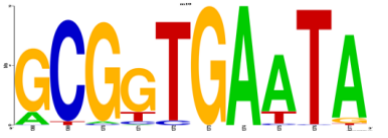
Table 4.7. Nucleotide composition of sequences in the rRNA multiple sequence alignments (SSU and LSU combined) compared to the nucleotide composition of the X circular code.

Nucleotide frequencies (%)	A	C	G	T
rRNA sequences	24.6	23.7	31.1	20.5
X circular code	25.0	25.0	25.0	25.0

Table 4.8. Trinucleotide composition of sequences in the SSU and LSU rRNA alignments. Trinucleotides belonging to the X circular code are highlighted in red. The mean frequency of X trinucleotides (1.58) is not significantly greater than the mean frequency of non-X trinucleotides (1.55).

Trinucleotide frequencies (%)					
AAA	1.9	CTA	1.1	AAC	1.7
AAG	2.4	CTT	1.1	AAT	1.4
ACA	1.1	GCA	1.5	ACC	1.6
ACG	1.4	GCG	2.0	ATC	1.1
ACT	1.2	GCT	1.5	ATT	1.0
AGA	1.6	GGA	2.3	CAG	1.5
AGC	1.9	GGG	3.3	CTC	1.2
AGG	2.4	GTG	2.1	CTG	1.7
AGT	1.6	TAA	1.7	GAA	2.6
ATA	1.0	TAG	1.4	GAC	1.4
ATG	1.3	TAT	0.7	GAG	2.2
CAA	1.3	TCA	1.0	GAT	1.5
CAC	1.0	TCC	1.4	GCC	2.0
CAT	0.9	TCG	1.3	GGC	2.2
CCA	1.2	TCT	1.0	GGT	2.2
CCC	1.9	TGA	1.9	GTA	1.6
CCG	2.3	TGC	1.3	GTC	1.4
CCT	1.5	TGG	2.1	GTT	1.4
CGA	1.8	TGT	1.2	TAC	1.0
CGC	1.5	TTA	1.1	TTC	1.0
CGG	2.3	TTG	1.4		
CGT	1.4	TTT	1.0		

Table 4.9. Nucleotide composition of the 13 *uX* motifs in the SSU rRNA multiple sequence alignments (Table 4.1). Each row represents one *uX* motif and each column represents one position within the *uX* motif. Each cell contains the number (percentage) of species with a given nucleotide at each position in the *uX* motifs.

		1	2	3	4	5	6	7	8	9	10	11	12	13	
a	A	131(98)	0(0)	133(100)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)					
	C	0(0)	133(100)	0(0)	133(100)	130(98)	0(0)	133(100)	133(100)	132(99)					
	G	0(0)	0(0)	0(0)	0(0)	0(0)	133(100)	0(0)	0(0)	0(0)					
	T	2(2)	0(0)	0(0)	0(0)	3(2)	0(0)	0(0)	0(0)	1(1)					
b	A	133(100)	133(100)	0(0)	0(0)	0(0)	1(1)	1(1)	132(99)	133(100)	0(0)				
	C	0(0)	0(0)	1(1)	0(0)	133(100)	0(0)	0(0)	1(1)	0(0)	133(100)				
	G	0(0)	0(0)	131(98)	0(0)	0(0)	132(99)	0(0)	0(0)	0(0)	0(0)				
	T	0(0)	0(0)	1(1)	133(100)	0(0)	0(0)	132(99)	0(0)	0(0)	0(0)				
c	A	133(100)	0(0)	0(0)	0(0)	108(81)	8(6)	0(0)	1(1)	0(0)	0(0)	133(100)	0(0)		
	C	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	132(99)	122(92)	3(2)	0(0)	0(0)	25(19)		
	G	0(0)	133(100)	133(100)	0(0)	2(2)	95(71)	0(0)	2(2)	130(98)	0(0)	0(0)	97(73)		
	T	0(0)	0(0)	0(0)	133(100)	23(17)	30(23)	1(1)	8(6)	0(0)	133(100)	0(0)	11(8)		
d	A		133(100)	133(100)	0(0)	2(2)	0(0)	96(72)	0(0)	0(0)	0(0)				
	C		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	132(99)	0(0)	133(100)				
	G		0(0)	0(0)	0(0)	0(0)	133(100)	37(28)	0(0)	133(100)	0(0)				
	T		0(0)	0(0)	133(100)	131(98)	0(0)	0(0)	1(1)	0(0)	0(0)				
e	A	0(0)	133(100)	0(0)	133(100)	0(0)	133(100)	0(0)	0(0)	3(2)					
	C	22(17)	0(0)	0(0)	0(0)	0(0)	0(0)	133(100)	132(99)	103(77)					
	G	0(0)	0(0)	133(100)	0(0)	4(3)	0(0)	0(0)	0(0)	27(20)					
	T	111(83)	0(0)	0(0)	0(0)	129(97)	0(0)	0(0)	1(1)	0(0)					
f	A	22(17)	0(0)	3(2)	2(2)	0(0)	0(0)	133(100)	106(80)	0(0)	121(92)				
	C	1(1)	128(96)	2(2)	7(5)	6(5)	0(0)	0(0)	3(2)	1(1)	0(0)				
	G	110(83)	0(0)	128(96)	102(77)	0(0)	133(100)	0(0)	0(0)	0(0)	8(6)				
	T	0(0)	5(4)	0(0)	22(17)	126(95)	0(0)	0(0)	23(17)	131(99)	3(2)				

Nucleotide and trinucleotide composition of uX motifs

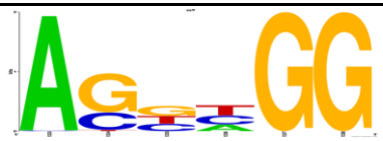
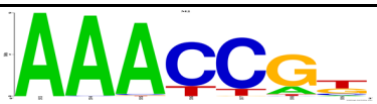
g	A	133(100)	0(0)	0(0)	99(74)	0(0)	0(0)							
	C	0(0)	0(0)	133(100)	34(26)	2(2)	133(100)							
	G	0(0)	133(100)	0(0)	0(0)	131(98)	0(0)							
	T	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)							
h	A	0(0)	0(0)	0(0)	0(0)	0(0)	131(98)	133(100)	0(0)	103(77)	0(0)			
	C	0(0)	133(100)	0(0)	0(0)	0(0)	0(0)	0(0)	15(11)	0(0)	132(99)			
	G	133(100)	0(0)	133(100)	133(100)	1(1)	2(2)	0(0)	3(2)	0(0)	1(1)			
	T	0(0.0)	0(0)	0(0)	0(0)	132(99)	0(0)	0(0)	115(86)	30(23)	0(0)			
i	A		13(10)	14(11)	69(52)	0(0)	42(32)	133(100)	0(0)	30(23)	30(23)	0(0)	131(98)	64(48) 
	C		1(1)	0(0)	0(0)	74(56)	9(7)	0(0)	103(77)	0(0)	0(0)	113(85)	0(0)	0(0)
	G		119(89)	118(89)	62(47)	0(0)	78(59)	0(0)	0(0)	103(77)	36(27)	0(0)	0(0)	69(52)
	T		0(0)	0(0)	2(2)	59(44)	4(3)	0(0)	30(23)	0(0)	67(50)	20(15)	2(2)	0(0)
j	A		132(99)	0(0)	0(0)	34(26)	0(0)	0(0)						
	C		1(1)	40(30)	39(29)	47(35)	0(0)	0(0)						
	G		0(0)	90(68)	52(39)	0(0)	133(100)	133(100)						
	T		0(0)	3(2)	42(32)	52(39)	0(0)	0(0)						
k	A	0(0)	77(59)	0(0)	38(29)	12(9)	0(0)	0(0)	1(1)	40(30)				
	C	5(4)	2(2)	0(0)	1(1)	9(7)	0(0)	0(0)	18(14)	2(2)				
	G	0(0)	6(5)	133(100)	0(0)	0(0)	133(100)	124(93)	6(5)	91(68)				
	T	128(96)	46(35)	0(0)	94(71)	112(84)	0(0)	9(7)	108(81)	0(0)				
l	A	0(0)	0(0)	0(0)	130(98)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)			
	C	26(20)	12(9)	103(78)	0(0)	0(0)	92(69)	3(2)	106(80)	0(0)	0(0)			
	G	103(78)	1(1)	0(0)	3(2)	133(100)	0(0)	0(0)	26(20)	133(100)	0(0)			
	T	3(2)	120(90)	30(23)	0(0)	0(0)	41(31)	130(98)	1(1)	0(0)	133(100)			
m	A	25(19)	1(1)	133(100)	133(100)	0(0)	1(1)	129(97)	29(22)	0(0)				
	C	0(0)	90(68)	0(0)	0(0)	132(99)	1(1)	0(0)	0(0)	133(100)				
	G	98(74)	1(1)	0(0)	0(0)	0(0)	131(98)	4(3)	104(78)	0(0)				
	T	10(8)	41(31)	0(0)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)				

Table 4.10. Nucleotide composition of the 19 *uX* motifs in the LSU rRNA multiple sequence alignments (Table 4.2). Each row represents one *uX* motif and each column represents one position within the *uX* motif. Each cell contains the number (percentage) of species with a given nucleotide at each position in the *uX* motifs.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
A	A	0(0)	0(0)	0(0)	130(98)	0(0)	3(2)										
	C	0(0)	132(99)	0(0)	0(0)	132(99)	36(27)										
	G	32(24)	1(1)	133(100)	3(2)	0(0)	94(71)										
	T	101(76)	0(0)	0(0)	0(0)	1(1)	0(0)										
B	A	103(77)	0(0)	6(5)	0(0)	3(2)	0(0)	131(98)	0(0)	0(0)	0(0)	0(0)	0(0)	3(2)	0(0)	0(0)	
	C	26(20)	133(100)	93(70)	3(2)	129(97)	0(0)	2(2)	3(2)	0(0)	0(0)	132(99)	0(0)	0(0)	113(85)	0(0)	
	G	0(0)	0(0)	0(0)	0(0)	0(0)	132(99)	0(0)	0(0)	133(100)	0(0)	0(0)	133(100)	125(94)	0(0)	0(0)	
	T	4(3)	0(0)	34(26)	130(98)	1(1)	1(1)	0(0)	130(98)	0(0)	133(100)	1(1)	0(0)	5(4)	20(15)	133(100)	
C	A	27(20)	0(0)	133(100)	7(5)	0(0)	28(21)	0(0)	6(5)	0(0)	4(3)						
	C	65(49)	100(75)	0(0)	4(3)	133(100)	104(78)	0(0)	1(1)	10(8)	33(25)						
	G	33(25)	0(0)	0(0)	2(2)	0(0)	0(0)	0(0)	102(77)	123(92)	76(57)						
	T	8(6)	33(25)	0(0)	120(90)	0(0)	1(1)	133(100)	24(18)	0(0)	20(15)						
D	A	1(1)	0(0)	0(0)	133(100)	0(0)	5(4)	0(0)	0(0)	0(0)	1(1)	0(0)	0(0)	0(0)			
	C	2(2)	36(27)	0(0)	0(0)	0(0)	125(94)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	46(35)			
	G	130(98)	0(0)	133(100)	0(0)	133(100)	0(0)	0(0)	133(100)	133(100)	132(99)	0(0)	0(0)	0(0)			
	T	0(0)	97(73)	0(0)	0(0)	0(0)	3(2)	133(100)	0(0)	0(0)	0(0)	133(100)	133(100)	87(65)			
E	A	133(100)	10(8)	130(98)	67(50)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)						
	C	0(0)	2(2)	0(0)	66(50)	133(100)	0(0)	1(1)	133(100)	0(0)	11(8)						
	G	0(0)	121(91)	0(0)	0(0)	0(0)	132(99)	0(0)	0(0)	133(100)	0(0)						
	T	0(0)	0(0)	3(2)	0(0)	0(0)	1(1)	132(99)	0(0)	0(0)	122(92)						
F	A		30(23)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	33(25)	0(0)	0(0)	32(24)	100(75)			
	C		0(0)	82(62)	0(0)	0(0)	10(8)	9(7)	133(100)	0(0)	133(100)	133(100)	75(56)	0(0)			
	G		103(77)	3(2)	0(0)	133(100)	0(0)	0(0)	0(0)	100(75)	0(0)	0(0)	10(8)	8(6)			
	T		0(0)	48(36)	133(100)	0(0)	123(92)	124(93)	0(0)	0(0)	0(0)	0(0)	16(12)	25(19)			
G	A	1(1)	0(0)	0(0)	133(100)	132(99)	90(68)										
	C	2(2)	64(48)	0(0)	0(0)	1(1)	28(21)										
	G	126(95)	0(0)	133(100)	0(0)	0(0)	14(11)										
	T	4(3)	69(52)	0(0)	0(0)	0(0)	1(1)										
H	A	133(100)	132(99)	129 (97)	0(0)	0(0)	24(18)	8(6)									
	C	0(0)	1(1)	2(2)	108(81)	116(87)	1(1)	15(11)									
	G	0(0)	0(0)	1(1)	0(0)	1(1)	102(77)	42(32)									
	T	0(0)	0(0)	1(1)	25(19)	16(12)	6(5)	68(51)									
I	A	0(0)	0(0)	1(1)	133(100)	9(7)	0(0)	46(35)	0(0)	0(0)							
	C	2(2)	0(0)	0(0)	0(0)	0(0)	0(0)	43(33)	118(89)	125(94)							
	G	0(0)	133(100)	132(99)	0(0)	124(93)	133(100)	7(5)	0(0)	0(0)							
	T	131(98)	0(0)	0(0)	0(0)	0(0)	0(0)	36(27)	15(11)	8(6)							

Nucleotide and trinucleotide composition of uX motifs

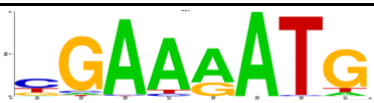
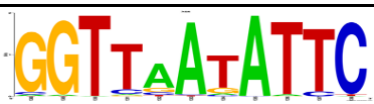
J	A	131(98)	8(6)	4(3)	27(20)	0(0)	2(2)	133(100)	133(100)	5(4)	1(1)			
	C	0(0)	0(0)	0(0)	8(6)	18(14)	0(0)	0(0)	0(0)	0(0)	73(55)			
	G	2(2)	125(94)	129(97)	52(39)	0(0)	130(98)	0(0)	0(0)	128(96)	39(30)			
	T	0(0)	0(0)	0(0)	46(35)	114(86)	0(0)	0(0)	0(0)	0(0)	19(14)			
K	A	5(4)	1(1)	133(100)	133(100)	0(0)	3(2)	130(98)	2(2)	50(38)				
	C	0(0)	24(18)	0(0)	0(0)	106(80)	1(1)	1(1)	25(19)	18(14)				
	G	70(53)	13(10)	0(0)	0(0)	0(0)	129(97)	0(0)	64(48)	33(25)				
	T	58(44)	95(71)	0(0)	0(0)	27(20)	0(0)	2(2)	42(32)	32(24)				
L	A	129(97)	0(0)	0(0)	0(0)	5(4)	0(0)	0(0)	4(3)	1(1)				
	C	4(3)	0(0)	133(100)	0(0)	6(5)	0(0)	0(0)	0(0)	132(99)				
	G	0(0)	130(98)	0(0)	2(2)	121(91)	133(100)	0(0)	0(0)	0(0)				
	T	0(0)	3(2)	0(0)	131(98)	1(1)	0(0)	133(100)	129(97)	0(0)				
M	A		39(29)	1(1)	0(0)	102(77)	133(100)	133(100)	0(0)	4(3)	0(0)	0(0)		
	C		11(8)	10(8)	2(2)	6(5)	0(0)	0(0)	121(91)	0(0)	10(8)	133(100)		
	G		77(58)	122(92)	0(0)	23(17)	0(0)	0(0)	10(8)	129(97)	123(92)	0(0)		
	T		6(5)	0(0)	131(98)	2(2)	0(0)	0(0)	2(2)	0(0)	0(0)	0(0)		
N	A		126(95)	0(0)	3(2)	0(0)	0(0)	133(100)						
	C		5(4)	22(17)	19(14)	0(0)	28(21)	0(0)						
	G		0(0)	69(52)	111(83)	0(0)	34(26)	0(0)						
	T		2(2)	42(32)	0(0)	133(100)	71(53)	0(0)						
O	A	131(98)	0(0)	0(0)	133(100)	0(0)	0(0)	133(100)	2(2)	133(100)				
	C	0(0)	0(0)	1(1)	0(0)	132(99)	0(0)	0(0)	0(0)	0(0)				
	G	1(1)	133(100)	0(0)	0(0)	0(0)	133(100)	0(0)	131(98)	0(0)				
	T	1(1)	0(0)	132(99)	0(0)	1(1)	0(0)	0(0)	0(0)	0(0)				
P	A		4(3)	4(3)	130(98)	118(89)	94(71)	133(100)	0(0)	10(8)				
	C		71(53)	5(4)	2(2)	6(5)	0(0)	0(0)	2(2)	1(1)				
	G		24(18)	121(91)	0(0)	0(0)	38(29)	0(0)	0(0)	108(82)				
	T		34(26)	3(2)	1(1)	9(7)	0(0)	0(0)	131(98)	12(9)				
Q	A	100(75)	1(1)	0(0)	0(0)	7(5)	17(13)							
	C	33(25)	4(3)	1(1)	0(0)	0(0)	0(0)							
	G	0(0)	97(73)	132(99)	0(0)	51(38)	110(83)							
	T	0(0)	31(23)	0(0)	133(100)	75(56)	6(5)							
R	A		0(0)	0(0)	130(98)	25(19)	0(0)	133(100)	130(98)	4(3)	28(21)			
	C		121(92)	0(0)	0(0)	13(10)	0(0)	0(0)	3(2)	0(0)	0(0)			
	G		0(0)	128(97)	0(0)	6(5)	133(100)	0(0)	0(0)	129(97)	105(79)			
	T		11(8)	4(3)	2(2)	89(67)	0(0)	0(0)	0(0)	0(0)	0(0)			
S	A		3(2)	2(2)	1(1)	3(2)	98(74)	129(97)	5(4)	132(99)	3(2)	0(0)	1(1)	
	C		5(4)	0(0)	0(0)	12(9)	11(8)	2(2)	5(4)	1(1)	0(0.0)	10(8)	129(97)	
	G		124(93)	130(98)	0(0)	3(2)	18(14)	0(0)	18(14)	0(0)	1(1)	0(0)	0(0)	
	T		1(1)	1(1)	132(99)	115(86)	6(5)	2(2)	105(79)	0(0)	129(97)	123(92)	3(2)	

Table 4.11. Trinucleotide composition of the 13 universal X motifs (uX motifs) in the SSU rRNA alignments (Table 4.1). Each row represents an X trinucleotide and each column represents a uX motif. Each cell contains the number (percentage) of trinucleotides observed in each uX motif.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*
AAC	0(0)	1(0)	8(2)	0(0)	0(0)	1(0)	0(0)	15(3)	46(9)	32(11)	0(0)	0(0)	155(40)
AAT	0(0)	0(0)	0(0)	0(0)	0(0)	3(1)	0(0)	113(23)	24(5)	2(1)	0(0)	0(0)	1(0)
ACC	130(50)	0(0)	7(2)	0(0)	131(34)	0(0)	0(0)	35(7)	0(0)	0(0)	0(0)	0(0)	1(0)
ATC	0(0)	0(0)	6(1)	0(0)	1(0)	0(0)	0(0)	0(0)	3(1)	0(0)	0(0)	1(0)	0(0)
ATT	0(0)	0(0)	1(0)	131(49)	2(1)	3(1)	0(0)	30(6)	11(2)	0(0)	1(0)	0(0)	0(0)
CAG	0(0)	0(0)	0(0)	0(0)	22(6)	1(0)	194(74)	38(8)	50(10)	0(0)	0(0)	101(21)	4(1)
CTC	2(1)	1(0)	0(0)	0(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	1(0)	90(18)	0(0)
CTG	0(0)	0(0)	2(0)	0(0)	55(14)	5(1)	1(0)	0(0)	1(0)	31(11)	5(1)	0(0)	0(0)
GAA	0(0)	83(24)	0(0)	8(3)	0(0)	105(28)	0(0)	9(2)	0(0)	114(40)	0(0)	0(0)	26(7)
GAC	0(0)	1(0)	0(0)	95(35)	0(0)	4(1)	0(0)	1(0)	84(16)	1(0)	8(2)	0(0)	3(1)
GAG	0(0)	1(0)	1(0)	0(0)	4(1)	0(0)	0(0)	0(0)	16(3)	34(12)	18(5)	0(0)	131(34)
GAT	0(0)	0(0)	0(0)	0(0)	149(38)	22(6)	0(0)	2(0)	65(13)	9(3)	30(8)	0(0)	23(6)
GCC	130(50)	0(0)	88(20)	0(0)	4(1)	3(1)	68(26)	0(0)	0(0)	8(3)	0(0)	25(5)	0(0)
GGC	0(0)	0(0)	2(0)	35(13)	0(0)	4(1)	0(0)	0(0)	45(9)	8(3)	17(5)	1(0)	29(7)
GGT	0(0)	0(0)	72(17)	1(0)	0(0)	97(26)	0(0)	129(26)	36(7)	2(1)	107(30)	26(5)	5(1)
GTA	0(0)	131(38)	217(50)	0(0)	0(0)	0(0)	0(0)	55(11)	21(4)	0(0)	56(16)	25(5)	3(1)
GTC	0(0)	130(37)	0(0)	0(0)	22(6)	0(0)	0(0)	0(0)	114(22)	29(10)	7(2)	182(37)	2(1)
GTT	0(0)	0(0)	0(0)	0(0)	0(0)	20(5)	0(0)	0(0)	0(0)	12(4)	92(26)	25(5)	7(2)
TAC	0(0)	0(0)	2(0)	0(0)	0(0)	109(29)	0(0)	41(8)	3(1)	0(0)	0(0)	0(0)	0(0)
TTC	0(0)	1(0)	24(6)	0(0)	0(0)	3(1)	0(0)	30(6)	0(0)	0(0)	17(5)	12(2)	0(0)

Table 4.12. Trinucleotide composition of the 19 uX motifs in the LSU rRNA multiple sequence alignments (Table 4.2). Each row represents an X trinucleotide and each column represents a uX motif. Each cell contains the number (percentage) of trinucleotides observed in each uX motif.

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*
AAC	0(0)	0(0)	5(2)	0(0)	64(21)	0(0)	6(2)	14(4)	0(0)	0(0)	107(19)	0(0)	121(25)	16(5)	0(0)	1(0)	0(0)	2(0)	4(1)
AAT	0(0)	0(0)	0(0)	0(0)	0(0)	8(2)	0(0)	7(2)	0(0)	2(0)	29(5)	0(0)	2(0)	47(15)	0(0)	92(25)	1(0)	0(0)	90(20)
ACC	6(2)	71(10)	8(3)	0(0)	66(21)	33(7)	28(10)	99(30)	35(9)	1(0)	2(0)	0(0)	3(1)	9(3)	0(0)	3(1)	1(0)	3(1)	5(1)
ATC	101(34)	6(1)	116(40)	2(0)	0(0)	0(0)	2(1)	14(4)	2(1)	1(0)	20(4)	3(1)	0(0)	7(2)	0(0)	1(0)	0(0)	11(3)	10(2)
ATT	3(1)	0(0)	24(8)	0(0)	0(0)	54(11)	53(19)	12(4)	0(0)	2(0)	52(9)	1(0)	0(0)	0(0)	0(0)	9(2)	0(0)	0(0)	109(24)
CAG	0(0)	3(0)	0(0)	43(9)	43(14)	0(0)	1(0)	12(4)	0(0)	101(22)	2(0)	14(4)	2(0)	0(0)	19(7)	1(0)	0(0)	0(0)	61(13)
CTC	0(0)	90(13)	8(3)	0(0)	0(0)	1(0)	0(0)	1(0)	8(2)	0(0)	2(0)	79(22)	4(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
CTG	2(1)	0(0)	100(35)	127(27)	0(0)	73(15)	6(2)	8(2)	40(10)	0(0)	11(2)	123(34)	0(0)	19(6)	0(0)	1(0)	0(0)	0(0)	13(3)
GAA	0(0)	0(0)	4(1)	0(0)	0(0)	0(0)	120(42)	125(37)	51(13)	131(29)	15(3)	0(0)	20(4)	64(21)	2(1)	113(30)	12(3)	174(42)	3(1)
GAC	128(43)	8(1)	0(0)	0(0)	0(0)	0(0)	0(0)	1(0)	44(11)	1(0)	24(4)	0(0)	0(0)	0(0)	0(0)	10(3)	30(9)	19(5)	1(0)
GAG	4(1)	1(0)	3(1)	125(27)	0(0)	16(3)	0(0)	1(0)	116(29)	1(0)	59(11)	0(0)	0(0)	0(0)	130(45)	0(0)	53(15)	7(2)	4(1)
GAT	2(1)	127(18)	1(0)	1(0)	0(0)	8(2)	0(0)	0(0)	2(1)	25(6)	53(9)	3(1)	0(0)	3(1)	0(0)	49(13)	30(9)	84(20)	14(3)
GCC	0(0)	0(0)	9(3)	0(0)	1(0)	91(19)	3(1)	2(1)	41(10)	30(7)	0(0)	0(0)	6(1)	0(0)	0(0)	34(9)	1(0)	40(10)	6(1)
GGC	27(9)	171(24)	4(1)	0(0)	0(0)	0(0)	18(6)	2(1)	10(3)	17(4)	73(13)	0(0)	189(39)	0(0)	0(0)	3(1)	92(26)	51(12)	1(0)
GGT	2(1)	23(3)	5(2)	124(27)	0(0)	5(1)	34(12)	2(1)	0(0)	74(16)	10(2)	0(0)	23(5)	80(26)	0(0)	6(2)	54(15)	2(0)	17(4)
GTA	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	11(3)	9(2)	4(1)	66(12)	4(1)	101(21)	1(0)	6(2)	47(13)	8(2)	0(0)	4(1)
GTC	13(4)	137(19)	0(0)	0(0)	131(42)	0(0)	0(0)	2(1)	33(8)	13(3)	12(2)	0(0)	6(1)	27(9)	0(0)	2(1)	0(0)	11(3)	11(2)
GTT	11(4)	2(0)	0(0)	43(9)	0(0)	69(14)	13(5)	21(6)	3(1)	47(10)	1(0)	124(34)	2(0)	22(7)	0(0)	0(0)	70(20)	3(1)	102(22)
TAC	0(0)	32(5)	0(0)	0(0)	3(1)	0(0)	1(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)	10(3)	132(46)	1(0)	0(0)	0(0)	2(0)
TTC	0(0)	37(5)	0(0)	0(0)	1(0)	123(26)	0(0)	0(0)	1(0)	0(0)	20(4)	11(3)	0(0)	0(0)	0(0)	0(0)	0(0)	3(1)	2(0)

4.9. Identification of *uX* motifs in the primordial proto-ribosome

We discussed previously the universality of the ribosome, particularly the identification of a rRNA common core which is conserved over all domains of life. This common core must have been present at the time of LUCA before the origin of the three domains of life. It is believed that LUCA must have had a pre-developed translation mechanism (proto-ribosome) capable of carrying out the important functions needed for the development of cellular life. This primordial translation system was sophisticated enough with the ribosomal catalytic center PTC, the small subunit containing the decoding center, proto-mRNAs and proto-tRNAs interacting with each other to produce essential amino acids and carrying out important ribosomal functions. It is generally assumed that the two ribosomal subunits initially existed independently and their interaction gave rise to sophisticated translation mechanisms; however there is some debate as to whether the LSU or the SSU emerged first (Kunnev & Gospodinov, 2018; Opron & Burton, 2018). Based on comparative structural analyses, proto-LSU (Agmon, 2017; Bokov & Steinberg, 2009; Hsiao et al., 2009, 2013; Petrov et al., 2015; Smith et al., 2008) and proto-SSU (Agmon, 2018; Petrov et al., 2015) models have been proposed (Figure 4.12). Here we will discuss in detail the identification of uX motifs in the primordial proto-ribosome.

The primordial proto-ribosome (Agmon, 2018) was indeed able to translate the genetic information coded in an RNA template (proto-mRNAs) into a polypeptide chain. In order to do so, it must have been composed of the two important structures, i.e. one which accommodates the mRNA molecule including the decoding center (proto-SSU) and another that accommodates the peptidyl transferase center PTC (proto-LSU). The proto-LSU model (Figure 4.12) corresponds to the PTC, a symmetrical region present deep within the large subunit of the modern ribosome, where new amino acids are incorporated into the growing peptide chain during translation (1.1.4). This region is structurally and phylogenetically conserved across all domains of life, suggesting that it was present before the origin of the three domains of life. In order to represent the ancestral translation system, this region has been generally modeled using the contemporary *E. coli* sequence. The dimeric proto-LSU (Agmon, 2017) can be divided into two L-shaped RNA molecules A- and P-monomers corresponding to the modern A-tRNA and P-tRNA sites, suggesting its origin from the proto-tRNAs. It consists of around 120 nucleotides, forming a pocket-like structure which could have accommodated two random amino acids, thereby providing positional catalysis and producing short peptides with random composition. We mapped the uX motifs to this two-dimensional proto-LSU model and found a total of 40 nucleotides (30%) which are present in uX motifs identified in the LSU rRNA multiple sequence alignment. These motifs are located almost entirely in the A-monomer corresponding to the modern A-tRNA site, with 35 (58%) of the 60 A-monomer nucleotides present in uX motifs. In addition to the universal regions, many of the nucleotides that make up the two halves

of the PTC cavity are composed of *X* trinucleotides (shown in bold) and these trinucleotides have been found to have a high degree of complementarity in different ancient bacteria (Agmon, 2017). Interestingly these trinucleotides reflect the important self-complementary property of the *X* circular code (Definition 2.9). This complementarity was suggested to demonstrate a simple and effective mode of replication. In simple terms, each monomer strand could have served as a template for synthesizing its counterpart, thereby suggesting that the proto-LSU may have been a self-replicating ribozyme (Agmon, 2017).

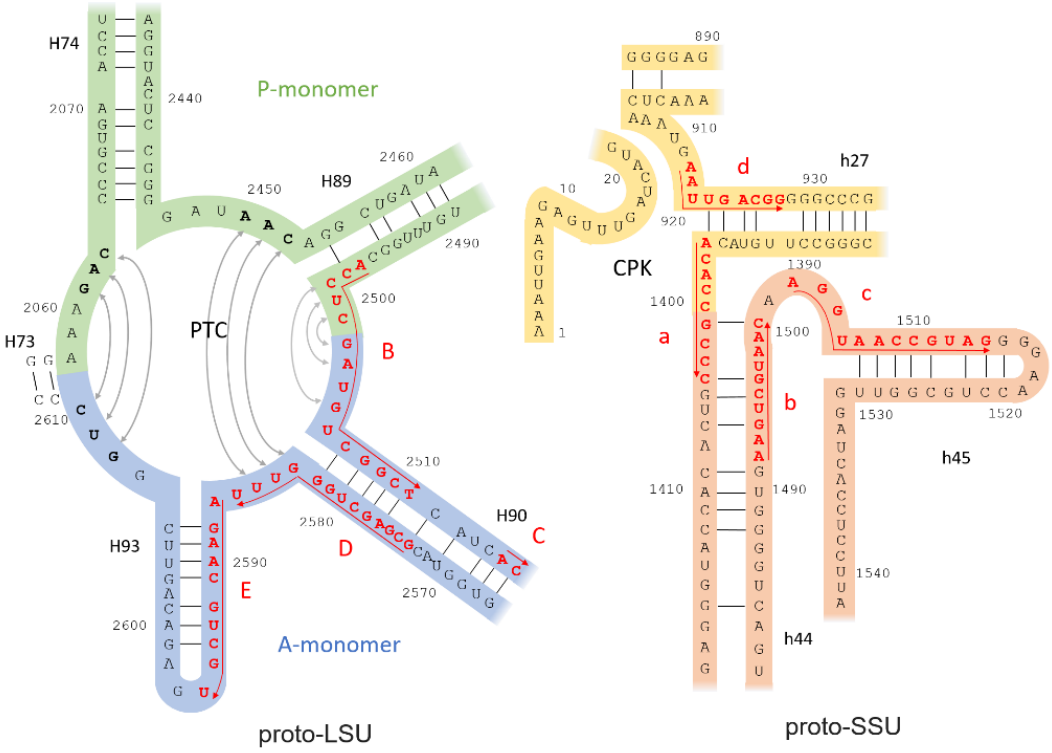


Figure 4.12. Proto-LSU and proto-SSU, with nucleotides and numbering from the contemporary *E.coli* 23S and 16S rRNA. *uX* motifs are highlighted in red and labeled according to the accretion model, with 5′ – 3′ direction indicated by red arrows. Sequence complementarity of nucleotides building the conserved PTC walls in bacterial ribosomes is indicated by gray arrows in the PTC loop (connecting *X* trinucleotides shown in bold).

On the other hand, the proto-SSU model is more complicated. As mentioned above, it should have accommodated the proto-mRNA binding along with the decoding center. The proposed models of the ancestor of the SSU correspond to the contemporary central pseudoknot (CPK) in the decoding center (Noller, 2012). Contrary to what is observed in the LSU, there is no single self-folding segment in the modern 16S RNA that encompasses the majority of the decoding site rRNA. A number of short disjoint segments with a total length of around 150 nucleotides have been considered ancestral (Agmon, 2018; Petrov et al., 2015). We mapped the *uX* motifs to this proto-SSU model and found that a total of 40 nucleotides (27%) are present in *uX* motifs identified in the SSU rRNA multiple sequence alignment. Notably, these

nucleotides in the proto-SSU include the future A-site (A1492-A1493) and P-site (C1402-C1403, U1498-A1499) tRNA binding sites which are essential for protein translation.

It should be noted that the combined models of the proto-ribosome, which incorporate the active sites of both ribosomal subunits, cover less than 6% of the modern prokaryotic rRNA, yet they integrate 80 (27%) of the 296 rRNA nucleotides found in uX motifs. These uX motifs are highlighted in red and labeled according to the accretion model (Petrov at al., 2015) : 4 (B, C, D, E) out of 19 uX motifs identified in LSU rRNA multiple sequence alignment (Table 4.2) and 4 (a, b, c, d) out of 13 uX motifs identified in SSU rRNA multiple sequence alignment (Table 4.1) are identified in the primordial proto-ribosome (Figure 4.12).

It has been suggested that an RNA molecule with a peptidyl transferase activity existed before the full sequential three-base decoding (Polacek & Mankin, 2005). This early non-coded proto-ribosome could have catalyzed the association of arbitrary amino acids, producing short peptides of random sequences. Here, we showed that the models of both proto-LSU and proto-SSU are enriched in uX motifs, with 30% of the nucleotides found in uX motifs. Concerning the LSU, we observed more uX motifs in the A-monomer than in the P-monomer, based on the *E. coli* sequence that was used in the model (Figure 4.12). This may reflect an inherent asymmetry of the proto-LSU, or it may be due to a stronger conservation of the A-site in evolution. We believe that these uX motifs must have played some functional role in the primitive systems; without any functional importance these would have been lost or replaced after various rounds of evolution. The universality of these X motifs and most importantly their enrichment in the rRNA common core and the proposed models of proto-ribosome allows us to suggest that these motifs are involved in important functions of the modern ribosome, which we will address next.

4.10. Identification of universal X motifs in functional centers of modern ribosome

In the previous section we discussed in detail the identification of uX motifs in the proposed models of primitive translation systems. Generally, the modern translation mechanisms for protein synthesis are quite similar in archaeal, bacterial, and eukaryotic organisms with some variations. However, the main functions of the ribosome are found to be conserved in all the three domains of life (Opron & Burton, 2018). Here we examine in detail the location of the 32 universal X motifs (13 uX motifs in SSU and 19 uX motifs in LSU) identified in modern ribosomes. We carried out structural analyses to relate these uX motifs to the known functional regions of the modern ribosome.

4.10.1. Functional centers of the modern ribosome

Before we talk about the results, we would like to recall briefly the translation mechanism and the functional regions which are involved. At the initiation stage of protein translation, the SSU binds the mRNA at the start site and, together with the tRNA, is responsible for the maintenance of translational fidelity. This means that the SSU, along with the tRNA, ensures the correct base pairing between the codons and anticodons in the decoding center. The tRNA, which is L-shaped, carries the cognate amino acid at one end (acceptor stem) and has an anticodon loop at the other end. The anticodon loop that pairs with the mRNA determines the amino acid attached at the acceptor end. The LSU binds the acceptor ends of the A-site and P-site tRNAs and catalyzes peptide bond formation at the peptidyl transferase center (PTC). This process continues until a stop codon is encountered. The synthesized polypeptide chain then passes through the exit tunnel that begins at the PTC and exits from the back of the LSU. Both the SSU and the LSU, with the help of other enzymes and tRNAs, are responsible for translating the mRNA template into the corresponding chain of amino acids to successfully synthesize the correct nascent protein (protein before folding). In this highly complex process, a single error can have a drastic effect on the synthesized protein. For example, an incorrect reading of the mRNA template can result either in the formation of a non-functional protein or even early termination of translation elongation if a stop codon is encountered. Translocation is the process in which the ribosome moves by one trinucleotide in each cycle. After the formation of a peptide bond at the PTC, the A-tRNA and the P-tRNA at the A and P sites respectively, translocate to the P and E sites respectively. Both the SSU and LSU are actively involved in translocating the mRNA by one trinucleotide in each cycle. During protein translation the ribosomal subunits are involved in three kinds of movement: swivel and/or tilting motion between the head and body of the small subunit for translocation of the tRNAs along with the mRNA acting as a translocation ratchet and a ratchet movement where both subunits rotate reversibly relative to one another (Jenner et al., 2010; Belardinelli et al., 2016; Opron & Burton, 2018). These movements are considered crucial for the initiation and translocation steps of protein translation.

4.10.2. Structural analysis of universal *X* motifs

Our analysis was based on a representative 3D ribosome structure from the bacteria *T. Thermophilus*, as it contains mRNA nucleotides and three deacylated tRNAs in the A, P and E sites. We carried out structural analyses of the 32 uX motifs in order to identify their interactions with different biomolecules, including mRNA, tRNA and ribosomal proteins. In Table 4.13 and Table 4.14 we summarize the interactions of uX motifs with different biomolecules along with their corresponding locations in the functional centers of the modern ribosome. This approach

using the 3D structure analysis revealed important interactions of *uX* motifs which compels us to suggest that they might be involved in important ribosomal functions.

Table 4.13. Contacts (<5 Å) of the 13 *uX* motifs in the SSU rRNA alignment (Table 4.1), with other *uX* motifs, mRNA, tRNA or ribosomal proteins. *uX* motifs are labeled according to the accretion model of Petrov et al., 2015. A, P, E in the tRNA column indicate contacts with the A-site, P-site and E-site tRNAs respectively.

uX motif	Contacts				Functional site
	uX motif	mRNA	tRNA	Protein	
a	*b*	+	P	S5	P-site; Ratchet pawl
b	*a*	+	A,P	S12	A-site; P-site
c	-	+		-	Ratchet pawl
d	-	+	P	S5	P-site; Head swivel hinge
e	-	+	P,E	S11	P-site; E-site
f	-			S7,S9,S10,S14	
g	*h*			S12	
h	*g*	+	A	S3,S12	A-site
i	*l*	+		S3,S5,S9,S10,S14	
j	-			-	PE loop
k	-			S17	
l	*i,m*			S2,S3,S5	Head swivel hinge
m	*l*			S2,S3	

Table 4.14. Contacts of the 19 *uX* motifs in the LSU rRNA alignment (Table 4.2), with other *uX* motifs, tRNA or ribosomal proteins. Contacts are defined as <5 Å unless specified otherwise. *uX* motifs are labeled according to the accretion model of Petrov et al., 2015. A, P, E in the tRNA column indicate contacts with the A-site, P-site and E-site tRNAs respectively. *indicates bacteria specific ribosomal proteins.

uX motif	Contacts			Functional site
	uX motif	tRNA	Protein	
A	-	A	L16	
B	*D*	P	L3,L32*	PTC (<10 Å), exit tunnel
C	-	A	-	PTC (<30 Å)
D	*B*	A,P	L3,L32*	PTC (<10 Å), exit tunnel
E	-		L2	PTC (<30 Å), exit tunnel
F	*F*	A	L14	PTC (<10 Å), exit tunnel
G	-		L22,L32*	Exit tunnel
H	-		L20*,L22,L32*	Exit tunnel
I	*J*		L2	
J	*I*		L2	
K	-		L2	
L	-		L4,L15,L20*	Exit tunnel
M	-		L2	
N	-	E	-	
O	-		L6	SRL
P	-		L3,L13,L36*	L11 stalk
Q	-		-	L11 stalk - GAC
R	-		L34*	
S	-		L23	Exit tunnel

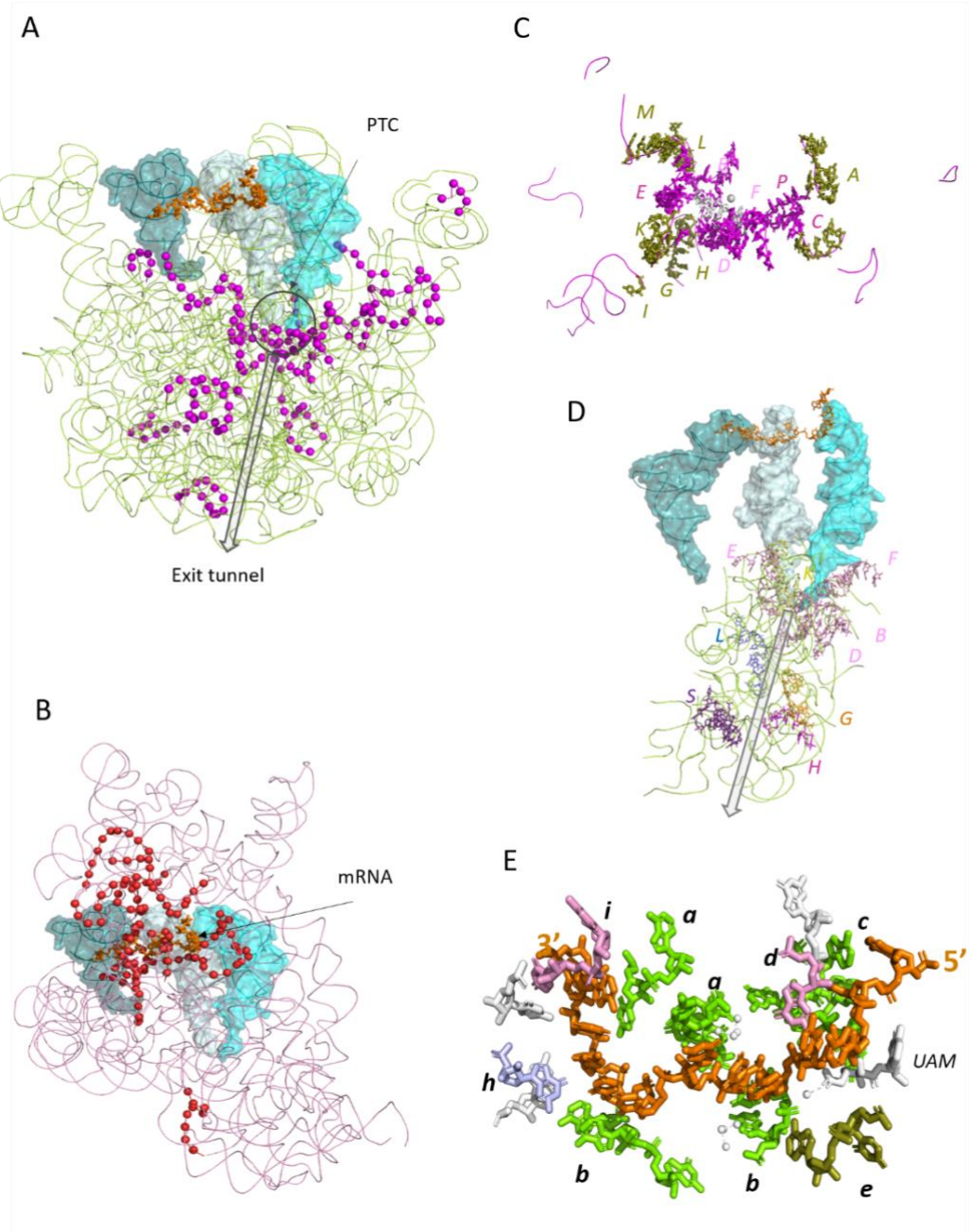


Figure 4.13. 3D structures of *uX* motifs in the rRNA of *T. thermophilus*.

A. LSU rRNA (green ribbon) with mRNA (orange sticks) and surface representations of tRNAs in the A-site (cyan), P-site (light blue) and E-site (deep teal). Nucleotides of the *uX* motifs are shown as magenta spheres. The PTC is identified by a black circle and the exit tunnel by a black arrow.

B. SSU rRNA (pink ribbon) with tRNA colored as in A. Nucleotides of the *uX* motifs are shown as red spheres.

C. Nucleotides in *uX* motifs close to the PTC (<10 Å in white sticks, <30 Å in magenta sticks, <50 Å in olive sticks). The distances were measured from atom N4 of CYT 2573 (white sphere). All *uX* motifs are shown as magenta ribbons.

D. All rRNA nucleotides (green ribbon) within 20 Å of the exit tunnel (black arrow) (defined by Dao Duc et al., 2019): nucleotides in *uX* motifs are colored according to rRNA domains, magenta for domain I, blue for domain II, violet for domain III, orange for domain 0, yellow for domain IV, pink for domain V (Table 4.6). tRNA are colored as in A.

E. SSU rRNA nucleotides in contact with mRNA (<5 Å): nucleotides in *uX* motifs are colored according to rRNA domains, light blue for domain 5', olive for the central domain, pink for 3'M and green for 3'm domains (Table 4.6), other nucleotides and amicoumacin A (UAM) are white. Magnesium ions and their coordinated water molecules are represented by white spheres.

In Figure 4.13, we show the 3D representation of the 19 *uX* motifs identified in the LSU rRNA multiple sequence alignment and the 13 *uX* motifs identified in the SSU rRNA multiple sequence alignment. We used various color representations to visualize the ribosomal functional centers and the *uX* motifs interacting with them in a close vicinity (distance measured in Å unit).

4.10.3. Mapping *uX* motifs to functional centers of modern ribosome

Here we discuss in detail the interactions of *uX* motifs mapping them to the functional regions involved in the translation machinery. Previously we identified *uX* motifs in the proposed model of proto-ribosome and also in the rRNA common core which is conserved across all domains of life. These results indicate that *uX* motifs were present in the early stages of evolution of the translation machinery in primitive systems. Also, both the SSU and LSU play a role in the complex processes leading to the synthesis of proteins in the ribosome. We evaluate the enrichment of *uX* motifs in both subunits separately; to recall it is generally believed that both the subunits of the ribosome have evolved independently.

In the LSU, the PTC is the most conserved functional site where amino acids are polymerized onto the growing nascent polypeptide chain. The proto-LSU is believed to contain the PTC, around which other functional segments were added as evolution progressed in phases. In the "ribosome as an onion" model (Hsiao et al., 2009), the authors explain how the different phases of ribosome evolution can be organized in concentric shells (each with a thickness of 10 Å) with the PTC occupying the center of origin. In addition, their results suggest the evolution of the PTC with Mg^{+2} ions stabilizing the early RNA conformations, which decreases as one moves from the center to periphery of the LSU. Moreover, the magnesium ions appear to have been substituted at the periphery by ribosomal proteins. The ribosomal proteins are found to be present with the highest density in the periphery of LSU in comparison to the region near PTC where they are found to be absent. According to our analysis, the majority of the identified *uX* motifs are found to be clustered around the PTC (Figure 4.13C) with 3 motifs (*B, D, F*) within a radius of 10 Å and 6 motifs (*B, C, D, E, F, P*) within a radius of 30 Å. In total, 13 out of the 19 *uX* motifs are located within a radius of 50 Å (*A, B, C, D, E, F, G, H, I, L, K, M, P*) in the LSU. Of the 175 nucleotides covered by *uX* motifs in the LSU, 105 (60 %) are located within a distance of 50 Å from the PTC. Also, we identified several *uX* motifs to be in direct contact with tRNAs: nucleotides G2553, U2555 (motif *F*) and G2583, U2585 (motif *D*) are in contact

with the A-site tRNA; U2585 (motif *D*) and U2506 (motif *B*) are in contact with the P-site tRNA; and G1850-A1853 (motif *N*) are in contact with the E-site tRNA. One motif (*A*) is found in helix H89, which is known to be involved in the accommodation of the A-site tRNA in the PTC (Opron & Burton, 2018). Another important structure in the LSU is the polypeptide exit tunnel that extends from the PTC to the surface of the ribosome. The tunnel shape is more conserved in the upper part close to the PTC, whereas in the lower part it is observed to be much narrower in eukaryotes than in bacteria (Dao Duc et al., 2019). In Figure 4.13D, we show the eight *uX* motifs (*B, D, E, F, H, G, L, S*) that are close to the exit tunnel. Finally, two *uX* motifs (*Q, O*) were found in regions involved in interactions with GTPase proteins during translation initiation and elongation: motif *Q* is in the GTP Associated Center (GAC) and motif *O* is in the sarcin-ricin loop. The remaining four *uX* motifs (*I, J, M, R*) in the LSU are not associated with known ribosomal functions to our knowledge.

In the SSU, the decoding center and the mRNA binding site are considered to be the most functional sites. According to our analysis, 7 (*a, b, c, d, e, h, i*) of the 13 *uX* motifs identified in the SSU are in contact with the mRNA (at a distance of <5 Å) (Figure 4.13E). Remarkably, only 3 of the 25 rRNA nucleotides in contact with the mRNA are not found in *uX* motifs. The *uX* motifs also include many of the rRNA contacts with tRNAs, such as the A-site conserved nucleotides A1492-A1493 (motif *b*) and G530 (motif *h*); the P-site G926 (motif *d*), A790 (motif *e*), U1498 (motif *b*), and C1400 (motif *a*); and the E-site C795 (motif *e*) (Khade and Joseph, 2010). Interactions of these functional sites with tRNAs, ribosomal proteins and mRNA are considered important for tRNA selection and translocation among other functions related to initiation and elongation phases of protein synthesis. Another important feature of the SSU is the dynamic swiveling of the SSU head (3'M domain) relative to the body (5' domain) during translation elongation. The movement originates from flexing at two hinge points, one in the middle of helix h28 at G926, and one in the linker between h34 and h35. Both of these hinges are found in *uX* motifs (*d, l* respectively). Rotation of the SSU head has also been linked to the opening and closing of a 13-Å constriction or 'gate' between the head and body domains between the P and E sites, presenting a steric block to the movement of the P-site tRNA. The gate involves G1338 (motif *j*) situated in the stable ridge that sterically separates the P and E sites, and A790 (motif *e*) located on the opposite side of the constriction (Achenbach & Nierhaus, 2015). The C1397 (motif *a*) and A1503 (motif *c*) have also been considered to be 'ratchet pawls' that intercalate with mRNA bases during reverse rotation of the head (Achenbach & Nierhaus, 2015). Three *uX* motifs (*f, g, k*) in the SSU are not associated with known functions to our knowledge.

To recall, the ribosome is composed of rRNAs and ribosomal proteins specific to both the subunits. Moreover, the number of proteins found in the ribosome differs across the three domains of life. In bacteria, there are 24 proteins in the SSU and 34 proteins in the LSU; in

archaea 28 proteins in the SSU and 40 proteins in the LSU, whereas in eukaryotes 32 proteins in the SSU and 46 proteins in the LSU. However, among the 102 known ribosomal protein families a total of 34 which includes 15 in the SSU and 19 in the LSU are represented in all three domains of life (Smith et al., 2008), hence they are considered universal. Many of these universal ribosomal proteins were found to be crucial for ribosome assembly, formation of inter-subunit bridges, and interactions with the tRNAs or the polypeptide exit channel (Lecompte et al., 2002). Notably, many of the *uX* motifs identified are also in contact with ribosomal proteins: 11 out of 13 *uX* motifs in the SSU (Table 4.13) and 16 out of 19 *uX* motifs in the LSU (Table 4.14). Our analysis also revealed that nearly all the proteins in contact with *uX* motifs are universal ribosomal proteins; in *T. thermophilus*, all 10 proteins in contact with the SSU *uX* motifs are universal, and 10 out of 14 proteins in contact with the LSU *uX* motifs are universal (Table 4.15).

Table 4.15. Ribosomal proteins represented in all three domains of life and classified according to their known 3D structure (Smith et al., 2008). Extensions refer to protein segments that extend away from the more compact or globular part of the protein for a significant distance. Ribosomal proteins in contact with *uX* motifs are shown in bold.

	Mainly globular	Globular domain with a long unstructured extension	Hairpin extension	Helical and hairpin extension
SSU	**S2, S3,** S4**, S14,** S15	**S7**, **S9**, **S11**, **S12**, S13, S19	**S5**, S8, **S10**	**S17**
LSU	L1, **L6**, L7, L11, L12, **L23**, L29, L30	**L2**, **L15**, **L16**, L18, L24	**L3**, **L4**, L5, **L13**, **L14**, **L22**	

These results revealed various interactions of the *uX* motifs which can be mapped to crucial ribosomal functions, which in effect enables us to assess their involvement in the diverse functions of the ribosome. Next, we will discuss the accretion of *uX* motifs according to the proposed model of accretion of the ribosome in various stages.

4.11. Accretion of *uX* motifs : transition from proto-ribosome to modern ribosome

In the previous sections we demonstrated that *uX* motifs identified in the rRNA multiple sequence alignments can be mapped to the functional regions of the ribosome. Moreover, they are found to be in contact with other *uX* motifs , mRNA, tRNAs and ribosomal proteins. We also identified *uX* motifs in the proposed proto-ribosome models, the primitive translation system that is believed to be present at the time of LUCA. Here we discuss in detail the accretion of *uX* motifs in the transition from the proto-ribosome to the modern ribosome.

Taking into account the complexity of the modern ribosome, it is highly unlikely to have appeared spontaneously (Hsiao et al., 2009; Petrov et al., 2015; Opron & Burton, 2018). Various studies on the evolution of the ribosome indicate that it developed in phases/stages over the evolutionary time period. Petrov and co-authors suggested that the proto-ribosome evolved into the modern rRNA common core through the recursive accumulation of "ancestral expansion segments" (AES). They also suggested an "accretion model" of the rRNA evolution divided into six major phases representing the successive steps involved in ribosome sophistication. The accretion model of ribosome evolution is shown in Figure 4.14 with the location of the *uX* motifs in red boxes. The *uX* motifs are labeled as per their presumed ancestry: *a-m* for the SSU *uX* motifs (Table 4.1) and *A-S* for the LSU *uX* motifs (Table 4.2).

After mapping the *uX* motifs to the accretion model, we can differentiate them into two subsets: those already present in the proto-ribosome model (phases 1 and 2) shown above and those gained during the subsequent phases of ribosome evolution (phases 3 to 6). As described above, 4 motifs (*B, C, D, E*) out of the 19 *uX* motifs (LSU) were present in the proto-LSU with 2 additional motifs (*A,F*) located in the vicinity of the slightly extended ancestral region, and 4 motifs (*a, b, c, d*) out of the 13 *uX* motifs (SSU) were present in the proto-SSU. According to the accretion model, in phases 1 and 2 the primordial RNAs interacted with metal cations in order to fold into stable structures; the exit tunnel of proto-LSU is formed. At this stage, the proto-LSU was able to synthesize nonspecific amino acids (oligomers) with the help of proto-tRNAs (CCA-tail) delivering activated substrates to the PTC. The function of the proto-SSU is not clear, which may have involved association with single-stranded RNA. However, there was no observed correlation between the two proto-subunits, suggesting independent evolution of proto-SSU and proto-LSU.

The remaining *uX* motifs correspond to the following phases:

- Phase 3: 7 motifs (*G, H, I, J, K, L, M*) out of the 19 *uX* motifs in the LSU are incorporated in this phase. Interestingly, motifs *G, H, I, J, K, L* are located near the extended exit tunnel and motifs *K,M* in the LSU–SSU interface. According to the accretion model, this phase leads to better catalytic efficiency and the production of longer protein chains. This is achieved by the new interactions between the SSU-LSU, proto-mRNAs (single stranded oligomers) and tRNAs. Supposedly, this phase led to the co-evolution of the LSU, the SSU, proto-mRNA and tRNAs.

- Phase 4: the SSU-LSU interactions are enhanced (motif *e*), the A-site and P-site tRNA binding pockets are formed with interactions in the 3′M domain (motif *f*) and the 5′ domain (motif *g*) respectively. According to the accretion model, interactions between the SSU and proto-mRNA enables tRNAs to be positioned and stabilized at the A- and P-site of the PTC. The exit tunnel is further extended and rigidified in the LSU. The

evolutionary pathways of both subunits are strongly linked together with the proto-mRNAs and tRNAs.

- Phase 5: the ribosome decoding system gains specificity and the ratcheting system is acquired to coordinate movement of the mRNA and tRNAs through the ribosome. In the LSU, binding sites for elongation factors G and Tu are established (motif *O*), together with the L11 stalk (motifs *P,Q*). In the SSU, the P-site tRNA pocket is further stabilized (motifs *i,j*) and the central pseudoknot is completed (motif *h*). According to the accretion model, this phase marks the ribosome's transformation into an energy-driven, ratcheting, translocating, decoding system. Specific codon-anticodon interactions between the mRNA (now polymeric with a defined sequence) and tRNAs begin to take place. This phase also leads to the expansion of the genetic code. The evolution of the ribosome is facilitated by the interactions between the RNA and ribosomal proteins.

- Phase 6: the newly acquired AES (motifs *R, S, k, l, m*) serve mainly as binding sites for the globular domains of ribosomal proteins. At the end of this phase, the "rRNA common core" of the contemporary ribosomes is established, composed of 3 rRNA and over 50 mature protein chains. The genetic code is also believed to be optimized at this stage.

From the accretion model of ribosome evolution, we understand how the ribosome may have developed in stages. Before proteins, metal ions stabilized the RNA structures in addition to providing catalytic features and avoiding chemical degradation. Following the involvement of polypeptides (early proteins) in the translation machinery, the evolution of the ribosome was driven by the interactions between RNA and proteins, which in turn also evolved the polypeptide assemblies into globular domains in a stepwise manner (Kovacs et al., 2017). In this extended analysis by Petrov and co-authors, the universal ribosomal proteins mentioned above were also integrated into the accretion model. They based the evolution of proteins on the assumption that the age of a given segment of protein is the same as that of the rRNA with which it interacts.

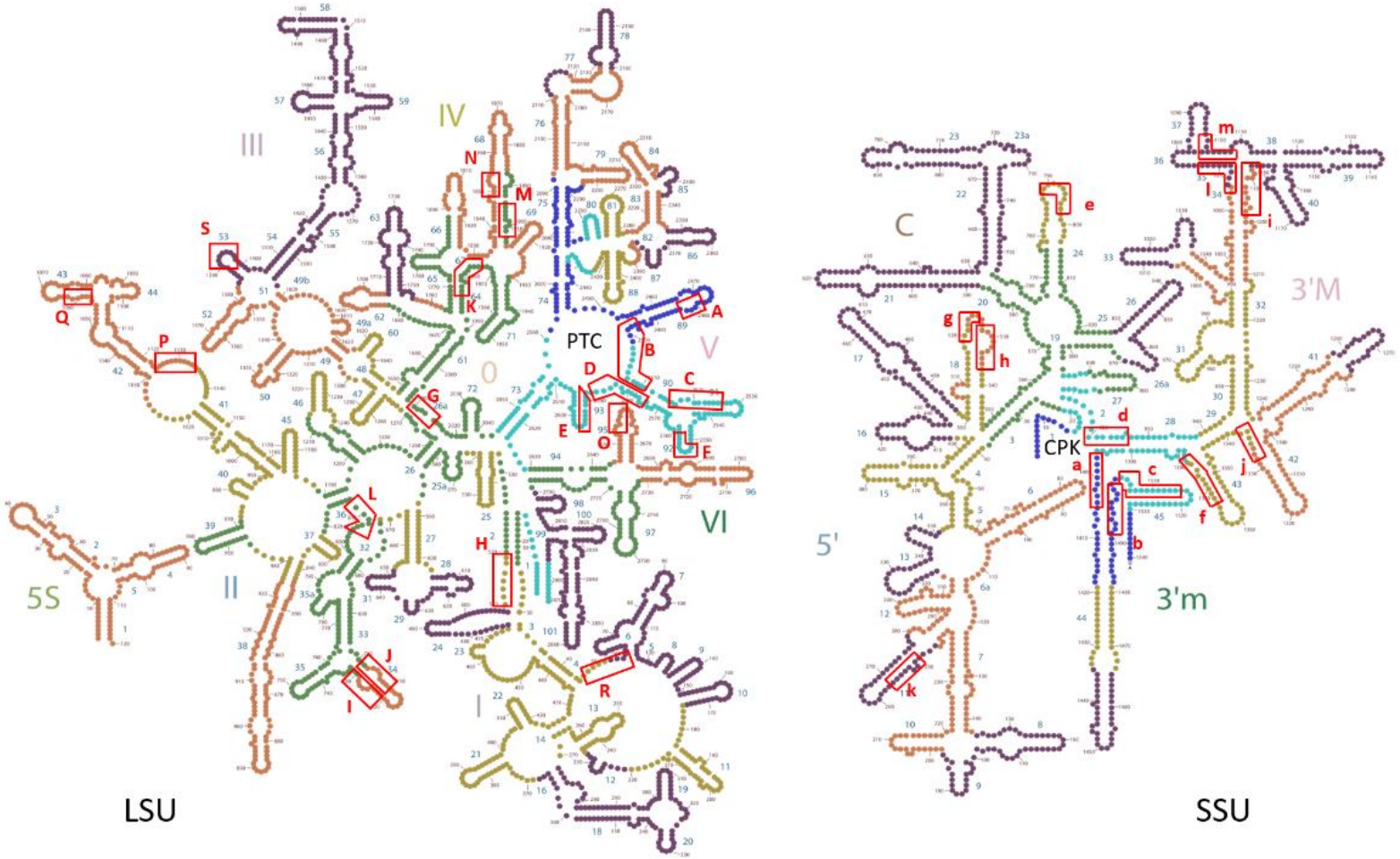


Figure 4.14. Secondary structure schema of the LSU and SSU rRNA (*E. coli*); coloured according to the six phases of the accretion model (Petrov et al., 2015) of ribosome evolution (phase 1, blue; phase 2, cyan; phase 3, green; phase 4, sepia; phase 5, brown; phase 6, purple). *uX* motifs (red boxes) are labelled according to their order of accretion in the different phases (Table 4.1 and Table 4.2).

In phases 1 and 2, it is generally assumed that only short random peptides are present in the proto-ribosome system. In phases 3 and 4, uX motifs (*A-M* in the LSU, *a-g* in the SSU) interact with 7 of the 19 universal proteins in the LSU (Table 4.14) and 7 of the 15 universal proteins in the SSU (Table 4.13). Many of these proteins are known to interact with the PTC (L2, L3, L4, L14) or have contacts to the tRNA binding site and/or the mRNA (S7, S9, S11, S12) mainly *via* their non-globular extensions (Smith et al., 2008). In phase 5, uX motifs (*O-Q, h-j*) contact globular domain proteins, including L6, L13, L36, and S3. In phase 6, most of the newly incorporated proteins are on the surface of the ribosome, and the uX motifs (*R-S, k-m*) contact only a few of them: L23, S2 and S17. Next we will discuss how the comma-free codes and circular codes may have helped in the evolution of the genetic code.

4.12. Coevolution model of the genetic code and translation system

Based on the results of our analyses of uX motifs in the proto-ribosome and the accretion model of ribosome evolution, we propose a coevolution model of the genetic code and the translation system in four stages (Figure 4.15). We suggest that comma-free codes and circular codes were the predecessors of the modern genetic code and were used to map the first trinucleotides to amino acids.

Recent studies suggests that RNA and peptides co-evolved from the beginning, or at least that the proto-ribosome building blocks were able to bind amino acids or small peptides very early (Kunnev & Gospodinov, 2018; Lupas & Alva, 2017). The interactions between peptides and RNA in extant organisms suggests an ancient origin and functional coevolution in the early stages of life on earth (Frenkel-Pinter et al., 2020). These interactions between peptides and RNA were extremely crucial for their mutual existence: increased lifetimes for peptides avoiding chemical degradation and stabilized structures for RNA. The first peptides were most likely of abiotic origin which included glycine and alanine, and binding would have been non-specific. However, we believe that natural selection must have favored products encoded and synthesized by early nucleic acids (proto-RNAs). Therefore, we propose that the first encoding system was based on a comma-free code, such as $\{GGC, GCC\}$, which would have allowed coding for the early amino acids and also maintaining the reading frame within a single code. At this time, the LSU and SSU would have evolved independently; the self-replicating ribozyme proto-LSU with a PTC function and aided by proto-tRNAs to synthesize amino acids, whereas the proto-SSU binding proto-mRNA from the available pool of single stranded RNAs. This marks the beginning of mutual existence and coevolution of the early biological polymers.

Increased interactions between LSU and SSU along with proto-mRNA and proto-tRNAs would have driven the synthesis of early coded products. Assembly of the two subunits with the intermediate tRNA would have given rise to the first ribosomes capable of coding longer and

more specific peptides. The ribosome and the genetic code would have co-evolved from this time on (Vitas & Dobovišek, 2018). It is generally believed that the repertoire of coded amino acids increased eventually with the expansion of the genetic code. With the addition of new amino acids, comma-free codes were no longer viable and the genetic code would have evolved towards the circular codes, possibly with a smaller number of amino acids initially. For example, it has been shown previously (Michel et al., 2017) that an X' circular code exists with 10 trinucleotides capable of coding 8 of the 10 hypothesized 'early amino acids' (Koonin, 2017). The peptides synthesized by the early ribosomes may have functioned as primordial ribosome cofactors to increase rRNA binding/stability in the prebiotic environment (Frenkel-Pinter et al., 2020; Lupas & Alva, 2017).

At the early/intermediate stages, in addition to coding for amino acids, circular codes would have allowed the detection and/or maintenance of the reading frame before the emergence of complex start/stop codon recognition systems, thereby allowing to code for the first simple proteins. Ribosome translation errors at the early stages of development of the translation machinery would have been a barrier to the genetic code optimization producing non-functional products and loss of resources. Therefore, we propose that the X circular code may have been the first error detection/correction system, avoiding reading the mRNA in the wrong frame in the primitive systems. The X circular code codes for 12 out of the 20 amino acids specified by the standard genetic code.

As mentioned above, no circular codes can include more than 20 trinucleotides. Therefore, the circular code property was not sufficient when more amino acids were incorporated into the genetic code. The standard genetic code requires a specific start codon to initiate the translation process, and sophisticated ratchet mechanisms (ribosome) to maintain the reading frame during translation elongation. Intriguingly, we have identified uX motifs in the functional regions of the modern ribosomes such as the ratchet pawls, the PTC and the decoding center. This compels us to suggest that the circular codes played a functional role in the coevolution of the ribosome and the genetic code at the early/intermediate stages of evolution. In addition to encoding the amino acids, comma-free codes and circular codes present the important synchronization property that would have allowed detection and maintenance of the reading frame in primordial and less sophisticated translation systems. These error-correcting codes provided an error-detection mechanism in the primitive systems, thereby allowing the primitive translation machinery to expand the repertoire of amino acids and optimize the genetic code simultaneously.

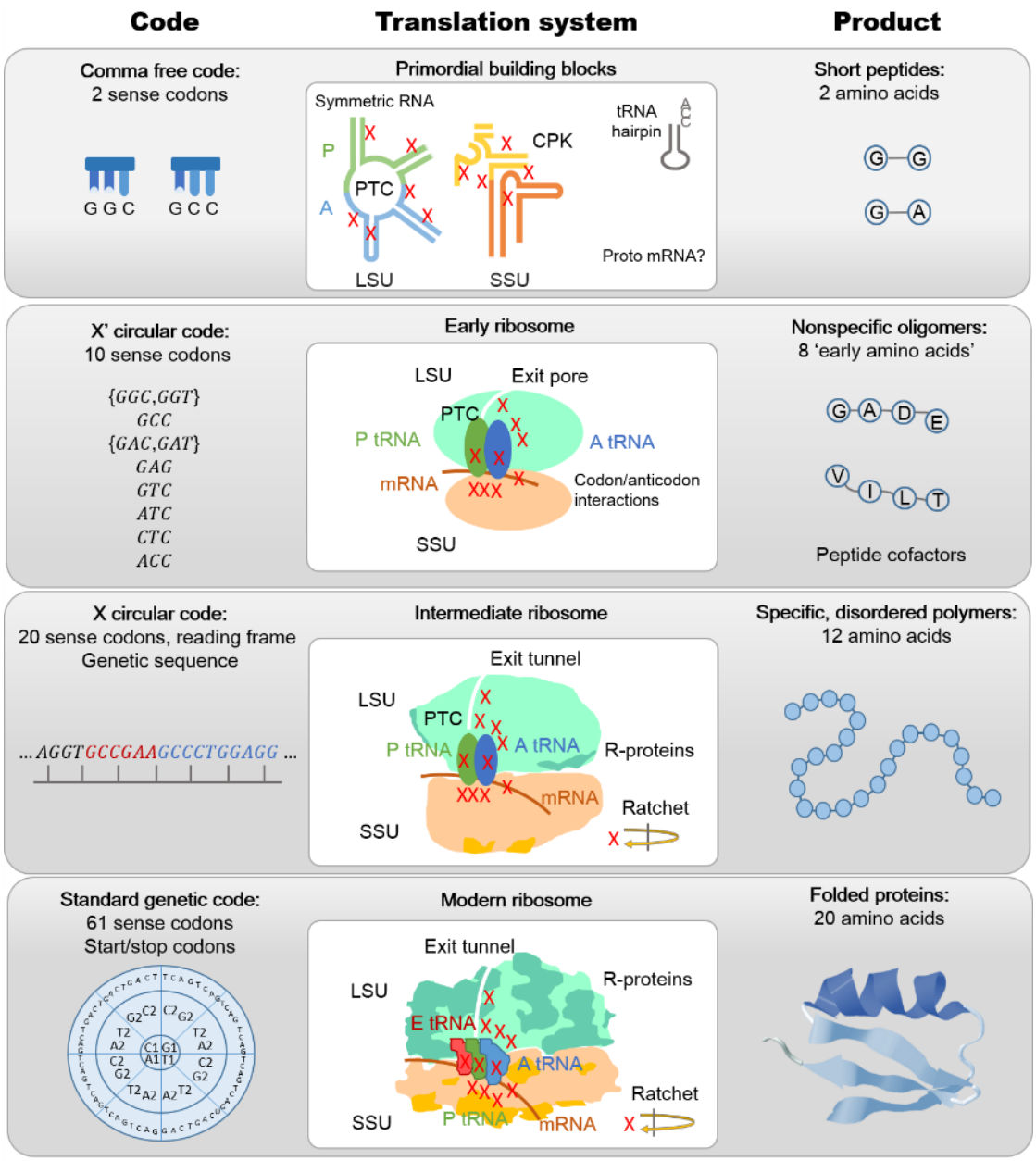


Figure 4.15. The proposed model of the evolution of the genetic code; associating codes, translation systems, and peptide products at the different stages from the primordial translation building blocks to the modern ribosome which is believed to be present at the time of LUCA.

The emergence of the translation machinery is an enigma, since proteins are synthesized in the ribosome but the ribosome also needs proteins to carry out various functions. In the RNA-peptide world scenario, the RNA polymers of the proto-ribosome served as templates to directly bind amino acids or short peptides. Cognate RNA triplets could have then evolved to act as anticodons in tRNAs and codons in mRNAs (Yarus, 2017). It has been observed previously that the early prebiotic amino acids are coded by G/C-rich codons, whereas engagement of new amino acids required more of A and T to be included in the codons (Gospodinov & Kunnev, 2020; Polyansky et al., 2013). Initially, the comma-free code $\{GGC, GCC\}$ was used to code Ala

and Gly, which was then expanded to an ancestral circular code X' containing 10 codons with a composition of 66% G/C and 33% A/T, and coded for 8 out of the 10 identified early amino acids (Koonin, 2017). The translation machinery kept on integrating new amino acids into the code, and in return was aided by new products (enzymes) in the sophistication of the translation process. Interactions between RNA and peptides paved the way for DNA and proteins to take over the predominant role of carrying and passing on genetic information.

4.13. Summary

The results presented in this chapter are crucial to understanding the complex evolutionary process and the possible role played by circular codes in the development of the translation machinery and the genetic code. The complexity of the genetic code suggests that it may not have emerged spontaneously. Various theories suggest the gradual increase of the amino acid repertoire (Gospodinov & Kunnev, 2020) and thereby the code encoding the signals for their synthesis. Also, the integration of amino acids into the translation machinery depends on their availability in prebiotic conditions. Some basic amino acids are believed to be present in the early stages of the evolution of life. As the primitive systems became more and more complex, the structures involved in important functions also evolved into complex structures. The translation mechanism started to use other amino acids, when it was able to manufacture them. However, there has been an ongoing debate on the order of incorporation of amino acids into the biological mechanism based on different analyses. It is speculated that the coding process began with a set of primitive amino acids and that others were added up to a total of 20 (Chatterjee & Yadav, 2019); glycine (Gly) and alanine (Ala) being the first two (Gospodinov & Kunnev, 2020).

Most studies on the origin and evolution of the genetic code have focused on mapping codons to amino acids (Ikehara, 2002; Hartman & Smith, 2014; Koonin, 2017), yet the origin of reading frame maintenance has not been addressed before. We have investigated the hypothesis that the standard genetic code originated from simpler comma-free codes *via* circular codes. The increase of the amino acid repertoire and the transition from the production of random peptides to the coding of specific protein sequences require more sophisticated mechanisms for codon recognition, but also identification of the reading frame. Circular codes represent an efficient means to synchronize the reading frame within a short window, before the evolution of a start codon and the modern translation initiation system. It is therefore tempting to suggest that base-pairing between the X motifs of the mRNA and those of the tRNA (Michel, 2013) and the rRNA would have given rise to the first coded ribosome apparatus. Traces of such interactions can be found in 3D structures of modern ribosomes, where we have shown that most of the uX motifs in the rRNA are in contact with the mRNA or the A, P and E

site tRNAs. From the analysis of rRNA sequences from the three domains of life, we identified 32 universal X motifs which are conserved in sequences that are not conserved in terms of nucleotides. For the first, we have shown the conservation of motifs in terms of trinucleotides (or codons) in the rRNAs. Furthermore, a circular code periodicity 0 modulo 3 was identified in the 16S rRNA, covering the region that corresponds to the primordial proto-ribosome decoding center and containing numerous sites that interact with the tRNA and mRNA during translation (Michel & Thompson, 2020). RNAs presumably synthesized proteins in the primitive systems, which suggests the origin of periodicity in RNAs lies in the primitive earth.

The enrichment of rRNAs with uX motifs is statistically significant and most of the motifs are clustered around important functional sites, including the PTC and the exit tunnel in the LSU, and the decoding center and ratchet mechanisms in the SSU. We propose that they represent the observable remnants of a primordial code used during the emergence of the RNA- or RNA-peptide world. Furthermore, it has been suggested that the X circular code arose from selection for non-redundant overlap coding in short nucleotide sequences (Demongeot & Seligmann, 2020; Michel, 2019). This is consistent with the hypothesis that the primordial genes maximized the number of coded amino acids over the shortest length in the process of evolution (Demongeot & Seligmann, 2019); in addition, these primordial genes, also known as RNA rings, are shown to be biased towards codons belonging to the X circular code. Moreover, we have shown that universally conserved X motifs are present at each evolutionary stage up to the common core of the modern ribosome. The question remains whether circular code X motifs have a functional role in modern translation systems, which we will discuss in the next chapter.

Chapter 5

5. Circular codes and ribosomal frameshift errors

5.1. Introduction

We have stressed the dependence of life forms on proteins and nucleic acids throughout this thesis. The genetic information stored in DNA is transferred into mRNA by transcription, which is then translated by the ribosome to synthesize proteins. During these processes, errors can occur which can lead to adverse effects on the protein product. In this chapter, we will discuss on the effects of errors in the translation process, and the mechanisms involved in minimizing their effects.

Biochemical and statistical studies have shown that the standard genetic code (SGC) is optimized to reduce the impact of errors caused by incorporation of wrong amino acids during translation. This is achieved by mapping codons that differ by only one nucleotide to the same amino acid or one with similar biochemical properties, so that if misincorporation occurs, the structure and function of the translated protein remain relatively unaltered. The most prominent cause of translation errors is the incorrect reading of a codon and the resulting incorporation of the wrong amino acid, known as missense errors. The per-codon missense error rate has been estimated to be between 10^{-4} and 10^{-3} (Garofalo et al., 2019). According to the adaptive theory (Woese, 1965), the SGC is optimized to minimize the effects of errors during transcription and translation. First, base changes at the third position of a codon, known as the wobble position, are generally synonymous, i.e. they code for the same amino acid. Second, amino acids with similar physicochemical properties are located in close proximity in the genetic code table and differ usually by only one substitution. For example, hydrophobic amino acids are usually coded by codons with thymine (T) in the second position and hydrophilic amino acids by those with adenine (A) in this position. Another important source of translation errors is ribosomal frameshifting, which occurs with an error rate of around 10^{-5} (Drummond & Wilke, 2009). Such errors can cause the premature termination of the translation process if a stop codon is encountered out-of-frame, or even if the process continues it can dramatically alter the amino acids being incorporated into the growing polypeptide chain. It has been shown that the SGC outperforms other theoretical alternative codes in terms of minimizing the effects of missense errors, when amino acid similarity is measured in terms of polarity (Freeland & Hurst, 1998; Haig & Hurst, 1991; Kumar & Saini, 2016), polarity and volume (Wnętrzak et al., 2019), or using empirical data of substitution frequencies (Freeland et al., 2000).

Here, we carried out an analysis on the ribosomal frameshift errors by taking into account various physicochemical properties of amino acids. We compare the optimality of the SGC with a set of circular codes, and in particular the circular code X in minimizing the effects of such errors. The results obtained during this analysis allowed us to evaluate further our hypothesis that the circular code X may have had a role in the evolution of the SGC. Next, we will introduce the ribosomal frameshift errors.

5.2. Ribosomal frameshift errors

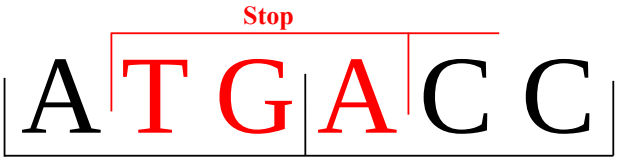
As previously mentioned, proteins have a well-defined 3D structure and their functions rely heavily on these structures. Ribosomal frameshift errors can lead to the synthesis of truncated products or misfolded proteins, causing an overall loss in protein function. The non-functional protein produced can affect the cellular function, even leading to diseases. So, it is very important for the ribosome to decode correctly the programmed sequence of amino acids coded in the mRNA according to the SGC. The translation of a nucleotide sequence into a protein sequence begins at the start codon (generally ATG) and terminates when a stop codon (generally TAA, TAG and TGA) is encountered. If the ribosome shifts on the nucleotide sequence by only one or two bases in either direction, the protein sequence can change dramatically (illustrated in Table 5.1). For example, the comparison of average sequence identity of wild-type proteins in humans with their $+1$ frameshifted counterparts shows a similarity of only 6% (Bartonek et al., 2020). It can be observed that a $+1$ shift gives the same read out of codons as for a -2 shift, which is also the case for -1 and $+2$ shifts. Therefore, we will only consider $+1$ and -1 frameshifts as the two different classes of ribosomal frameshifts in our analysis.

Table 5.1. The four different types of ribosomal frameshift errors, the incorrect base is denoted by N, where N denotes any nucleotide on $B = \{A, C, G, T\}$. Start codon is shown in green and stop codons in red.

	Frameshift	Trinucleotide sequence
Reading frame	0	ATG AAC GTC GGC
Forward 1 base shift	+1	TGA ACG TCG GCN
Forward 2 base shift	+2	GAA CGT CGG CNN
Backward 1 base shift	-1	NAT GAA CGT CGG
Backward 2 base shift	-2	NNA TGA ACG TCG

The "ambush hypothesis" (Seligmann & Pollock, 2004) suggests that out-of-frame stop codons (Figure 5.1), also known as hidden stops, allow rapid termination of frameshifted translations and are selected for (Itzkovitz & Alon, 2007; Seligmann, 2019). Moreover, it is suggested that

the codon usage in some organisms is often biased towards codons that can form a stop codon after a frameshift. Interestingly, the stop codons TAA, TAG and TGA of the SGC do not overlap either with themselves or each other no matter how they are frameshifted. Moreover, most of the abundant codons coding for an amino acid produce a stop codon upon a frameshift, which suggests that the SGC is highly optimized to minimize the effects related to frameshift errors. SGC's ability to terminate the translation process upon a frameshift can be related to its ability to encode multiple overlapping signals or "auxiliary information" (Itzkovitz & Alon, 2007). In the "refined ambush hypothesis" (Abrahams & Hurst, 2018), the authors suggested that genomes follow one of two approaches to counter the effects of frameshift: avoiding frameshifts (GC-rich genomes) or allowing frameshifts but reducing their impacts by early detection (AT-rich genomes). It has been suggested that the SGC is also optimized to reduce the effects of frameshift errors when no out-of-frame stop codon is encountered (Bartonek et al., 2020; Geyer & Madany Mamlouk, 2018).



Out-of-frame stop codon in case of +1 frameshift.

Figure 5.1. Two sense codons producing a stop codon after a +1 frameshift.

Therefore, to minimize the costs of errors, organisms may have evolved either by implementing "increased accuracy" or "increased robustness". It is still unclear how the optimization process was carried out during evolution and most importantly which mechanisms were involved. The robustness of the SGC to frameshift errors also represents an attractive problem from a coding theory point of view. The comma-free codes proposed by Crick was one of the first solutions to the problem of frameshift, where in addition to coding for the 20 amino acids, the comma-free codes had the ability to retrieve and maintain the reading frame in genes (Crick et al., 1957). However, it was shown later that the genetic code could not be a comma-free code.

Another possible solution to this problem can be identified as the circular codes. Like comma-free codes they also have the synchronization property, i.e. to retrieve and maintain the reading frame in genes by using an appropriate window of nucleotides. Based on previous results and the enrichment of X motifs in the functional centers of ribosome involved in protein translation, we have suggested that the circular code X was an ancestor of the SGC. This motivated us to investigate the ability of the X circular code to minimize frameshift errors, in terms of conserving the physicochemical properties of the encoded amino acid sequence.

5.3. Physicochemical properties of amino acids

The biological function of a protein is linked to its 3D structure, molecular dynamics, and the physicochemical properties of the amino acids it contains. For example, a protein's stability and solubility in an aqueous environment depends on the hydrophobicity profiles, size and ionization properties of the constituent amino acids. In Figure 1.4, we showed the chemical structure and various basic profiles on which the 20 amino acids coded by the SGC are divided. As mentioned above, these physicochemical profiles influence the stability, structure and function of proteins.

In a recent study (Bartonek et al., 2020) involving biologically realistic sequences in multiple organisms (*M. jannaschii, Thermococcus kodakarensis, E. coli, Pseudomonas aeruginosa, M. musculus, and H. sapiens*), the authors compared the wild-type protein sequences with their frameshifted counterparts in terms of 604 different amino acid properties classified under 5 categories (alpha: α and turn propensity; beta: β propensity; hydro: hydrophobicity; nuc: nucleobase affinity; other). Their results suggest that, even though frameshifts result in altered protein sequences, some physicochemical properties (hydrophobicity profiles, nucleobase affinity and structural disorder) are quite similar to that of the original protein sequences. As mentioned above, other studies evaluating the optimality of SGC were carried out with fewer amino acid properties (polarity and volume).

In our analysis, we used a larger set of 13 amino acid properties $\mathbb{P}$: charge $\mathbb{P}_C$, hydrophobicity $\mathbb{P}_H$, isoelectric point $\mathbb{P}_{IP}$, melting point $\mathbb{P}_{MP}$, molecular weight $\mathbb{P}_{MW}$, optical rotation $\mathbb{P}_{OR}$, polarity $\mathbb{P}_{Pr}$, polarizability $\mathbb{P}_{Pz}$, size $\mathbb{P}_{Si}$, steric $\mathbb{P}_{St}$, volume $\mathbb{P}_V$, alpha-helix $\mathbb{P}_\alpha$ and beta-sheet conformation $\mathbb{P}_\beta$. To our knowledge, this is the most extensive set of amino acid properties used to evaluate the optimality to translation errors. The amino acid properties were extracted from the AAindex database (Kawashima & Kanehisa, 2000); details are shown in Table 5.2. In the AAindex database, a physicochemical property $\mathbb{P}$ is defined by a set of 20 numerical values representing the absolute or relative value of the property for each amino acid Table 5.3.

For the diverse set of physicochemical properties, we used amino acid substitution matrices to calculate the effect of frameshift errors. We evaluated the effect of frameshift errors by calculating the absolute difference between the physicochemical values of the amino acid encoded by the codon in the reading frame and the amino acid encoded by the frameshifted codons in frames $+1$ and -1.

Table 5.2. An extensive set of thirteen amino acid indices representing various physicochemical properties taken from the AAindex database at http://www.genome.ad.jp/aaindex/.

Property $\mathbb{P}$	AAindex name	Reference
Charge $\mathbb{P}_C$	KLEP840101	(Klein et al., 1984)
Hydrophobicity $\mathbb{P}_H$	FASG890101	(Fasman, 1989)
Isoelectric point $\mathbb{P}_{IP}$	ZIMJ680104	(Zimmerman et al., 1968)
Melting point $\mathbb{P}_{MP}$	FASG760102	(Fasman, 1976)
Molecular weight $\mathbb{P}_{MW}$	FASG760101	(Fasman, 1976)
Optical rotation $\mathbb{P}_{OR}$	FASG760103	(Fasman, 1976)
Polarity $\mathbb{P}_{Pr}$	ZIMJ680103	(Zimmerman et al., 1968)
Polarizability $\mathbb{P}_{Pz}$	CHAM820101	(Charton & Charton, 1982)
Size $\mathbb{P}_{Si}$	DAWD720101	(Dawson, 1972)
Steric $\mathbb{P}_{St}$	CHAM810101	(Charton, 1981)
Volume $\mathbb{P}_V$	BIGC670101	(Bigelow, 1967)
Alpha-helix $\mathbb{P}_\alpha$	CHOP780201	(Chou & Fasman, 1978)
Beta-sheet $\mathbb{P}_\beta$	CHOP780202	(Chou & Fasman, 1978)

We will mathematically explain the methods with the help of examples.

Definition 5.1. Let us denote an AAindex vector as $\mathbf{V_{1\times 20}}(\mathbb{P})$ for a physicochemical property $\mathbb{P} = \{\mathbb{P}_C, \mathbb{P}_H, \mathbb{P}_{IP,}\mathbb{P}_{MP}, \mathbb{P}_{MW}, \mathbb{P}_{OR}, \mathbb{P}_{Pr}, \mathbb{P}_{Pz}, \mathbb{P}_{Si}, \mathbb{P}_{St}, \mathbb{P}_V, \mathbb{P}_\alpha, \mathbb{P}_\beta\}$, where each element $\boldsymbol{v_i}(\mathbb{P})$ of the vector $\mathbf{V}$ is associated with an amino acid $i \in AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$.

Example 13. In Table 5.3, if we take the volume property $\mathbb{P}_V$, the physicochemical value for the amino acid glycine (G) is $v_G(\mathbb{P}_V) = 36.3$.

Definition 5.2. For any physicochemical property $\mathbb{P}$, we construct an amino acid substitution matrix $\mathbf{M}_{20\times 20}(\mathbb{P})$ of absolute differences $m_{ij}(\mathbb{P})$ between the physicochemical values $v_i(\mathbb{P})$ (Definition 5.1) of the amino acid i and $v_j(\mathbb{P})$ of the amino acid j:

$$m_{ij}(\mathbb{P}) = \left|v_i(\mathbb{P}) - v_j(\mathbb{P})\right| \quad (5)$$

where $v_i(\mathbb{P})$ and $v_j(\mathbb{P})$ are the physicochemical values of the amino acids i and j, $i, j \in AA$. The matrices $\mathbf{M}(\mathbb{P})$ are symmetric with diagonal elements equal to zero; if $i = j$, then $m_{ij}(\mathbb{P}) = 0$.

Example 14. In Table 5.3, if we take the volume property $\mathbb{P}_V$, $i = A$ and $j = C$, then the substitution value for the amino acids alanine A and cysteine C is equal to

$$m_{AC}(\mathbb{P}_V) = |v_A(\mathbb{P}_V) - v_C(\mathbb{P}_V)| = |52.6 - 68.3| = 15.7$$

The amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ is provided in Table 5.4.

Remark 11. Different amino acid properties have different scales (Table 5.3); the mean and standard deviation of the 20 amino acids for the melting point property $\mathbb{P}_{Mp}$ are 262.7 and 43.6 respectively, whereas for the optical rotation property $\mathbb{P}_{Or}$ they are -10.6 and 24.3 respectively.

Definition 5.3. In order to make comparisons between the various amino acid properties, each amino acid substitution matrix $\mathbf{M}_{20\times20}(\mathbb{P})$ is normalized by dividing each element of the given matrix by the sum of the whole matrix, leading to the normalized amino acid substitution matrix $\widehat{\mathbf{M}}_{20\times20}(\mathbb{P})$:

$$\widehat{m}_{ij}(\mathbb{P}) = \frac{1000}{\sum_{i=1}^{20}\sum_{j=1}^{20} m_{ij}(\mathbb{P})} m_{ij}(\mathbb{P}) \quad (6)$$

where $m_{ij}(\mathbb{P})$ is defined in Definition 5.2 for the amino acids i and j, $i, j \in AA$.

The matrices $\widehat{\mathbf{M}}(\mathbb{P})$ are also symmetric with diagonal elements equal to zero.

Example 15. The normalized amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ is provided in Table 5.5, where the normalized substitution value for the amino acids glycine G and proline P is equal to $\widehat{m}_{GP}(\mathbb{P}_V) = \widehat{m}_{PG}(\mathbb{P}_V) = \frac{1000}{\sum_{i=1}^{20}\sum_{j=1}^{20} m_{ij}(\mathbb{P})} m_{PG}(\mathbb{P}_V) = \frac{1000}{10790.8} 37.3 = 3.5$.

Table 5.3. Amino acid property vectors for the indices mentioned in Table 5.2, where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ denotes the 20 amino acid alphabet.

Property $\mathbb{P}$	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Charge $\mathbb{P}_C$	0	0	-1	-1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Hydrophobicity $\mathbb{P}_H$	-0.2	-6.0	1.4	2.3	-4.7	0.0	-1.2	-4.8	3.9	-4.7	-3.7	1.0	0.8	1.5	2.1	1.7	0.8	-3.5	-3.3	-1.0
Isoelectric point $\mathbb{P}_{IP}$	6.0	5.1	2.8	3.2	5.5	6.0	7.6	6.0	9.7	6.0	5.7	5.4	6.3	5.7	10.8	5.7	5.7	6.0	5.9	5.7
Melting point $\mathbb{P}_{MP}$	297	178	270	249	284	290	277	284	224	337	283	236	222	185	238	228	253	293	282	344
Molecular weight $\mathbb{P}_{MW}$	89.1	121.2	133.1	147.1	165.2	75.1	155.2	131.2	146.2	131.2	149.2	132.1	115.1	146.2	174.2	105.1	119.1	117.2	204.2	181.2
Optical rotation $\mathbb{P}_{OR}$	1.8	-16.5	5.1	12.0	-34.5	0.0	-38.5	12.4	14.6	-11.0	-10.0	-5.6	-86.2	6.3	12.5	-7.5	-28.0	5.6	-33.7	-10.0
Polarity $\mathbb{P}_{Pr}$	0.0	1.5	49.7	49.9	0.4	0.0	51.6	0.1	49.5	0.1	1.4	3.4	1.6	3.5	52.0	1.7	1.7	0.1	2.1	1.6
Polarizability $\mathbb{P}_{Pz}$	0.05	0.13	0.11	0.15	0.29	0.00	0.23	0.19	0.22	0.19	0.22	0.13	0.13	0.18	0.29	0.06	0.11	0.14	0.41	0.30
Size $\mathbb{P}_{Si}$	2.5	3.0	2.5	5.0	6.5	0.5	6.0	5.5	7.0	5.5	6.0	5.0	5.5	6.0	7.5	3.0	5.0	5.0	7.0	7.0
Steric $\mathbb{P}_{St}$	0.52	0.62	0.76	0.68	0.70	0.00	0.70	1.02	0.68	0.98	0.78	0.76	0.36	0.68	0.68	0.53	0.50	0.76	0.70	0.70
Volume $\mathbb{P}_V$	52.6	68.3	68.4	84.7	113.9	36.3	91.9	102.0	105.1	102.0	97.7	75.7	73.6	89.7	109.1	54.9	71.2	85.1	135.4	116.2
Alpha-helix $\mathbb{P}_\alpha$	1.42	0.7	1.01	1.51	1.13	0.57	1	1.08	1.16	1.21	1.45	0.67	0.57	1.11	0.98	0.77	0.83	1.06	1.08	0.69
Beta-sheet $\mathbb{P}_\beta$	0.83	1.19	0.54	0.37	1.38	0.75	0.87	1.6	0.74	1.3	1.05	0.89	0.55	1.1	0.93	0.75	1.19	1.7	1.37	1.47

Table 5.4. Amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$, where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ denotes the 20 amino acid alphabet. The matrix **M** is symmetric with diagonal elements equal to zero.

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
A	0	15.7	15.8	32.1	61.3	16.3	39.3	49.4	52.5	49.4	45.1	23.1	21.0	37.1	56.5	2.3	18.6	32.5	82.8	63.6
C	15.7	0	0.1	16.4	45.6	32.0	23.6	33.7	36.8	33.7	29.4	7.4	5.3	21.4	40.8	13.4	2.9	16.8	67.1	47.9
D	15.8	0.1	0	16.3	45.5	32.1	23.5	33.6	36.7	33.6	29.3	7.3	5.2	21.3	40.7	13.5	2.8	16.7	67.0	47.8
E	32.1	16.4	16.3	0	29.2	48.4	7.2	17.3	20.4	17.3	13.0	9.0	11.1	5.0	24.4	29.8	13.5	0.4	50.7	31.5
F	61.3	45.6	45.5	29.2	0	77.6	22.0	11.9	8.8	11.9	16.2	38.2	40.3	24.2	4.8	59.0	42.7	28.8	21.5	2.3
G	16.3	32.0	32.1	48.4	77.6	0	55.6	65.7	68.8	65.7	61.4	39.4	37.3	53.4	72.8	18.6	34.9	48.8	99.1	79.9
H	39.3	23.6	23.5	7.2	22.0	55.6	0	10.1	13.2	10.1	5.8	16.2	18.3	2.2	17.2	37.0	20.7	6.8	43.5	24.3
I	49.4	33.7	33.6	17.3	11.9	65.7	10.1	0	3.1	0.0	4.3	26.3	28.4	12.3	7.1	47.1	30.8	16.9	33.4	14.2
K	52.5	36.8	36.7	20.4	8.8	68.8	13.2	3.1	0	3.1	7.4	29.4	31.5	15.4	4.0	50.2	33.9	20.0	30.3	11.1
L	49.4	33.7	33.6	17.3	11.9	65.7	10.1	0.0	3.1	0	4.3	26.3	28.4	12.3	7.1	47.1	30.8	16.9	33.4	14.2
M	45.1	29.4	29.3	13.0	16.2	61.4	5.8	4.3	7.4	4.3	0	22.0	24.1	8.0	11.4	42.8	26.5	12.6	37.7	18.5
N	23.1	7.4	7.3	9.0	38.2	39.4	16.2	26.3	29.4	26.3	22.0	0	2.1	14.0	33.4	20.8	4.5	9.4	59.7	40.5
P	21.0	5.3	5.2	11.1	40.3	37.3	18.3	28.4	31.5	28.4	24.1	2.1	0	16.1	35.5	18.7	2.4	11.5	61.8	42.6
Q	37.1	21.4	21.3	5.0	24.2	53.4	2.2	12.3	15.4	12.3	8.0	14.0	16.1	0	19.4	34.8	18.5	4.6	45.7	26.5
R	56.5	40.8	40.7	24.4	4.8	72.8	17.2	7.1	4.0	7.1	11.4	33.4	35.5	19.4	0	54.2	37.9	24.0	26.3	7.1
S	2.3	13.4	13.5	29.8	59.0	18.6	37.0	47.1	50.2	47.1	42.8	20.8	18.7	34.8	54.2	0	16.3	30.2	80.5	61.3
T	18.6	2.9	2.8	13.5	42.7	34.9	20.7	30.8	33.9	30.8	26.5	4.5	2.4	18.5	37.9	16.3	0	13.9	64.2	45.0
V	32.5	16.8	16.7	0.4	28.8	48.8	6.8	16.9	20.0	16.9	12.6	9.4	11.5	4.6	24.0	30.2	13.9	0	50.3	31.1
W	82.8	67.1	67.0	50.7	21.5	99.1	43.5	33.4	30.3	33.4	37.7	59.7	61.8	45.7	26.3	80.5	64.2	50.3	0	19.2
Y	63.6	47.9	47.8	31.5	2.3	79.9	24.3	14.2	11.1	14.2	18.5	40.5	42.6	26.5	7.1	61.3	45.0	31.1	19.2	0

Table 5.5. Normalized amino acid substitution matrix $\widehat{\mathbf{M}}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$, where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ denotes the 20 amino acid alphabet. The matrix $\widehat{\mathbf{M}}$ is symmetric with diagonal elements equal to zero.

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
A	0	1.5	1.5	3.0	5.7	1.5	3.6	4.6	4.9	4.6	4.2	2.1	1.9	3.4	5.2	0.2	1.7	3.0	7.7	5.9
C	1.5	0	0.0	1.5	4.2	3.0	2.2	3.1	3.4	3.1	2.7	0.7	0.5	2.0	3.8	1.2	0.3	1.6	6.2	4.4
D	1.5	0.0	0	1.5	4.2	3.0	2.2	3.1	3.4	3.1	2.7	0.7	0.5	2.0	3.8	1.3	0.3	1.5	6.2	4.4
E	3.0	1.5	1.5	0	2.7	4.5	0.7	1.6	1.9	1.6	1.2	0.8	1.0	0.5	2.3	2.8	1.3	0.0	4.7	2.9
F	5.7	4.2	4.2	2.7	0	7.2	2.0	1.1	0.8	1.1	1.5	3.5	3.7	2.2	0.4	5.5	4.0	2.7	2.0	0.2
G	1.5	3.0	3.0	4.5	7.2	0	5.2	6.1	6.4	6.1	5.7	3.7	3.5	4.9	6.7	1.7	3.2	4.5	9.2	7.4
H	3.6	2.2	2.2	0.7	2.0	5.2	0	0.9	1.2	0.9	0.5	1.5	1.7	0.2	1.6	3.4	1.9	0.6	4.0	2.3
I	4.6	3.1	3.1	1.6	1.1	6.1	0.9	0	0.3	0.0	0.4	2.4	2.6	1.1	0.7	4.4	2.9	1.6	3.1	1.3
K	4.9	3.4	3.4	1.9	0.8	6.4	1.2	0.3	0	0.3	0.7	2.7	2.9	1.4	0.4	4.7	3.1	1.9	2.8	1.0
L	4.6	3.1	3.1	1.6	1.1	6.1	0.9	0.0	0.3	0	0.4	2.4	2.6	1.1	0.7	4.4	2.9	1.6	3.1	1.3
M	4.2	2.7	2.7	1.2	1.5	5.7	0.5	0.4	0.7	0.4	0	2.0	2.2	0.7	1.1	4.0	2.5	1.2	3.5	1.7
N	2.1	0.7	0.7	0.8	3.5	3.7	1.5	2.4	2.7	2.4	2.0	0	0.2	1.3	3.1	1.9	0.4	0.9	5.5	3.8
P	1.9	0.5	0.5	1.0	3.7	3.5	1.7	2.6	2.9	2.6	2.2	0.2	0	1.5	3.3	1.7	0.2	1.1	5.7	3.9
Q	3.4	2.0	2.0	0.5	2.2	4.9	0.2	1.1	1.4	1.1	0.7	1.3	1.5	0	1.8	3.2	1.7	0.4	4.2	2.5
R	5.2	3.8	3.8	2.3	0.4	6.7	1.6	0.7	0.4	0.7	1.1	3.1	3.3	1.8	0	5.0	3.5	2.2	2.4	0.7
S	0.2	1.2	1.3	2.8	5.5	1.7	3.4	4.4	4.7	4.4	4.0	1.9	1.7	3.2	5.0	0	1.5	2.8	7.5	5.7
T	1.7	0.3	0.3	1.3	4.0	3.2	1.9	2.9	3.1	2.9	2.5	0.4	0.2	1.7	3.5	1.5	0	1.3	5.9	4.2
V	3.0	1.6	1.5	0.0	2.7	4.5	0.6	1.6	1.9	1.6	1.2	0.9	1.1	0.4	2.2	2.8	1.3	0	4.7	2.9
W	7.7	6.2	6.2	4.7	2.0	9.2	4.0	3.1	2.8	3.1	3.5	5.5	5.7	4.2	2.4	7.5	5.9	4.7	0	1.8
Y	5.9	4.4	4.4	2.9	0.2	7.4	2.3	1.3	1.0	1.3	1.7	3.8	3.9	2.5	0.7	5.7	4.2	2.9	1.8	0

5.4. Parameters for measuring frameshift optimality

In this extensive analysis for measuring frameshift optimality of different codes, we defined two different score parameters: frameshift code score and frameshift dicodon score. To measure optimality, the frameshift code score takes into account all the codons of a given code Y in comparison with the frameshifted codons from its two permutated codes Y_1 and Y_2. In the case of maximal C^3 self-complementary circular codes ($\mathbb{X}$), all 60 codons i.e. 20 codons each of Y, Y_1 and Y_2 are considered to calculate the frameshift code score. Therefore, this approach can be viewed as a codon score. The frameshift dicodon score takes into account a pair of codons or a dicodon (in the reading frame) from a given code Y, where the frameshift is analyzed according to 1 or 2 base shifts (not in the reading frame) in the dicodon. Therefore, this approach can be viewed as a dicodon score.

The different codes Y used in this analysis are: (i) the maximal C^3 self-complementary trinucleotide circular code X identified in genes (Definition 2.9); (ii) the 215 maximal C^3 self-complementary trinucleotide circular codes $\mathbb{X}\backslash X$; and (iii) the standard genetic code SGC. Both frameshift parameters are based on the average differences in the various physicochemical properties between the amino acids (AA) in the original reading frame and those after a frameshift ($+1$ and -1). From a biological point of view, forward ($+1$) and backward (-1) frameshifts are fundamentally different events (Abrahams & Hurst, 2018). Forward frameshifts are assumed to be the more frequent form of accidental ribosomal slippage. As translation occurs in the 5' to 3' direction, the molecular mechanics required to halt and reverse the direction of translation during a backward frameshift are likely to be more complex and require greater energy than for a ribosome to skip to the $+1$ frame in the same direction.

Therefore, we defined two frameshift optimality scores: one for the shifted frame $f = 1$ and one for the shifted frame $f = -1$. We already defined the amino acid substitution matrices for the diverse set of physicochemical properties used to calculate these parameters. Next, we define these frameshift optimality parameters in detail.

5.4.1. Frameshift code score

The frameshift code score takes into account frameshift errors from a code Y point of view. A codon $c = l_0 l_1 l_2$ of a code $Y \subseteq B^3$ is associated with the reading frame $f = 0$, the shifted codon $\mathcal{P}(c) = l_1 l_2 l_0$ of the code $Y_1 = \mathcal{P}(Y) \subseteq B^3$ is associated with the shifted frame $f = 1$ ($+1$) and the shifted codon $\mathcal{P}^2(c) = l_2 l_0 l_1$ of the code $Y_2 = \mathcal{P}^2(Y) \subseteq B^3$ is associated with the shifted frame $f = -1$ ($+2$).

Therefore, the code Y is associated with the reading frame $f = 0$, the shifted code Y_1 is associated with the shifted frame $f = 1$ and the shifted code Y_2 is associated with the shifted

frame $f = -1$. The frameshift code score is defined by the average difference for a given amino acid property $\mathbb{P}$ when all codons of a given code Y are substituted into all shifted codons of a shifted code Y_1 or Y_2. As we have mentioned above, these parameters are based on the expansive set of physicochemical properties of amino acid. Therefore, only the sense codons (i.e. codons coding for an amino acid) are considered in a code Y; we exclude the three stop codons from the calculations. Frameshift optimality is measured separately for each of the three classes of codes Y defined above.

Let us define the frameshift code score mathematically to understand the method used to evaluate the frameshift optimality.

Definition 5.4. The three stop codons $S = \{TAA, TAG, TGA\}$, which do not code for any amino acid are excluded from the analysis. We define here the two permutation sets of the stop codons *S*:

$S_1 = \mathcal{P}(S) = \{AAT, AGT, GAT\}$ and $S_2 = \mathcal{P}^2(S) = \{ATA, ATG, GTA\}$.

Definition 5.5. The frameshift code score in a $+1$ frameshift of a code Y is denoted as $CS_{+1}(Y)$ and defined by

$$CS_{+1}(Y, \mathbb{P}) = \frac{1}{|Y \backslash (S \cup S_2)|} \sum_{c \in Y \backslash (S \cup S_2)} \widehat{m}_{ij}(\mathbb{P}) \tag{7}$$

where the codon $c \in Y \backslash (S \cup S_2)$ belongs to the code Y excluding the stop codons S and the codons S_2 (as S_2 in frame 0 leads to $\mathcal{P}(S_2) = S$ in $+1$ frameshift), $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Definition 5.3) of an AA property $\mathbb{P}$ where *i* and *j* are the amino acids coded by the codons $c \in Y$ and $\mathcal{P}(c) \in Y_1 = \mathcal{P}(Y)$ (we recall that the matrix $\widehat{\mathbf{M}}$ is symmetric).

Similarly, in a -1 frameshift of a code Y, the frameshift code score $CS_{-1}(Y)$ is defined by

$$CS_{-1}(Y, \mathbb{P}) = \frac{1}{|Y \backslash (S \cup S_1)|} \sum_{c \in Y \backslash (S \cup S_1)} \widehat{m}_{ij}(\mathbb{P}) \tag{8}$$

where the codon $c \in Y \backslash (S \cup S_1)$ belongs to the code Y excluding the stop codons S and the codons S_1 (as S_1 in frame 0 leads to $\mathcal{P}^2(S_1) = S$ in -1 frameshift), $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Definition 5.3) of an AA property $\mathbb{P}$ where *i* and *j* are the amino acids coded by the codons $c \in Y$ and $\mathcal{P}^2(c) \in Y_2 = \mathcal{P}^2(Y)$.

Remark 12. For the circular code $Y = X$, $X \cap S = \emptyset$ (X has 20 sense codons, defined in (1)), $X \cap S_2 = \{GTA\}$ (X_1 has 19 sense codons and one stop codon $\mathcal{P}(\{GTA\}) = \{TAG\}$, defined in (3)) and $X \cap S_1 = \{AAT, GAT\}$ (X_2 has 18 sense codons and two stop codons

$\mathcal{P}^2(\{AAT, GAT\}) = \{TAA, TGA\}$, defined in (4)). Thus, for Equation (4), $X\backslash(S \cup S_2) = X\backslash\{GTA\}$ and $|X\backslash\{GTA\}| = 20 - 1 = 19$ and for Equation (8), $X\backslash(S \cup S_1) = X\backslash\{AAT, GAT\}$ and $|X\backslash\{AAT, GAT\}| = 20 - 2 = 18$.

Remark 13. For the standard genetic code $Y = \mathrm{SGC} = B^3$, $Y \cap S = S$ (Y has 61 sense codons and three stop codons S), $Y \cap S_2 = S_2$ (Y_1 has 61 sense codons and three stop codons $\mathcal{P}(S_2) = S$) and $X \cap S_1 = S_1$ (Y_2 has 61 sense codons and three stop codons $\mathcal{P}^2(S_1) = S$). Thus, for Equation (7), $Y\backslash(S \cup S_2) = B^3\backslash\{ATA, ATG, GTA, TAA, TAG, TGA\}$ and $|Y\backslash(S \cup S_2)| = 64 - 6 = 58$ and for Equation (8), $Y\backslash(S \cup S_1) = B^3\backslash\{AAT, AGT, GAT, TAA, TAG, TGA\}$ and $|Y\backslash(S \cup S_1)| = 64 - 6 = 58$.

Remark 14. For each of the 215 maximal C^3 self-complementary trinucleotide circular codes $\mathbb{X}\backslash X$, we use the same approach to calculate the frameshift optimality score. The codes having none, one or several stop codons are analysed similarly by considering only the sense codons.

5.4.2. Frameshift dicodon score

The frameshift dicodon score takes into account the frameshift errors from a code motif point of view, precisely a motif with two consecutive trinucleotides (dicodon) from a code Y. A codon $c = l_0 l_1 l_2$ of a code $Y \subseteq B^3$ is associated with the reading frame $f = 0$. The shifted frames $f = 1$ (+1) and $f = -1$ (+2) are obtained from the dicodons.

Let us denote a dicodon as $c \cdot c' = l_0 l_1 l_2 \cdot l'_0 l'_1 l'_2$, such that the codon $c' = l'_0 l'_1 l'_2$ also belongs to the code $Y \subseteq B^3$. Let the map $Q: B^3 \times B^3 \to B^3$. Then, the shifted codon $Q(c \cdot c') = l_1 l_2 l'_0$ is associated with the shifted frame $f = 1$ and the shifted codon $Q^2(c \cdot c') = l_2 l'_0 l'_1$ is associated with the shifted frame $f = -1$. The frameshift dicodon score is defined by the average difference for a given amino acid property $\mathbb{P}$ when all codons $c = l_0 l_1 l_2$ of all dicodons $c \cdot c' = l_0 l_1 l_2 \cdot l'_0 l'_1 l'_2$ of a given code Y are "substituted" into the shifted codons $Q(c \cdot c') = l_1 l_2 l'_0$ or $Q^2(c \cdot c') = l_2 l'_0 l'_1$. As with the code score, only the sense codons are considered in the dicodons of a code Y and frameshift optimality is measured separately for each of the three classes of codes Y.

Let us define the frameshift dicodon score mathematically to understand the method used to evaluate the frameshift optimality.

Definition 5.6. Let us denote the set of dicodons containing a stop codon as $DS = \{c \cdot c'\}$, where $c \in S$ or $c' \in S$. The two sets of dicodons that result in a stop codon are:

$DS_1 = \{NTA.ANN, NTA.GNN, NTG.ANN\}$ for the $+1$ frameshift and

$DS_2 = \{NNT.AAN, NNT.AGN, NNT.GAN\}$ for the -1 frameshift, N being any letter on B^3.

Definition 5.7. The frameshift dicodon score in a $+1$ frameshift of a code Y is denoted as $DS_{+1}(Y)$ and defined by

$$DS_{+1}(Y, \mathbb{P}) = \frac{1}{|Y^2 \backslash (DS \cup DS_1)|} \sum_{c \cdot c' \in Y^2 \backslash (DS \cup DS_1)} \widehat{m}_{ij}(\mathbb{P}) \tag{9}$$

where the dicodon $c \cdot c'$ belong to the code Y^2 excluding the stop codons DS and DS_1, $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Definition 5.3) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $Q(c \cdot c')$.

Similarly, the frameshift dicodon score $DS_{-1}(Y)$ in a -1 frameshift of a code Y is defined by

$$DS_{-1}(Y, \mathbb{P}) = \frac{1}{|Y^2 \backslash (DS \cup DS_2)|} \sum_{c \cdot c' \in Y^2 \backslash (DS \cup DS_2)} \widehat{m}_{ij}(\mathbb{P}) \tag{10}$$

where the dicodon $c \cdot c'$ belong to the code Y^2 excluding the stop codons DS and DS_2, $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Equation (6)) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $Q^2(c \cdot c')$.

Remark 15. In contrast to the frameshift code score, the shifted codon $Q(c \cdot c')$ does not necessarily belong to the code $Y_1 = \mathcal{P}(Y) \subseteq B^3$ and the shifted codon $Q^2(c \cdot c')$ does not necessarily belong to the code $Y_2 = \mathcal{P}^2(Y) \subseteq B^3$ (see Example 16).

Example 16. In the case of code $Y =$ the maximal C^3 self-complementary trinucleotide circular code X, where

$$\begin{aligned} X &= \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\ &\quad GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}, \\ \mathcal{P}(X) = X_1 &= \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, \\ &\quad GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \text{ and} \\ \mathcal{P}^2(X) = X_2 &= \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, \\ &\quad CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}. \end{aligned}$$

If we take the dicodon $c \cdot c' = AAT \cdot CTC$, where $c, c' \in X$, then the $+1$ frameshifted codon $Q(c \cdot c') = Q(AAT \cdot CTC) = ATC \notin X_1$ and the -1 frameshifted codon $Q^2(c \cdot c') = (AAT.CTC) = TCT \notin X_2$.

5.4.3. Multi-objective score parameter

Here, we define a multi-objective parameter to compare the frameshift optimality of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$. This parameter is based on either the frameshift code score or the frameshift dicodon score, and takes into account several amino acid properties simultaneously. We calculate a multi-objective score for each code in $\mathbb{X}$, representing the number of AA properties for which it has a better frameshift optimality when compared to the circular code X. We will explain this parameter mathematically.

Definition 5.8. In order to compare the frameshift optimality of $|\mathbb{X}| = 216$ maximal C^3 self-complementary circular codes $\mathbb{X}$ when a combination of the $|\mathbb{P}|$ (=13 AA physicochemical properties $\mathbb{P}$) is taken into account, we have calculated the number N_i, for $i = 0, \ldots, |\mathbb{P}|$, of AA properties that were optimized better with the codes x, $x \in \mathbb{X}\backslash\mathrm{X}$, than with the circular code X. Hence, for $i = 0, \ldots, |\mathbb{P}|$,

$$N_i(\mathcal{S}) = \sum_{x \in \mathbb{X}} \Delta_i \left(\sum_{j=1}^{|\mathbb{P}|} \delta\left(x, \mathbb{P}_j\right) \right) \tag{11}$$

where

$$\delta\left(x, \mathbb{P}_j\right) = \begin{cases} 1 & \text{if } \mathcal{S}\left(x, \mathbb{P}_j\right) \leq \mathcal{S}\left(X, \mathbb{P}_j\right) \\ 0 & \text{otherwise} \end{cases},$$

$$\Delta_i(k) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases},$$

the code score $\mathcal{S} \in \{CS_{+1}, CS_{-1}, DS_{+1}, DS_{-1}\}$ and

$j \in \mathbb{P} = \left\{\mathbb{P}_C, \mathbb{P}_H, \mathbb{P}_{IP}, \mathbb{P}_{MP}, \mathbb{P}_{MW}, \mathbb{P}_{OR}, \mathbb{P}_{Pr}, \mathbb{P}_{Pz}, \mathbb{P}_{Si}, \mathbb{P}_{St}, \mathbb{P}_V, \mathbb{P}_\alpha, \mathbb{P}_\beta\right\}$.

Remark 16. If $x = X$ then $\delta\left(x, \mathbb{P}_j\right) = 1$ for any $\mathbb{P}_j$, thus $\sum_{j=1}^{|\mathbb{P}|} \delta\left(x, \mathbb{P}_j\right) = |\mathbb{P}|$ and $N_{|\mathbb{P}|}(\mathcal{S}) \geq 1$.

Remark 17. If $N_{|\mathbb{P}|}(\mathcal{S}) = 1$ then the X circular code is optimal among its combinatorial class of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$.

Remark 18. $\sum_{i=0}^{|\mathbb{P}|} N_i(\mathcal{S}) = |\mathbb{X}|$.

We explained the two different approaches to evaluate the effects of frameshift errors. We have defined the various parameters used in this analysis with the expansive set of AA physicochemical properties and their substitution matrices. Next, we will discuss the results obtained from this analysis.

5.5. Frameshift optimality of circular code X and the standard genetic code

In order to evaluate the frameshift optimality, we first estimated the capacity of the circular code X to reduce the effects of a frameshift error, and compared it to the capacity of the standard genetic code (SGC). To estimate the effects of either a $+1$ or -1 frameshift error on the encoded amino acids (AA), we calculated the frameshift optimality scores defined in section 5.4. A smaller score thus implies a smaller effect of the frameshift error, and hence suggests a better frameshift optimality. We will discuss the results obtained from the comparison of frameshift code score and comparison of frameshift dicodon score separately.

5.5.1. Comparison of frameshift code score

First, we compared the frameshift code scores (Definition 5.5) of the circular code X and the SGC, after a frameshift error of either $+1$ or -1. We computed the frameshift code scores $CS_{+1}(Y)$ after a $+1$ frameshift and $CS_{-1}(Y)$ after a -1 frameshift of a code Y, where $Y = X$ for the circular code X and $Y = \text{SGC}$ for the standard genetic code, for a set of 13 fundamental AA properties (Table 5.2). To recall, these scores are based on the absolute difference between the physicochemical properties for the AA coded by the non-shifted codons of Y and the shifted codons of Y_1 for the $+1$ frameshift and of Y_2 for the -1 frameshift. The results for the circular code X and the standard genetic code SGC are shown in Figure 5.2 and Figure 5.3, for $+1$ and -1 frameshift errors respectively. The 13 AA properties $\mathbb{P}$ are ordered according to the difference between the code scores for the SGC and X for $+1$ frameshift, and retained throughout the results for comparison purposes.

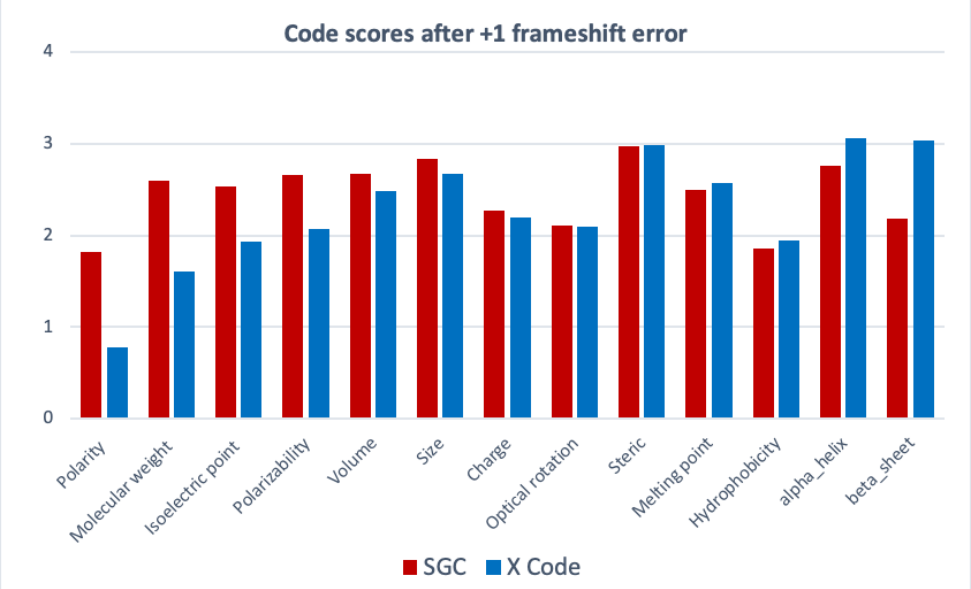


Figure 5.2. Frameshift code score CS_{+1} (Equation (7)) after $+1$ frameshift error for the circular code X and the standard genetic code SGC.

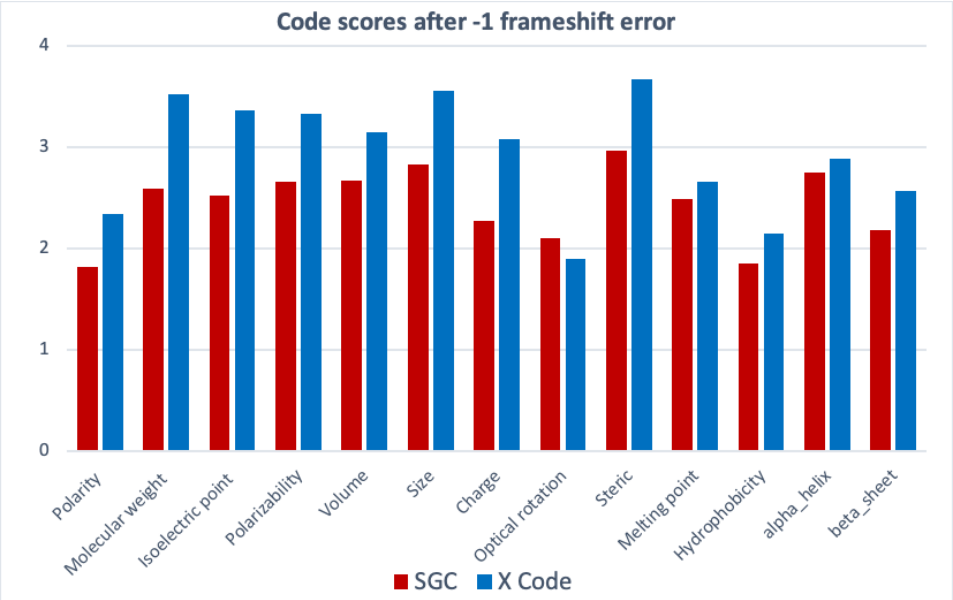


Figure 5.3. Frameshift code score CS_{-1} (Equation (8)) after -1 frameshift error for the circular code X and the standard genetic code SGC.

We observe that in the case of the SGC, the code scores obtained for the 13 AA properties are equal for $+1$ and -1 frameshifts. This equality of scores for both $+1$ and -1 frameshifts can be explained by the symmetry property of the 64 codons of SGC. Thus, for all $\mathbb{P}$, $CS_{+1}(\mathrm{SGC}, \mathbb{P}) = CS_{-1}(\mathrm{SGC}, \mathbb{P})$.

However, the code scores obtained for the circular code X are clearly different for $+1$ and -1 frameshifts, i.e. for all $\mathbb{P}$, $CS_{+1}(X, \mathbb{P}) \neq CS_{-1}(X, \mathbb{P})$. In the case of $+1$ frameshift, the code scores obtained for polarity $\mathbb{P}_{Pr}$, molecular weight $\mathbb{P}_{MW}$, isoelectric point $\mathbb{P}_{IP}$, polarizability $\mathbb{P}_{Pz}$, volume $\mathbb{P}_{V}$, size $\mathbb{P}_{Si}$ and charge $\mathbb{P}_{C}$ are smaller for X than for SGC (Figure 5.2), i.e. for $\mathbb{P} \in \{\mathbb{P}_{Pr}, \mathbb{P}_{MW}, \mathbb{P}_{IP}, \mathbb{P}_{Pz}, \mathbb{P}_{V}, \mathbb{P}_{Si}, \mathbb{P}_{C}\}$,

$$CS_{+1}(X, \mathbb{P}) < CS_{+1}(\mathrm{SGC}, \mathbb{P}). \quad (12)$$

For the remaining properties the code scores obtained are larger for X than for SGC (Figure 5.2), i.e. for $\mathbb{P} \in \{\mathbb{P}_{OR}, \mathbb{P}_{St}, \mathbb{P}_{MP}, \mathbb{P}_{H}, \mathbb{P}_{\alpha}, \mathbb{P}_{\beta}\}$,

$$CS_{+1}(X, \mathbb{P}) > CS_{+1}(\mathrm{SGC}, \mathbb{P}). \quad (13)$$

In contrast, for -1 frameshift, the code scores obtained for most of the properties $\mathbb{P}$ are larger for X than for SGC (Figure 5.3), with the exception of optical rotation $\mathbb{P}_{OR}$, i.e. for $\mathbb{P} \neq \mathbb{P}_{OR}$,

$$CS_{-1}(X, \mathbb{P}) > CS_{-1}(\mathrm{SGC}, \mathbb{P}). \quad (14)$$

To summarize the comparison of the code scores, in the case of $+1$ frameshift, the circular code X has a better frameshift optimality than the standard genetic code SGC for 7 AA

physicochemical properties, namely polarity $\mathbb{P}_{Pr}$, molecular weight $\mathbb{P}_{MW}$, isoelectric point $\mathbb{P}_{IP}$, polarizability $\mathbb{P}_{Pz}$, volume $\mathbb{P}_{V}$, size $\mathbb{P}_{Si}$ and charge $\mathbb{P}_{C}$. But, in the case of -1 frameshift error, the SGC has better frameshift optimality than the circular code X for all AA properties except for optical rotation $\mathbb{P}_{OR}$.

5.5.2. Comparison of frameshift dicodon score

While introducing circular codes, we highlighted their role in reading frame maintenance. Circular codes have the ability to retrieve and synchronize the reading frame using an appropriate window of nucleotides. In fact, for the circular code X, a window of at most 13 consecutive nucleotides is sufficient to successfully identify the reading frame in genes. This led us to consider the frameshift optimality for the same AA properties at the motif level, and specifically for a dicodon. Also, we mentioned previously that some dicodons are found to be linked with the expression level of genes; more precisely, dicodons that belong to the circular code X.

Here, we compare the frameshift dicodon scores (Definition 5.7) of the circular code X and the SGC, after a frameshift error of either $+1$ or -1. We computed the frameshift dicodon scores $DS_{+1}(Y)$ after $+1$ frameshift (Figure 5.4) and $DS_{-1}(Y)$ after -1 frameshift (Figure 5.5) of a code Y, where $Y = X$ for the circular code X and $Y = \mathrm{SGC}$ for the standard genetic code, for the same set of 13 fundamental AA properties (Table 5.2).

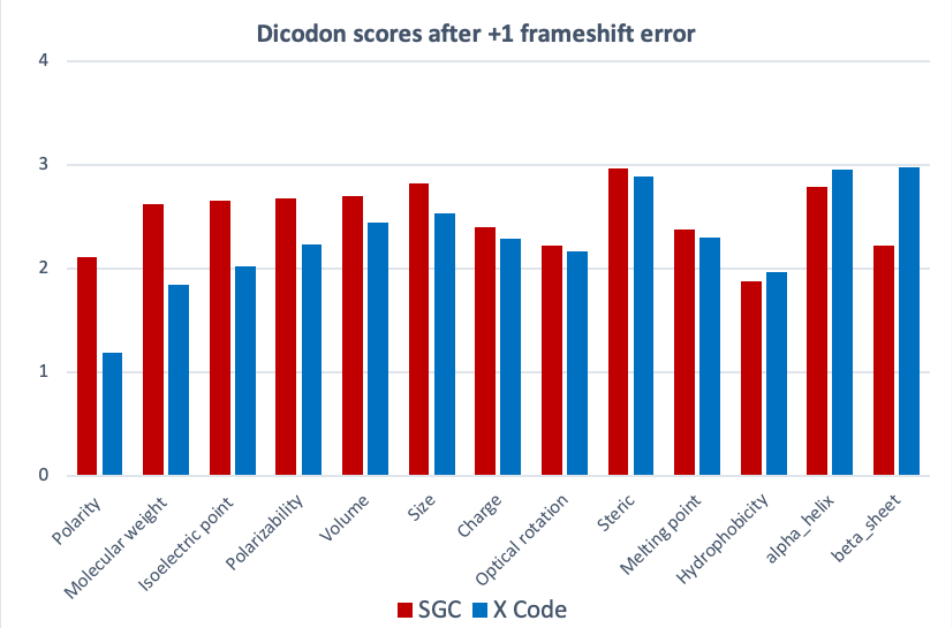


Figure 5.4. Frameshift dicodon score DS_{+1} (Equation (9)) after $+1$ frameshift error for the circular code X and the standard genetic code SGC.

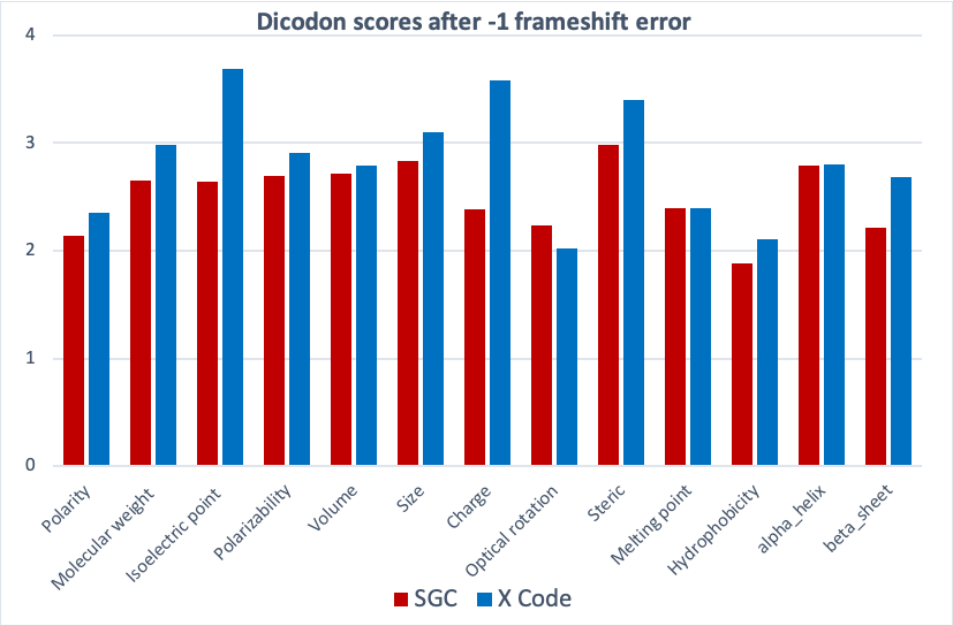


Figure 5.5. Frameshift dicodon score DS_{-1} (Equation (10)) after -1 frameshift error for the circular code X and the standard genetic code SGC.

Again, we observe that the dicodon scores in case of the SGC are same for the $+1$ and -1 frameshifts with the 13 AA properties $\mathbb{P}$. As expected, $DS_{+1}(\mathrm{SGC}, \mathbb{P}) = DS_{-1}(\mathrm{SGC}, \mathbb{P})$. The dicodon scores of X are clearly different for $+1$ and -1 frameshifts, i.e. for all $\mathbb{P}$, $DS_{+1}(X, \mathbb{P}) \neq DS_{-1}(X, \mathbb{P})$.

In case of $+1$ frameshift, the circular code X has smaller scores than the SGC for all AA properties except hydrophobicity $\mathbb{P}_H$, alpha-helix $\mathbb{P}_\alpha$ and beta-sheet $\mathbb{P}_\beta$ (Figure 5.4), i.e. for $\mathbb{P} \neq \{\mathbb{P}_H, \mathbb{P}_\alpha, \mathbb{P}_\beta\}$

$$DS_{+1}(X, \mathbb{P}) < DS_{+1}(\mathrm{SGC}, \mathbb{P}). \quad (15)$$

In contrast, in case of -1 frameshift, the SGC achieves smaller scores than the circular code X for all AA properties (Figure 5.5), except for the optical rotation $\mathbb{P}_{OR}$ and the melting point $\mathbb{P}_{MP}$, i.e. for $\mathbb{P} \neq \{\mathbb{P}_{OR}, \mathbb{P}_{MP}\}$

$$DS_{-1}(X, \mathbb{P}) > DS_{-1}(\mathrm{SGC}, \mathbb{P}). \quad (16)$$

To summarize the comparison of the dicodon scores, in the case of $+1$ frameshift, the circular code X is better optimized than the standard genetic code SGC for 10 AA properties (except hydrophobicity $\mathbb{P}_H$, alpha-helix $\mathbb{P}_\alpha$ and beta-sheet $\mathbb{P}_\beta$). Whereas, in the case of -1 frameshift error, the SGC has better frameshift optimality than the circular code X for all AA properties except for optical rotation $\mathbb{P}_{OR}$ and the melting point $\mathbb{P}_{MP}$.

5.6. Frameshift optimality of the 216 maximal circular codes

Since the discovery of the circular code X (1996) in the reading frame of genes, it is still unclear why this particular code was chosen among its combinatorial class $\mathbb{X}$ of 216 maximal self-complementary C^3 circular codes. Previous studies, based on combinatorics and graph theory, did not provide any answers to explain this particular preference. Interestingly, transformations of the circular code X by letter invariance with respect to complementarity lead to circular codes in $\mathbb{X}$ with combinatorial properties identical to that of X. Our analysis on the frameshift optimality unexpectedly revealed that this particular preference is of biological and biochemical importance. The results obtained for the circular code X and SGC suggest that the circular code X has a better frameshift optimality to minimize the effects of $+1$ frameshift errors, but this is not the case for -1 frameshift errors. Here in the second part of the results, we will evaluate the frameshift optimality of the 216 maximal self-complementary C^3 circular codes $\mathbb{X}$.

We implemented the same approach that we used to compare the frameshift optimality of the circular code X and the standard genetic code (SGC). We calculated the frameshift optimality scores for each of the 216 circular codes in $\mathbb{X}$, and the code which has the lowest score for each of the 13 AA properties is taken as optimal for the respective property. To recall, a smaller score indicates better optimality after frameshift errors.

5.6.1. Comparison of frameshift code score

First, we compared the frameshift code scores (Definition 5.5) to evaluate frameshift optimality. We computed the frameshift code scores $CS_{+1}(Y)$ after a $+1$ frameshift and $CS_{-1}(Y)$ after a -1 frameshift of a code Y, where $Y = \mathbb{X}$ for the 216 maximal self-complementary C^3 circular codes, for the set of 13 fundamental AA properties (Table 5.2). For each of the given properties, the optimal code, i.e. the circular code having the lowest score is denoted as $\min_{216} CS_{+1/-1}(\mathbb{X}, \mathbb{P})$ named $\min 216$ for simplicity. The results for the circular code X and $\min 216$ are shown in Figure 5.6 and Figure 5.7, for $+1$ and -1 frameshift errors respectively.

In the case of both $+1$ and -1 frameshifts, we observe that there is a circular code among its combinatorial class which has a better frameshift optimality than the circular code X. For all 13 AA properties, i.e. for

$$\mathbb{P} = \{\mathbb{P}_C, \mathbb{P}_H, \mathbb{P}_{IP,}\mathbb{P}_{MP}, \mathbb{P}_{MW}, \mathbb{P}_{OR}, \mathbb{P}_{Pr}, \mathbb{P}_{Pz}, \mathbb{P}_{Si}, \mathbb{P}_{St}, \mathbb{P}_V, \mathbb{P}_\alpha, \mathbb{P}_\beta\},$$

$$CS_{+1}(X, \mathbb{P}) > CS_{+1}(\min 216\,, \mathbb{P}) \tag{17}$$

and

$$CS_{-1}(X, \mathbb{P}) > CS_{-1}(\min 216\,, \mathbb{P}). \tag{18}$$

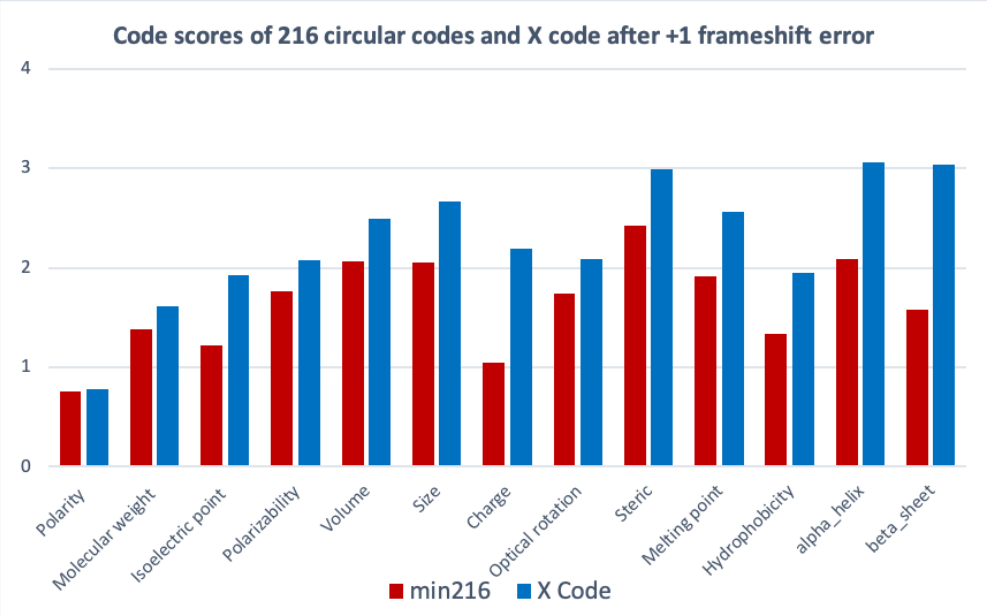


Figure 5.6. Frameshift code score CS_{+1} (Equation (7)) after $+1$ frameshift error for the circular code X and the optimal code among the 216 maximal self-complementary C^3 circular codes denoted by min 216.

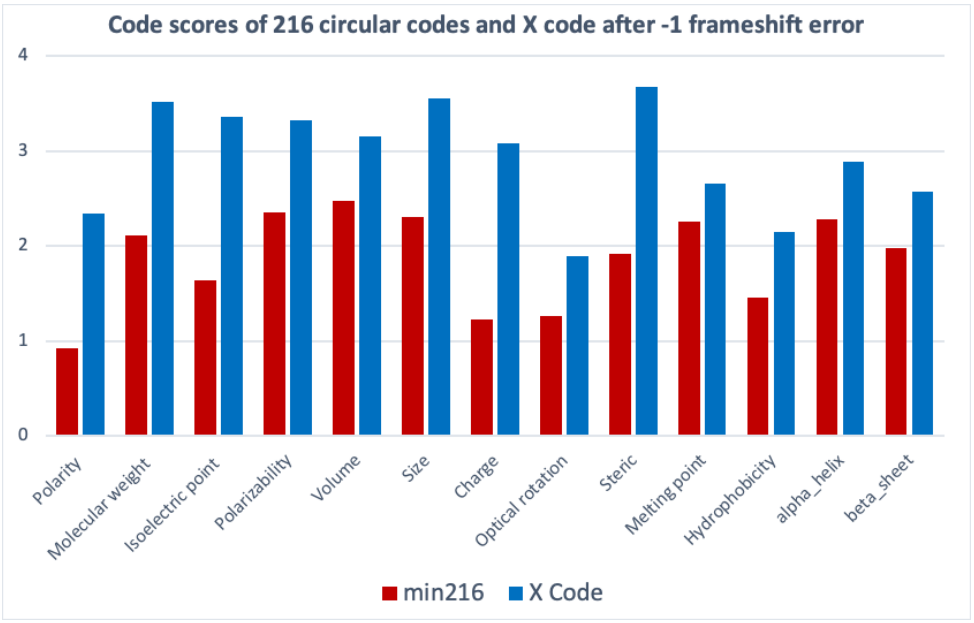


Figure 5.7. Frameshift code score CS_{-1} (Equation (8)) after -1 frameshift error for the circular code X and the optimal code among the 216 maximal self-complementary C^3 circular codes denoted by min 216.

5.6.2. Comparison of frameshift dicodon score

Here, we compare the frameshift dicodon scores (Definition 5.7) to evaluate frameshift optimality. We computed the frameshift dicodon scores $DS_{+1}(Y)$ after a $+1$ frameshift and

$DS_{-1}(Y)$ after a -1 frameshift of a code Y, where $Y = \mathbb{X}$ for the 216 maximal self-complementary C^3 circular codes, for the set of 13 fundamental AA properties (Table 5.2). For each of the given properties, the optimal code, i.e. the circular code having the lowest score is denoted as $\min_{216} DS_{+1/-1}(\mathbb{X}, \mathbb{P})$ named $\min 216$ for simplicity. The results for the circular code X and $\min 216$ are shown in Figure 5.8 and Figure 5.9, for $+1$ and -1 frameshift errors respectively.

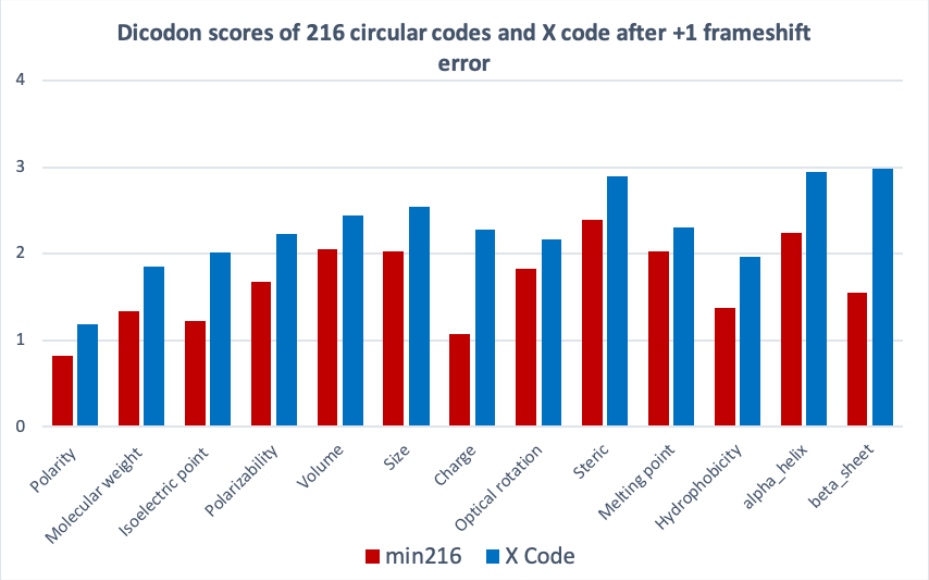


Figure 5.8. Frameshift dicodon score DS_{+1} (Equation (9)) after $+1$ frameshift error for the circular code X and the optimal code among the 216 maximal self-complementary C^3 circular codes denoted by $\min 216$.

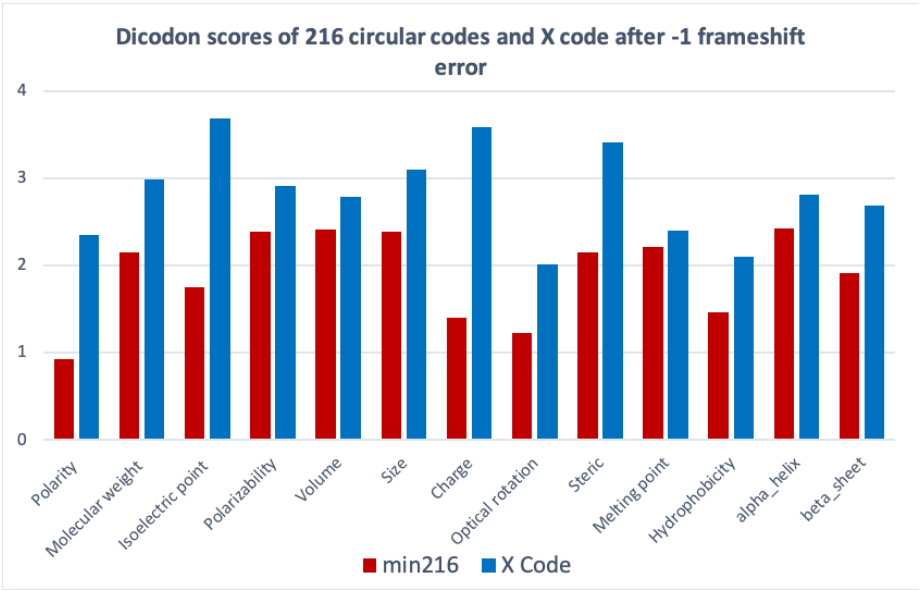


Figure 5.9. Frameshift dicodon score DS_{-1} (Equation (10)) after -1 frameshift error for the circular code X and the optimal code among the 216 maximal self-complementary C^3 circular codes denoted by $\min 216$.

Again, in the case of both $+1$ and -1 frameshifts, there is a circular code among its combinatorial class which has a better frameshift optimality than the circular code X. For all 13 AA properties, i.e. for $\mathbb{P} = \{\mathbb{P}_C, \mathbb{P}_H, \mathbb{P}_{IP,}\mathbb{P}_{MP}, \mathbb{P}_{MW}, \mathbb{P}_{OR}, \mathbb{P}_{Pr}, \mathbb{P}_{Pz}, \mathbb{P}_{Si}, \mathbb{P}_{St}, \mathbb{P}_V, \mathbb{P}_\alpha, \mathbb{P}_\beta\}$,

$$DS_{+1}(X, \mathbb{P}) > DS_{+1}(\min 216\,, \mathbb{P}) \quad (19)$$

and

$$DS_{-1}(X, \mathbb{P}) > DS_{-1}(\min 216\,, \mathbb{P}). \quad (20)$$

From the comparison of both code and dicodon scores for the 216 maximal self-complementary C^3 circular codes, we conclude that there exists a circular code other than the circular code X with a better frameshift optimality, when the 13 AA properties are considered independently. Different circular codes were found to be optimal for different AA properties. However, we will show the optimality of the circular code X when several AA properties are taken simultaneously.

5.6.3. Multi-objective score results

Here, we evaluate the frameshift optimality of the 216 maximal self-complementary C^3 circular codes $\mathbb{X}$ (including the circular code X) with the help of a multi-objective parameter (Definition 5.8). We take into account the frameshift code scores and the frameshift dicodon scores, for the given set of AA properties. As mentioned in the previous section, these scores measure differences between the physicochemical properties for the AA and therefore a smaller score indicates a smaller effect of the frameshift error, and hence a better optimality of the code. Previously, we showed that for each individual AA property measured either with the code score or with the dicodon score in the $+1$ or -1 frameshifts (Figure 5.6, Figure 5.7, Figure 5.8 and Figure 5.9), there exists a different circular code x with better optimality than the circular code X.

Since there are different circular codes that are more optimized for individual AA properties, we decided to test the hypothesis that the circular code X is able to optimize a combination of AA properties $\mathbb{P}$ rather than a single one. In order to do this, a multi-objective score N_i (Definition 5.8), where i corresponds to the number of AA properties optimized more effectively than the circular code X, was calculated for each of the 216 maximal self-complementary C^3 circular codes x. In simpler terms, this parameter assigns a multi-objective score (number of AA properties optimized) to each of the codes in $\mathbb{X}$ to evaluate the combination of AA properties they are optimal for. For example, if a particular code x ($x \in \mathbb{X}$) has a multi-objective score of 5, then the code x is said to optimize a combination of 5 AA properties (out of 13 AA properties). Therefore, the circular codes with a multi-objective score of 13 will be considered optimal overall, as they optimize the combination of all 13 AA properties.

Taking into account the frameshift code scores (Definition 5.5) for the 216 codes $\mathbb{X}$, we computed the multi-objective scores in the case of $+1$ frameshift and -1 frameshift errors. In the case of $+1$ frameshift errors (Figure 5.10), we observe that a significant number of circular codes in $\mathbb{X}$ optimize a combination of up to 8 AA properties $\mathbb{P}$ taken together (i.e. circular codes in $\mathbb{X}$ for $N_i(CS_{+1})$, $i \leq 8$). However, when more than 8 AA properties $\mathbb{P}$ are taken into account, the circular code X is among the 12 best circular codes (top 5% of the 216 codes $\mathbb{X}$ for $N_i(CS_{+1})$, $i > 8$). Furthermore, no other circular code in $\mathbb{X}$ achieves the same optimality as the circular code X for 12 or 13 AA properties $\mathbb{P}$ taken together ($N_{13}(CS_{+1}) = 1$ and $N_{12}(CS_{+1}) = 0$).

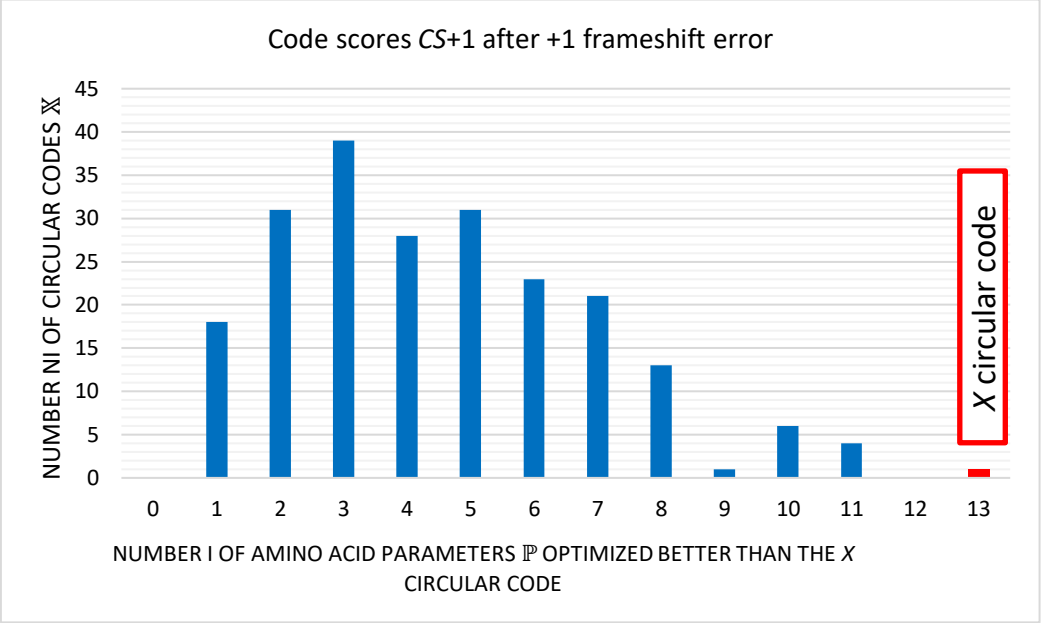


Figure 5.10. Multi-objective score taking into account the $+1$ frameshift code scores (Figure 5.6). The multi-objective score $N_i(CS_{+1})$ (Equation (11)) of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ that gives the number of codes in $\mathbb{X}$ which optimize a combination of AA properties $\mathbb{P}$ better than or equal to the circular code X, for a number i of amino acid properties varying from 0 to 13.

In the case of -1 frameshift errors (Figure 5.11), we clearly see a difference in the distribution. A significant number of circular codes in $\mathbb{X}$ optimize a combination of up to 12 AA properties $\mathbb{P}$ taken together ($N_{10}(CS_{-1}) = 45$, $N_{11}(CS_{-1}) = 50$ and $N_{12}(CS_{-1}) = 27$). However, when all 13 AA properties $\mathbb{P}$ are taken together, the circular code X along with one other code x achieves the best optimality ($N_{13}(CS_{-1}) = 2$). The code x consists of the following 20 trinucleotides:

$$x = \{ATC, CAA, CAC, CAG, CTG, GAA, GAC, GAT, GCC, GGA, GGC, GTA, GTC, GTG, TAA, TAC, TCC, TTA, TTC, TTG\} \quad (21)$$

and codes the stop codon *TAA* and the 12 following amino acids:

$$\{Ala, Asp, Gln, Glu, Gly, His, Ile, Leu, Phe, Ser, Tyr, Val\}.$$

However, this maximal circular code x cannot exist in the reading frame of genes as it contains a stop codon.

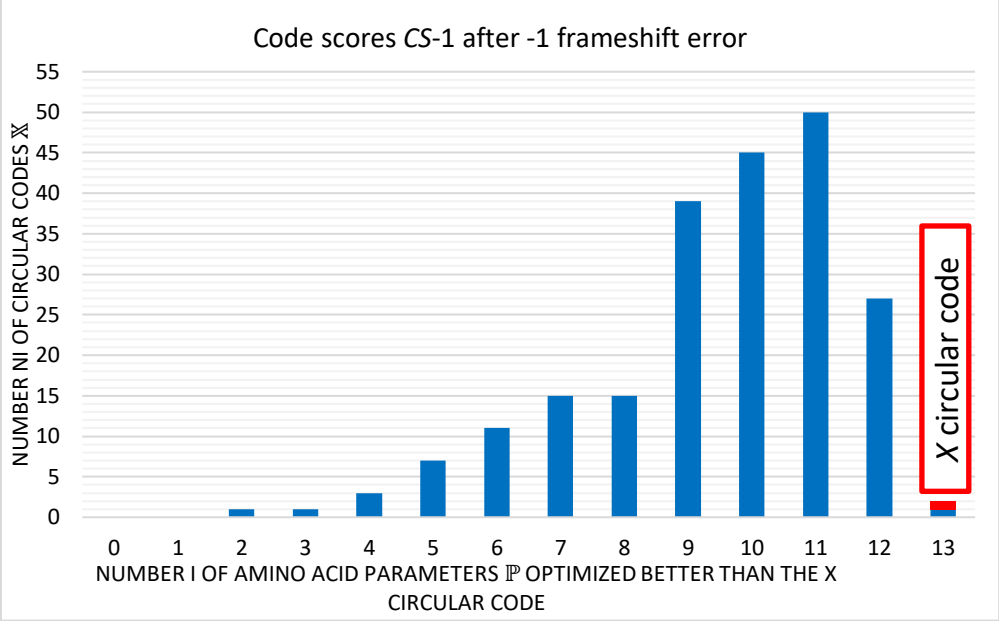


Figure 5.11. Multi-objective score taking into account the -1 frameshift code scores (Figure 5.7). The multi-objective score $N_i(CS_{-1})$ (Equation (11)) of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ that gives the number of codes in $\mathbb{X}$ which optimize a combination of AA properties $\mathbb{P}$ better than or equal to the circular code X, for a number i of amino acid properties varying from 0 to 13.

Next, taking into account the frameshift dicodon scores (Definition 5.7) for the 216 codes $\mathbb{X}$, we computed the multi-objective scores in the case of $+1$ and -1 frameshift errors. As observed above for the code scores, the multi-objective parameter with frameshift dicodon scores also gives similar distributions of optimal circular codes after $+1$ or -1 frameshift errors. In the case of $+1$ frameshift errors (Figure 5.12), we observe that a significant number of circular codes in $\mathbb{X}$ optimize a combination of up to 9 AA properties $\mathbb{P}$ taken together (i.e. circular codes in $\mathbb{X}$ for $N_i(DS_{+1})$, $i \leq 9$). However, when more than 9 AA properties $\mathbb{P}$ are taken into account, the circular code X is among the 13 best circular codes (top 6% of the 216 codes $\mathbb{X}$ for $N_i(DS_{+1})$, $i > 9$). Only 3 of the 216 circular codes $\mathbb{X}$ (1%) optimize a combination of up to 12 AA properties $\mathbb{P}$ taken together ($N_{12}(DS_{+1}) = 3$). Furthermore, no other circular code in $\mathbb{X}$ achieves the same optimality as the circular code X when all 13 AA properties $\mathbb{P}$ are taken together ($N_{13}(DS_{+1}) = 1$).

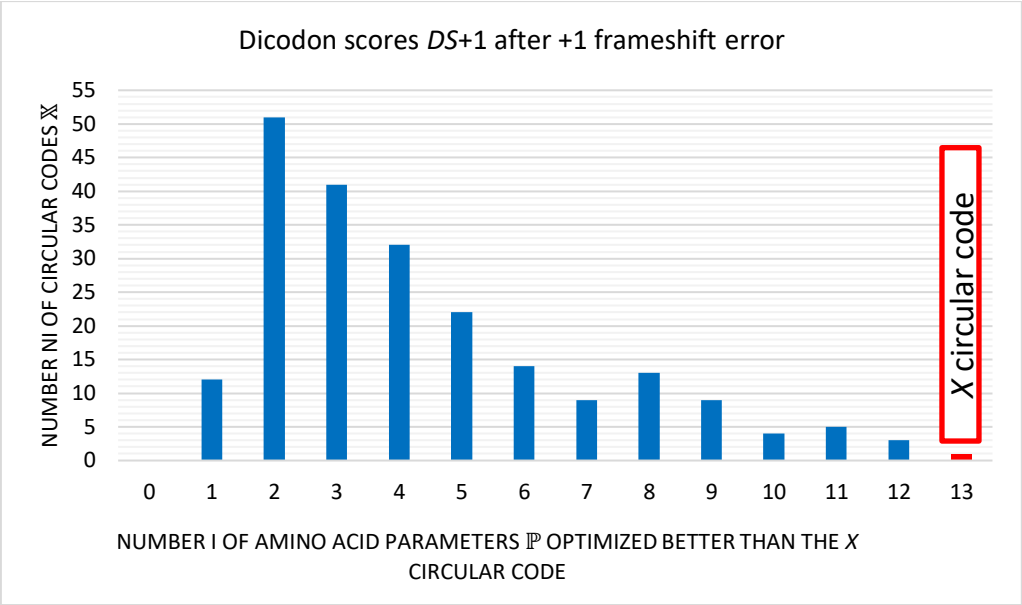


Figure 5.12. Multi-objective score taking into account the $+1$ frameshift dicodon scores (Figure 5.8). The multi-objective score $N_i(DS_{+1})$ (Equation (11)) of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ that gives the number of codes in $\mathbb{X}$ which optimize a combination of AA properties $\mathbb{P}$ better than or equal to the circular code X, for a number i of amino acid properties varying from 0 to 13.

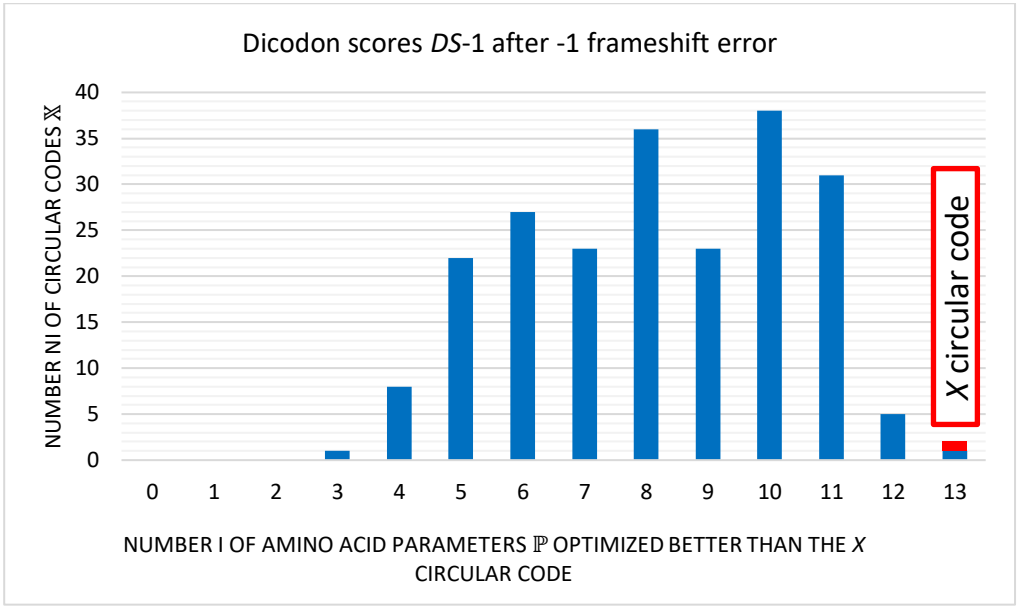


Figure 5.13. Multi-objective score taking into account the -1 frameshift dicodon scores $\mathbb{X}$ (Figure 5.9). The multi-objective score $N_i(DS_{-1})$ (Equation (11)) of the 216 maximal C^3 self-complementary circular codes that gives the number of codes in $\mathbb{X}$ which optimize a combination of AA properties $\mathbb{P}$ better than or equal to the circular code X, for a number i of amino acid properties varying from 0 to 13.

In the case of -1 frameshift errors (Figure 5.13), a significant number of circular codes in $\mathbb{X}$ optimize a combination of up to 11 AA properties $\mathbb{P}$ taken together ($N_8(DS_{-1}) = 36$,

$N_9(DS_{-1}) = 23$, $N_{10}(DS_{-1}) = 38$ and $N_{11}(DS_{-1}) = 31$). Only 5 of the 216 circular codes $\mathbb{X}$ (2%) optimize a combination of up to 12 AA properties $\mathbb{P}$ taken together ($N_{12}(DS_{-1}) = 5$). And, when all 13 AA properties $\mathbb{P}$ are taken together, the circular code X along with one other code x, the same code described by Equation (18), achieves the best optimality ($N_{13}(DS_{-1}) = 2$). As, this code cannot be present in the reading frame of genes the circular code X could be considered optimal.

To summarize the results from the multi-objective parameter using both code scores and dicodon scores, we conclude that the circular code X is the best circular code among its combinatorial class to minimize the overall effects of $+1$ and -1 frameshift events on the translated AA sequence.

5.7. Summary

Translation of mRNA sequences into proteins is one of the most error-prone processes during protein synthesis. Translation errors affect all the domains of life and are shown to reduce the fitness of an organism (Wilke, 2015). Therefore, to minimize the costs of errors, organisms have evolved complex mechanisms for either error prevention by reducing the frequency of errors leading to increased translational accuracy, or error mitigation by minimizing the consequences of errors leading to increased robustness (Drummond & Wilke, 2009). It is widely accepted that the standard genetic code (SGC) is optimized to reduce the impact of translation errors.

In this chapter, we have mainly focused on the optimality of the circular codes, especially the circular code X in minimizing the effects of frameshift errors during translation. We carried out an extensive analysis taking into account the physicochemical properties of amino acids, to compare the frameshift optimality of the circular code X with the SGC and other circular codes. We performed a comprehensive evaluation of the optimality of different circular codes, and measured the differences in the amino acid (AA) sequences produced after a frameshift. We defined an extensive set of 13 AA properties to provide a better picture of the potential changes to the physicochemical properties of the translated protein sequence. In addition, the AA properties are associated with the fundamental chemistry of the amino acid considered to be an elementary unit, i.e. chemical properties that would have acted in a primitive environment (Earth, solar and extrasolar planets, etc.). We introduced various parameters to estimate the frameshift optimality of the codes: a frameshift code score, a frameshift dicodon score, and a multi-objective parameter. The code score is designed to evaluate the optimality of different codes as it takes into account the permutations of a code in the case of frameshift events. The dicodon score is designed to investigate the effects of frameshifts in a DNA sequence motif. We restricted the sequence motif to a length of two codons, but in the future

this could be extended to longer motifs. The multi-objective parameter evaluates the optimality of the circular codes when more than one AA properties are taken simultaneously. We also considered the events of forward $(+1)$ and backward (-1) frameshifts separately, since it is known that the biological mechanisms involved in the two types of frameshift are very different. While translating the mRNA sequence, the ribosome is more likely to slip forward missing one nucleotide rather than slipping backward. Therefore, $+1$ frameshifts are more energy efficient and are generally much more frequent than -1 frameshifts. Using both code-level and dicodon-level scores, we have shown that the circular code X is more optimized than the SGC to reduce the effects of $+1$ frameshifts, in particular with respect to the AA volume, size and molecular weight, as well as the polarity, isoelectric point, polarizability, and charge properties. In contrast, in case of a -1 frameshift, the SGC was generally more optimized than the X circular code. Furthermore, we have shown that the code X is the most optimized out of the 216 maximal C^3 self-complementary circular codes in the case of both $+1$ and -1 frameshifts, when all the 13 AA properties are taken together. Based on these results, we suggest that, in addition to its frameshift synchronization property, the circular code X may have a functional role also in the error mitigation of the more frequent $+1$ frameshift events.

We also discussed other proposed mechanisms to reduce the impact of frameshift errors, such as presence of out-of-frame stop codons in the coding sequences acting as a frameshift catch and break mechanism, which is suggested to minimize the impact of frameshift errors by terminating the translation process after frameshift events. However, there are complex mechanisms involved in the stop codon detection, which includes various protein release factors to detect and stop the translation process (Adio et al., 2018). Earlier, we put forward the hypothesis that circular codes represented an important step in the emergence of the modern genetic code, allowing simultaneous coding of amino acids as well as synchronization of the reading frame in primitive translation systems, prior to the advent of more sophisticated mechanisms. Here, we extend our hypothesis based on the results obtained from the frameshift optimality of circular codes and the SGC. Since the circular code X does not contain any stop codons, it might have played a functional role in the primitive systems, allowing the detection and mitigation of frameshift errors, prior to the evolution of the stop codon recognition machinery.

Chapter 6

6. Conclusion and perspectives

6.1. Error correcting codes

Transmission of information in the biological and communication sciences have similarities in their efficiency and error control capabilities. Prevention of errors is the most effective technique in an error-tolerant design, also known as poka-yoke meaning “mistake-proofing” or “inadvertent error-prevention”, i.e. constraints that prevent errors or incorrect operation by the user. For example, the shape of a battery unit is often designed so that it cannot be inserted incorrectly.

Unfortunately, errors cannot always be prevented. For example, the Western Electric crossbar systems had failure rates of two hours per forty years, and therefore were highly fault resistant. However, when a fault did occur they stopped operating completely, and therefore were not fault tolerant. In the 1950s, John Von Neumann pioneered the concept of adding redundancy to increase the efficiency of an error control system (Von Neumann, 1956). Redundancy implies that, if errors cannot be prevented, a system should fail-safe or fail-secure or fail gracefully, generally by performing at a reduced level in case of danger. Thus, another effective technique in error-tolerant design is the mitigation or limitation of the effects of errors after they have been made.

These two different approaches can be combined to achieve a robust, highly available system: fault-avoidance (prevention of errors) and fault-tolerance (mitigation of errors). An example is Google's use of spell checking on searches performed through their search engine. The spell checking minimizes the problems caused by incorrect spelling, not only by highlighting the error to the user, but also by providing a link to search using the correct spelling instead. Searches like this are commonly performed using a combination of edit distance, soundex, and metaphone calculations.

6.2. The genetic code and errors in translation of protein-coding genes

In the introduction, we discussed the redundancy of the standard genetic code (SGC), and the idea that the SGC is optimized to minimize the effects of errors. Indeed, protein translation is one of the most error-prone processes affecting all domains of life. Effective transmission of information from DNA to proteins by the ribosome therefore requires error prevention (reduce

the rate of occurrence of errors) and/or error mitigation (limit deleterious effects after occurrence of errors) strategies (Drummond & Wilke, 2009).

During protein synthesis, the most frequent errors are misincorporation of non-cognate tRNAs causing missense substitutions. There are specific enzymes at work to prevent misincorporation, and this is known as the proofreading step during protein translation. To achieve translational accuracy, codons with low error rates are selected. Also, there is a highly significant tendency for preferred codons to be associated with evolutionary conserved sites and sites important for protein structure and function (Zhou et al., 2009). Despite the error-preventing mechanisms in place, errors do occur during these complex processes. By optimizing translational robustness to particular errors, the adverse effects can be minimized so that the end product is similar in composition; e.g. proteins can fold and function properly even if they are mistranslated (Drummond & Wilke, 2009). Notably, base changes at the wobble position are generally synonymous, i.e. they code for the same amino acid. Amino acids with similar physicochemical properties are coded by codons that differ usually by only one substitution. For example, hydrophobic amino acids are usually coded by codons with thymine (*T*) in the second position and hydrophilic amino acids by those with adenine (*A*) in this position. Other robust error-mitigation strategies include the presence of intronic stop codons (in eukaryotes) or the presence of stop codons in the alternative reading frames to reduce the cost of synthesis after frameshift errors (Abrahams & Hurst, 2018; Jaillon et al., 2008; Seligmann & Pollock, 2004). Codons which are more likely to form hidden stops or off-frame stops have a higher usage frequency and bias in their favour among the synonymous codons (Warnecke & Hurst, 2011).

6.3. Origin and evolution of the genetic code

The observed error-mitigating properties of the SGC raise the question of how the genetic code emerged and became optimized during evolution. The origin and evolution of the genetic code is undoubtedly coupled with the origin and evolution of the translation machinery. The mystery is deeply rooted in the prebiotic world, when simple biomolecules acquired the ability to synthesize proteins and utilise them as cofactors in performing their activities. There is a continued debate on what must have originated first: proteins or nucleic acids, although a consensus is emerging that proteins (proto-peptides) and nucleic acids (proto-nucleic acids) co-existed and evolved due to mutual stabilization. Recent experiments suggest that cationic proto-peptides can increase the thermal stability of folded RNA structures, and in turn their lifetimes are increased due to interaction with RNA (Frenkel-Pinter et al., 2020). This study highlights the idea of "*RNA and protein are Molecules in Mutualism*" (Lanier et al., 2017). Primitive translation systems might have translated a simple genetic code with a small number of amino

acids, and the genetic code would then have co-evolved into the modern ribosome and the SGC (Bowman et al., 2020). However, it is extremely difficult to recreate the steps that led to the origin of life or to trace back through billions of years of evolution.

One hypothesis is that earlier genetic codes included some kind of error control mechanism, in addition to the amino acid encoding. Due to their frame retrieval properties, circular codes might have provided solutions to the problems of translation errors in primitive systems, before the standard genetic code came into being. The enrichment in motifs of the circular code X identified in modern protein-coding genes (mRNA) and rRNAs might be the remnants of a primordial genetic code, reducing the number of frameshift errors and/or mitigating the effects of the errors.

The main objective of the thesis was to shed light on the questions: do X motifs represent remnants of a primordial code based on error-correcting codes, and/or do they still have a role in the error-correction mechanisms of extant organisms? Therefore, we analysed the evolutionary conservation of X motifs in protein-coding genes and ribosomal RNAs of species from all three domains of life, viz. archaea, bacteria and eukaryotes. We also investigated whether the circular code X presented a frameshift optimality in comparison to the SGC and other maximal circular codes.

6.4. Circular codes are potential ancestors of the modern genetic code

In the analysis involving protein-coding genes (mRNA sequences), we selected two different sets of organisms: four mammals and nine yeasts. The organisms chosen represent a large phylogenetic distribution, and a wide variety of gene structures, ranging from the simple single exon genes of *S. cerevisiae* to the highly complex intron/exon structure of human genes. Moreover, the mammals represent the evolution of closely related species (sharing a common ancestor nearly 300 million years ago), whereas the yeasts represent a more divergent evolution (sharing a common ancestor nearly 1 billion years ago). We constructed multiple gene alignments for both mammals and yeasts separately. We set the minimum length $l \geq 12$ to identify X motifs in the multiple gene alignments, so that each motif is able to retrieve the reading frame with a probability of 100%. We identified a strong enrichment of X motifs (both number and length) in both mammal and yeast multiple gene alignments, thus confirming the previous studies on enrichment of X motifs in protein-coding genes. With the help of various parameters of evolutionary conservation, we showed that the X motifs are more conserved compared to the rest of the gene sequences, with a lower ratio of non-synonymous to synonymous substitutions, indicating a purifying selection. We also carried out an in-depth investigation of synonymous substitutions in X motifs. The results obtained suggest two types of evolutionary selection pressures in the gene segments corresponding to X motifs: first to

preserve the amino acids of the respective proteins encoded by the genes and second to preserve the X motifs, thereby suggesting that the X motifs may represent functional elements of extant genomes. In support of this hypothesis, we demonstrated a strong correlation between protein expression levels and the enrichment of X motifs in genes. In the future, this could applied as a new strategy for efficient gene optimization.

To further investigate the potential role of X motifs in translation of protein-coding genes, we decided to search for the presence of X motifs in the gene translation machinery, namely the ribosome. In this analysis involving ribosomal RNAs (rRNAs), we selected an extensive set of 133 species representing the three domains of life (32 eukaryotes, 65 bacteria, and 36 archaea). We constructed multiple sequence alignments for the SSU rRNAs and the LSU rRNAs of the ribosome, separately. As the rRNA sequences are shorter and more conserved than most mRNAs, we set the minimum length $l \geq 8$ nucleotides to identify X motifs in the rRNA multiple sequence alignments. We then looked for 'universal X motifs' (uX motifs) that are conserved in all the species studied. We identified 32 uX motifs (13 in the SSU and 19 in the LSU), most of which are located in regions involved in important ribosome functions, notably the peptidyl transferase center (PTC) and the decoding center that are supposed to form the primordial "proto-ribosome". Intriguingly, although the X motif property is conserved in the rRNAs, the sequences are not conserved in terms of nucleotides. We also carried out structural analyses of the uX motifs, which revealed that most of the uX motifs are in direct contact with different biomolecules, including mRNA, tRNA and ribosomal proteins. Finally, many of these interactions of the uX motifs can be mapped to other crucial functions of the modern ribosome, including the exit tunnel and ratchet pawls. Building on the existing accretion models for ribosome evolution, we proposed that circular codes represented an important step in the emergence of the standard genetic code (SGC), allowing encoding of the amino acid sequence and at the same time providing a mechanism for retrieval of the correct reading frame.

In order to investigate other potential functions of the circular code X, we carried out an extensive analysis taking into account the physicochemical properties of amino acids, to compare the frameshift optimality of the circular code X with the SGC and other maximal self-complementary C^3 circular codes. We performed a comprehensive evaluation of the frameshift optimality of different codes, and measured the differences in the amino acid (AA) sequences produced after a frameshift. We defined an extensive set of 13 AA properties, providing a better picture of the potential changes to the physicochemical properties of the translated protein sequence. We considered the events of forward $(+1)$ and backward (-1) frameshifts separately, since it is known that the biological mechanisms involved in the two types of frameshift are very different. From this analysis, we identified a new functionality of the circular code X in minimizing the effects of ribosomal translation errors after the more frequent $+1$ frameshift

events. Moreover, the circular code X achieves the best frameshift optimality among its combinatorial class of 216 circular codes after $+1$ and -1 frameshift events.

These results enabled us to put forward a hypothesis of the evolution of the standard genetic code, where the circular code X provided error-prevention and/or error-mitigation mechanisms in the earlier stages of evolution of the translation machinery. In a recent study (Demongeot & Henrion-Caude, 2020), investigating the origin of life through the study of hypothetical RNA rings, the authors speculated that the ”primary informational and functional molecule” was an “22-nucleotide ancestral hairpin/ring” called as the ancestral loop (AL). The AL was shown to be the most thermodynamically stable hairpin and the smallest possible RNA sequence consistent with the production of a wide range of peptides, and shows proximity to ribozymes, tRNAs and rRNAs. Interestingly, their analysis also revealed that 14 codons from the circular code X are found in the AL-hairpin, and that 12 of these 14 codons code for all 12 amino acids coded by the circular code X. The enrichment of codons/motifs from the circular code X in the putative models of primitive nucleic acids strengthens our hypothesis that primordial genetic codes might have been based on error-correcting circular codes.

Finally, it has been suggested recently that the molecular machinery used to translate the genetic code may have evolved from tRNAs (de Farias & José, 2020; Kim et al., 2019; Eigen & Winkler-Oswatitsch, 1981b, 1981a). Interestingly, it has also been shown previously (El Soufi & Michel, 2015; Michel, 2013) that X motifs are present in the 5' and 3' regions of some tRNAs. Based on our results showing that uX motifs identified in the functional regions of the ribosome are in contact with tRNAs, it would be interesting to further investigate the conservation of X motifs in tRNAs using the much larger set of genomic sequences available today. It would also be interesting to identify the interactions between X motifs in tRNAs, mRNAs and rRNAs in the context of the available 3D structural data available.

6.5. Role of the circular code X in modern genes

Evolution is said to be myopic: an event/mechanism will not be selected for, just because it can turn out to be advantageous millions of years later. If we follow this assumption, we can suggest that the presence of X motifs in modern genes cannot merely represent evolutionary remnants, since only functional elements will be positively selected and conserved. Thus, we hypothesize that X motifs continue to play a role in error-correction processes in extant organisms. Unfortunately, the molecular mechanisms underlying these processes are not yet known. However, in our preliminary studies, we have demonstrated a correlation between the enrichment of X motifs in genes and their expression levels. This correlation could be integrated in a prediction method for gene expression, or could be used for re-engineering of genes to increase/decrease protein expression levels depending on the context of use. In the future, it

would be very interesting to perform more comprehensive studies, for example using the high throughput data produced by ribosomal profiling techniques.

We could also study other systems that may reflect the primordial world more closely. If we consider the evolutionary time frame, we as humans have not been present on Earth for a long time. But from the very beginning, viruses are present as inalienable components. A recent study suggests that the LUCA was not a homogenous microbial population, as generally believed (Krupovic et al., 2020). Instead, the LUCA was a population of diverse microorganisms, with a shared gene pool that was inherited by all life forms. It also included a diversified pangenome with genes involved in virus-host interactions (defence strategies). Although RNA-viruses might have been the first to emerge among viruses, a significant diversity of DNA-viruses were already present pre-LUCA. Recently (Michel et al., 2020), X motifs were used to predict accessory genes in coronavirus genomes. Thus, the enrichment of motifs from the circular code X in the genes of viruses can be used efficiently to identify open reading frames (ORFs) to predict functional genes. It might be insightful to examine the role of X motifs in the human virome to identify/predict defence strategies

Finally, if X motifs do play a functional role in modern gene translation, it is possible that the disruption of the X motifs will have a significant effect on expression levels. One way of estimating this effect would to perform systematic mutation experiments in a well-studied organism such as *S. cerevisiae*. Alternatively, we could analyse the ever-increasing human genome data and the genetic variants that are associated with human disorders to investigate whether the presence/absence of X motifs might help to explain their pathogenicity.

Publications in international peer-reviewed journals

Dila G., Michel C. J., Poch O., Ripp R., Thompson J.D. (2019a). Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. *BioSystems* 175, 57-74. https://doi.org/10.1016/j.biosystems.2018.10.014.

Dila G., Ripp R., Mayer C., Poch O., Michel C.J., Thompson J.D. (2019b). Circular code motifs in the ribosome: a missing link in the evolution of translation? *RNA* 25, 1714-1730. https://doi.org/10.1261/rna.072074.119.

Dila G., Michel C.J., Thompson J.D. (2020). Optimality of circular codes versus the genetic code after frameshift errors. *Biosystems* 195, 104134, 1-11. https://doi.org/10.1016/j.biosystems.2020.104134.

International conferences attended

Poster presentation titled, “Circular codes and the genetic code”, in 48th European Mathematical Genetics Meeting (EMGM), Lausanne, April 16-17, 2020.

Bibliography

[Abrahams & Hurst, 2018] Abrahams, L., & Hurst, L. D. (2018). Refining the Ambush Hypothesis: Evidence That GC- and AT-Rich Bacteria Employ Different Frameshift Defence Strategies. *Genome Biology and Evolution*, *10*(4), 1153–1173. https://doi.org/10.1093/gbe/evy075

[Achenbach & Nierhaus, 2015] Achenbach, J., & Nierhaus, K. H. (2015). The mechanics of ribosomal translocation. *Biochimie*, *114*, 80–89. https://doi.org/10.1016/j.biochi.2014.12.003

[Adio et al., 2018] Adio, S., Sharma, H., Senyushkina, T., Karki, P., Maracci, C., Wohlgemuth, I., Holtkamp, W., Peske, F., & Rodnina, M. V. (2018). Dynamics of ribosomes and release factors during translation termination in *E. coli*. *ELife*, *7*, e34252. https://doi.org/10.7554/eLife.34252

[Agmon, 2017] Agmon, I. (2017). Sequence complementarity at the ribosomal Peptidyl Transferase Centre implies self-replicating origin. *FEBS Letters*, *591*(20), 3252–3258. https://doi.org/10.1002/1873-3468.12781

[Agmon, 2018] Agmon, I. (2018). Hypothesis: Spontaneous Advent of the Prebiotic Translation System via the Accumulation of L-Shaped RNA Elements. *International Journal of Molecular Sciences*, *19*(12). https://doi.org/10.3390/ijms19124021

[Ahmed at al., 2010] Ahmed, A., Frey, G., & Michel, C. J. (2010). Essential molecular functions associated with the circular code evolution. *Journal of Theoretical Biology*, *264*(2), 613–622. https://doi.org/10.1016/j.jtbi.2010.02.006

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/

[Arquès and Michel, 1996] Arquès, D. G., & Michel, C. J. (1996). A complementary circular code in the protein coding genes. *Journal of Theoretical Biology*, *182*(1), 45–58. https://doi.org/10.1006/jtbi.1996.0142

[Avery et al., 1944] Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *Journal of Experimental Medicine*, *79*(2), 137–158. https://doi.org/10.1084/jem.79.2.137

[Banwell et al., 2018] Banwell, E. F., Piette, B. M. A. G., Taormina, A., & Heddle, J. G. (2018). Reciprocal Nucleopeptides as the Ancestral Darwinian Self-Replicator. *Molecular Biology and Evolution*, *35*(2), 404–416. https://doi.org/10.1093/molbev/msx292

[Baralle et al., 2019] Baralle, F. E., Singh, R. N., & Stamm, S. (2019). RNA structure and splicing regulation. *Biochimica et Biophysica Acta. Gene Regulatory Mechanisms*, *1862*(11–12), 194448.

https://doi.org/10.1016/j.bbagrm.2019.194448

[Baralle & Baralle, 2018] Baralle, M., & Baralle, F. E. (2018). The splicing code. *Bio Systems*, *164*, 39–48. https://doi.org/10.1016/j.biosystems.2017.11.002

[Bartonek et al., 2020] Bartonek, L., Braun, D., & Zagrovic, B. (2020). Frameshifting preserves key physicochemical properties of proteins. *Proceedings of the National Academy of Sciences*, *117*(11), 5907–5912. https://doi.org/10.1073/pnas.1911203117

[Belardinelli et al., 2016] Belardinelli, R., Sharma, H., Caliskan, N., Cunha, C. E., Peske, F., Wintermeyer, W., & Rodnina, M. V. (2016). Choreography of molecular movements during ribosome progression along mRNA. *Nature Structural & Molecular Biology*, *23*(4), 342–348. https://doi.org/10.1038/nsmb.3193

[Bergman & Tuller, 2020] Bergman, S., & Tuller, T. (2020). Widespread non-modular overlapping codes in the coding regions. In *Physical Biology*. https://doi.org/10.1088/1478-3975/ab7083

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. In *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.235

[Bernier et al., 2014] Bernier, C. R., Petrov, A. S., Waterbury, C. C., Jett, J., Li, F., Freil, L. E., Xiong, X., Wang, L., Migliozzi, B. L. R., Hershkovits, E., Xue, Y., Hsiao, C., Bowman, J. C., Harvey, S. C., Grover, M. A., Wartell, Z. J., & Williams, L. D. (2014). RiboVision suite for visualization and analysis of ribosomes. *Faraday Discussions*, *169*(0), 195–207. https://doi.org/10.1039/C3FD00126A

[Bigelow, 1967] Bigelow, C. C. (1967). On the average hydrophobicity of proteins and the relation between it and protein structure. *Journal of Theoretical Biology*, *16*(2), 187–211. https://doi.org/10.1016/0022-5193(67)90004-5

[Bokov & Steinberg, 2009] Bokov, K., & Steinberg, S. V. (2009). A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, *457*(7232), 977–980. https://doi.org/10.1038/nature07749

[Bowman et al., 2015] Bowman, J. C., Hud, N. V, & Williams, L. D. (2015). The Ribosome Challenge to the RNA World. *Journal of Molecular Evolution*, *80*(3), 143–161. https://doi.org/10.1007/s00239-015-9669-9

[Bowman et al., 2020] Bowman, J. C., Petrov, A. S., Frenkel-Pinter, M., Penev, P. I., & Williams, L. D. (2020). Root of the Tree: The Significance, Evolution, and Origins of the Ribosome. In *Chemical Reviews* (Vol. 120, Issue 11, pp. 4848–4878). American Chemical Society. https://doi.org/10.1021/acs.chemrev.9b00742

[Brar, 2016] Brar, G. A. (2016). Beyond the Triplet Code: Context Cues Transform Translation. *Cell*, *167*(7), 1681–1692. https://doi.org/10.1016/j.cell.2016.09.022

[Brule & Grayhack, 2017] Brule, C. E., & Grayhack, E. J. (2017). Synonymous codons: Choose wisely for expression. *Trends in Genetics : TIG*, *33*(4), 283–297. https://doi.org/10.1016/j.tig.2017.02.001

[Bussoli et al., 2012] Bussoli, L., Michel, C. J., & Pirillo, G. (2012). *On Conjugation Partitions of Sets of Trinucleotides*. *2012*.

https://doi.org/10.4236/am.2012.31017

[Carter & Wills, 2018] Carter, C. W., & Wills, P. R. (2018). Interdependence, Reflexivity, Fidelity, Impedance Matching, and the Evolution of Genetic Coding. *Molecular Biology and Evolution*, *35*(2), 269–286. https://doi.org/10.1093/molbev/msx265

[Charton, 1981] Charton, M. (1981). Protein folding and the genetic code: An alternative quantitative model. *Journal of Theoretical Biology*, *91*(1), 115–123. https://doi.org/10.1016/0022-5193(81)90377-5

[Charton & Charton, 1982] Charton, Marvin, & Charton, B. I. (1982). The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, *99*(4), 629–644. https://doi.org/10.1016/0022-5193(82)90191-6

[Chatterjee & Yadav, 2019] Chatterjee, S., & Yadav, S. (2019). The Origin of Prebiotic Information System in the Peptide/RNA World: A Simulation Model of the Evolution of Translation and the Genetic Code. *Life (Basel, Switzerland)*, *9*(1). https://doi.org/10.3390/life9010025

[Chevance & Hughes, 2017] Chevance, F. F. V, & Hughes, K. T. (2017). Case for the genetic code as a triplet of triplets. *Proceedings of the National Academy of Sciences*, *114*(18), 4745–4750. https://doi.org/10.1073/pnas.1614896114

[Chou & Fasman, 2006] Chou, P. Y., & Fasman, G. D. (2006). Prediction of the Secondary Structure of Proteins From Their Amino Acid Sequence. In *Advances in Enzymology and Related Areas of Molecular Biology* (Vol. 47, pp. 45–148). Wiley Blackwell. https://doi.org/10.1002/9780470122921.ch2

[Crick et al., 1961] Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General Nature of the Genetic Code for Proteins. *Nature*, *192*(4809), 1227–1232. https://doi.org/10.1038/1921227a0

[Crick et al., 1976] Crick, F. H. C., Brenner, S., Klug, A., & Pieczenik, G. (1976). A speculation on the origin of protein synthesis. *Origins of Life*, *7*(4), 389–397. https://doi.org/10.1007/BF00927934

[Crick et al., 1957] Crick, F. H. C., Griffith, J. S., & Orgel, L. E. (1957). Codes Without Commas. *Proceedings of the National Academy of Sciences*, *43*(5), 416–421. https://doi.org/10.1073/pnas.43.5.416

[Dahm, 2005] Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, *278*(2), 274–288. https://doi.org/10.1016/j.ydbio.2004.11.028

[Dao Duc et al., 2019] Dao Duc, K., Batra, S. S., Bhattacharya, N., Cate, J. H. D., & Song, Y. S. (2019). Differences in the path to exit the ribosome across the three domains of life. *Nucleic Acids Research*, *47*(8), 4198–4210. https://doi.org/10.1093/nar/gkz106

[Darwin, 1859] Darwin, C. (2004). On the Origin of Species, 1859. In *On the Origin of Species, 1859*. https://doi.org/10.4324/9780203509104

[Dawson, 1972] Dawson, D. M. (1972). *In:The Biochemical Genetics of Man (Brock, D. J. H., and Mayo, O., eds.)* (pp. 1–38). Academic Press.

[de Farias & José, 2020] de Farias, S. T., & José, M. V. (2020). Transfer RNA: The molecular demiurge in the origin of biological systems. *Progress in Biophysics and Molecular Biology*, *153*, 28–34. https://doi.org/10.1016/j.pbiomolbio.2020.02.006

[Demongeot & Henrion-Caude, 2020] Demongeot, J., & Henrion-Caude, A. (2020).

Footprints of a Singular 22-Nucleotide RNA Ring at the Origin of Life. *Biology*, *9*(5), 88. https://doi.org/10.3390/biology9050088

[Demongeot & Seligmann, 2019] Demongeot, J., & Seligmann, H. (2019). Spontaneous evolution of circular codes in theoretical minimal RNA rings. *Gene*, *705*, 95–102. https://doi.org/10.1016/j.gene.2019.03.069

[Demongeot & Seligmann, 2020] Demongeot, J., & Seligmann, H. (2020). Theoretical minimal RNA rings mimick molecular evolution before tRNA-mediated translation: codon-amino acid affinities increase from early to late RNA rings. *Comptes Rendus. Biologies*, *343*(1), 111–122. https://doi.org/10.5802/crbiol.1

[Diambra, 2017] Diambra, L. A. (2017). Differential bicodon usage in lowly and highly abundant proteins. *PeerJ*, *5*, e3081. https://doi.org/10.7717/peerj.3081

[Drummond & Wilke, 2009] Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews. Genetics*, *10*(10), 715–724. https://doi.org/10.1038/nrg2662

[Eigen & Schuster, 1978] Eigen, M., & Schuster, P. (1978). The Hypercycle. *The Science of Nature*, *65*(7), 341–369. https://doi.org/10.1007/BF00439699

[Eigen & Winkler-Oswatitsch, 1981a] Eigen, M., & Winkler-Oswatitsch, R. (1981a). Transfer-RNA, an early gene? *Naturwissenschaften*, *68*(6), 282–292. https://doi.org/10.1007/BF01047470

[Eigen & Winkler-Oswatitsch, 1981b] Eigen, M., & Winkler-Oswatitsch, R. (1981b). Transfer-RNA: The early adaptor. *Naturwissenschaften*, *68*(5), 217–228. https://doi.org/10.1007/BF01047323

[El Soufi & Michel, 2014] El Soufi, K., & Michel, C. J. (2014). Circular code motifs in the ribosome decoding center. *Computational Biology and Chemistry*, *52*, 9–17. https://doi.org/10.1016/j.compbiolchem.2014.08.001

[El Soufi & Michel, 2015] El Soufi, K., & Michel, C. J. (2015). Circular code motifs near the ribosome decoding center. *Computational Biology and Chemistry*, *59*, 158–176. https://doi.org/10.1016/j.compbiolchem.2015.07.015

[El Soufi & Michel, 2016] El Soufi, K., & Michel, C. J. (2016). Circular code motifs in genomes of eukaryotes. *Journal of Theoretical Biology*, *408*, 198–212. https://doi.org/10.1016/j.jtbi.2016.07.022

[El Soufi & Michel, 2017] El Soufi, K., & Michel, C. J. (2017). Unitary circular code motifs in genomes of eukaryotes. *Biosystems*, *153–154*, 45–62. https://doi.org/10.1016/j.biosystems.2017.02.001

[Farabaugh & Björk, 1999] Farabaugh, P. J., & Björk, G. R. (1999). How translational accuracy influences reading frame maintenance. *The EMBO Journal*, *18*(6), 1427–1434. https://doi.org/10.1093/emboj/18.6.1427

[Fasman, 1976] Fasman, G D. (1976). *Handbook of Biochemistry and Molecular Biology. Proteins – Volume 1, 3rd ed. CRC Press, Cleveland.* CRC Press.

[Fasman, 1989] Fasman, Gerald D. (1989). *Practical Handbook of Biochemistry and Molecular Biology*. CRC Press. https://books.google.fr/books?id=TQ6Q99anaSMC

[Ferus et al., 2017] Ferus, M., Pietrucci, F., Saitta, A. M., Knížek, A., Kubelík, P.,

Ivanek, O., Shestivska, V., & Civiš, S. (2017). Formation of nucleobases in a Miller–Urey reducing atmosphere. *Proceedings of the National Academy of Sciences*.
https://doi.org/10.1073/pnas.1700010114

[Fournier et al., 2010] Fournier, G. P., Neumann, J. E., & Gogarten, J. P. (2010). Inferring the Ancient History of the Translation Machinery and Genetic Code via Recapitulation of Ribosomal Subunit Assembly Orders. *PLoS ONE*, *5*(3).
https://doi.org/10.1371/journal.pone.0009437

[Freeland & Hurst, 1998] Freeland, S. J., & Hurst, L. D. (1998). The Genetic Code Is One in a Million. *Journal of Molecular Evolution*, *47*(3), 238–248.
https://doi.org/10.1007/PL00006381

[Freeland et al., 2000] Freeland, S. J., Knight, R. D., Landweber, L. F., & Hurst, L. D. (2000). Early Fixation of an Optimal Genetic Code. *Molecular Biology and Evolution*, *17*(4), 511–518.
https://doi.org/10.1093/oxfordjournals.molbev.a026331

[Frenkel-Pinter et al., 2020] Frenkel-Pinter, M., Haynes, J. W., Mohyeldin, A. M., C, M., Sargon, A. B., Petrov, A. S., Krishnamurthy, R., Hud, N. V, Williams, L. D., & Leman, L. J. (2020). Mutually stabilizing interactions between proto-peptides and RNA. *Nature Communications*, *11*(1), 3137.
https://doi.org/10.1038/s41467-020-16891-5

[Frey & Michel, 2003] Frey, G., & Michel, C. J. (2003). Circular codes in archaeal genomes. *Journal of Theoretical Biology*, *223*(4), 413–431.
https://doi.org/10.1016/S0022-5193(03)00119-X

[Frey & Michel, 2006] Frey, G., & Michel, C. J. (2006). Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. *Computational Biology and Chemistry*, *30*(2), 87–101.
https://doi.org/10.1016/j.compbiolchem.2005.11.001

[Gabius & Roth, 2017] Gabius, H.-J., & Roth, J. (2017). An introduction to the sugar code. *Histochemistry and Cell Biology*, *147*(2), 111–117.
https://doi.org/10.1007/s00418-016-1521-9

[Gamble et al., 2016] Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S., & Grayhack, E. J. (2016). Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell*, *166*(3), 679–690.
https://doi.org/10.1016/j.cell.2016.05.070

[Garofalo et al., 2019] Garofalo, R., Wohlgemuth, I., Pearson, M., Lenz, C., Urlaub, H., & Rodnina, M. V. (2019). Broad range of missense error frequencies in cellular proteins. *Nucleic Acids Research*, *47*(6), 2932–2945.
https://doi.org/10.1093/nar/gky1319

[Geyer & Madany Mamlouk, 2018] Geyer, R., & Madany Mamlouk, A. (2018). On the efficiency of the genetic code after frameshift mutations. *PeerJ*, *6*, e4825.
https://doi.org/10.7717/peerj.4825

[Gilbert, 1986] Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, *319*(6055), 618.
https://doi.org/10.1038/319618a0

[Gospodinov & Kunnev, 2020] Gospodinov, A., & Kunnev, D. (2020). Universal Codons with Enrichment from GC to AU Nucleotide Composition Reveal a Chronological Assignment from Early to Late Along with LUCA Formation. *Life*, *10*(6), 81.
https://doi.org/10.3390/life10060081

[Haig & Hurst, 1991] Haig, D., & Hurst, L. D. (1991). A quantitative measure of error

minimization in the genetic code. *Journal of Molecular Evolution*, *33*(5), 412–417. https://doi.org/10.1007/BF02103132

[Hartman & Smith, 2014] Hartman, H., & Smith, T. F. (2014). The evolution of the ribosome and the genetic code. *Life*. https://doi.org/10.3390/life4020227

[Hsiao et al., 2013] Hsiao, C., Lenz, T. K., Peters, J. K., Fang, P.-Y., Schneider, D. M., Anderson, E. J., Preeprem, T., Bowman, J. C., O'Neill, E. B., Lie, L., Athavale, S. S., Gossett, J. J., Trippe, C., Murray, J., Petrov, A. S., Wartell, R. M., Harvey, S. C., Hud, N. V, & Dean Williams, L. (2013). Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucleic Acids Research*, *41*(5), 3373–3385. https://doi.org/10.1093/nar/gkt023

[Hsiao et al., 2009] Hsiao, C., Mohan, S., Kalahar, B. K., & Williams, L. D. (2009). Peeling the Onion: Ribosomes Are Ancient Molecular Fossils. *Molecular Biology and Evolution*, *26*(11), 2415–2425. https://doi.org/10.1093/molbev/msp163

[Ikehara, 2002] Ikehara, K. (2002). Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *Journal of Biosciences*, *27*(2), 165–186. https://doi.org/10.1007/BF02703773

[Itzkovitz & Alon, 2007] Itzkovitz, S, & Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, *17*(4), 405–412. https://doi.org/10.1101/gr.5987307

[Itzkovitz et al., 2010] Itzkovitz, Shalev, Hodis, E., & Segal, E. (2010). Overlapping codes within protein-coding sequences. *Genome Research*, *20*(11), 1582–1589. https://doi.org/10.1101/gr.105072.110

[Jaillon et al., 2008] Jaillon, O., Bouhouche, K., Gout, J. F., Aury, J. M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B. M., Ségurens, B., Le Mouël, A., Lepère, G., Schächter, V., Bétermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L., & Meyer, E. (2008). Translational control of intron splicing in eukaryotes. *Nature*, *451*(7176), 359–362. https://doi.org/10.1038/nature06495

[Jenner et al., 2010] Jenner, L., Demeshkina, N., Yusupova, G., & Yusupov, M. (2010). Structural rearrangements of the ribosome at the tRNA proofreading step. *Nature Structural & Molecular Biology*, *17*(9), 1072–1078. https://doi.org/10.1038/nsmb.1880

[Jenuwein & Allis, 2001] Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science (New York, N.Y.)*, *293*(5532), 1074–1080. https://doi.org/10.1126/science.1063127

[Johnson & Wang, 2010] Johnson, D. B. F., & Wang, L. (2010). Imprints of the genetic code in the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(18), 8298–8303. https://doi.org/10.1073/pnas.1000704107

[Kawashima & Kanehisa, 2000] Kawashima, S., & Kanehisa, M. (2000). AAindex: Amino Acid index database. *Nucleic Acids Research*, *28*(1), 374. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102411/

[Kim et al., 2019] Kim, Y., Opron, K., & Burton, Z. F. (2019). A tRNA- and Anticodon-Centric View of the Evolution of Aminoacyl-tRNA Synthetases, tRNAomes, and the Genetic Code. *Life*, *9*(2).

https://doi.org/10.3390/life9020037

[Klein et al., 1984] Klein, P., Kanehisa, M., & DeLisi, C. (1984). Prediction of protein function from sequence properties: Discriminant analysis of a data base. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *787*(3), 221–226. https://doi.org/10.1016/0167-4838(84)90312-1

[Komander & Rape, 2012] Komander, D., & Rape, M. (2012). The ubiquitin code. *Annual Review of Biochemistry*, *81*, 203–229. https://doi.org/10.1146/annurev-biochem-060310-170328

[Koonin, 2017] Koonin, E. V. (2017). Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code. *Life*, *7*(2), 22. https://doi.org/10.3390/life7020022

[Koonin & Novozhilov, 2009] Koonin, E. V, & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: The universal enigma. *IUBMB Life*, *61*(2), 99–111. https://doi.org/10.1002/iub.146

[Kovacs et al., 2017] Kovacs, N. A., Petrov, A. S., Lanier, K. A., & Williams, L. D. (2017). Frozen in Time: The History of Proteins. *Molecular Biology and Evolution*, *34*(5), 1252–1260. https://doi.org/10.1093/molbev/msx086

[Krupovic et al., 2020] Krupovic, M., Dolja, V. V., & Koonin, E. V. (2020). The LUCA and its complex virome. *Nature Reviews Microbiology*, 1–10. https://doi.org/10.1038/s41579-020-0408-x

[Kumar & Saini, 2016] Kumar, B., & Saini, S. (2016). Analysis of the optimality of the standard genetic code. *Molecular BioSystems*, *12*(8), 2642–2651. https://doi.org/10.1039/C6MB00262E

[Kunnev & Gospodinov, 2018] Kunnev, D., & Gospodinov, A. (2018). Possible Emergence of Sequence Specific RNA Aminoacylation via Peptide Intermediary to Initiate Darwinian Evolution and Code Through Origin of Life. *Life (Basel, Switzerland)*, *8*(4). https://doi.org/10.3390/life8040044

[Lanier et al., 2017] Lanier, K. A., Petrov, A. S., & Williams, L. D. (2017). The Central Symbiosis of Molecular Biology: Molecules in Mutualism. *Journal of Molecular Evolution*, *85*(1–2), 8–13. https://doi.org/10.1007/s00239-017-9804-x

[Lecompte et al., 2002] Lecompte, O., Ripp, R., Thierry, J.-C., Moras, D., & Poch, O. (2002). Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Research*, *30*(24), 5382–5390. https://doi.org/10.1093/nar/gkf693

[Lupas & Alva, 2017] Lupas, A. N., & Alva, V. (2017). Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *Journal of Structural Biology*, *198*(2), 74–81. https://doi.org/10.1016/j.jsb.2017.04.007

[Ma, 2010] Ma, W. (2010). The scenario on the origin of translation in the RNA world: in principle of replication parsimony. *Biology Direct*, *5*(1), 65. https://doi.org/10.1186/1745-6150-5-65

[Maier et al., 2013] Maier, U.-G., Zauner, S., Woehle, C., Bolte, K., Hempel, F., Allen, J. F., & Martin, W. F. (2013). Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes. *Genome Biology and Evolution*, *5*(12), 2318–2329.

https://doi.org/10.1093/gbe/evt181

[Maraia & Iben, 2014] Maraia, R. J., & Iben, J. R. (2014). Different types of secondary information in the genetic code. In *RNA*. https://doi.org/10.1261/rna.044115.113

[Melnikov et al., 2012] Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., & Yusupov, M. (2012). One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology*, *19*(6), 560–567. https://doi.org/10.1038/nsmb.2313

[Mendel, 1865] Mendel, G. (1865). Experiments in plants hybridization. *Scholarly Publishing*.

[Michel et al., 2017] Michel, C., Ngoune, V. N., Poch, O., Ripp, R., & Thompson, J. (2017). Enrichment of Circular Code Motifs in the Genes of the Yeast Saccharomyces cerevisiae. *Life*, *7*(4), 52. https://doi.org/10.3390/life7040052

[Michel & Thompson, 2020] Michel, C. J., & Thompson, J. D. (2020). Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? *RNA Biology*. https://doi.org/10.1080/15476286.2020.1719311

[Michel, 2008] Michel, C. J. (2008). A 2006 review of circular codes in genes. *Computers & Mathematics with Applications*, *55*(5), 984–988. https://doi.org/10.1016/j.camwa.2006.12.090

[Michel, 2012] Michel, C. J. (2012). Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes. *Computational Biology and Chemistry*, *37*, 24–37. https://doi.org/10.1016/j.compbiolchem.2011.10.002

[Michel, 2013] Michel, C. J. (2013). Circular code motifs in transfer RNAs. *Computational Biology and Chemistry*, *45*, 17–29. https://doi.org/10.1016/j.compbiolchem.2013.02.004

[Michel, 2014] Michel, C. J. (2014). A genetic scale of reading frame coding. *Journal of Theoretical Biology*, *355*, 83–94. https://doi.org/10.1016/j.jtbi.2014.03.029

[Michel, 2017] Michel, C. J. (2017). The Maximal C3 Self-Complementary Trinucleotide Circular Code *X* in Genes of Bacteria, Archaea, Eukaryotes, Plasmids and Viruses. *Life*, *7*(2). https://doi.org/10.3390/life7020020

[Michel, 2019] Michel, C. J. (2019). Single-Frame, Multiple-Frame and Framing Motifs in Genes. *Life*, *9*(1), 18. https://doi.org/10.3390/life9010018

[Michel & Pirillo, 2013] Michel, C. J, & Pirillo, G. (2013). A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. *Journal of Theoretical Biology*, *319*, 116–121. https://doi.org/10.1016/j.jtbi.2012.11.023

[Michel et al., 2020] Michel, C. J., Mayer, C., Poch, O., & Thompson, J. D. (2020). Characterization of accessory genes in coronavirus genomes. *Virology Journal*, *17*(1), 131. https://doi.org/10.1186/s12985-020-01402-1

[Miller, 1953] Miller, S. L. (1953). A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science*, *117*(3046), 528–529. https://doi.org/10.1126/science.117.3046.528

[Nei & Gojobori, 1986] Nei, M., & Gojobori, T. (1986). Simple methods for estimating

the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

[Nirenberg & Matthaei, 1961] Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.47.10.1588

[Noller, 2012] Noller, H. F. (2012). Evolution of Protein Synthesis from an RNA World. *Cold Spring Harbor Perspectives in Biology*, *4*(4). https://doi.org/10.1101/cshperspect.a003681

[Nutman et al., 2016] Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J., & Chivas, A. R. (2016). Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature*, *537*(7621), 535–538. https://doi.org/10.1038/nature19355

[Opron & Burton, 2018] Opron, K., & Burton, Z. F. (2018). Ribosome Structure, Function, and Early Evolution. *International Journal of Molecular Sciences*, *20*(1). https://doi.org/10.3390/ijms20010040

[Parker, 1989] Parker, J. (1989). Errors and alternatives in reading the universal genetic code. *Microbiological Reviews*, *53*(3), 273–298. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC372737/

[Pelc & Welton, 1966] Pelc, S. R., & Welton, M. G. E. (1966). Stereochemical Relationship Between Coding Triplets and Amino-Acids. *Nature*, *209*(5026), 868–870. https://doi.org/10.1038/209868a0

[Petrov et al., 2015] Petrov, A. S., Gulen, B., Norris, A. M., Kovacs, N. A., Bernier, C. R., Lanier, K. A., Fox, G. E., Harvey, S. C., Wartell, R. M., Hud, N. V, & Williams, L. D. (2015). History of the ribosome and the origin of translation. *Proceedings of the National Academy of Sciences*, *112*(50), 15396–15401. https://doi.org/10.1073/pnas.1509761112

[Piette & Heddle, 2020] Piette, B. M. A. G., & Heddle, J. G. (2020). A Peptide-Nucleic Acid Replicator Origin for Life. *Trends in Ecology & Evolution*, *35*(5), 397–406. https://doi.org/10.1016/j.tree.2020.01.001

[Plankensteiner et al., 2005] Plankensteiner, K., Reiner, H., & Rode, B. M. (2005). Prebiotic Chemistry: The Amino Acid and Peptide World. *Current Organic Chemistry*, *9*(12), 1107–1114. https://doi.org/10.2174/1385272054553640

[Polacek & Mankin, 2005] Polacek, N., & Mankin, A. S. (2005). The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Critical Reviews in Biochemistry and Molecular Biology*, *40*(5), 285–311. https://doi.org/10.1080/10409230500326334

[Polyansky et al., 2013] Polyansky, A. A., Hlevnjak, M., & Zagrovic, B. (2013). Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biology*, *10*(8), 1248–1254. https://doi.org/10.4161/rna.25977

[Quast et al., 2013] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids*

Research.
https://doi.org/10.1093/nar/gks1219

[Rackovsky, 1993] Rackovsky, S. (1993). On the nature of the protein folding code. *Proceedings of the National Academy of Sciences*, *90*(2), 644–648. https://doi.org/10.1073/pnas.90.2.644

[Redies & Takeichi, 1996] Redies, C., & Takeichi, M. (1996). Cadherins in the developing central nervous system: an adhesive code for segmental and functional subdivisions. *Developmental Biology*, *180*(2), 413–423. https://doi.org/10.1006/dbio.1996.0315

[Root-Bernstein & Root-Bernstein, 2019] Root-Bernstein, R., & Root-Bernstein, M. (2019). The Ribosome as a Missing Link in Prebiotic Evolution III: Over-Representation of tRNA- and rRNA-Like Sequences and Plieofunctionality of Ribosome-Related Molecules Argues for the Evolution of Primitive Genomes from Ribosomal RNA Modules. *International Journal of Molecular Sciences*, *20*(1), 140. https://doi.org/10.3390/ijms20010140

[Schimmel et al., 1993] Schimmel, P., Giegé, R., Moras, D., & Yokoyama, S. (1993). An operational RNA code for amino acids and possible relationship to genetic code. *Proceedings of the National Academy of Sciences*, *90*(19), 8763–8768. https://doi.org/10.1073/pnas.90.19.8763

[Segal et al., 2006] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., & Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, *442*(7104), 772–778. https://doi.org/10.1038/nature04979

[Seligmann, 2019] Seligmann, H. (2019). Localized Context-Dependent Effects of the "Ambush" Hypothesis: More Off-Frame Stop Codons Downstream of Shifty Codons. *DNA and Cell Biology*, *38*(8), 786–795. https://doi.org/10.1089/dna.2019.4725

[Seligmann & Pollock, 2004] Seligmann, H., & Pollock, D. D. (2004). The Ambush Hypothesis: Hidden Stop Codons Prevent Off-Frame Gene Reading. *DNA and Cell Biology*, *23*(10), 701–705. https://doi.org/10.1089/dna.2004.23.701

[Sharma & Hiller, 2017] Sharma, V., & Hiller, M. (2017). Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. *Nucleic Acids Research*, *45*(14), 8369–8377. https://doi.org/10.1093/nar/gkx554

[Shepherd, 1981] Shepherd, J. C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(3), 1596–1600. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC319178/

[Smith et al., 2008] Smith, T. F., Lee, J. C., Gutell, R. R., & Hartman, H. (2008). The origin and evolution of the ribosome. *Biology Direct*, *3*(1), 16. https://doi.org/10.1186/1745-6150-3-16

[Spielman & Wilke, 2015] Spielman, S. J., & Wilke, C. O. (2015). The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, *32*(4), 1097–1108. https://doi.org/10.1093/molbev/msv003

[Szathmáry, 1999] Szathmáry, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends in Genetics: TIG*, *15*(6), 223–229.

https://doi.org/10.1016/s0168-9525(99)01730-8

[Thompson et al., 1994] Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

[Trifonov, 1999] Trifonov, E. N. (1999). Elucidating Sequence Codes: Three Codes for Evolution. *Annals of the New York Academy of Sciences*, *870*(1 MOLECULAR STR), 330–338. https://doi.org/10.1111/j.1749-6632.1999.tb08894.x

[van der Gulik & Speijer, 2015] van der Gulik, P. T. S., & Speijer, D. (2015). How amino acids and peptides shaped the RNA world. *Life (Basel, Switzerland)*, *5*(1), 230–246. https://doi.org/10.3390/life5010230

[Verhey & Gaertig, 2007] Verhey, K. J., & Gaertig, J. (2007). The tubulin code. *Cell Cycle (Georgetown, Tex.)*, *6*(17), 2152–2160. https://doi.org/10.4161/cc.6.17.4633

[Vitas & Dobovišek, 2018] Vitas, M., & Dobovišek, A. (2018). In the Beginning was a Mutualism - On the Origin of Translation. *Origins of Life and Evolution of Biospheres*, *48*(2), 223–243. https://doi.org/10.1007/s11084-018-9557-6

[Von Neumann, 1956] Von Neumann, J. (1956). Probabilistic Logics and Synthesis of Reliable Organisms from Unreliable Components. *In Shannon, C.E. and McCarthy, J., Eds., Automata Studies, in Annals of Mathematical Studies, No. 34, Princeton University Press, Princeton, 43-98.* https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/ReferencesPapers.aspx?ReferenceID=1392686

[Warnecke & Hurst, 2011] Warnecke, T., & Hurst, L. D. (2011). Error prevention and mitigation as forces in the evolution of genes and genomes. *Nature Reviews. Genetics*, *12*(12), 875–881. https://doi.org/10.1038/nrg3092

[Watson & Crick, 1953] Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

[Weatheritt & Babu, 2013] Weatheritt, R. J., & Babu, M. M. (2013). Evolution. The hidden codes that shape protein evolution. *Science (New York, N.Y.)*, *342*(6164), 1325–1326. https://doi.org/10.1126/science.1248425

[Wilke, 2015] Wilke, C. O. (2015). Evolutionary paths of least resistance. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 112, Issue 41, pp. 12553–12554). National Academy of Sciences. https://doi.org/10.1073/pnas.1517390112

[Wills & Carter, 2018] Wills, P. R., & Carter, C. W. (2018). Insuperable problems of the genetic code initially emerging in an RNA World. *Bio Systems*, *164*, 155–166. https://doi.org/10.1016/j.biosystems.2017.09.006

[Wnętrzak et al., 2019] Wnętrzak, M., Błażej, P., & Mackiewicz, P. (2019). Optimization of the standard genetic code in terms of two mutation types: Point mutations and frameshifts. *Biosystems*, *181*, 44–50. https://doi.org/10.1016/j.biosystems.2019.04.012

[Woese, 1965] Woese, C. R. (1965). On the evolution of the genetic code. *Proceedings of the National Academy of Sciences*, *54*(6), 1546–1552. https://doi.org/10.1073/pnas.54.6.1546

[Wong, 1975] Wong, J. T.-F. (1975). A Co-Evolution Theory of the Genetic Code. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(5), 1909–1912. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC432657/

[Wu et al., 2007] Wu, G., Zheng, Y., Qureshi, I., Zin, H. T., Beck, T., Bulka, B., & Freeland, S. J. (2007). SGDB: a database of synthetic genes re-designed for optimizing protein over-expression. *Nucleic Acids Research*, *35*(Database issue), D76–D79. https://doi.org/10.1093/nar/gkl648

[Yarus, 2017] Yarus, M. (2017). The Genetic Code and RNA-Amino Acid Affinities. *Life*, *7*(2), 13. https://doi.org/10.3390/life7020013

[Yarus et al., 2009] Yarus, M., Widmann, J. J., & Knight, R. (2009). RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code. *Journal of Molecular Evolution*, *69*(5), 406. https://doi.org/10.1007/s00239-009-9270-1

[Yu et al., 2015] Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., & Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular Cell*, *59*(5), 744–754. https://doi.org/10.1016/j.molcel.2015.07.018

[Zhou et al., 2009] Zhou, T., Weems, M., & Wilke, C. O. (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution*, *26*(7), 1571–1580. https://doi.org/10.1093/molbev/msp070

[Zimmerman et al., 1968] Zimmerman, J. M., Eliezer, N., & Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology*, *21*(2), 170–201. https://doi.org/10.1016/0022-5193(68)90069-6

APPENDIX

APPENDIX

Table I. For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, codon substitution matrix $\mathbf{A}(m(X,\mathbb{H}))$ of X motifs $m(X,\mathbb{H})$ (Section 3.4). For each codon, the encoded amino acid is given.

		N	*N*	*T*	*I*	*I*	*Q*	*L*	*L*	*E*	*D*	*E*	*D*	*A*	*G*	*G*	*V*	*V*	*V*	*Y*	*F*
		AAC	*AAT*	*ACC*	*ATC*	*ATT*	*CAG*	*CTC*	*CTG*	*GAA*	*GAC*	*GAG*	*GAT*	*GCC*	*GGC*	*GGT*	*GTA*	*GTC*	*GTT*	*TAC*	*TTC*
K	*AAA*	521	569	81	33	22	378	11	23	1449	126	494	125	21	39	21	35	10	7	14	3
N	*AAC*	7493	20002	767	186	53	92	36	10	105	1728	200	569	199	471	129	3	47	16	178	19
K	*AAG*	650	744	109	31	37	2148	15	96	423	130	2316	133	45	40	42	23	9	9	14	3
N	*AAT*	16365	56175	200	40	194	131	8	7	147	477	221	1539	58	112	275	5	8	46	53	5
T	*ACA*	110	103	5453	143	139	61	11	124	147	16	112	15	387	20	14	264	63	56	3	17
T	*ACC*	905	268	67464	1434	369	22	126	50	26	149	45	37	3797	184	44	27	516	103	24	79
T	*ACG*	57	64	5203	126	99	61	8	120	22	9	98	14	292	24	10	59	42	31	1	5
T	*ACT*	204	698	12412	333	1202	23	43	22	30	48	27	94	943	60	100	28	143	327	8	23
R	*AGA*	108	101	33	22	13	169	3	20	244	33	108	38	15	94	58	47	13	5	3	2
S	*AGC*	4025	1342	1741	344	112	64	47	8	57	460	68	151	412	3307	669	5	119	38	61	30
R	*AGG*	104	123	80	17	22	602	17	86	69	15	297	25	19	113	99	16	12	2	5	2
S	*AGT*	1020	3570	328	72	297	49	9	10	49	136	69	392	70	724	1386	7	35	54	18	4
I	*ATA*	16	16	93	2981	2224	23	79	339	53	4	35	5	36	6	3	1421	236	137	4	28
I	*ATC*	236	45	1429	95708	19908	5	1531	226	11	47	8	10	413	90	16	252	5416	909	32	522
M	*ATG*	75	58	370	854	862	120	178	2474	28	13	164	11	132	13	12	302	223	212	3	54
I	*ATT*	63	167	286	16366	54789	8	241	149	12	9	9	32	96	32	42	140	869	2577	7	94
Q	*CAA*	95	61	9	3	4	14988	56	182	909	71	339	43	12	18	15	13	2	1	46	28
H	*CAC*	798	320	52	24	4	1695	360	79	55	412	113	116	51	164	31	2	18	5	1241	82
Q	*CAG*	130	131	24	8	7	148989	119	1289	566	136	2483	110	32	43	25	13	9	15	102	12
H	*CAT*	226	653	17	9	24	1223	92	49	49	114	93	279	21	34	77	3	4	20	331	30
P	*CCA*	5	10	59	10	9	672	122	1091	65	4	33	6	119	7	7	31	8	13	3	17
P	*CCC*	72	38	740	52	29	222	685	289	11	30	15	11	788	49	18	9	88	18	59	108
P	*CCG*	4	4	48	7	3	721	90	1000	16	3	66	6	68	10	3	7	3	4	3	15
P	*CCT*	17	47	233	21	89	157	246	284	12	12	14	48	279	14	19	7	17	49	22	40
R	*CGA*	15	8	1	1	1	608	21	69	63	10	30	3	0	30	9	9	1	1	8	5
R	*CGC*	72	37	27	7	3	413	283	62	8	65	21	7	36	279	38	0	9	3	90	45
R	*CGG*	17	13	3	2	4	3737	48	427	22	7	137	6	5	37	20	5	20	9	12	4
R	*CGT*	25	99	3	7	11	240	58	29	3	14	11	31	8	57	77	1	3	11	29	12
L	*CTA*	10	3	10	72	49	130	2150	13178	16	0	6	0	15	2	2	227	33	23	3	97
L	*CTC*	42	15	126	1521	351	89	74979	10241	22	17	18	14	158	33	10	41	887	155	74	1850
L	*CTG*	12	19	28	340	242	1221	11984	163737	14	5	101	4	59	9	7	187	181	116	14	403
L	*CTT*	11	28	28	268	913	79	10926	6044	3	8	7	19	32	11	17	22	140	427	20	297
E	*GAA*	177	163	35	6	4	579	13	22	80704	1875	25592	1845	91	118	96	154	24	39	27	10
D	*GAC*	2014	716	147	56	30	170	17	6	2386	90508	3351	23213	659	1402	334	16	181	52	173	15
E	*GAG*	244	226	30	16	15	2591	21	86	27179	3295	152612	3028	158	207	174	57	52	46	26	11
D	*GAT*	605	1944	61	15	44	94	8	7	1769	20719	2606	70821	141	396	893	17	46	147	47	2
A	*GCA*	24	32	265	60	35	52	15	145	508	83	395	59	5329	69	74	696	127	128	2	13
A	*GCC*	170	75	3897	416	127	27	131	68	128	621	225	177	86827	1100	199	132	1969	392	25	111
A	*GCG*	11	13	248	49	28	91	11	149	145	44	485	34	4952	88	49	145	135	105	2	10
A	*GCT*	62	135	912	86	303	16	27	39	62	150	126	365	17318	280	437	106	432	1191	5	22
G	*GGA*	19	43	17	7	8	71	8	15	1247	125	434	102	70	4063	2444	126	25	19	2	3
G	*GGC*	452	183	195	91	27	47	31	14	110	1304	197	389	940	70299	10181	19	321	69	47	22
G	*GGG*	39	66	30	11	10	239	6	97	334	115	1478	123	119	4951	3298	75	40	38	10	4
G	*GGT*	91	343	44	27	59	24	8	1	91	330	125	862	200	10850	24402	16	61	199	17	4
V	*GTA*	1	5	28	212	145	12	27	261	142	9	69	11	128	12	4	12620	1378	974	1	31
V	*GTC*	50	26	577	6081	1244	19	937	173	34	153	57	51	2028	234	50	1824	47252	6691	29	481
V	*GTG*	17	13	171	989	767	139	188	2165	102	23	498	26	562	57	49	8757	6074	4909	2	71
V	*GTT*	18	49	113	869	3036	7	127	96	20	43	30	138	437	86	195	1159	5514	24443	10	71
*	*TAA*	3	2	1	0	0	29	1	4	43	2	15	2	1	0	0	2	0	0	41	2
Y	*TAC*	161	63	21	34	6	86	82	21	16	170	35	55	29	39	11	2	22	4	68411	941
*	*TAG*	2	5	0	1	1	207	3	35	12	1	89	1	0	1	0	0	0	0	28	1
Y	*TAT*	55	209	13	7	26	76	26	17	16	73	26	169	5	3	32	0	11	10	16017	273
S	*TCA*	16	10	134	4	5	74	23	203	44	3	37	5	152	6	3	54	15	8	15	75
S	*TCC*	108	26	1174	57	25	30	187	52	13	51	27	23	1690	51	7	8	108	26	261	725
S	*TCG*	7	12	103	12	8	71	10	151	10	4	39	5	138	7	4	9	6	2	9	51
S	*TCT*	20	83	325	23	67	28	49	66	6	14	20	51	470	37	54	11	32	57	103	258
*	*TGA*	1	1	0	0	0	27	1	12	14	3	10	1	1	12	4	4	1	0	3	3
C	*TGC*	55	21	51	25	3	36	83	20	9	46	9	19	54	338	56	1	26	5	566	269
W	*TGG*	3	6	5	2	8	431	11	197	9	6	46	3	1	28	6	4	2	1	25	45
C	*TGT*	35	72	20	12	24	55	20	16	13	12	29	33	9	109	195	4	5	11	185	93
L	*TTA*	0	2	10	43	39	20	291	1886	16	2	10	1	5	1	0	204	19	28	15	382
F	*TTC*	18	10	72	505	105	13	1918	278	10	25	4	6	114	37	2	26	397	83	1042	92119
L	*TTG*	7	1	23	75	80	130	918	15979	14	2	43	0	35	7	6	97	59	41	17	591
F	*TTT*	15	22	23	86	350	15	366	234	4	10	7	12	29	8	8	18	80	197	287	16156

Table II. For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, codon substitution matrix $\mathbf{A}(m(X,\mathbb{C}))$ of X motifs $m(X,\mathbb{C})$ (Section 3.4). For each codon, the encoded amino acid is given

		N	*N*	*T*	*I*	*I*	*Q*	*L*	*L*	*E*	*D*	*E*	*D*	*A*	*G*	*G*	*V*	*V*	*V*	*Y*	*F*
		AAC	*AAT*	*ACC*	*ATC*	*ATT*	*CAG*	*CTC*	*CTG*	*GAA*	*GAC*	*GAG*	*GAT*	*GCC*	*GGC*	*GGT*	*GTA*	*GTC*	*GTT*	*TAC*	*TTC*
K	*AAA*	1235	1785	475	304	592	769	107	187	2233	890	1083	1517	357	284	432	250	221	409	277	187
N	*AAC*	9852	10627	507	238	394	477	82	136	1752	1456	764	2375	319	418	805	209	193	346	300	245
K	*AAG*	1512	2008	612	364	595	959	117	227	2653	991	1280	1665	467	320	551	289	289	579	298	243
N	*AAT*	7143	10014	516	288	532	490	90	150	1931	1473	878	2699	319	453	821	251	186	338	304	214
T	*ACA*	574	900	1923	320	462	215	70	105	771	344	420	638	278	141	251	267	283	451	130	178
T	*ACC*	622	880	3844	405	545	240	81	136	889	418	402	690	345	131	318	301	405	605	186	163
T	*ACG*	281	471	929	175	263	119	36	67	365	201	190	307	132	55	119	142	115	237	77	83
T	*ACT*	796	1084	3676	392	657	255	90	149	1055	454	475	805	381	170	353	351	402	661	207	171
R	*AGA*	491	677	171	140	225	276	50	94	622	233	334	471	119	114	211	123	116	172	155	115
S	*AGC*	581	823	274	118	185	160	42	62	699	391	338	649	187	179	278	120	95	180	103	104
R	*AGG*	174	218	61	60	77	129	21	26	255	95	120	184	38	37	60	26	40	61	47	44
S	*AGT*	698	1041	405	149	216	208	54	84	816	509	379	818	305	247	432	126	127	261	119	112
I	*ATA*	157	287	150	2179	3768	97	227	406	305	125	152	229	105	55	101	684	665	1091	137	284
I	*ATC*	287	331	292	7454	10862	131	350	601	398	192	221	327	214	75	127	1072	1659	2450	248	538
M	*ATG*	371	552	256	908	1677	214	357	776	590	199	245	421	201	100	161	408	402	751	275	644
I	*ATT*	356	516	391	8368	15453	194	471	869	640	222	316	429	260	109	190	1608	2023	3469	343	715
Q	*CAA*	1011	1519	380	228	459	3493	94	183	2517	752	1100	1405	330	220	413	210	177	359	223	171
H	*CAC*	535	664	125	89	177	185	43	50	472	241	247	414	95	85	120	72	82	110	411	188
Q	*CAG*	793	1139	284	227	357	2678	87	159	1835	561	894	1012	224	174	268	202	121	256	185	154
H	*CAT*	640	928	167	116	227	270	60	73	622	323	309	583	92	127	181	118	96	143	454	225
P	*CCA*	416	637	216	156	290	185	52	73	767	388	330	672	261	115	271	127	131	239	103	109
P	*CCC*	253	383	133	89	174	94	34	62	484	239	242	382	179	76	165	92	87	172	65	67
P	*CCG*	127	252	60	38	91	58	14	34	238	134	109	230	75	40	82	53	52	73	47	39
P	*CCT*	381	598	190	151	268	159	59	66	737	364	365	677	264	107	245	126	124	245	122	118
R	*CGA*	193	292	69	63	106	157	30	43	258	134	156	204	53	54	87	46	47	85	68	51
R	*CGC*	113	166	36	32	76	77	21	23	154	95	90	127	39	30	46	24	23	43	35	22
R	*CGG*	99	136	46	39	53	64	3	19	120	66	72	110	23	22	50	22	19	38	42	18
R	*CGT*	211	299	75	55	106	127	22	39	319	160	133	250	51	54	102	47	42	82	68	43
L	*CTA*	105	147	58	353	595	68	391	797	174	74	89	132	65	23	45	181	170	279	98	241
L	*CTC*	127	187	128	734	1021	89	573	1274	264	98	140	163	97	54	92	259	294	508	177	451
L	*CTG*	231	368	135	555	883	132	431	1076	390	177	190	298	141	83	111	253	230	410	144	379
L	*CTT*	230	338	166	959	1513	118	808	1749	368	166	182	270	150	76	105	397	418	709	248	572
E	*GAA*	1790	2638	701	450	770	1160	137	241	23870	3467	9443	6507	673	493	810	429	404	652	352	246
D	*GAC*	1796	2674	382	226	343	460	64	103	4714	9211	2002	14480	308	431	701	166	147	288	210	145
E	*GAG*	1246	1802	459	316	550	861	103	174	15702	2460	6558	4270	451	314	554	285	256	493	250	219
D	*GAT*	2463	3664	533	306	438	616	119	151	6351	11838	2842	21898	413	569	929	267	237	413	275	228
A	*GCA*	481	651	364	284	445	207	69	114	879	348	426	707	2571	281	519	321	318	536	131	169
A	*GCC*	528	719	447	285	482	216	88	158	937	393	447	627	4330	303	796	319	408	610	166	166
A	*GCG*	180	283	116	114	193	103	38	61	333	135	167	267	946	114	233	142	112	199	53	79
A	*GCT*	681	986	584	431	711	305	101	187	1339	590	621	926	5124	437	979	463	486	851	202	229
G	*GGA*	654	980	189	110	226	160	47	52	717	499	401	905	350	3102	8881	96	114	154	134	116
G	*GGC*	490	745	149	67	128	112	40	47	505	370	291	655	219	1922	4921	68	66	104	73	89
G	*GGG*	194	318	67	47	59	43	12	21	216	161	109	262	104	829	2096	28	27	51	29	37
G	*GGT*	807	1045	212	134	201	146	57	54	769	548	371	889	517	4173	21358	83	127	209	116	143
V	*GTA*	166	238	193	808	1464	86	113	221	357	115	174	235	166	43	94	1160	1414	2377	119	214
V	*GTC*	237	340	360	1648	2406	104	146	286	438	168	220	253	356	77	133	1897	3774	5672	158	304
V	*GTG*	276	344	337	1577	2623	165	181	346	528	187	231	320	337	104	171	2142	2734	4596	199	337
V	*GTT*	334	517	505	2487	4180	177	237	431	771	264	321	494	500	127	199	3025	5325	9415	260	481
*	*TAA*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	*TAC*	338	479	166	270	444	151	121	179	410	179	193	358	90	72	121	141	154	303	10095	1871
*	*TAG*	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Y	*TAT*	326	453	144	267	451	119	116	172	420	204	212	374	83	67	130	159	161	252	6896	1430
S	*TCA*	654	1071	451	173	272	254	61	96	918	488	428	883	410	217	454	167	133	287	151	149
S	*TCC*	635	951	558	149	249	201	46	74	849	470	402	760	558	228	424	165	166	248	170	125
S	*TCG*	547	828	418	136	206	223	42	88	714	406	397	683	351	187	365	176	106	209	119	105
S	*TCT*	981	1513	725	239	414	322	64	118	1430	740	636	1272	724	312	620	240	203	414	223	213
*	*TGA*	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
C	*TGC*	73	85	79	98	170	22	20	37	73	37	29	44	141	31	69	120	115	173	63	75
W	*TGG*	85	125	50	86	164	24	52	71	115	51	62	80	49	23	47	60	45	108	473	481
C	*TGT*	83	157	151	176	230	36	33	79	85	63	43	82	248	51	86	189	247	343	69	141
L	*TTA*	194	318	154	855	1465	112	894	1970	332	167	181	252	146	52	110	411	374	694	197	575
F	*TTC*	219	340	133	522	759	102	242	407	271	136	146	266	131	79	107	246	282	484	1598	11263
L	*TTG*	384	587	316	1793	2826	231	1667	4327	615	266	279	436	264	101	190	745	740	1265	426	1171
F	*TTT*	299	453	162	579	1064	130	301	505	397	200	188	336	149	100	155	306	272	557	1725	10020

APPENDIX

Table III. For the mammal gene multiple alignments with respect to the human reference genes $s_1 = \mathbb{H}$, codon substitution submatrices of $\mathrm{B}(m(X, \mathbb{H}))$ (in %) of X motifs $m(X, \mathbb{H})$ (Section 3.4) for the 12 amino acids $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ coded by the circular code X (1).

A				
	GCA	*GCC*	*GCG*	*GCT*
GCA				
GCC		66.1		
GCG				
GCT				
Sum		66.1		
Mean		66.1		

D		
	GAC	*GAT*
GAC	72.9	22.0
GAT	16.7	67.1
Sum	89.6	89.1
Mean	89.4	

E		
	GAA	*GAG*
GAA	67.3	13.0
GAG	22.7	77.7
Sum	90.0	90.7
Mean	90.3	

F		
	TTC	*TTT*
TTC	78.9	
TTT		
Sum	78.9	
Mean	78.9	

G				
	GGA	*GGC*	*GGG*	*GGT*
GGA				
GGC		69.6		21.9
GGG				
GGT		10.7		52.4
Sum		80.3		74.3
Mean		77.3		

I		
	ATA	*ATC*	*ATT*
ATA			
ATC	73.1		22.4
ATT	12.5		61.8
Sum	85.6		84.2
Mean	84.9		

L						
	CTA	*CTC*	*CTG*	*CTT*	*TTA*	*TTG*
CTA						
CTC		68.1		4.6		
CTG		10.9		73.0		
CTT						
TTA						
TTG						
Sum		79.0		77.6		
Mean		78.3				

N		
	AAC	*AAT*
AAC	71.1	22.2
AAT	15.5	62.3
Sum	86.6	84.5
Mean	85.6	

Q		
	CAA	*CAG*
CAA		
CAG		80.7
Sum		80.7
Mean		80.7

T				
	ACA	*ACC*	*ACG*	*ACT*
ACA				
ACC		63.5		
ACG				
ACT				
Sum		63.5		
Mean		63.5		

V				
	GTA	*GTC*	*GTG*	*GTT*
GTA	42.7	1.9		2.1
GTC	6.2	64.2		14.8
GTG				
GTT	3.9	7.5		53.9
Sum	52.8	73.6		70.9
Mean		65.7		

Y		
	TAC	*TAT*
TAC	76.1	
TAT		
Sum	76.1	
Mean	76.1	

Table IV. For the yeast gene multiple alignments with respect to the *S. cerevisiae* reference genes $s_1 = \mathbb{C}$, codon substitution submatrices of $\mathrm{B}(m(X,\mathbb{C}))$ (in %) of X motifs $m(X,\mathbb{C})$ (Section 3.4) for the 12 amino acids $p \in \mathcal{X} = \{A, D, E, F, G, I, L, N, Q, T, V, Y\}$ coded by the circular code X (1).

A	*GCA*	*GCC*	*GCG*	*GCT*
GCA				
GCC	16.1			
GCG				
GCT				
Sum	16.1			
Paac	16.1			

D	*GAC*	*GAT*
GAC	19.9	17.9
GAT	25.6	27.1
Sum	45.4	45.1
Paac	45.3	

E	*GAA*	*GAG*
GAA	26.3	23.6
GAG	17.3	16.4
Sum	43.6	39.9
Mean	41.8	

F	*TTC*	*TTT*
TTC	29.9	
TTT		
Sum	29.9	
Mean	29.9	

G	*GGA*	*GGC*	*GGG*	*GGT*
GGA				
GGC	10.1		9.1	
GGG				
GGT	21.9		39.7	
Sum	32.0		48.8	
Mean	40.4			

I	*ATA*	*ATC*	*ATT*
ATA			
ATC	18.5	16.3	
ATT	20.7	23.1	
Sum	39.2	39.4	
Mean	39.3		

L	*CTA*	*CTC*	*CTG*	*CTT*	*TTA*	*TTG*
CTA						
CTC	5.6	6.2				
CTG	4.2	5.2				
CTT						
TTA						
TTG						
Sum	9.9	11.4				
Mean	10.6					

N	*AAC*	*AAT*
AAC	20.6	16.2
AAT	15.0	15.3
Sum	35.6	31.5
Mean	33.5	

Q	*CAA*	*CAG*
CAA		
CAG	13.6	
Sum	13.6	
Mean	13.6	

T	*ACA*	*ACC*	*ACG*	*ACT*
ACA				
ACC	14.9			
ACG				
ACT				
Sum	14.9			
Mean	14.9			

V	*GTA*	*GTC*	*GTG*	*GTT*
GTA	5.2	5.0	5.0	
GTC	8.4	13.3	11.9	
GTG				
GTT	13.5	18.8	19.7	
Sum	27.1	37.2	36.6	
Mean	33.6			

Y	*TAC*	*TAT*
TAC	33.0	
TAT		
Sum	33.0	
Mean	33.0	

APPENDIX

Table V. 100 random codes R generated with similar properties to the X circular code, except its circular code property, for comparison purposes. We only provide 25 random codes.

R	Codons belonging to the random code R																			
1	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC
2	AAC	AAT	ACA	ACT	AGG	ATA	ATT	CAA	CAG	CTC	CTG	GCC	GCG	GCT	GGC	GTA	GTC	GTT	TGC	TGT
3	ACT	AGG	AGT	ATA	ATG	CAA	CAG	CCA	CCG	CTC	GAA	GAG	GCT	GGC	TAC	TAT	TCC	TCT	TGG	TTG
4	AAT	ACC	AGA	ATT	CCT	CGA	CGG	CTA	CTC	CTG	GAA	GAC	GAT	GCC	GCG	GGT	GTG	TAC	TAT	TTA
5	ACC	AGA	AGG	ATA	ATC	CCG	CCT	CGC	CTA	CTC	CTG	CTT	GAA	GAT	GCA	GGA	GTA	GTG	TGT	TTA
6	AAC	AAG	ACA	AGC	CCA	CGA	CTG	CTT	GAG	GCA	GCC	GGC	GTA	GTT	TAC	TAT	TCA	TCT	TGG	TGT
7	AAC	ACG	AGA	AGG	ATA	CCT	CGC	CGG	CGT	CTA	CTG	GAA	GAC	GCT	GTA	TAT	TCC	TGC	TTA	TTG
8	AAT	ACT	AGT	ATC	CAA	CCT	CGA	CTG	GAC	GAG	GAT	GCA	GCG	GCT	GTA	TAC	TAT	TCC	TCG	TGG
9	AAG	ACG	AGG	ATA	ATG	CAT	CCA	CGG	CGT	CTT	GAC	GCA	GCC	GCT	GGC	TAC	TAT	TCA	TTA	TTG
10	AAG	ACG	ACT	AGG	AGT	ATC	CAA	CAG	CCT	CGA	CGC	CTG	CTT	GCA	GGT	GTG	GTT	TAC	TAT	TCA
11	ACC	ACT	AGA	AGT	ATG	CAG	CAT	CCA	CGC	CTT	GAA	GAT	GCA	GCT	GGA	GGC	GTA	TCT	TGT	TTC
12	AAC	AAT	ACA	ACT	AGG	AGT	ATA	CAT	CGC	CTA	CTC	GAC	GCA	GCC	GGC	GTC	GTT	TGG	TGT	TTG
13	AAG	ACG	ACT	AGG	ATA	CAA	CAC	CAG	CCT	CGT	CTA	GAG	GCA	GCC	GTG	GTT	TAT	TCG	TCT	TTG
14	AAC	AAT	AGA	AGC	ATT	CAG	CCT	CGC	CGG	CTA	GAC	GAT	GCC	GGC	GGT	GTA	TAT	TCA	TTC	TTG
15	ACA	ACC	ACG	AGA	AGG	ATC	ATT	CAC	CAT	CCT	CGT	GAG	GCA	GGA	GGC	GTT	TGC	TGT	TTA	TTC
16	AAC	ACA	ACC	AGC	AGG	ATA	CAT	CGC	CGG	CTT	GCA	GGA	GTA	GTG	TAC	TCG	TCT	TGG	TTA	TTC
17	AAC	ACC	ACG	ACT	AGA	AGG	ATA	ATC	ATG	CAA	CCG	CCT	CGT	GCG	GTG	TCG	TGG	TTA	TTC	TTG
18	AAC	AAG	ACA	ACG	ACT	AGC	AGG	CAT	CGA	CTA	CTC	GAG	GGA	GTT	TCC	TCT	TGC	TGG	TTC	TTG
19	AAT	ACT	AGC	AGG	ATA	ATC	ATG	CAA	CGC	CTA	CTC	GAG	GCC	GCT	GTA	GTC	GTT	TAC	TCG	TGG
20	AAC	AGA	AGG	AGT	ATA	CAG	CCA	CGA	CGG	CTA	CTG	CTT	GAC	GGA	GTC	GTT	TAT	TCC	TCG	TTC
21	AAT	AGA	AGG	CAA	CAG	CCA	CGG	CTC	CTG	GAC	GCA	GCT	GTA	GTC	GTG	TAC	TAT	TCA	TTC	TTG
22	AAG	ACA	ACC	ACT	ATA	ATG	ATT	CAA	CAT	CGA	CGG	CTG	CTT	GAG	GCG	GCT	GGC	GTG	TCT	TTC
23	ACA	AGG	AGT	ATG	ATT	CAT	CCG	CGA	CTC	CTT	GAA	GAC	GAG	GCA	GCG	GTA	TAC	TCT	TGC	TTC
24	AAC	ACA	ACT	AGC	ATA	CAA	CAT	CCG	CTT	GAA	GAC	GCG	GCT	GGT	GTA	GTC	TCG	TCT	TGG	TGT
25	ACA	AGA	AGT	ATA	ATC	CAG	CAT	CCT	CGC	CTA	CTC	GAG	GCG	GGA	GTA	TAC	TCG	TCT	TGG	TGT

Table VI. List of the 133 organisms included in the multiple sequence alignments of LSU rRNAs (23S/28S and 5S) and SSU rRNAs (16S/18S).

Bacteria	Archaea	Eukaryota
Acinetobacter sp.	*Aeropyrum pernix*	*Adineta vaga*
Agrobacterium tumefaciens	*Archaeoglobus fulgidus*	*Aedes albopictus*
Anabaena variabilis	*Caldivirga maquilingensis*	*Anolis carolinensis*
Azoarcus sp.	*Haloarcula marismortui*	*Arabidopsis thaliana*
Bacillus anthracis	*Halobacterium sp.*	*Caenorhabditis briggsae*
Bacteroides thetaiotaomicron	*Haloferax volcanii*	*Caenorhabditis elegans*
Bartonella henselae	*Haloquadratum walsbyi*	*Cryptosporidium hominis*
Bifidobacterium longum	*Halorubrum lacusprofundi*	*Cyanidioschyzon merolae*
Blochmannia floridanus	*Hyperthermus butylicus*	*Danio rerio*
Bradyrhizobium japonicum	*Ignicoccus hospitalis*	*Dictyostelium discoideum*
Buchnera aphidicola	*Metallosphaera sedula*	*Drosophila melanogaster*
Burkholderia sp.	*Methanocaldococcus jannaschii*	*Eremothecium gossypii*
Caulobacter crescentus	*Methanococcoides burtonii*	*Gallus gallus*
Chlamydophila caviae	*Methanococcus aeolicus*	*Guillardia theta*
Chlorobium tepidum	*Methanocorpusculum labreanum*	*Homo sapiens*
Coxiella burnetii	*Methanoculleus marisnigri*	*Latimeria chalumnae*
Crocosphaera watsonii	*Methanopyrus kandleri*	*Leishmania major*
Cytophaga hutchinsonii	*Methanoregula boonei*	*Monodelphis domestica*
Dechloromonas aromatica	*Methanosaeta thermophila*	*Mus musculus*
Dehalococcoides ethenogenes	*Methanosarcina acetivorans*	*Oryza sativa*
Deinococcus radiodurans	*Methanosarcina barkeri*	*Pan troglodytes*
Escherichia coli	*Methanosarcina mazei*	*Plasmodium falciparum*
Fusobacterium nucleatum	*Methanosphaera stadtmanae*	*Rattus norvegicus*
Geobacillus kaustophilus	*Methanospirillum hungatei*	*Saccharomyces cerevisiae*
Geobacter sulfurreducens	*Methanothermobacter thermautotrophicus*	*Schizosaccharomyces pombe*
Gloeobacter violaceus	*Nanoarchaeum equitans*	*Tetrahymena thermophila*
Gluconobacter oxydans	*Natronomonas pharaonis*	*Thalassiosira pseudonana*
Haemophilus influenzae	*Picrophilus torridus*	*Trypanosoma brucei*
Helicobacter hepaticus	*Pyrobaculum calidifontis*	*Yarrowia lipolytica*
Legionella pneumophila	*Pyrococcus furiosus*	*Xenopus laevis*
Leifsonia xyli	*Staphylothermus marinus*	
Listeria monocytogenes	*Sulfolobus acidocaldarius*	
Magnetococcus sp.	*Sulfolobus tokodaii*	
Magnetospirillum magnetotacticum	*Thermococcus kodakarensis*	
Mesoplasma florum	*Thermofilum pendens*	
Mycobacterium leprae	*Thermoplasma volcanium*	
Neisseria gonorrhoeae		
Nitrosomonas europaea		
Novosphingobium aromaticivorans		
Oceanobacillus iheyensis		
Photorhabdus luminescens		
Polaromonas sp.		
Porphyromonas gingivalis		
Propionibacterium acnes		
Pseudomonas aeruginosa		
Ralstonia eutropha		
Rhodobacter sphaeroides		
Rhodopirellula baltica		
Shewanella oneidensis		
Sinorhizobium meliloti		
Staphylococcus aureus		
Streptococcus pneumoniae		
Streptomyces coelicolor		
Symbiobacterium thermophilum		
Synechococcus sp.		
Synechocystis sp.		
Thermoanaerobacter tengcongensis		
Thermosynechococcus elongatus		
Thermotoga maritima		
Thermus thermophilus		
Treponema pallidum		
Tropheryma whipplei		
Wolbachia_endosymbiont		
Xanthomonas_axonopodis		
Yersinia_pestis		
Zymomonas_mobilis		

Gopal Krishna DILA

Motifs de codes circulaires dans les gènes codant les protéines et les ARN ribosomaux

Résumé

La thèse porte sur les motifs du code circulaire *X*, un code correcteur d'erreurs trouvé dans les gènes, qui ont la capacité de trouver le cadre de lecture. Nous avons étudié la conservation des motifs *X* dans les gènes de différentes espèces et identifié des pressions sélectives spécifiques pour les maintenir. Nous avons aussi identifié des motifs *X* universels dans l'ARN ribosomique, situés dans des régions fonctionnelles importantes du ribosome et suggérant que les codes circulaires ont représenté une étape importante dans l'émergence du code génétique standard (SGC). Ensuite, nous avons étudié le rôle fonctionnel des motifs *X* dans la traduction moderne et identifié une forte corrélation entre l'enrichissement des motifs *X* et le niveau de traduction des gènes. Enfin, nous avons comparé l'optimalité du code *X* avec le SGC et d'autres codes circulaires maximaux, et identifié une nouvelle fonctionnalité de *X* dans la minimisation des effets des erreurs après un décalage du cadre.

Mots-clé : motifs du code circulaire, code correcteur d'erreurs, code génétique standard, l'ARN ribosomique, cadre de lecture, gènes codant les protéines, décalage du cadre, traduction erreurs

Summary

The thesis focuses on motifs of the circular code *X*, an error-correcting code found in protein-coding genes, which have the ability to synchronize the reading frame. We first investigated the evolutionary conservation of *X* motifs in genes of different species and identified specific selective pressures to maintain them. We also identified a set of universal *X* motifs in ribosomal RNAs, which are located in important functional regions of the ribosome and suggest that circular codes represented an important step in the emergence of the standard genetic code (SGC). Then, we investigated the functional role of *X* motifs in modern translation processes and identified a strong correlation between *X* motif enrichment in genes and translation levels. Finally, we compared the frameshift optimality of the circular code *X* with the SGC and other maximal circular codes, and identified a new functionality of the code *X* in minimizing the effects of translation errors after frameshift events.

Keywords: circular code motifs, error-correcting code, standard genetic code, ribosomal RNAs, reading frame, protein-coding genes, frameshift, translation errors