UNIVERSITY OF SIENA

Department in Biotechnology, Chemistry and Pharmacy

Machine learning methods for the prediction of translation speed

PhD Candidate: Giorgia Giacomini Giorgia Giacomini

PHD SCHOOL IN BIOCHEMISTRY AND MOLECULAR BIOLOGY BiBim 2.0

Cycle XXXIII
Coordinator of PhD school: Prof. Lorenza Trabalzini
ING-INF/05

Supervisor

Prof. Monica Bianchini

Dr. Sherine Awad

Academic year: 2019/2020

Abstract

Ribosomes carry out protein synthesis from mRNA templates by a highly regulated process called translation. Translational control plays a key role in the regulation of gene expression, under physiological and pathological conditions. How translation is regulated under different conditions and what factors greatly influence the translation speed remains open questions in molecular biology.

In recent years, Ribosome profiling technique (Ribo-seq) has emerged as a powerful method for globally monitoring the translation process in vivo at single nucleotide resolution [1]. Ribo-seq is based on deep sequencing of mRNA fragments covered by ribosomes, called Ribosome Protected Fragments (RPFs). Sequencing of RPFs allows to record the precise position of the ribosomes at the time in which the translation was blocked. However, the exploitation of the full power of this technique is hindered by notable weaknesses (e.g. a low signal to noise ratio), influencing the reproducibility of Ribo-seq experiment. [2]. The aim of this thesis is the development of a newly designed statistical approach integrated with machine learning methodologies for a comprehensive understanding of the information contained in Ribosome Profiling data and for prediction of translation speed. Our data analysis approach consists of a systematic comparison of Ribo-seq profiles referring to several publically available Ribo-seq datasets generated in different laboratories, in different time but under the same experimental conditions. In the *E.coli* case studio, the analysis of 3588 Ribo-seq profiles across eight independent datasets revealed that only 40 profiles are significantly reproducibles.

The identification of reproducible Ribo-seq profiles allows us to build consensus sequences which highlighted the nucleotides located within fast and slow regions. The density of the RPFs along the mRNAs reflects the different time spent by ribosomes in translating each part of the ORF. Therefore slow regions,

extremely rich of ribosomes, and fast regions, characterized by few ribosomes, can be easily identified by Ribo-seq. We analysed the occurrences of nucleotides, dinucleotides, and codons of consensus sequences in order to conjecture the existence (or not) of signals in the sequence that could modulate the speed of translation. To this aim, we implemented different neural network architectures that let us classify the translation speed of the previously identified consensus sequences with high accuracy. Although the limited amount of data, the results clearly demonstrate that the models can extract useful information. Furthermore, we used the significantly reproducible profiles as a reference for comparative analyses aimed at detecting whether modifications in experimental conditions (heat shock stress and aminoacid starvation) could affect the reproducibility of our Ribo-seq workflow and thus influence the translation control. A preliminary analysis on Ribo-seq human data suggests that our method provides a rich resource for further in-depth studies about translation control of gene expression in all kind of Ribo-seq datasets, including those related to highly differentiated organisms like humans.

Contents

List of Figures

List of Tables

Thesis structure

The thesis is organized as follows:

- Chapter 1 gives a broad introduction to all the biological concepts this dissertation is based on. In particular, a brief overview of the gene expression regulation is provided, highlighting the main mechanisms of transcriptional and post-transcriptional control in both prokaryotes and eukaryotes organisms. Molecular mechanisms of translational regulation are exploited, by focusing on known processes like codon usage bias. Finally, the biological role of translation control in human tumorigenesis and how its mechanisms lead to the phenotypic hallmarks of cancer are illustrated.

- Chapter 2 describes the Ribosome Profiling technique (Ribo-seq) used to study the translation of gene expression and to investigate the factors that influence the velocity of the ribosomes during translation.
  Its advantages and its limits are reported.

- Chapter 3 introduces our data analysis approach to identify the reproducible Ribo-seq profile from the comparison of independent Ribo-seq experiments. The implementation of a novel data analysis method to address the limitations of Ribo-seq and to recover its full resolution is proposed. *Based on [3].*

- In Chapter 4 our method is applied to different *E.coli*'s Ribo-seq datasets performed in different laboratories under the same conditions. The results obtained will be analyzed in the following chapters, through statistical and machine learning approaches to deepen the content of the consensus sequences of the reproducible Ribo-seq profiles.

- Chapter 5 explores the nucleotide/dinucleotide and codon composition of the consensus sequences through statistical methods in order to conjecture the existence of signals that could modulate the speed of translation.
- Chapter 6 provides a brief presentation of Artificial Neural Network (ANN) models and their application in our study (Multilayer Perceptron model and Convolutional Neural network). In addition, we discuss some possible extensions for the results presented and suggest eventual areas of future research.
- Chapter 7 describes how our approach is used to perform comparative analysis aimed at detecting other conditions potentially affecting translational control, using *E.coli* datasets under stress condition (i.e. heat shock and amino acid starvation). The significantly reproducible profiles identified in Chapter 4 are used as a reference for these comparative experiments.
- Chapter 8 examines the preliminary results of Human Ribo-seq profiles of paired liver tumours and adjacent noncancerous normal liver tissues from 10 patients with hepatocellular carcinoma (HCC).
- Our results are discussed in the Chapter 9, accompanied by supplementary material in the Appendix section.
- Chapter 10 examines the research carried out during the PhD period, but not directly covered by this thesis.

Chapter 1

Background

1.1 From DNA to RNA

The central dogma of molecular biology defines the basic flow of genetic information within a biological system. In every cell, the information encoded in a segment of DNA (*gene*) is transcribed into an RNA molecule, which can then be translated into a linear sequence of amino acids. The information stored in the DNA sequences is expressed in a highly selective, differentiated, and regulated manner depending on internal conditions and external stimuli. When a particular protein is needed by the cell, the gene is transcript into another type of nucleic acid—RNA (ribonucleic acid). The RNA polymerase moves stepwise along the DNA and the single growing RNA chain is synthesized in 5' => 3' direction, using an exposed DNA strand as a template. The resulting RNA (*mRNA*) is then used to carries genetic information from the nucleus to cytoplasm for protein synthesis (See Figure 1.1).

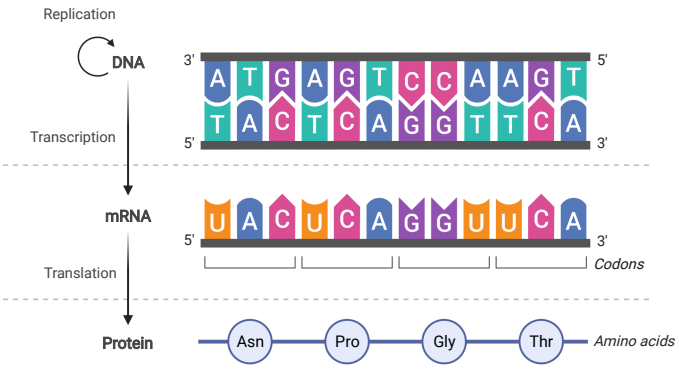


Figure 1.1: Central dogma of molecular biology. Schematic representation showing the main steps involved in protein synthesis. The information stored in a segment of DNA is transcribed into an RNA molecule (transcription). The mRNA is then used as template for the synthesis of an amino acidic chain (translation). Adapted from "Central Dogma", by BioRender.com (2020).

In addition to the mRNAs, a transcriptome also contains many RNA transcripts which do not encode for proteins (i.e. *noncoding RNAs*).
Prokaryotic genome is characterized by a low percentage of noncoding DNA (average of 12%). In contrast, the noncoding sequences account for a large portion of eukaryotic genome. It is estimated that only around 2% of the human genome contains coding genes, i.e DNA sequences that encode for protein products [4]. In accordance with several studies, the importance of non-coding RNAs (*ncRNAs*) has been recognized [5]. They have emerged as pivotal molecules that are related to several biological processes, including mRNA stability, RNA processing and transcriptional regulation[6].
Figure 1.2 summarizes the most common types of RNA based on their function within the cell. Some of them have important roles in regulating gene expression at multiple levels (e.g. microRNAs and small nucleolar RNA), while others like transfer RNAs (*tRNA*) and ribosomal RNAs (*rRNA*) are directly involved in translation process.

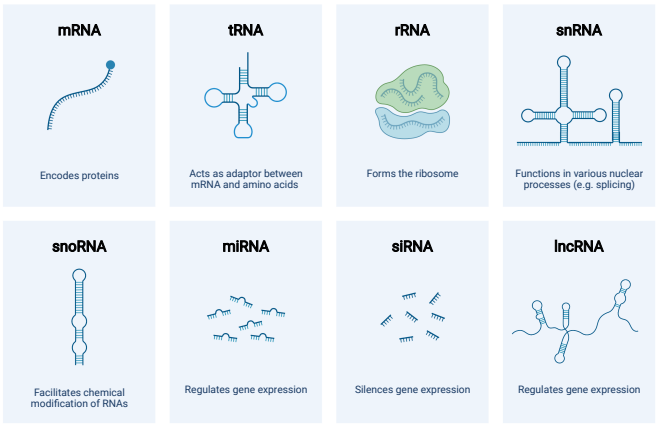


Figure 1.2: Types of RNA Produced in Cells The diagram shows the structure of the different types of RNA and their main functions. Adapted from BioRender.com (2020).

1.1.1 Main principles of genes regulation

Both prokaryotic and eukaryote cells carry out the control of gene expression at different levels by modulating the amount and type of protein. Although the genome is far more complex in eukaryotes than in bacteria, basic similarities in regulation of gene expression exist between them. Transcription is the first

stage of gene expression and its regulation can occur in every step, including initiation, elongation and termination. Some genes, namely housekeeping genes, are expressed continuously across different tissues, because theirs products are constantly needed to maintain essential cellular functions.
The constant expression not subjected to regulation is defined *constitutive gene expression*. Differently, other genes are expressed at different quantitative levels in different cell types (e.g. the genes encoding for insulin which is secreted exclusively by the pancreas cells). The genes whose expression increase in response to external stimuli are called *inducible genes* (e.g genes that encode for enzymes involved in DNA reparation in response to DNA damage). Instead, the genes whose expression decrease in response to certain molecular stimuli are known as *repressible genes*. Although the constitutive genes are expressed at a constant level, the proteins they encode are present in variable amounts. For these genes, the RNA polymerase-promotor interaction influences (affects) considerably the timing of transcription initiation [7].

Transcription overview in Prokaryotes

Initiation of transcription requires the assembly of the pre-initiation complex at the transcription start site.
In bacteria, the transcription process is catalyzed by a single type of RNA polymerase. Specifically, in *E.coli* the complete enzyme or holoenzyme has a molecular weight of 460 kD and it consists of two components: the core enzyme with its five subunits ($\alpha_2\beta\beta\omega$) performs the elongation reaction of RNA polymerization and the σ factor, which is involved in promoter recognition [8]. In *E.coli*, the expression of housekeeping genes depends on the expression of σ^{70}, the first sigma factor discovered. It is able to recognize specifically two conserved sequences in the promoter, at 10 and at 35 base pairs upstream of the transcription start site (5 'TATAAT 3 'and 5 'TTGACA 3', respectively) [9]. *E.coli* encodes seven alternative σ factors which allow RNA polymerase to bind different promoter consensus sequences and regulate distinct classes of genes. The availability of different σ factors allows to modify the gene expression pattern in response to environmental changes such as heat shock stress. In the case of heat shock stress, the transcription initiation is mediated by a specialized sigma-factor namely σ^{32}, which directs the core
RNA polymerase to recognize the promoters for heat shock genes[8]. In addition, some *E.coli* promoters have a third sequence (UP) located upstream of the 35 region which binds C-terminal domain of the subunit α.
Transcription initiation involves several defined steps: 1)Firstly, RNAP holoenzyme, leading by the factor σ, binds to the promoter sequence to form

the closed "preinitiation" complex; 2) then unwinds the DNA around the initiation site (12-14 bases of DNA) to form the open-promoter complex. The single stranded DNA is available as a template for transcription; 3) Once RNAP has added about the first ten nucleotides, the σ factor dissociates from the core polymerase which leaves the promoter, moves along the template DNA and elongates the growing RNA chain in the 5'-to-3' direction. The last step of the transcription phase is represented by *termination*. It is a crucial step in the regulation of gene expression since it modulates the relative levels of various genes and controls the transcription response to a metabolic or regulatory signal. Two classes of termination signals have been identified: one of them involves a nascent RNA-dependent helicase (Rho), while the other relies on specific *Rho-independent* sequences in the DNA template strand [10]. Typically, these sequences consist of a GC-rich sequence followed by approximately seven A residues and let the formation of a stable stem-loop structure by complementary base pairing in the transcribed mRNA.
While basic similarities in gene transcription exist between prokaryotes and eukaryotes—including the fact that RNA polymerase binds upstream of the gene on its promoter to initiate the process of transcription —multicellular eukaryotes control cell differentiation through more complex and precise temporal and spatial regulation of gene expression.

Regulation of gene expression: *E.coli* model

Because of their relative simplicity, bacteria are ideal models for studying many fundamental aspects of control of gene expression.
Generally, in prokaryotic cells the control of gene expression occurs mainly at the transcription level, through the activation or inactivation of genes. One of the key characteristics of bacterial chromosome is that that functionally-related genes are organized in clusters and they are transcribed together into a single mRNA molecule (polycistronic mRNA). A group of functionally-related genes controlled by the same promoter and other regulatory sequences constitute the prototype of genic organization namely *operon*. Figure 1.3 illustrates the structure of the well-known inducible *lac operon* which encodes to enzymes required to metabolize sugar lactose, a sugar used as a source of carbon and energy. The transcription of lac genes is regulated by the binding of a repressor, encoded by the lacI gene, to a specific DNA sequence overlapping the promoter. The repressor is constitutively expressed and turns off transcription in the absence of lactose by a mechanism referred to negative regulation. The lac operon is also subjected to a positive regulation that relies on glucose availability. When glucose is available, enzymes involved in the catabolism of

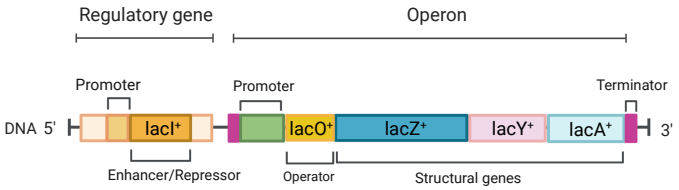


Figure 1.3: Structure of Operon lac. The lac operon encodes for enzyme proteins involved in lactose metabolism. The lac operon consists of three structural genes controlled by the same promoter, lacZ (encoding β-galactosidase), lacY (encoding permease), and lacA (encoding transacetylase). Created with BioRender.com (2020).

lactose are not expressed. In contrast, if glucose is low, cyclic AMP levels are high, and it readily binds to catabolite activator protein (CAP) and stimulates its binding to regulatory sequences of lac operon to increase the expression of β-galactosidase. This global mechanism is termed *catabolite repression* (CCR) and allows bacteria to selectively use the preferred substrates (glucose), inhibiting the expression and functions related to secondary carbon sources, such as lactose [11]. Another regulatory feature found in bacteria is attenuation which causes premature termination of transcription. This regulatory mechanism represses genes in the presence of their own products and it is typically used to regulate amino acid synthesis. Attenuation involves the 5'-cis-acting regulatory regions (attenuators) that fold into alternative RNA structures (stem-loop mRNA), acting as terminators of transcription [12]. The frequency at which the transcription is attenuated is based on the availability of amino acids in order to prevent unregulated and unnecessary gene expression.
One of the main aspect of control of gene expression in prokaryotes is represented by *regulon* in which multiple operons and single genes are under the same type of transcriptional control. An example of bacterial regulon in E.coli is the *response SOS* which allows the simultaneous and coordinated activation of several genes responsible for DNA damage repair [13].
Once a cell accumulates a large amount of DNA damage, the replication of DNA is arrested and the number of single strand breaks increases. RecA, a master protein in homologous recombination, binds to single-stranded DNA (ssDNA), it becomes activated and its co-protease activity is induced facilitating the excision of LexA repressor and allowing access to all promoters, operons, and genes involved in DNA damage response. Interestingly, when the repressor is inactivated, the transcript levels of recA gene increase 50-100 times compared

to normal conditions.

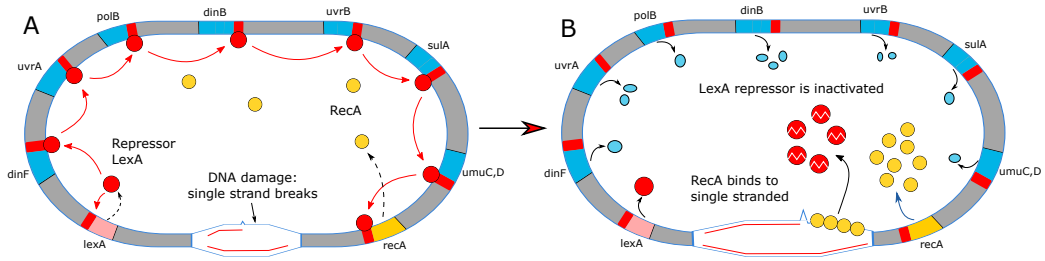


Figure 1.4: SOS response in E.coli. A)When the DNA is severely damaged, DNA duplication is blocked and the number of single-chain breaks increases. The RecA protein binds to single-stranded damaged DNA and acquires protease activity. B) RecA protein facilitates the self-cleavage and inactivation of the LexA repressor. Then, the SOS genes are transcribed. Created with BioRender.com (2020).

1.1.2 Transcription overview in Eukaryotes

The initiation of transcription is a pivotal moment of gene expression in both prokaryotes and eukaryotes. Both in bacteria and in eukaryotes, promoters are DNA sequences where RNA polymerase binds and initiate transcription of specific gene.
In eukaryotes, *transcription factors*(TFs) are DNA binding proteins equivalent to bacterial repressors and activators. Typically, transcription activators and repressors contain a single DNA-binding domain and one or a few activation or repression domains, respectively. The most common structural motifs found in the DNA-binding domains are the C_2H_2 zinc finger, homeodomain, helix-turn-helix (HTH), and basic zipper (leucine zipper). Differently from prokaryotes, transcription control elements in eukaryotes are often located tens of thousands of bases far away from the promoter that they regulate. These cis-acting elements which allow to stimulate or repress eukaryotic promoter are namely *enhancers* and *silencers*, respectively. The cooperative binding of multiple activators close to an enhancer forms a multiprotein complex called *enhanceosome* [14]. Different from bacteria, eukaryotic cells have three different nuclear RNA polymerases, named I, II and III, that transcribe distinct classes of genes:

- RNA polymerase I (RNAPI): transcribes genes encoding precursor rRNA (pre-rRNA), the three largest species of rRNAs (28S, 18S, and 5.8S).
- RNA polymerase II (RNAPII): transcribes all protein-coding genes, microRNAs (miRNAs) and long noncoding RNAs (lncRNAs) involved in regulation of gene expression.
- RNA polymerase III (RNAPIII): synthesizes tRNAs, the smallest species of ribosomal RNA (5s rRNA) and other short and stable RNAs involved in splicing and protein transport.

Although eukaryote transcription is more complex than prokaryote transcription, RNA polymerase II exhibits striking structural similarities [15]. This suggests that the mechanism used to transcribe DNA to RNA is highly conserved among different species. Structural domain analysis within RNAPII revealed the presence of a major subunit with a carboxy-terminal domain (CTD) consisting of 52 repeats of 7 amino acids (consensus sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser). During the transcription cycle, the aforementioned domain CTD is subject to extensive post-translational modification which regulates its activity[16]. The promoters of many genes transcribed by polymerase II contain a TATA box (consensus sequence TATAA) located 25-30 nucleotides upstream of the transcription start site.

The TATA box is the most common core promoter element and is prevalent in rapidly transcribed genes [17].
This consensus sequence is recognized by transcription factor TFIID which is constituted of multiple subunits, including 38-kDa TATA-binding protein (TBP) and thirteen TBP associated factors (TAFs).
The complex TFIID-TATA acts as catalizator by adding the other transcriptional factors sequentially. The preinitiation complex (PIC) includes RNA polymerase II and six general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH.
The ATP-dependent helicase activity of the TFIIH subunit allows to separates the template strands at the start site in most promoters [18].
Figure 1.5 summarizes the stepwise of transcriptional Initiation in eukaryotic cells.

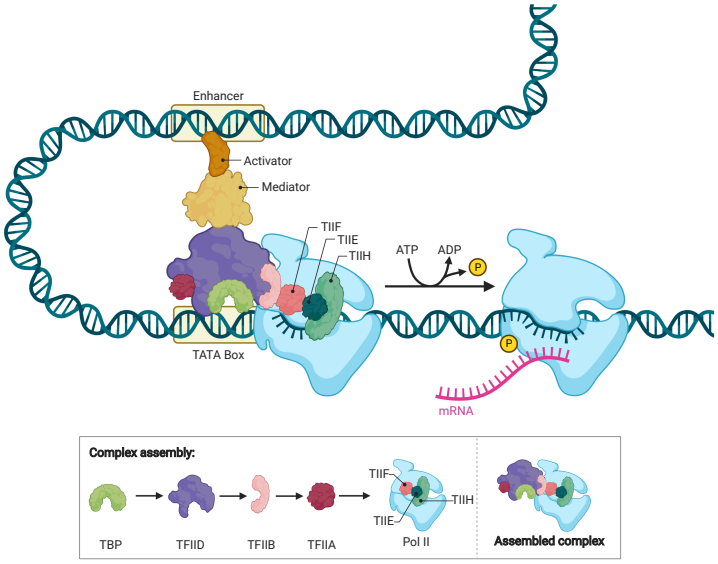


Figure 1.5: Transcriptional Initiation. Assembly of eukaryotic preinitiation complex (PIC) is stimulated in response to activator binding to enhancer, which in turn recruits mediator, which interacts directly with CTD of RNAPII and TFs. Adapted from BioRender.com (2020).

Before the start of transcription, the CTD domain of RNA polymerase II is not phosphorylated but associated with a large protein complex (mediators) which serves as a functional link between TFs bound to enhancer and the basal transcriptional complex. Once the complex is complete, the RNA polymerase II begins transcription and is phosphorylated on the Ser 5 residues of the CTD by Transcription Factor IIH (TFIIH). After mediator interacts with the

CTD of RNAPII, it undergoes a conformational change and is released when elongation phase begins. Mediator functions are related to chromatin structure and enhancer-promoter contacts [19]. It also has been observed that mutation in Saccharomyces Cerevisiae genes product belonging to the mediator complex causes block of transcription (SRB loci) [10].
The initial phosphorylation of CTD recruit DRB Sensivity Inducing Factor (DSIF), which in turn recruits NELF (Negative Elongation Factor), which forces RNA polymerase to stop transcription; First, the capping complex interacts with DSIF and then with the phoshorylated CTD. At this time, the cap is added at the 5 'end; DSIF turned into a positive elongation factor upon phosphorylation by positive transcription elongation factor (P-TEFb), which is recruited by the capping complex. P-TEFb phosphorylates both the Ser 2 on the CTD and DSIF. Through the last modification, the RNA polymerase II stall complex disassembles, and it can thus restart the elongation of the transcript.
These phases can be further dissected into distinct biochemical steps, each of which can become a regulatory level. To understand how regulation occurs at the level of a gene, it is necessary to identify which steps represent the "rate-limiting steps " and analyze how activators and repressors act on them.

Regulation of transcription: eukaryotic model

Transcription of eukaryotic genes is controlled by proteins that bind to regulatory sequences, which can be located either near promoters or in distant enhancers. Both activators and repressors can regulate transcription at the level of formation of a preinitiation complex, e.g. by binding with the mediator of transcription complex who then binds to RNAPII and directly regulates assembly of transcription preinitiation complex. Eukaryotic transcription is regulated at *elongation* step as well as initiation steps, by direct modulation of RNAP activity and by effects on chromatin structure.
Several studies have shown that the recruitment of RNAPII and its stalling from 20 to 50 nucleotides downstream of the transcriptional start site (TSS) of genes are key steps for the transcriptional regulation. Then, the pause of the polymerase II downstream of the TSS is not only functional in determining the capping of nascent transcript, but represents a first fine mechanism for controlling the level of transcription of the genes [20]. The escape of paused Pol II into productive elongation is tightly regulated in relation to the response to an environmental stimulus or to embryonic differentiation, in which a very high response speed is required [21]. Transcription elongation is dependent on activity of P-TEFb and its recruitment represents the major regulatory step

in the control of this phase [22]. As mentioned above, in eukaryotic cells the gene expressions are also largely controlled by chromatin-regulating proteins. Eukaryotic DNA is packaged into chromatin and its basic structural unit is the nucleosome. It contains 147 bp of DNA wrapped tightly around a central octamer composed of two copies of each of the four core histones H2A, H2B, H3, and H4 [23]. The transcriptionally active regions tend to have few H1 histones and it is rich in the histone variants H3.3 and H2AZ. About 10% of the chromatin is found in a more condensed form and is transcriptionally inactive. In order for the RNA polymerase and TFs to bind the gene's regulatory regions, they must be accessible. The chromatin structural changes are generated by a process called chromatin remodeling. Acetylation and methylation of histones are part of this processes. During transcription, H3 is methylated at Lysine 4 at the 5 'end of the coding region and at lysine 36 present in the coding region. These methylations facilitate the binding of HAT (histone acetyltransferase), enzymes that acetylate lysine residues. Acetylation of the side chains of specific lysine residues is crucial for the interaction of nucleosomes with other proteins. The acetyl groups are negatively charged and neutralize the positively charged histones that slowly lose affinity for the negatively charged DNA, by relaxing chromatin structure. When the transcription of a gene is no longer required, the degree of acetylation of adjacent nucleosomes is reduced by the action of HDACs (histone deacetylase), returning chromatin to a transcriptionally inactive state.

1.1.3 Translation overview

Translation is the process by which polypeptide chains are produced using a molecule of mRNA as a template. Although the machinery complex of translation is highly conserved, several differences are known between prokaryotic and eukaryotic cells. In particular, differences have been detected in the signals that determine the positions at which synthesis of a polypeptide chain is started on an mRNA template [24]. mRNA is composed of two untranslated regions (UTRs) placed at the extremities of the mRNA and of a central region, the coding sequence (CDS), that contains the information for synthesizing the new protein (See Figure 1.6). Translation is performed by a ribosome, consisting of ribosomal RNA (rRNA) and a set of distinct ribosomal proteins, arranged in two ribosomal subunits: small (SSU) and large (LSU). The bacterial ribosome is composed of about 65% of rRNA and the remaining 35% of proteins. The size of bacterial ribosome is 70S with the small subunit (30S) is composed of the 16S rRNA and 21 ribosomal proteins (RPs) and the large subunit (50S) consisting of the 23S and 5S rRNAs and 34 RPs (See Figure 1.7).

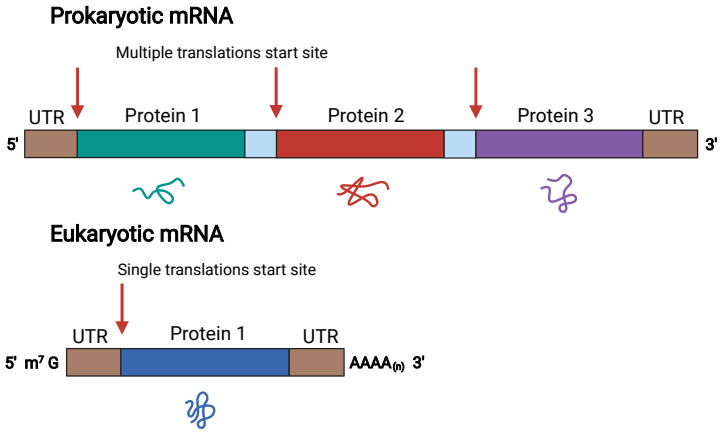


Figure 1.6: Prokaryotic and eukaryotic mRNAs.
Prokaryotic and eukaryotic mRNAs have untranslated regions (UTRs) at their 5'and 3'ends. Prokaryotic mRNAs are frequently polycistronic while eukaryotic mRNAs encodes a single protein. In addition, Eukaryotic mRNAs also contain 5'7-methylguanosine (m^7G) caps and 3'poly-A tails. Red arrows indicate translations start sites. (Created with Biorender).

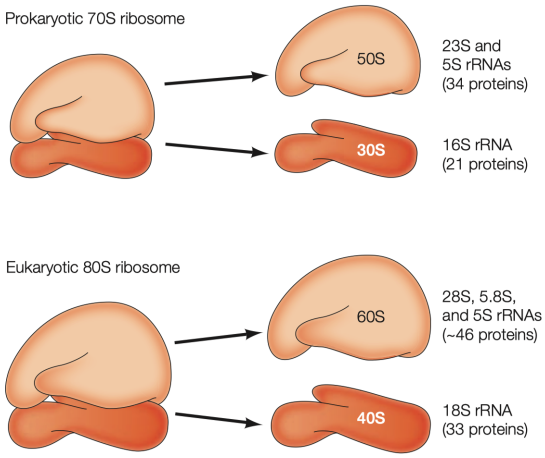


Figure 1.7: A comparison of the structures of prokaryotic and eukaryotic ribosomes.
Prokaryotic and eukaryotic ribosomes are commonly designated as 70S and 80S, respectively. The unit "S" stands for Svedberg, a coefficients measuring its sedimentation time during ultracentrifugation.

Differently, eukaryotic ribosome is a complex macromolecular machine formed of 4 rRNA species and 80 RPs. The mature ribosome is composed of the small

40S subunit, containing the 18S rRNA and 33 RPs and the large 60S subunit containing the 28S, 5.8S, and 5S rRNAs and 47 RPs (Figure 1.7).
The large subunit of the ribosome contains three active sites where the translation occurs, capable of binding tRNA molecules, carrying amino acids which will form a peptide. The aminoacyl-site (A-site) binds aminoacyl-tRNA (aa-tRNA, a tRNA with an amino acid attached to its 3 'end), the peptidyl bond between two amino acids is formed at the Peptidyl-site (P-site) while Exit site (E-site) binds free tRNA before it exits the ribosome. The peptide moves through the exit tunnel, which spans from the P-site to the cytoplasmic surface of the large subunit of the ribosome. Each site can accommodate a single tRNA. tRNA has two distinct ends, one of which binds to a specific amino acid, and the other which binds to the corresponding mRNA codon or triplets in the mRNA. The 3 'end of all tRNAs have the sequence CCA, and amino acids are covalently bound to the terminal adenosine by a family of enzymes namely aminoacyl tRNA synthetases. The relationship between triplets and amino acids is contained in the genetic code that is highly conserved among all organisms [25].
The translation process is composed of three main phases: initiation, elongation, termination. The initiation represents a rate-limited step of the translation process and will be discussed separately between prokaryotic and eukaryotic organisms, while the general steps of translation process, applicable for both prokaryotes and eukaryotes, are summarized in Figure 1.8.

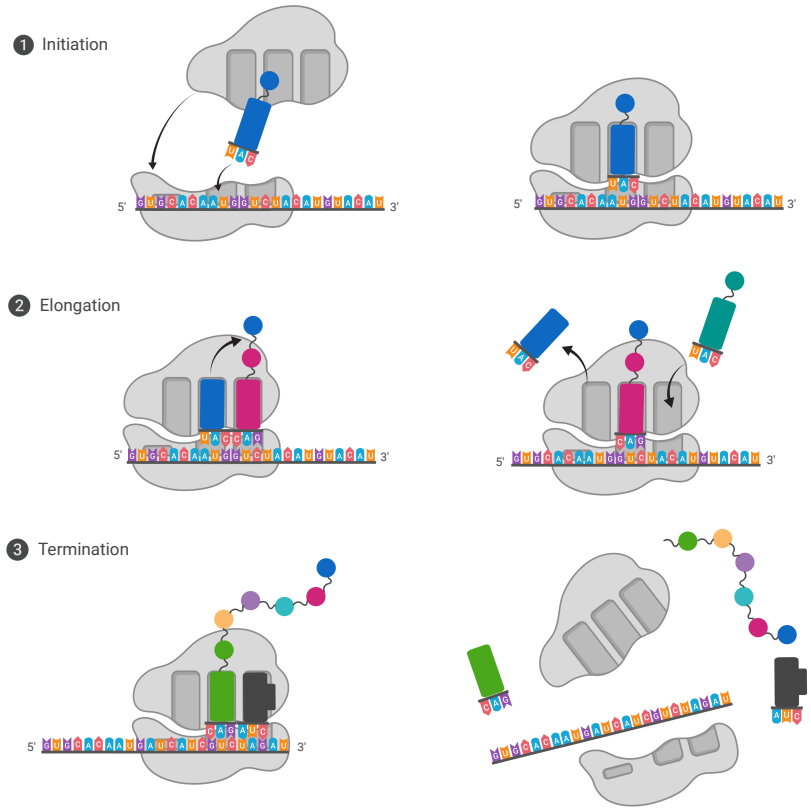


Figure 1.8: Phases of translation. 1) Initiation: mRNA binds to the small ribosome subunit. The ribosome moves along the mRNA in the 5 'to 3 'direction until it encounters the start codon. As the start codon is recognized, the ribosomal subunits are assembled together. 2) Elongation: a peptide bond is formed on the nascent chain in the P-site. The ribosomes then move one codon towards the 3 'direction. 3)Termination: Once the ribosome hits a stop codon, a release factor binds in the P-site. The polypeptide chain is released and the subunits of ribosome are disassembled.

Translation Initiation in Prokaryotes

In bacteria transcription and translation are coupled and a mRNA is being translate on ribosome before their transcription is complete, since no nuclear membrane separates these processes. In both prokaryotic and eukaryotic cells, the AUG start codon signal usually represents the start of translation. The selection of alternative start codons by the small ribosomal subunit can also occur in *E.coli* even though it causes a less efficient translation [26]. Initiation sites in prokaryotic mRNAs are characterized by the Shine-Dalgarno (SD)

sequence (5'AGGAG 3') 5-8 nt upstream from the initiation codon. It binds specifically to a complementary conserved sequence motif in the 3'end of the 16S rRNA, the so-called anti-Shine-Dalgarno sequence (3 'UCCUC 5') [27]. Bacterial ribosomes can start translation not only at the 5'end of an mRNA but also at the internal initiation sites of polycistronic messages, as depicted in Figure 1.6. The ribosomal small subunit complexes with three initiation factors (IF1, IF2 and IF3) and Met-tRNA, recruits the mRNA and recognizes the start codon, forming the 30S initiation complex. In particular, IF1 and IF3 act together to prevent premature association of 50S subunits. [28]. The start codon on the mRNA template is then recognized by the anticodon loop of tRNA. Once the complex joins the 50S ribosomal subunit, the IF2-GTP bound determines the hydrolysis of GTP to GDP and P, causing a conformational change in IF2 and the detachment of all three factors from the ribosome. Therefore, the correct binding of fMet-tRNA to the P site of the 70S initiation complex is ensured by the presence of at least three points of recognition: the codon-anticodon interaction; the interaction between the SD sequence and rRNA 16S; the binding interaction between the P site of the ribosome and fMet-tRNA. The result is the formation of a 70S initiation complex which is ready for the elongation phase of protein synthesis.

Translation Initiation in Eukaryotes

Before a mRNA is ready to be translated into a protein in eukaryotic cells, it has already been processed by (1) capping, a process involved in the attachment of 7-methylguanosine residue to the 5' terminal of the transcript (5'cap), (2) polyadenylation that allow the addition of poly-adenosine (Poly-A) tail to the 3'end of the transcript, (3) RNA splicing that refers to the removal of non-coding RNA introns and the joining of exons to form the mature mRNA and (4) optional modifications. The poly(A) tail is bound by multiple poly-A binding proteins (PABP), a protein family consisting of four RNA-recognition motifs (RRMs) and a C-terminal region containing a peptide binding region known as the PABC domain [29]. Besides to protect the mRNA from degradation by interacting with the poly (A), several studies have demonstrated that PABP proteins can interact through its PABC domain with other regulatory sequences and perform different functions inside the cell, e.g. control of mRNA stability, export, surveillance of transcripts, miRNA activity [30]. Interestingly, some eukaryotic and viral mRNAs (e.g. hepatitis C virus) have internal ribosome entry sites (IRESs) in which the translation can initiate independently of the 5 'cap, by direct engagement of the small subunit of ribosome [31]. In eukaryotes translation occurs in the cytoplasm, making this process uncoupled

from transcription. Eukaryotic initiation factors (eIFs) promote the binding of the mRNA and methionyl-tRNA to the small subunit (40S), giving rise to the pre-initiation complex 43S. In more detail, the 5 'cap of the mRNA is recognized by eIF4E. The pre-initation complex binds with a mature mRNA by addition of eIF4F and PABPs, resulting into the initiation complex 48S. During this phase, the small subunit scans the mRNA to identify the initiation codon. Once the 40S reaches the initiation start site, eIF5 triggers the hydrolysis of GTP bound to eIF2 inducing its release. Then, the 60S subunit joins the 48S complexes and 80S ribosome is assembled. After the initiation complex has formed, translation proceeds by elongation of the polypeptide chain.

Elongation

The elongation process involves the formation of a complex consisting of the aminoacyl-tRNA, *elongation factors* and GTPs. The ribosome reads the ORF moving towards its 3'end by three nucleotides at a time, adding at each step the correct amino acids to the nascent peptide chain. Then, it involves repetitive cycles of decoding, peptide bond formation, and translocation. The basic mechanism is very similar in bacteria and eukaryotes, it is facilitated by homologous elongation factors (EF-Tu, EF-G, EF-P, SelB for bacteria and eEF1α, eEF2, eIF5A, EFsec for eukaryotes). During the elongation phase, the correct amminoacyl-tRNA binds to the EF-Tu-GTP forming the so-called ternary complex. The resulting aminoacyl-tRNA-EF-Tu-GTP complex binds to the A site of the 70S initiation complex. This leads to a conformational change that induces hydrolysis of GTP bound to EF-Tu/eEF1α and release of elongation factor from the ribosome. Once EF-Tu/eEF1α has left the ribosome, a peptide bond is formed between the second aminoacyl-tRNA at the A site of initiator and methionyl tRNA at P-site, by peptidyl transferase activity. Generally, the selection of the correct aminoacyl tRNA for its incorporation into the growing polypeptide chain represents a crucial step to determines the efficiency of protein synthesis [32]. The final step of the elongation cycle is defined translocation. It allows the next codon to move into the decoding center. [33]. During translocation, the ribosome moves by one codon toward the 3 'end of the mRNA, positioning the next codon in an empty A site. This movement translocates the peptidyl tRNA to the P site and the uncharged tRNA to the E site, leaving an empty A site ready for addition of the next amino acid. Translocation is mediated by eEF2, coupled to GTP hydrolysis [34]. Finally, EF-G/eEF2 and the tRNA are released and another cycle can start. The speed of elongation phase is not uniform and multiple upstream factors can influence it (e.g. codon usage).

Termination

Elongation of the polypeptide chain continues until a stop codon (UAA, UAG, or UGA) is translocated into the A site of the ribosome. Then the *release factors* recognize these signals and terminate protein synthesis. In bacteria, two release factors termed RF1 and RF2 read UAG/UAA and UGA/UAA codons, respectively. A third factor, RF3, promotes turnover of the other two. In eukaryotes, all three stop codons are recognized by a single release factor namely eRF1 [35]. Subsequently, the ribosomes leaves the stop codon and may be recycled.
It is interestingly to highlighted that the process of translation is not limited to the conversion of mRNA into protein, it also regulates the effective composition of the proteame, in a coordinated and reactive way.

1.1.4 Translation control in prokaryotes

Once the mRNA is synthesized, its function can be modulated by a heterogeneous group of molecules called RNA regulators that also include small RNA (sRNA) and riboswitches. sRNAs are trans-acting factors that exert their regulatory function by binding to a specific sequence in the mRNA target inhibiting its translation. The majority of sRNAs regulate responses to environmental changes and a well-characterized example is represented by OxyS, oxidative stress response regulatory protein. OxyS functions to prevent the expression of unnecessary repair pathways by inhibiting the synthesis of rpoS (sigma factor of RNA polymerase) through antisense mechanism [36]. Transregulatory elements include RNA binding proteins (RPB) and non coding RNAs, increasing the complexity of regulatory mechanisms. Another class of RNA regulators is represented by riboswitches. Riboswitches are regulatory sequences encoded within the mRNA itself (usually at the 5' end of mRNA) that bind metabolites or metal ions and regulate mRNA expression by forming alternative structures in response to ligand binding. Since the regulatory sequence is encoded within the same mRNA molecule that also encode the gene which expression has been affected they are called cis-acting elements. The binding of a bio-switch to its specific ligand causes a conformational change in the mRNA and the inhibition of translation due to the stabilization of a premature termination structure [37].

1.1.5 Translation control in Eukaryotes

The latest findings underline the complexity of protein synthesis and show how translational regulatory mechanisms may acting both in cis through specific

sequences within the mRNA molecule, different usage of the codons and secondary stable structures formation and in trans through the binding of ncRNA and RNA binding proteins (RBPs) [38]. The initiation phase strongly depends on RBPs which play a key role in controlling various aspects of transcript fate and metabolism, including transcript degradation and its stability [39]. It has been shown that abnormalities in the expression of RBPs can promote malfunctions at the level of RNA stability in several types of diseases, including multiple sclerosis and cancer [40, 41]. In response to stress stimuli, eukaryotic cells activate an adaptive pathway termed *Integrated Stress Response pathway* (ISR) that inhibits pre-initiation complex formation. As evidenced in [42] all stress stimuli like amino acid deprivation and oncogene activation converge to eIF2α phosphorylation, leading to a decrease in global protein synthesis. Among the regulatory mechanisms described above, *codon usage bias* has been object of different studies due to its potential role in modulating gene expression levels in both eukaryotes and prokaryotes. The term *codon usage bias* refers to different frequency of synonymous codons which codify for the same aminoacid. Recently, the concept of "codon optimality" has been defined and allows to discriminate between optimal codons encoded with high speed, and non-optimal codons which are slowly translated [43]. Optimal codons are decoded by abundant tRNA and are found mainly in high expressed gene in *E.coli* and eukaryotes. Non-optimal codons are associated with secondary structures like as inter-domain linker regions and are recognized by less abundant tRNAs . The presence of rare codons represent an efficient mechanism able to stall the elongation phase, leading to the premature termination of translation [44]. In addition, codon usage can also influence splicing, and polyadenylation process [45]. A bioinformatics approach has been employed to investigate the role of codon usage in translation process. It reveals that optimal and non-optimal codons are clustered in different region of mRNA to optimize the translation efficiency. In particular, it has been evidenced that non optimal codons are widely used to slow translation in order to facilitate the correct folding of the proteins in regions where errors in co-translation folding are more costly. [46]. In addition, a recent study highlighted the effect of codon context on translation process, in *Salmonella enterics*. It has been demonstrates that the rate of translation of the UCA codon, encoding Serine, is also modulated by neighboring codon's position [43].

Although several studies have deepened the mechanisms by which codon usage affects translation speed, they remain still unclear and elusive [47].

The relative density of ribosomes and the speed at which they move along the mRNA template has been challenged by recent Ribosome profiling (Ribo-seq)

studies. The Ribo-seq technique will be described in the next chapter.

Translation deregulation during tumorigenesis

Translational control has a significant impact on eurkaryotic cellular functions and thus plays an important role in modulating the expression of many genes in response to stress conditions. Disregulation of translation control is implicated in a wide range of diseases, including cancer. Highly proliferating cancer cells required rapid and continuous protein synthesis that can be associated with 1) altered expression of genes encoding proto-oncogenes such as c-MYC, RAS, mTOR, 2) inhibition of tumor-suppressor genes such as TP53, PTEN, RB1 and 3) modification of translation initiation factors. Misregulation of translation initiation is the major contributing event in tumorigenesis [48]. The formation of the complex containing the initiation factors eIFs and the binding of the small ribosomal subunit to the 5'mRNA cap structure represents two crucial steps of regulation of translation. Normally, the eIF2 activity is regulated by a mechanism involving both guanine nucleotide exchange and phosphorylation. During stress condition, serine kinase proteins (PERK, PKR, HRI, and GCN2) phosphorylate eIF2α subunit at Serine 51. Guanine nucleotide exchange factor (GEF) eIF2B, can exchange GDP to GTP only when eIF2α is unphosporylated. Once phosporylated, eIF2α sequesters eIF2B and cannot return to their active state, thereby hindering the ternary complex.
Differently, a constitutively expression of elF2α is observed across different cancer types, including non-Hodgkins lymphomas [49]. At the same time, downregulation of elF2α kinase like eIF2α kinase heme-regulated inhibitor (HRI) can be implicated in tumori-genesis and promotion of cancer growth [50]. Other initiation factors are strongly involved in malignancy: For instance, high levels of eIF4E is an indicator of poor prognosis in luminal B breast cancers, suggesting that it could be a potential breast cancer biomarker and therapeutic target [51]. Its overexpression is associated with the upregulation of huge number of proto-oncogenes, like component of cell cycle machinery (c-Myc, Cyclin D1, CDK2), growth factors implicated in angiogenesis (VEGF, FGF-2, PDGF), and proteases involved in the process of tumor invasion (MMP-3 and MMP-9) [48] [52]. Although several studies have been done in this field, the responses caused by different regulatory mechanisms affecting mRNA are still unclear.
In the following chapter, we will focus on the Ribosome profiling approach, which allows assessing ribosome occupancy along the ORF, in order to investigate the translation status of different transcripts, how the translation is regulated, where it occurs, as well as the study of the roles of specific translation factors.

Chapter 2

Ribosome Profiling

2.1 Ribosome profiling

In this section, we present Ribosome Profiling (Ribo-seq) technique [53] which represents the most advanced tool able to exploits deep sequencing to study the translation of gene expression. The correlation between mRNA and protein levels is frequently poor due to the sophisticated regulation mechanisms of translation [54]. Ribo-seq approach investigates the translation status of different transcript at single nucleotide resolution, providing a global measurement of the translation in vivo. The general idea of this approach is based on the fact that each ribosome covers a short fragment around 28-30 nucleotides of translated mRNA [47]. Ribo-seq consists in the blockage of the translation process (elongation phase) in living cells, followed by nuclease digestion of the mRNA not covered by ribosomes. The remaining mRNA fragments, called Ribosome Protected Fragments (RPFs), are used to infer the ribosome's precise location. Deep sequencing are performed to characterizes the pool of RPFs and measure the abundance of different sequences. Once the sequences are processed and aligned to the reference transcriptome, Ribosome Profiles are generated and the number of reads that cover each nucleotide along the ORF is calculated. Aligning the sequenced reads back to the transcriptome produces a quantitative profile of ribosome occupancy. Therefore, Ribo-seq can reveal the composition and regulation of the expressed proteome by identifying transcripts undergoing active translation.

2.1.1 Protocol

A typical Ribo-seq experiment consists in the following steps (See Figure 2.1):
Lysis: Cells or tissue are mainly lysed using a lysis buffer containing Cycloheximide (CHX), which freezes ribosomes in the act of translation (other translational inhibitors are Harringtonine and Lactimidomycin [1].
Nuclease Footprinting: Nuclease digestion of the mRNA sequences unprotected by bound ribosomes is performed by using an endonuclease such as RNAse I. This process leaves ribosome intact and it is followed by recovery

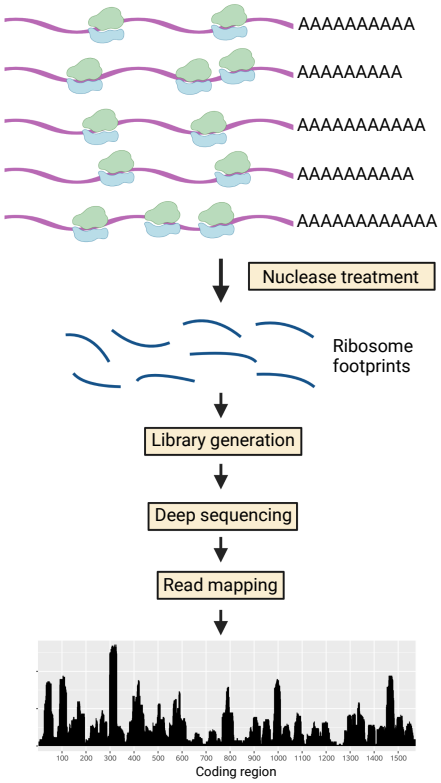


Figure 2.1: Ribosome Profiling overview. The positions of Ribosome on different mRNA templates are converted in ribosome footprints (RPFs) by nuclease treatment. These RPFs are converted into a DNA library and subjected to deep sequencing. RPFs typically cover coding DNA sequence. (Created with BioRender).

of the ribosome-protected fragments (RPFs).
Purification of protected fragment: RPFs can be purified using a sucrose density gradient or sucrose cushion ultracentrifugation.
rRNA depletion: Ribosomal RNA (rRNA) composes the majority of total RNA preparations. Different strategies can be applied to remove rRNA, including the use of specific beads with the RiboZero method. This protocol involves steps of washing and resuspention of magnetic beads that bind to removal probed hybridized to rRNA, allowing its removal. It produces an RNA sample ready for library preparation and for deep sequencing.
Size selection: Ribosomes leave 30 nt footprints when they are bound to mRNAs that can be extracted using PAGE (poly-acrylamide gel electrophoresis).
Library preparation and Sequencing: RPFs are recovered and converted into a DNA library. Reverse transcription (RT) and then PCR amplification

are performed. To add the adapter sequences, the 3'end of the fragments must be phosphorylated. After adapter ligation, cDNA libraries are analyzed by deep sequencing.
Computational analysis: The fragments are then mapped to the appropriate reference genome. Typically, the ribosome footprints show precise positioning between the start and the stop codon of a gene (coding sequence).
A modified protocol to isolate mitochondrial ribosomes has been also established [55]. The abundance of mtrRNA is extremely variable and it depends on cell type and stages of differentiation [56].

To globally measure translation in vivo at single nucleotide resolution, the original Ribo-seq protocol requires a high amount of RNA material, usually corresponding to tens of million cells. Recently, a new library construction strategy has been employed to generate mouse brain tissue data. It is based on skipping the adapter ligation step, using a much lower amount of input RNA material (1 ng of purified RNA footprints) [57]. Although the high amount of input data required and a lengthy protocol ($>$ 5 days), Ribo-seq experiments have been applied to different organisms, including bacteria, plants, viruses and human cells/tissues [1, 53, 58].
The application of this method to a different number of organisms subjected to different conditions, from deprivation of nutrients in bacterial cells to development of cancer in human cells, has allowed to investigate fundamental aspect of cell biology. Ribo-seq provides measurement for how the translation is regulated, what is being translated and where a specifics protein is translate. Several studies have highlighted the discovery of translation mechanisms and investigated the roles of specific translation factors, in different organism [59, 2]. For instance, the function of dom34 in yeast cells (a homologue of eukaryotic release factor 1) in freeing ribosomes from truncated transcripts in 3 'UTR [60]. Ribo-seq experiment has been adapted to identify non-canonical translation events, including upstream open reading frames (uORFs) which regulate the level of downstream protein coding genes [48]. In addition, Ribo-seq experiments have provided novel insights into molecular mechanisms of miRNAs.
For instance, they can affect mRNA abundance and translation of target genes by repression and inducing mRNA decay, as evidenced in [61]. Interestingly, Ribo-seq data reveal the density of the ribosomes at each position along the mRNA and events that can influence the translation dynamics (like as, e.g., tRNA modification, codon mutation) can be detected by producing a quantitative profile of ribosome occupancy. As mentioned above, the cells are treated with translation elongation inhibitors in order to capture the exact positions of the

ribosomes on the coding sequence.
A Ribo-seq experiment allows to reveal regions of higher and lower ribosome density along mRNA template. Local differences in the density of RPFs along the ORF reflect differences in the speed of translation elongation, determining regions where the translation is slower and faster (See Figure 2.2). This

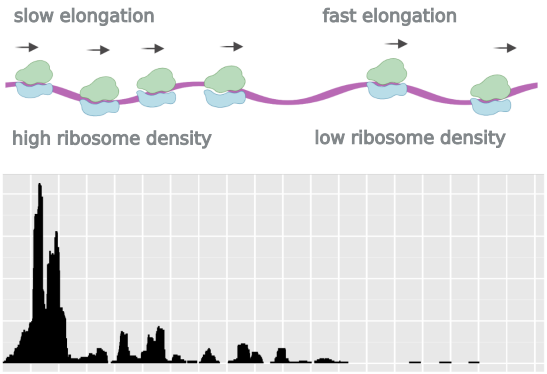


Figure 2.2: Ribosome footprint density along mRNA. The schematics distribution of translating ribosomes along mRNA (top) and their ribosome profiles (bottom). Ribo-seq data show differences in the density of ribosomes: regions of fast elongation accumulate fewer ribosomes (slow density) than the region of slow elongation (high density). (Created with BioRender).

schematic figure illustrates how the translation speed is not uniform, pointing out the differences in ribosome occupancy. This information are well visible in Ribo-seq profiling data and it can be used to infer how the codon usage, the protein sequences and other features can regulate the speed of translation [62].

Unfortunately, the reproducibility of Ribo-seq experiments can be affected by multiple variables due to the complexity of the experimental protocol and computational data analysis [63]. In particular, a critical issue is the choice of the translation elongation inhibitor and nuclease treatment. Indeed, it is known that an inhibitor such as cycloheximide can alter the local distribution of RPFs, causing spurious peaks near the initiation site. Nuclease such as RNAse I potentially compromises the stability of the ribosomal structure. Interestingly, it has been found that the use of this nuclease in Drosophila Melanogaster leads to a degradation of the ribosome while it not detected in budding yeast. Alternatives nucleases are considered, such as Nuclease S7 [64].
This thesis is inspired by the idea to exploit the full power of the Ribosome Profiling technique, trying to overcome the aforementioned limitations by introducing a newly designed statistical method.

In the following chapter, we will describe the fine-tuning of a novel data analysis approach for Ribo-seq data that will identify the reproducible Ribo-seq profiles. To this aim we will take advantages from publically available dataset of *E.coli* and we will compare different Ribo-seq datasets referring to experiments performed independently in different laboratories.

Chapter 3

Method

Our analysis of the ORF-specific Ribo-seq profiles consists of two phases, that were originally introduced in [3].

3.0.1 Upstream phase

The upstream phase allows us to compute the Ribo-seq profiles starting from the raw Ribo-seq data. The sequence for each read is provided in a fastaq file. The overall quality of our sequences is encoded in a Phred score (Q-score), which represents the estimated probability of an error, i.e. that the base is incorrect. The fastq data is filtered using CUTADAPT (release 1.8.3) in order to keep only high quality reads with Q-score ≥ 40. Moreover, the Ribo-seq protocol produces short RNA sequences and, the 3' adaptor sequence needs to be trimmed from the remaining reads, in order to obtain the exact footprinted RNA fragment. Furthermore, the reads which are shorter than 15 nucleotides are discarded to reduce the prevalence of multi-mapping errors. The reads originated from rRNAs and tRNAs sequences are identified and filtered out by aligning the reads to bacterial rRNA and tRNA sequences using Bowtie2 aligner (release 2.2.5) with no mismatch allowed [65]. To reconstruct the Ribo-seq profiles, the remaining reads are mapped against the whole set of coding sequences (CDS) in E. coli K12 MG, taken from the EnsemblBacteria database [66]. Among the reads that mapped on the reference ORFs we selected the ones those are mapped with the highest score possible for Bowtie2, using a mapping quality (MAPQ) equal to 42. To assess the quality of the reads before and after trimming we used fastqc tool [67]. After the mapping was complete, we get the alignments reported in the Sequence Alignment Map (SAM) format file, containing the genomic position where our reads are mapped and their mapping statistics, including mapping quality score which recorded in the fifth column. In our experiment, we extracted and counted the number of reads mapping to each gene/region from SAM/BAM alignment file using bedtools [68]. Therefore the genomics coordinates are stored in a BED file to build the Ribo-seq profiles representing the input of the subsequent analysis.

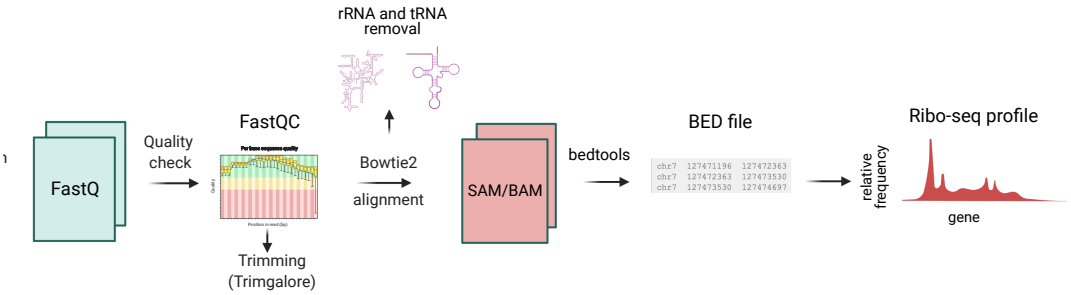


Figure 3.1: Bioinformatic pipeline used for processing Ribosome Profiling data in our study.

3.0.2 Downstream phase

The downstream phase is the most important step of our method because we can appraise the similarity and difference between Ribo-seq profile coming from different dataset with a statistical approach. All the procedures described are implemented through a custom script in the 'Python environment'.
Our method is articulated as follows:

Signal digitalisation strategy

To make the pairwise comparison of ORF-specific Ribo-seq profile coming from different datasets, we decided to proceed as follows. Each ORF can be associated to a specific Ribo-seq profile, an histogram that counts the number of reads that cover each nucleotide position. After calculating the median of the coverage values at each nucleotide, we assign +1 or -1 to the position having a coverage value higher or lower than the median, respectively. Each Ribo-seq profile is converted into the corresponding digitalised profile that is a vector of the length of the associated ORF, made by a sequence of -1 and +1. Figure 3.2 illustrates an example of Riboseq profile and the correspondent digitalised profile.

Comparison of the digital profiles

Digital profiles are used to quantify similarities and difference between riboseq profile of different dataset referring to the same ORF. A similarity score ($s_{i,k}$) is assigned to each pairwise comparison. It is computed by aligning each pair of digital profile and counting the number of times(n) that the same value (+1 or -1) appeared in the same position in both profiles and dividing the sum (n) by the length of the corresponding ORF. Mathematically, the similarity score can

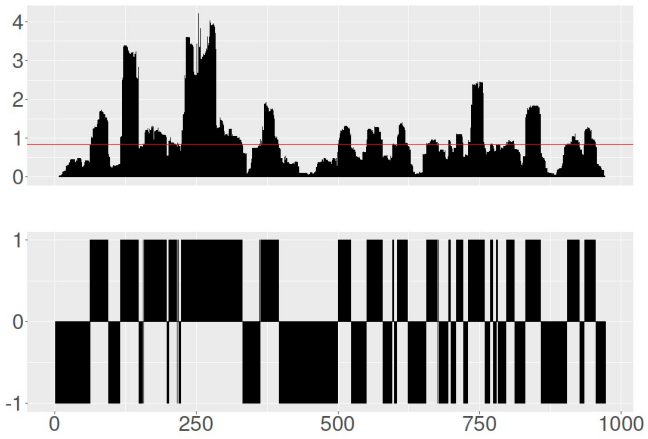


Figure 3.2: Example of a Ribo-seq profile (top figure) and the correspondent digitalised profile (bottom figure) for the gene ispB (EG10017) taken from E. coli dataset 1 (detailed in table 4.1). Red horizontal line (top): median of the Ribo-seq profile; y-axis (bottom): y-coordinate of the digitalised profile (+1: the corresponding coverage value is above the median; −1: the corresponding coverage value is below the median).

be between 0 and 1, but two random and independent profiles, in general, would give a score very close to 0.5. Figure 3.3 illustrates the pairwise comparison of two Ribo-seq profiles. Table 3.1 illustrates an example of a matrix similarity score. Similarity score has not statistically significant because each score has a certain probability of being obtained by chance. Then, we developed a method of two steps to give to similarity score a statistical significance.

	Dataset 1 vs Dataset 2	**Dataset 1 vs Dataset 3**	**...**
alr	0.5	0.6	...
modB	0.6	0.8	...
cysZ	0.7	0.5	...
dfp	0.5	0.7	...
fruB	1	0.6	...
...	...	...	...

Table 3.1: Illustrative example of a matrix similarity score. For the sake of readability, only two pair comparison between Dataset 1 and Dataset 2 are reported here.

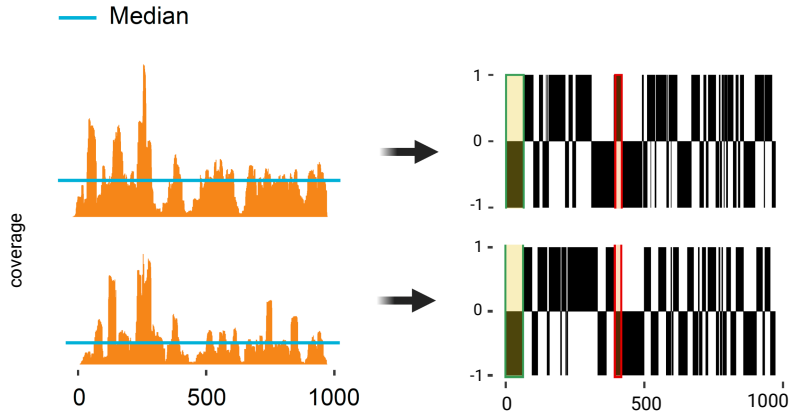


Figure 3.3: Pairwise comparison of two Ribo-seq profiles. Left: Two independent Ribo-seq profiles obtained by computing the coverage at each nucleotide position within the same ORF are compared to the median coverage to produce the digital ± 1 profiles on the right. Blue horizontal lines (left side): median coverage. Right: The digital profiles can be easily compared to detect matches (e.g. green rectangle) and mismatches (e.g. red rectangle). The ratio between the number of matches and the total number of nucleotides in the ORF gives the matching score.

A data-driven hypothesis test to assess the reproducibility of Ribo-seq profiles

Our strategy for assessing the significance of a given similarity score ($s_{i,k}$) consists of two steps:

- Construction of the null model:

Given a pair of Ribo-seq profiles referring to the same ORF and coming from two different experiments, e.g. the two profiles reported in Figure 3.2, the null model (H0) represents the distribution of the similarity scores as they would be if the matches and mismatches between the two profiles are due to randomness. To build such a distribution, we consider the two sets of reads those generated the Ribo-seq profiles in hand and we re-distribute them randomly on the respective ORF, thus generating a pair of random Ribo-seq profile. Starting from them and following the procedure used to create the ORF-specific digitalised profiles, it is then possible to compute a pair of digitalised random profiles. In turn, these profiles can be compared pairwise according to the method explained above, thus obtaining a random similarity

score.
Reiterating this process, we generated 10^4 pairs of random Ribo-seq profiles and an equal number of digitalised random profiles that, compared pairwise, yielded 10^4 random similarity scores. These scores are used to build a ORF-specific null distribution which allows us to estimate the probability of obtaining by chance each similarity score. It is worth to point out here that our way of building the null distribution through a data-driven random process allows us to formulate the null hypothesis taking solely into account the features of the data without any further hypotheses or approximations. The workflow to evaluate the significance of a given similarity score is shown in Figure 3.4.

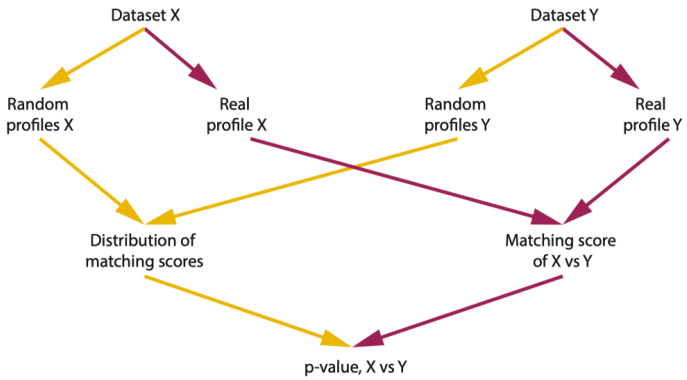


Figure 3.4: .
Workflow for the derivation of matching score significance. For each gene, two Ribo-seq profiles from independent dataset are digitalised and then compared to obtain the real matching score (purple arrows). The RPFs of each dataset are used to generate random profiles, which are digitalised and compared pairwise to obtain a distribution of matching scores (yellow arrows).

- Mapping the similarity score on the null distribution:

Given a pair of Ribo-seq profiles, the similarity score arising from their comparison is tested for significance by comparing it with the correspondent ORF-specific null distribution, as depicted in Figure 3.5. For each $s_{i,k}$ contained in the scores matrix and the corresponding null distribution, we computed a z-score $z_{i,k}$, mapping each similarity score on a standard normal distribution through the equation

$$z_{i,k} = \frac{s_{i,k} - \mu_{N_{i,k}}}{\sigma_{N_{i,k}}} \tag{3.1}$$

where $\mu_{Ni,k}$ and $\sigma_{Ni,k}$ are, respectively, the mean and standard deviation of the $N_{i,k}$ null distribution. Subsequently, we computed the p-value $p_{i,k}$, as the

integral:

$$p_{i,k} = \int_{z_{i,k}}^{+\infty} N_S(z)dz \tag{3.2}$$

where $N_S(z)$ is the standard normal distribution. The results of this process can be summarised into a matrix (call it p-values matrix, Table 3.2) containing all the computed p-values and composed by one column for each pairwise comparison and one row for each considered ORF. Each p i,k quantifies the probability of obtaining a similarity score at least as extreme as the corresponding $s_{i,k}$, given that the null hypothesis is true. In our context, the lower the p-value, the lower the probability that the similarity between the compared pairs of (digital) Ribo-seq profiles occur by chance.

	Dataset 1 vs Dataset 2	**Dataset 1 vs Dataset 3**	**Dataset 1 vs Dataset 4**	**...**
alr	0.769298564	0.122368427	0.632263895	...
modB	0.165522551	0.056591384	0.601754757	...
cysZ	0.005770742	0.00011569	0.2021111	...
dfp	0.002343099	0.000384015	0.093624025	...
fruB	0.566785395	0.85548442	0.381131384	...
...	...	...	...	...

Table 3.2: Representation of the p-value matrix. Each column corresponds to a pairwise comparison between two datasets while each row contains the gene ID. For the sake of readability, only three columns and 5 rows are reported here.

Identification of the significantly reproducible Ribo-seq profiles

Our strategy consists in inspecting each row of the p-values matrix. We define reproducible the Ribo-seq profiles referring to those rows featuring all the p-values below a chosen significance threshold. To cast our strategy into a more rigorous statistical framework, we exploited the False Discovery Rate (FDR) concept and the Benjamini-Hockberg (BH) method correction for multiple testing.

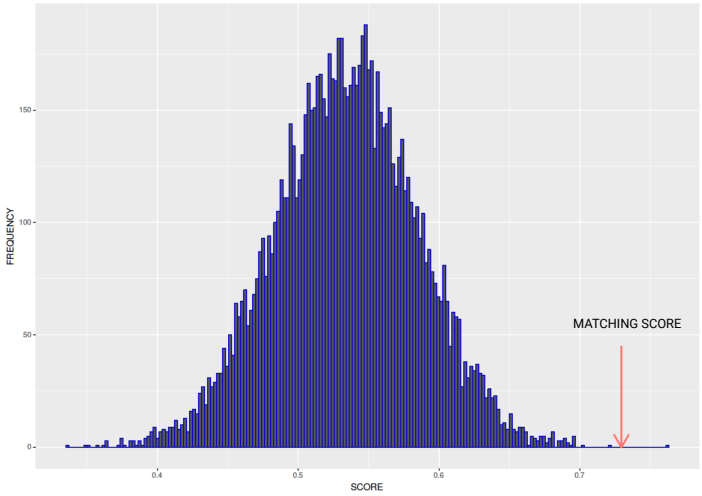


Figure 3.5: The null distribution for the gene ispB (EG10017) of *E.coli*. The red arrow indicates the value of the real matching score obtained comparing the Ribo-seq profiles.

Chapter 4

Analysing a broad-scale scenario: The E. coli case-study

To illustrate how our method works, we report here the analysis applied to E.coli's Ribo-seq profiles. To this aim, we relied on the data stored in the GEO repository [69]. The GEO coordinates for these datasets are reported in the fourth and fifth columns of Table 4.1.
Firstly, our analysis regarded a subset of nine samples, each belonging to a different series, that refer to experiments performed culturing wild-type *E.coli* strains under control conditions.
Specifically, this subset includes samples obtained through experiments characterised by K-12 MG1655 genotype and cultured in a MOPS-based medium.
Subsequently, in Chapter 7 we used the group of samples grown under normal conditions as a benchmark and then we compared it to datasets with different stress conditions like starvation or heat stress.
Table 4.1 summarizes the main features of the control series and the samples contained therein. More precisely, we compared the datasets under control conditions following our method previously described:

Dataset	Genotype	Culture's medium	GEO Series ID	GEO Sample ID	Ref
1			GSE64488	GSM1572266	[70]
2			GSE90056	GSM2396722	[71]
3			GSE72899	GSM1874188	[72]
4			GSE53767	GSM1300279	[73]
5	E.coli k-12 MG1655	MOPS, 0.2 % glucose	GSE51052	GSM1399615	[74]
6			GSE77617	GSM2055244	[75]
7			GSE35641	GSM872393	[76]
8			GSE88725	GSM2344796	[77]
9			GSE58637	GSM1415871	[78]

Table 4.1: The Samples chosen for our analysis belonging to different GEO Series. Column 1: ID Dataset. Column 2: Genotype. Column 3: Culture media. Columns 4 and 5: Samples coordinates (GEO Series ID and GEO Sample ID. Column 6: references.

Quantifying similarities between Ribo-seq profiles: a signal digitalisation strategy

Set up of the coverage matrix Firstly, we selected the 3534 ORFs in common between all the nine datasets highlighted in Table 4.1. For each ORF of each dataset, we generated a Ribo-seq profile which has collected into a matrix, named coverage matrix. In this case, each matrix is composed of 3534 rows which correspond to the genes in common between the dataset, and the number of columns corresponds to the length of the longest gene. In our case, the longest is yeeJ, a bacterial Ig-like protein (long 7077 bp). In more detail, each column corresponds to a single nucleotide position. Table 4.2 shows an example of coverage matrix.

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**...**
rodZ	0	1	2	3	4	8	45	46	47	47	...
arcB	0	0	0	3	3	3	3	6	6	6	...
dld	0	0	0	2	2	3	7	7	7	20	...
dnaX	8	8	8	10	10	12	12	16	18	18	...
fhuA	4	4	4	4	4	4	4	4	4	4	...
...	...	...	...	...	...	...	...	...	...	...	...

Table 4.2: Representation of the coverage matrix obtained from the Dataset 1. For the sake of readability, only ten columns (nucleotide positions) and five rows (genes ID) are reported here. The complete matrix is composed by 7077 columns and 3534 rows.

Elaboration of the digitalised profiles According to our method, each Ribo-seq profile is digitalised: the coverage is compared with the median value compute along the entire ORF. In this way, we obtained a vector containing a sequence of -1 and +1 for each ORF.

Comparison of the digitalised profiles Subsequently, we compared pairwise digital profiles in common between our subset of samples. We carried out 36 -$\binom{9}{2}$ comparison for each ORF. A matching score close to one reveals a high degree of similarity of pairwise comparison. Differently, a matching score around 0.5 could indicate those matches occurred by chance (random matches).

Assessment of similarity scores To check the statistical significance of the matching score and to avoid that it is obtained by chance, we generated a distribution of 10^4 random similarity scores, as previously explained. Each null distribution fits very closely a normal distribution. For each dataset, we computed the mean and standard deviations on 10^4 random coverage. Then, we mapped the real scores obtained by comparing the real coverage matrices

in the random distribution. The results of this process can be summarised into a summary matrix containing 1) computed mean, 2) standard deviation, 3) zscores and 4) pvalue for each pairwise comparison for the same ORF (See Table 4.3).

	Mean: Dataset 1 vs Dataset 2	**Std: Dataset 1 vs Dataset 2**	**Zscore: Dataset 1 vs Dataset 2**	**Pvalue: Dataset 1 vs Dataset 2**
rodZ	0.536424063	0.057725514	3.571721991	0.000177321
arcB	0.515414848	0.036041052	3.508046211	0.000225705
dld	0.520227622	0.041626933	2.25794901	0.011974419
dnaX	0.526314545	0.039817469	2.952919859	0.001573918
fhuA	0.513242112	0.035942448	3.92145903	4.40E-05
...	...	...	...	...

Table 4.3: Representation of the summary matrix. Each column correspond to a pairwise comparison between Dataset 1 and Dataset 2 (mean, standard deviation, z-score and p-value).

Identification of reproducible Ribo-seq profiles For any given row of the p-values matrix, we set an FDR threshold of 0.01. It means that we accept 1% of the reproducible profiles to be so by chance. Then, we counted in each row how many p-values resulted significant according to the BH method, and we defined reproducible those Ribo-seq profiles associated with the rows where 80% of the p-values are significant. Following this strategy, we found that out of 3534 genes that are in common to 9 datasets, the 25 genes listed in Table 4.4 have a significantly reproducible Ribo-seq profile.

Samples quality check

To identify the samples affected by unpredictable bias, we applied a jackknife approach on our dataset. Specifically, we repeated the entire reproducibility analysis nine times, excluding each time one of the control datasets. After this analysis, it turned out that when the GEO Sample ID GSM1415871 belonging to the Series GSE58637 is excluded, the number of reproducible Ribo-seq profiles raised from 25 to 40. We interpreted this result as originating from a peculiarity of this experiment. Indeed, it is interesting to note that the genes in common increase from 3534 to 3588, excluding the sample GSM1415871.
Thus, we decided to keep out this Sample from the subsequent analysis, and we considered the set of 40 genes as a benchmark. The list of reproducible genes

Genes ID	Annotation
rodZ	Cytoskeleton protein RodZ
dnaX	DNA polymerase III subunit tau
gltB	Glutamate synthase [NADPH] large chain
infB	Translation initiation factor IF-2
secY	Protein translocase subunit SecY
purL	Phosphoribosylformylglycinamidine synthase
rne	Ribonuclease E
sucA	2-oxoglutarate dehydrogenase E1 component
tufA	Elongation factor Tu 1
tufB	Elongation factor Tu 2
hokB	Toxic component of a type I toxin-antitoxin (TA) system
ubiJ	Ubiquinone biosynthesis protein UbiJ
lptD	LPS-assembly protein LptD
rpnC	Recombination-promoting nuclease RpnC
rpnA	Recombination-promoting nuclease RpnA
fdoG	Formate dehydrogenase-O major subunit
wbbH	O-antigen polymerase
wbbI	Beta-1,6-galactofuranosyltransferase WbbI
rpnE	Inactive recombination-promoting nuclease-like protein RpnE
lpoA	Penicillin-binding protein activator LpoA
intR	Putative transposase
rlmL	Ribosomal RNA large subunit methyltransferase K/L
rsxC	Electron transport complex subunit RsxC
yfcI	Recombination-promoting nuclease RpnB
gtrS	Uncharacterized protein YfdI; Putative ligase

Table 4.4: Genes with significantly reproducible Ribo-seq profiles using the nine dataset listed in table 4.1. Column 1: Genes ID. Column 2: Annotation

across eight datasets is presented in Table. 4.5. For one of these genes, namely ompC (EG10670), we show its profiles across all dataset as an illustrative example (See Figure 4.1).
OmpC, also known as outer membrane (OM) protein C, is a porin of gram-negative bacteria tightly associated with the peptidoglycan layer. It has been recognized to have a crucial role in the non-specific diffusion of small solutes such as sugars, ions and amino acids across the outer membrane or the cell [79].

The analysis of the control dataset confirmed the poor reproducibility of Ribo-seq: only 40 genes of 3588 profiles could be defined as reproducible when eight different datasets are considered. However, these analyses provide a small but reliable reference set that can be used as a benchmark for comparative studies, as we will illustrate in Chapter 7.

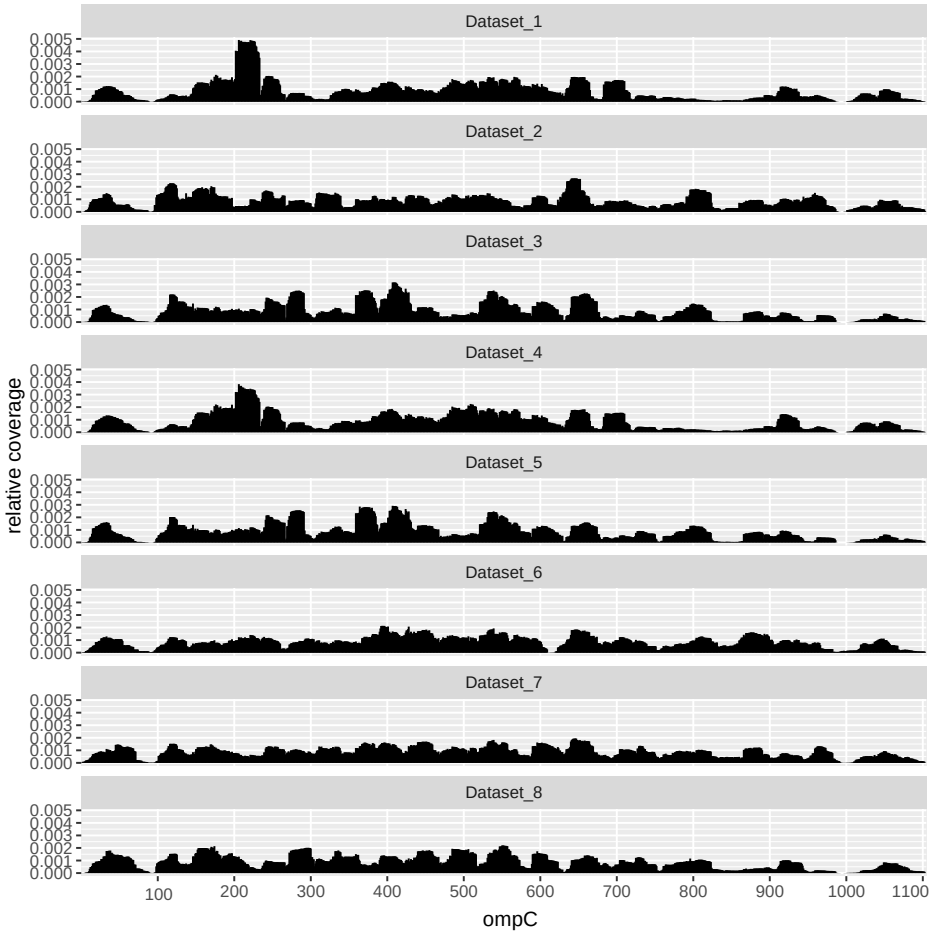


Figure 4.1: Illustrative example of a significantly reproducible Ribo-seq profile (gene ompC, EG10670)

Although is beyond the scope of this dissertation, we decided to deepen the role and the features of the genes corresponding to these reproducible Ribo-seq profiles. We investigated the functional implication of the 40 reproducible genes and classified them by gene ontology categories (i.e. molecular function and biological processes) using PANTHER gene Classification System [80]. "Catalytic activity" (GO:0003824) ranked the top in the molecular function ontology, with 18 genes. In the biological process ontology, cellular processes (GO:0009987) and metabolic processes (GO:0008152) possess the top two represented GO categories, with 22 and 21 genes, respectively.
Over-representation test and pathway analysis are performed but not significant results have been obtained. This could be related to the low number of input genes. Future works will be devoted to repeat the entire pipeline previously

described, choosing a less conservative false discovery rate. It could represents an immediate solution and highly alleviate this problem, increasing the number of reproducible Ribo-seq profile.

To highlight which specific regions within the Riboseq profiles are similar to each other we built a consensus sequence.
The consensus sequence is a character string representing the nucleotides of the reference ORF: In red are reported those position where a peak is present and the ribosome proceeds slower while in green where a valley is located and the ribosome proceeds faster (See Figure 4.2 below for a graphic representation of the consensus sequence).

...AGTGCGGTTATCCCGGCTGTCGCCCCTACGCGGAAGCCAA...

Figure 4.2: Part of a consensus sequence with fast and slow translation regions. The nucleotides located within fast regions are depicted in green while those located in the slow region in red.

The consensus sequences built from the 40 reproducible Ribo-seq profiles constitute the dataset on which the following statistic analysis and machine learning approaches are conducted.
In this dissertation, we focused on the labelled portions of the consensus sequences +1 and -1, i.e. the subsequences in which the ribosome proceeds at speed fast and slow, respectively.

Genes ID	Annotation
rodZ	Cytoskeleton protein RodZ
arcB	Aerobic respiration control sensor protein ArcB
dld	Quinone-dependent D-lactate dehydrogenase
dnaX	DNA polymerase III subunit tau
fhuA	Ferrichrome outer membrane transporter/phage receptor
glnA	Glutamine synthetase
gltB	Glutamate synthase NADPH large chain
hisS	Histidine–tRNA ligase
infB	Translation initiation factor IF-2
katG	Catalase-peroxidase
malF	Maltose transport system permease protein MalF
metG	Methionine–tRNA ligase
mukB	Chromosome partition protein MukB
ompC	Outer membrane protein C
parC	DNA topoisomerase 4 subunit A
secY	Protein translocase subunit SecY
purL	Phosphoribosylformylglycinamidine synthase
rne	Ribonuclease E
sucA	2-oxoglutarate dehydrogenase E1 component
tufA	Elongation factor Tu 1
tufB	Elongation factor Tu 2
leuA	2-isopropylmalate synthase
hokB	Toxin HokB; Toxic component of a type I toxin-antitoxin (TA) system.
acnA	Aconitate hydratase A
ubiJ	Ubiquinone biosynthesis protein UbiJ
lptD	LPS-assembly protein LptD
rpnC	Recombination-promoting nuclease RpnC
rpnA	Recombination-promoting nuclease RpnA
fdoG	Formate dehydrogenase-O major subunit
wbbH	O-antigen polymerase
wbbI	Beta-1,6-galactofuranosyltransferase WbbI
wbbK	Putative glycosyltransferase WbbK
rpnE	Inactive recombination-promoting nuclease-like protein RpnE
lpoA	Penicillin-binding protein activator LpoA
gspD	Putative type II secretion system protein D
yfjI	Uncharacterized protein YfjI; Phage or Prophage Related
rlmL	Ribosomal RNA large subunit methyltransferase K/L
rsxC	Electron transport complex subunit RsxC
yfcI	Recombination-promoting nuclease RpnB
gtrS	Uncharacterized protein YfdI; Putative ligase

Table 4.5: Genes with significantly reproducible Ribo-seq profiles after excluding the dataset GSM1415871. Column 1: Genes ID. Column 2: Annotation.

Chapter 5

Statistical data analysis

In this Chapter, the thesis will explore the composition of the consensus sequences, relating to the assigned labels: -1 for fast speed and $+1$ for slow translation speed. In general, descriptive statistics can be helpful to provide basic information about variables in our benchmark and highlight the potential relationships between variables (e.g. if exist a correlation between a high or low frequency of a specific nucleotide in one of the classes under investigation). Based on results obtained with our method, we defined the elements of consensus sequences labelled with -1 as fast subsequences, while the elements labelled with $+1$ as slow subsequences. In the following, we will briefly describe the statistical analyses performed in the dissertation.

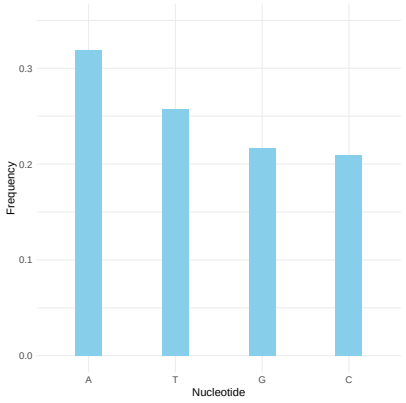
Descriptive analysis

Firstly, we computed the relative frequency for each nucleotide (A, T, G, C). Then, relative frequencies are also computed for all possible couples of nucleotides: the number of occurrences is calculated for each dinucleotide along the subsequences and then normalized by the total number of base pairs. Once the relative frequency for each nucleotide and base pairs are analyzed, we focused on the frequency distribution of all codons across our subsequences. As explained in Chapter 2, the codon usage bias refers to the concept that different organisms have divergences in the frequency of occurrence of the synonymous codons during translation. According to [81, 82, 83], it is widely assumed that codon choice has strong effects on protein expression in organism from E.coli to more complex like human. More recently, codon optimality has been shown to be an important regulatory mechanism involved in the kinetics of protein synthesis[84]. Based on this evidence, once the relative frequency of the four nucleotides is calculated we decide to explore the distribution of 64 codons which encode a pool of 20 amino acid and translation stop signal. Two distinct analyses are carried out: Firstly, only triplets containing all nucleotides with label -1/+1 are included in the computation; Subsequently, a "tolerance" is introduced and we considered fast/slow translated codons only those containing at least two nucleotides labelled with -1/+1.

In the following, we report the frequency analysis results obtained using fast and slow subsequences.

5.0.1 Analysis of "fast subsequences"

First, we report the statistical methods employed on fast subsequences, labelled with -1. To start the analysis, we computed the relative frequency for each nucleotide (A, T, G, C). In Figure 5.1, the nucleotide frequency is displayed along the bottom of the barplot. The table with frequency values for each nucleotide is shown in the right panel of Figure 5.1. As we can observe, adenine has a greater frequency than other nucleotides across the fast subsequences, with a relative frequency of approximately 0.32 while, cytosine has the lowest frequency (0.2).



Nucleotide	Frequency
A	0.318326031
T	0.25683638
G	0.21604472
C	0.208792869

Figure 5.1: Nucleotide relative frequency across fast subsequences

In Table 5.1 and in Figure 5.2 we report the number of occurrences calculated for each dinucleotide along the subsequences labelled with -1. As we note, the frequency distribution of the dinucleotides reflects the nucleotide frequency: the base pair AA has the highest frequency (0.12439954), while the GG and CC dinucleotides show the lowest relative frequencies (0.03014742 and 0.03693888, respectively). As evidenced in multiple studies [85] [86], an mRNA rich in A-residues at the beginning of the sequence can promote the translation initiation due to its unstructured domain in a wide variety of bacterial species, including *E.coli.* [85]. The results of the analysis of codon relative frequency are presented below 5.3. In Table (5.2), the first column displays the amino-acid encoded, the second lists the triplet which encodes the specific amino-acid, and the last two columns correspond to the codon relative frequency (computed

Dinucleotide	Frequency
AA	0.12439954
AT	0.0722213
AG	0.05565678
AC	0.05880404
TA	0.0601292
TT	0.07520292
TG	0.07172437
TC	0.05416598
GA	0.06609243
GT	0.05565678
GG	0.03014742
GC	0.06410469
CA	0.06178566
CT	0.05052178
CG	0.06244824
CC	0.03693888

Table 5.1: Dinucleotide relative frequency across fast subsequences

with and without tolerance, as explained above). The triplets are grouped based on the amino-acid encoded in order to observe which synonymous codons are used in the fast translated subsequences. Interestingly, the triplet AGG results absent across the fast subsequences. In *E.coli*, AGG is a rare arginine codon which occurs at a frequency of 0.14% [87]. It has been demonstrated that the triplet AGG can negatively influence the translation process by reducing protein synthesis [88].
Based on codon usage bias and the nucleotide relative frequency in *E.coli* genes [89], we can state that our fast subsequences have a similar frequency value. For example, codons with the highest frequency detailed in Table 5.2 correspond to the optimal codon chosen during the translation process.

AA	Codon	Frequency	Frequency (Tol)
Ala	GCT	0.02530964	0.022065728
Ala	GCC	0.02638665	0.023474178
Ala	GCA	0.01507808	0.015492958
Ala	GCG	0.02423263	0.022065728
Arg	AGA	0.00323102	0.002816901
Arg	AGG	0	0
Arg	CGG	0.00215401	0.001877934
Arg	CGA	0.00107701	0.000938967
Arg	CGC	0.01992461	0.017840376
Arg	CGT	0.03231018	0.033333333
Asn	AAT	0.02907916	0.030985915
Asn	AAC	0.0360797	0.03943662
Asp	GAT	0.03931072	0.038967136
Asp	GAC	0.01938611	0.019248826
Cys	TGT	0.00430802	0.003755869
Cys	TGC	0.00215401	0.002347418
Gln	CAA	0.02692515	0.027699531
Gln	CAG	0.03231018	0.030985915
Glu	GAA	0.06031233	0.062441315
Glu	GAG	0.01669359	0.015962441
Gly	GGT	0.01184707	0.010798122
Gly	GGC	0.01077006	0.009389671
Gly	GGA	0.00107701	0.001408451
Gly	GGG	0.00161551	0.001408451
His	CAT	0.00861605	0.010328638
His	CAC	0.00646204	0.006103286
Ile	ATA	0.00484653	0.004694836
Ile	ATT	0.02154012	0.021596244
Ile	ATC	0.02261712	0.022065728
Leu	TTA	0.01238557	0.014084507
Leu	TTG	0.01400108	0.015023474
Leu	CTT	0.00700054	0.007511737
Leu	CTC	0.00969305	0.008920188
Leu	CTA	0.00323102	0.003755869
Leu	CTG	0.04684976	0.042723005
Lys	AAA	0.07162089	0.076525822
Lys	AAG	0.01615509	0.015492958
Met	ATG	0.03177167	0.030046948
Phe	TTT	0.02746365	0.030046948
Phe	TTC	0.03446419	0.032394366
Pro	CCC	0.00323102	0.002816901
Pro	CCA	0.00538503	0.005633803
Pro	CCG	0.01507808	0.01314554
Pro	CCT	0.00376952	0.003755869
Ser	AGT	0.00538503	0.005633803
Ser	TCT	0.00969305	0.010798122
Ser	TCC	0.00484653	0.004694836
Ser	TCA	0.00538503	0.005164319
Ser	TCG	0.00323102	0.002816901
Ser	AGC	0.01023156	0.010798122
Thr	ACT	0.01023156	0.012676056
Thr	ACC	0.01292407	0.011737089
Thr	ACA	0.00376952	0.003286385
Thr	ACG	0.00430802	0.003755869
Trp	TGG	0.00538503	0.005164319
Tyr	TAT	0.02423263	0.029107981
Tyr	TAC	0.01507808	0.017370892
Val	GTT	0.02530964	0.025821596
Val	GTC	0.01561659	0.013615023
Val	GTA	0.01184707	0.013615023
Val	GTG	0.01184707	0.010798122
Stop	TAA	0.00969305	0.008920188
Stop	TAG	0	0
Stop	TGA	0.00323102	0.002816901

Table 5.2: Codon relative frequency (fast subsequences)

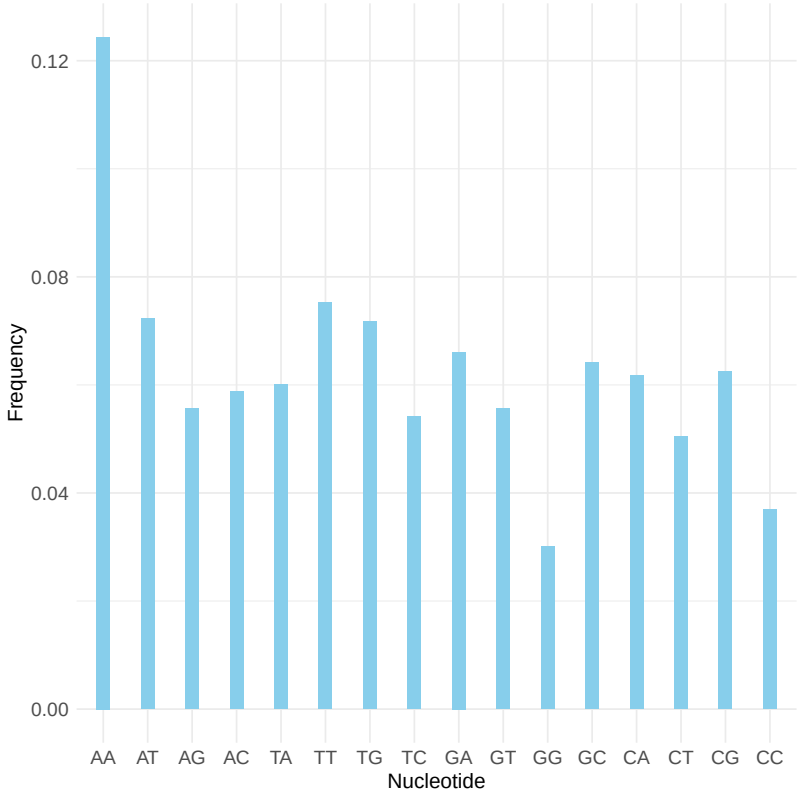


Figure 5.2: Relative frequency histogram of base pairs in fast sequences

5.0.2 Analysis of "slow subsequences"

We applied the same procedure of the fast subsequences analysis on the slow subsequences, labelled with +1. Figure 5.4 shows the frequency of each nucleotide. The table to the right panel reports the frequency values of each nucleotide. The number of occurrences calculated for each dinucleotide along subsequences with +1 are listed in Table 5.3 and shown in Figure 5.5. Looking at the data, both guanine and cytosine have the highest relative nucleotide frequency across the fast subsequences. In contrast to fast subsequences, the most frequent base pair in slow subsequences turns out GC, with a relative frequency of 0.088507266. In contrast to fast subsequences, the rare codon AGG is present and has a frequency of 0.00460678. According to the literature [90], non optimal codons are usually associated with protein domain linker regions along the mRNA sequence, where the ribosome tends to move slowly. Recent evidence suggests that these regions could be related to co-translational folding [91]. In histogram 5.6 both the relative frequencies of all fast codons and those of the triplets with at least two fast nucleotides have been reported. Table

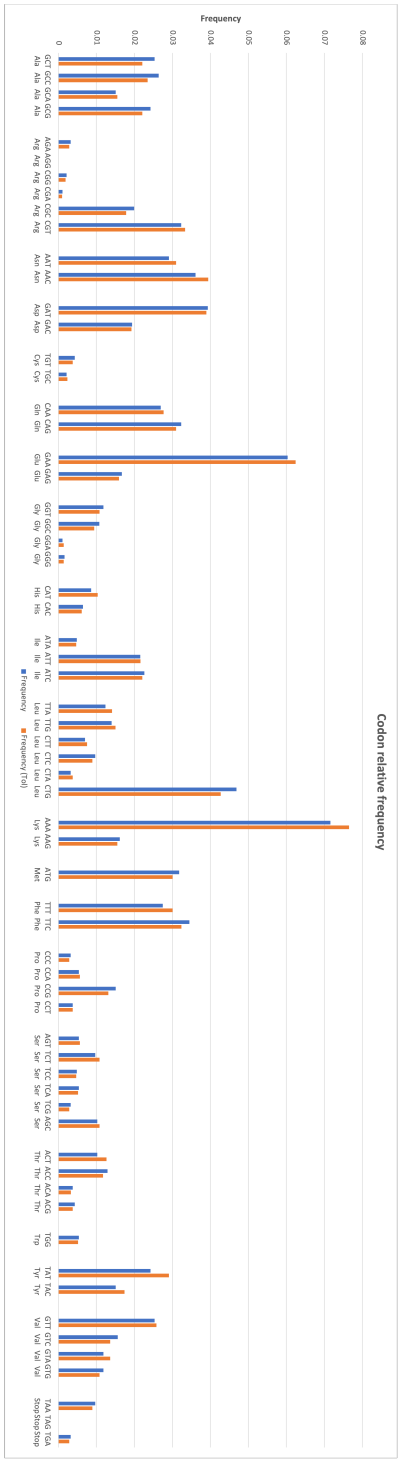
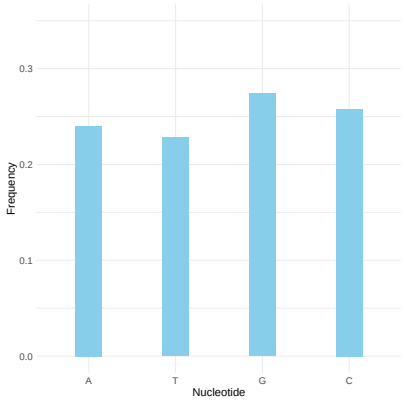


Figure 5.3: Relative frequency histogram of all 64 codons (fast subequences)



Nucleotide	Frequency
A	0.240073492
T	0.228284169
G	0.273706237
C	0.257936103

Figure 5.4: Nucleotide frequency across slow subsequences (label +1)

Dinucleotide	Frequency
AA	0.064464993
AT	0.056169089
AG	0.048929987
AC	0.062826948
TA	0.038097754
TT	0.055852048
TG	0.077886394
TC	0.053579921
GA	0.063989432
GT	0.055270806
GG	0.071228534
GC	0.088507266
CA	0.06002642
CT	0.057807133
CG	0.084174373
CC	0.061188904

Table 5.3: Dinucleotide relative frequency value across slow subsequences (label +1)

5.6 shows the codon frequency relative to all 64 codons in slow subsequences.

AA	Codon	Frequency	Frequency (Tol)
Ala	GCT	0.01908523	0.01813748
Ala	GCC	0.02599539	0.02438653
Ala	GCA	0.01842711	0.01813748
Ala	GCG	0.02862784	0.02743484
Arg	AGA	0.00460678	0.00426764
Arg	AGG	0.00246792	0.00228624
Arg	CGG	0.00361961	0.00335315
Arg	CGA	0.00279697	0.00259107
Arg	CGC	0.02286936	0.02179546
Arg	CGT	0.02221125	0.02149063
Asn	AAT	0.01135242	0.01082152
Asn	AAC	0.01694636	0.01981405
Asp	GAT	0.0266535	0.03200732
Asp	GAC	0.02435012	0.02834934
Cys	TGT	0.00477131	0.00472489
Cys	TGC	0.00723922	0.00685871
Gln	CAA	0.01233959	0.01204085
Gln	CAG	0.02879237	0.03139765
Glu	GAA	0.03948667	0.04039018
Glu	GAG	0.02105956	0.02210029
Gly	GGT	0.03208292	0.03094041
Gly	GGC	0.04162553	0.03993294
Gly	GGA	0.00806186	0.00746837
Gly	GGG	0.01036525	0.00960219
His	CAT	0.00839092	0.00838287
His	CAC	0.01135242	0.01326017
Ile	ATA	0.00691017	0.00670629
Ile	ATT	0.02566634	0.02560585
Ile	ATC	0.02648898	0.02819692
Leu	TTG	0.00987167	0.00960219
Leu	TTA	0.0088845	0.00929736
Leu	CTT	0.01233959	0.01204085
Leu	CTC	0.01020072	0.01036427
Leu	CTA	0.00361961	0.00365798
Leu	CTG	0.05100362	0.0501448
Lys	AAA	0.02862784	0.02773967
Lys	AAG	0.01135242	0.01112635
Met	ATG	0.0266535	0.02621552
Phe	TTT	0.01793353	0.01828989
Phe	TTC	0.0159592	0.01569883
Pro	CCC	0.00510036	0.00472489
Pro	CCA	0.00839092	0.00777321
Pro	CCG	0.03191839	0.02987349
Pro	CCT	0.00871997	0.00807804
Ser	AGT	0.00789733	0.00731596
Ser	TCT	0.01135242	0.01066911
Ser	TCC	0.01316222	0.01280293
Ser	TCA	0.00839092	0.00792562
Ser	TCG	0.00839092	0.00777321
Ser	AGC	0.01497203	0.01386984
Thr	ACT	0.01118789	0.01082152
Thr	ACC	0.03339914	0.03109282
Thr	ACA	0.00674564	0.00640146
Thr	ACG	0.01497203	0.01402225
Trp	TGG	0.01382034	0.01295534
Tyr	TAT	0.01135242	0.01234568
Tyr	TAC	0.01447845	0.01691815
Val	GTT	0.01826259	0.01874714
Val	GTC	0.01135242	0.01204085
Val	GTA	0.01118789	0.01249809
Val	GTG	0.02221125	0.0231672
Stop	TAA	0.00131622	0.00121933
Stop	TAG	0	0
Stop	TGA	0.00032906	0.00030483

Table 5.4: Relative frequency value of all 64 codons (slow subsequences)

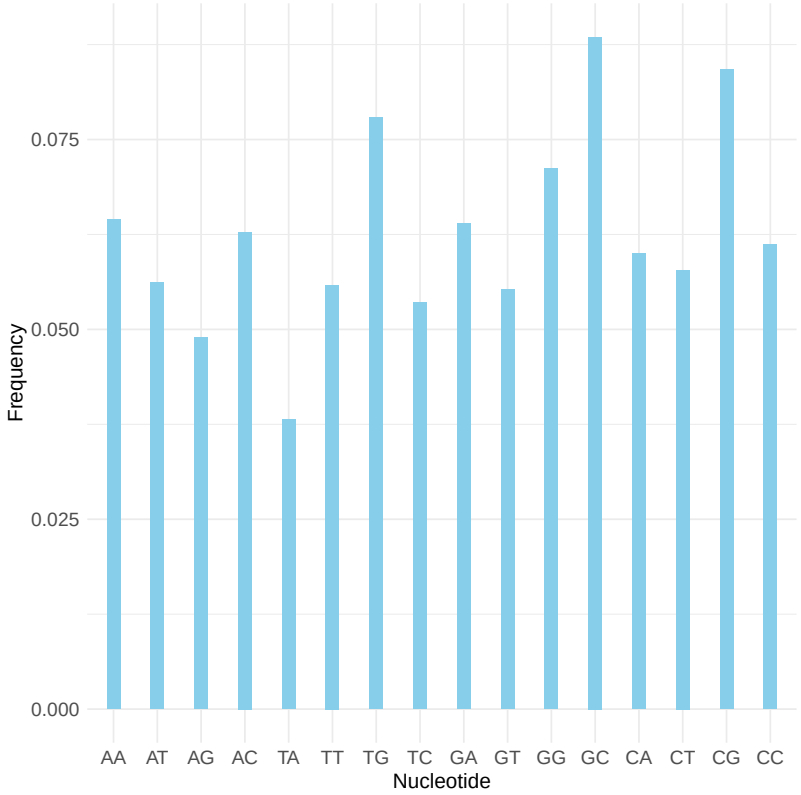


Figure 5.5: Relative frequencies histogram of base pairs in slow sequences (label +1)

Test of Significance of the Relative Frequency of Nucleotides

To assess the significance of our results and to find out if the subsequences obtained are specific characteristics of fast and slow subsequences or simply related to the chance, we built a statistical test.
This significance test is developed exclusively for the relative frequency at the nucleotide level. The null hypothesis is the following: the slow or fast subsequences are characterized by nucleotide composition equal to which one would occur if the nucleotides are arranged in random way.
The aim of the test is to assign a probability value (i.e. p-value) to the null hypothesis. Therefore, we build a random distribution of relative frequency for each nucleotide in order to test the null hypothesis of the randomness of the frequencies: 10^4 profiles are generated, each representing a dataset random created from the original dataset. Then, we used them to build a null distribution which allowed us to estimate the probability of obtaining by chance each nucleotide frequency.

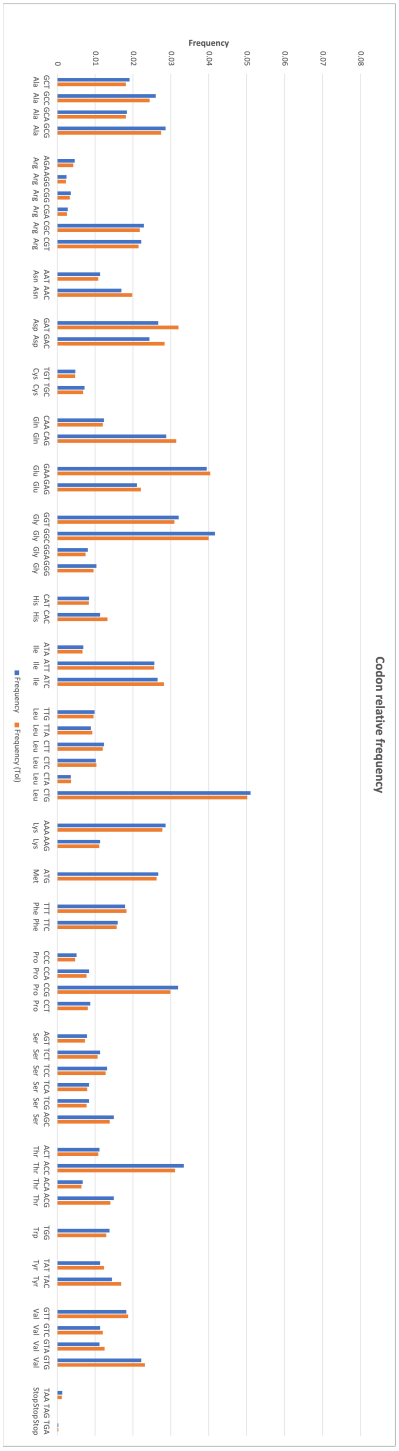


Figure 5.6: Relative frequency histogram of all 64 codons (slow sequences)

We performed random permutation inside each subsequence keeping the consecutive labels with all values equal to -1 or +1. In the randomization process, the length of the subsequences is kept fixed, as shown in Figure 5.7. The relative frequencies of the nucleotides are calculated on each dataset

Figure 5.7: Example of a sequence fragment with the labels of the original dataset (above) and the random one (below). In this case, we report the random permutation of slow subsequences (red nucleotides). The length of the red nucleotide string is kept fixed during randomization

and the results are represented by four histograms which correspond to the frequency distribution of each nucleotide along the random dataset (See Figure 5.8).
Once the null distributions has been achieved, we compared them with the frequency values computed in the original datasets.
Intuitively, the data is significant depending on how far it is from the mean of null distribution: if it is centered, then it is completely random while if it is found in the tails it is significant. The p-value is computed based on the frequency of the original position compared to the mean distribution. If the frequency is greater than the mean, we calculate the probability that a value greater than or equal to the real value can be found by chance. If this probability is under a specific cut-off (0.05), it attests that the subsequences have a nucleotide frequency greater than can occur simply by chance. Conversely, if the original frequency is lower than the mean, we calculate the probability that a value lower than or equal to the real value can be found by chance. In the case of rejection of the null hypothesis, it is established that the fast/slow subsequences are characterized by a nucleotide frequency lower than expected to find by chance.
Once we applied this statistical method to the relative frequency of slow and fast data, we obtained that all four nucleotides have p-values equal to 0. In our context, the lower the p-value represents the lower probability that the nucleotide frequency across subsequences fast and slow occurs by chance.
Therefore, we can assert that the subsequences with both label -1 and +1 are not characterized by a random distribution of the nucleotides.

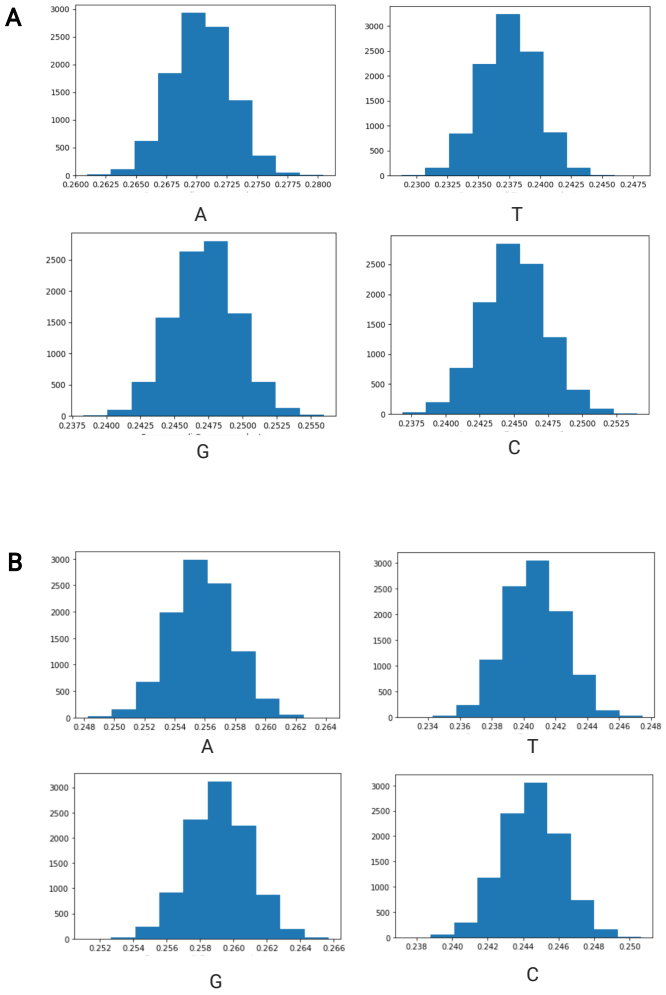


Figure 5.8: A) Null distribution of four nucleotides across fast subsequences with label -1. B) Null distribution of four nucleotides across slow subsequences with label +1. Each set of random profiles is computed distributing randomly the labels coming from fast subsequences (-1) and slow subsequences (+1).

5.0.3 Discussion

In conclusion, we can affirm that the nucleotide frequencies observed are statistically significant: it is unlikely to find them by chance. Based on the results obtained, we can state the frequency distributions are notably different for each nucleotide and dinucleotide across the subsequences, slow and fast. In particular, we can observe that A and T nucleotides have a higher frequency in

fast sequences than G and C. Differently, the frequency of G and C nucleotides is significantly higher in slow sequences than A and T. The frequency results on the pairs reflect those on nucleotides: the most frequent pairs are GC and CG, that one less frequent of AT. In this work, we do not address the possible causes of the origin of fast and slow regions along the ORF.
Further analysis is necessary to interpret their biological impact during the translation process. It would be interesting to analyse each of the 40 sequences, in order to deepen where effectively the region rich in GC and AT are located along the ORF.
However, our results suggest that the nucleotide composition of the subsequences constitutes useful information for recognizing and distinguishing the fast from the slow subsequences. Based on this evidence, we have thought it would be interesting to exploit this information to predict the translational speed using machine learning approaches.

5.1 Subsequences frequency distribution analysis

Finally, we analyzed the evolution of frequency distributions of nucleotides in order to verify whether the composition of subsequences is homogeneous. In this section, we report the evolution of frequency distribution experiment relating to fast subsequences. Figure 5.9 depicts the frequency distribution of fast subsequences with a minimum length of four nucleotides. The nucleotide composition seems almost homogeneous throughout the entire dataset, especially if we exclude the noise due to the very short subsequences.

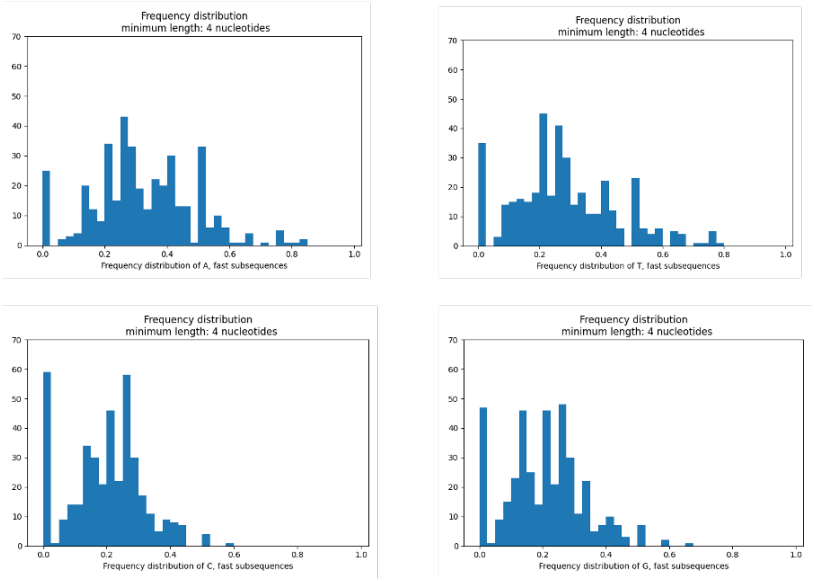


Figure 5.9: Frequency distribution in the subsequences with a minimum length of four nucleotides

We have reproduced the previous analysis selecting subsequences with a minimal length of 6 nucleotides and a maximum length of 18 nucleotides. In this way, we have removed the noise due to too short sequences and the vanish gradient caused by too long sequences [92]. As can be seen in Figures 5.10, 5.11, 5.12, 5.13 the frequency distributions turn out to be quite symmetrical and centred around the mean. 5.10. Based on the results obtained, we can state that the interval of observations where the frequencies are, is reduced. Thanks to our experiments, we can state that the optimal length of the subsequences for the classification analyses with the machine learning approach is 6 to 18 nucleotides. In the following analyses, we will use only the subsequences which satisfy the criteria chosen. From 1251, we selected 485 subsequences, whose 220 are fast and 265 are slow.

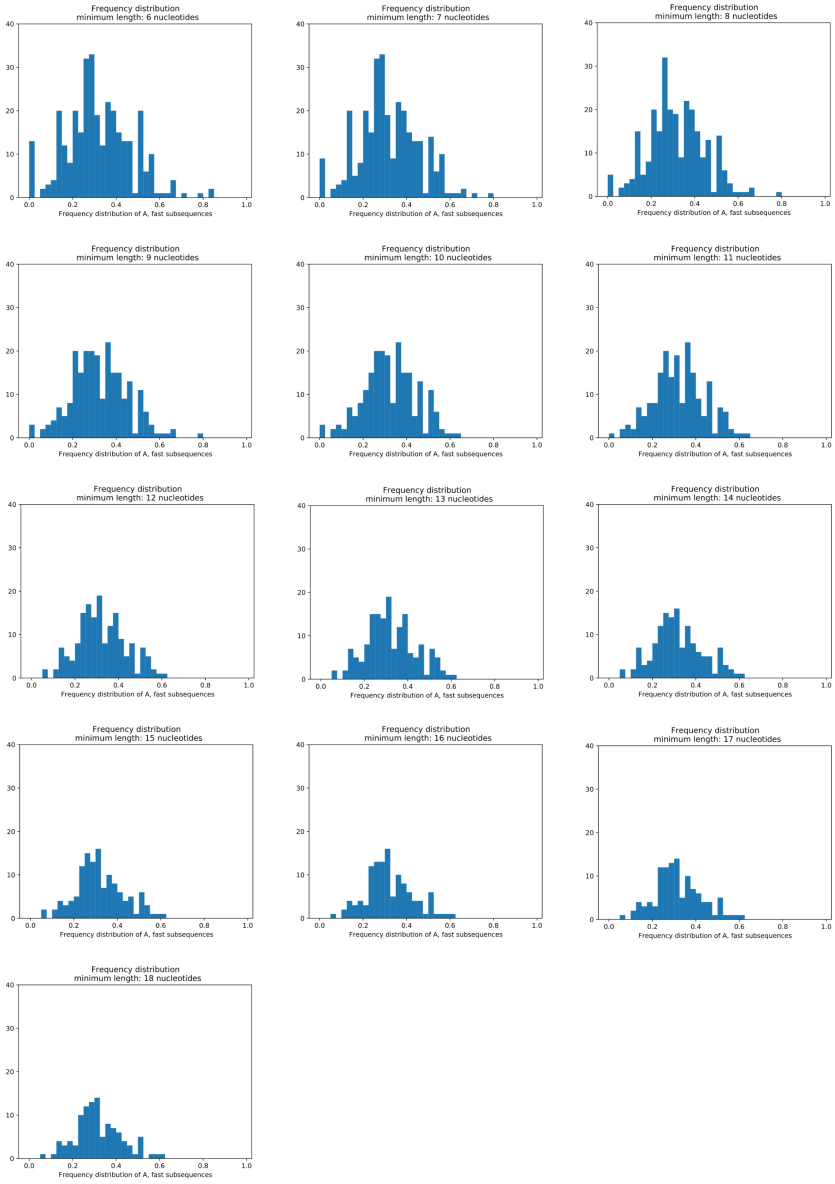


Figure 5.10: Frequency distribution of A nucleotide in the fast subsequences with a minimum length from 6 to 18 nucleotides

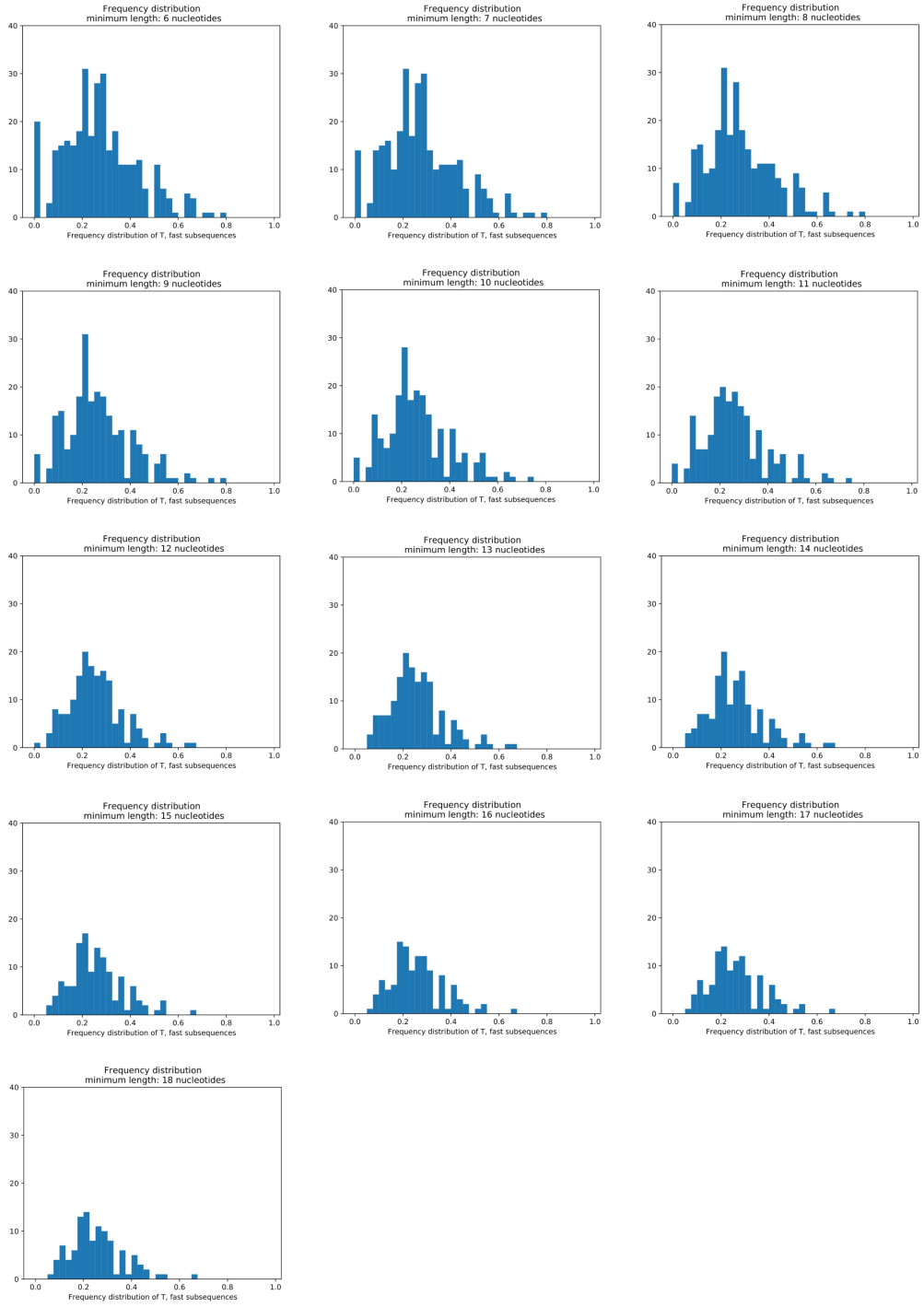


Figure 5.11: Frequency distribution of T nucleotide in the fast subsequences

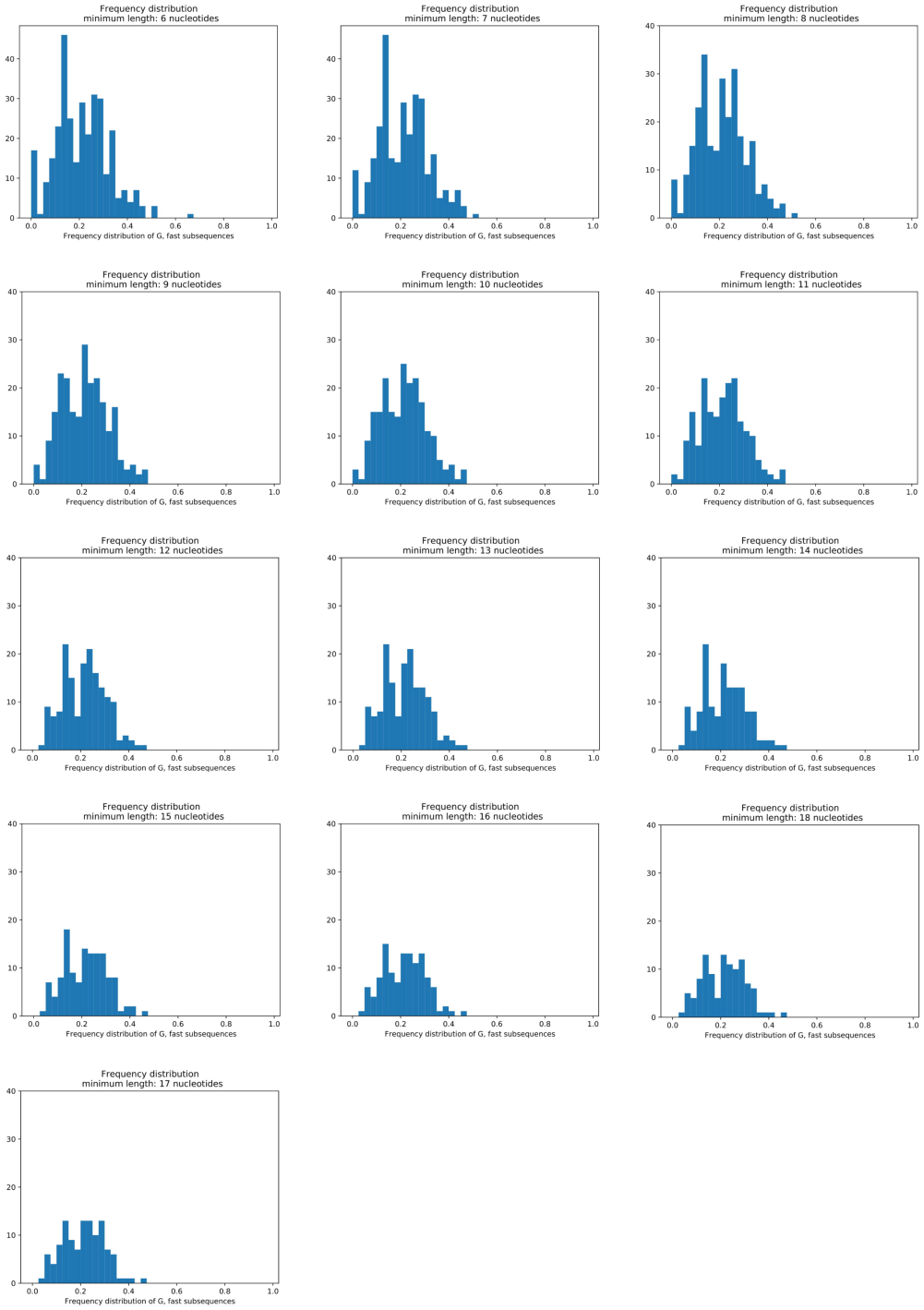


Figure 5.12: Frequency distribution of G nucleotide in the fast subsequences

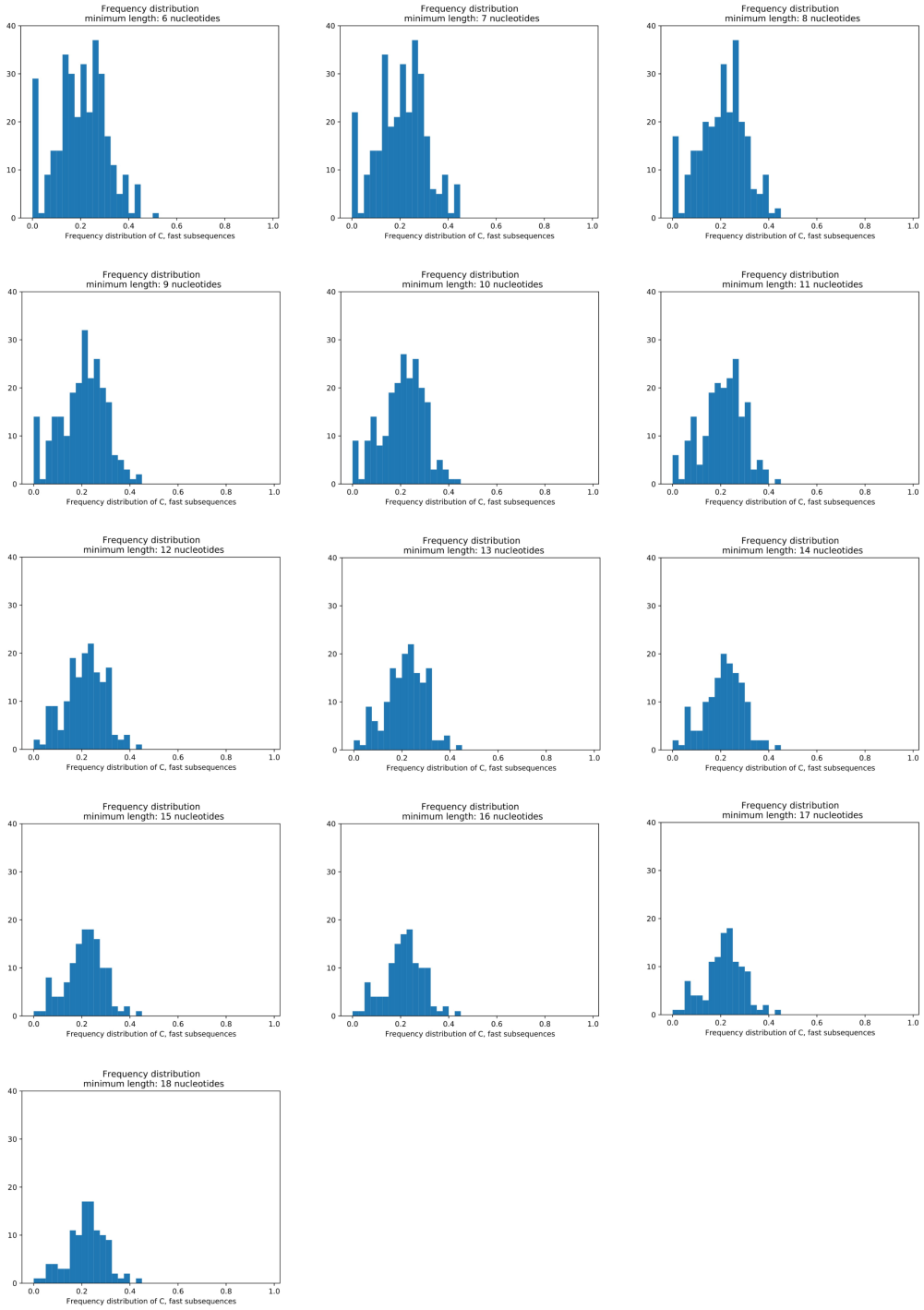


Figure 5.13: Frequency distribution of C nucleotide in the fast subsequences

Chapter 6

Artificial Neural Networks

This chapter provides a brief presentation of Artificial Neural Network (ANN) models and their application in our study. With the increasing availability of highly dimensional and complex data, the application of ANNs, especially deep architectures, has become more frequent. Thanks to their ability to make predictions and to address sequence classification problems, ANNs have been useful in the field of '-omics' research, including genomics and transcriptomics [93]. Indeed, ANN approaches represent a powerful tool for biological data processing. For instance, they have been used for medical image classification [94], for the prediction of structural properties of the protein surface, for variant calling [95], and in many other type of biological problems [96][97]. To provide the information necessary to interpret their meaning, we will introduce the theoretical framework of the model and comprehensively examine the basic methods used.

6.1 Biological neural networks

The relationship between biology and the field of machine learning is very complex.
The nervous system comprises neurons, excitable cells that process information, contributing to more significant cognitive functions. The human brain includes tens of billions of neurons, which represent the functional unit of substantia nigra (SN). Each of them is interconnected through cytoplasmic extensions to approximately ten thousands of other neurons.
As shown in Figure 6.1 a) , a typical neuron consists of:

- a **cell body (soma)**, the region containing the nucleus and most other cell organelles; it combines and integrates incoming signals;

- **dendrites** and an **axon**, which are two types of cytoplasmic extensions. *Dendrites* are thin fibers that extend like tendrils to receive many signals from neighboring neurons and conduct them towards the cell body; the *axon* is a long thin fiber that carries potential actions away from the cell body.

Neuronal communication occurs through synapses, specialized structures that allow to receive and transfer the signals from one neuron to another. Most synapses are chemical and enable communication via the release of chemical signaling molecules, called *neuro-transmitters*. Chemical agents are synthesized and packaged into vesicles by the presynaptic axon terminal. Firstly, they are released from the presynaptic cell into the space between the pre- and postsynaptic cells, known as *synaptic clefts* and then they bound to receptors on the membrane of the postsynaptic cell. The general structure of a chemical synapse is illustrated schematically in Figure 6.1 c)
If the total strength of the electrical signal exceeds a certain threshold limit, the signal will be sent down the axon to the synapses. The ability of neurons to conduct neural impulses is mainly related to the presence (or spreading) along the axons of three specific types of voltage-gated ion channels, which are able to generate the membrane potential (V_m). The charge separation gives rise to a difference of electrical potential, or voltage, across the membrane, called the membrane potential. V_m is defined as

$$V_m = V_{in} - V_{out}$$

where V_{in} is the potential on the inside of the cell and V_{out} the potential on the outside. The ion flow through a voltage channel induces a redistribution of charges on the two sides of the membrane, modifying the V_m value.

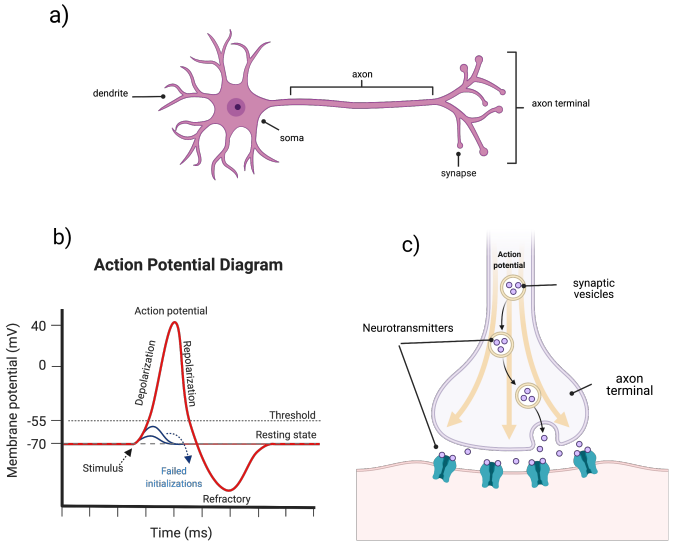


Figure 6.1: Model of the neuron: a) Structure of the biological neuron; b) Action Potential diagram; c) Signal transmission across the synaptic cleft.

Similar to neurons in the brain, an ANN is a collection of simple processing units, called artificial neurons, that are connected together through weighted edges, that resemble synapses.

6.2 Mathematical model of the neuron

The origins of neural networks are based on the construction of a model that mimics the functioning of the human brain. Indeed, understanding biological systems, such as the neural system, has been a major challenge for many researchers. Dating back to 1943, the mathematical model known as *Threshold Logic Unit* (TLU), proposed by McCulloch and Pitts [98], represents the first step in this direction: The TLU is a system capable of receiving n binary inputs producing an output based on them, able to realize any boolean function. This neuron is a finite-state machine that describes a propositional logic with quantifiers, allowing to formulate precise hypotheses on the nature of brain mechanisms.
Later on, in 1969, Frank Rosenblatt, inspired by the Hebbian theory of synaptic plasticity, developed the first type of artificial neuron — the *perceptron* — able to process real data [99]. This mathematical model is a binary classifier, constituted by a single neuron, able to correctly partition only linearly separable patterns. A historic application of the perceptron model can be found in [100], in which the identification of translational initiation sites in E. coli is carried out.
The perceptron model is depicted in Figure 6.2.

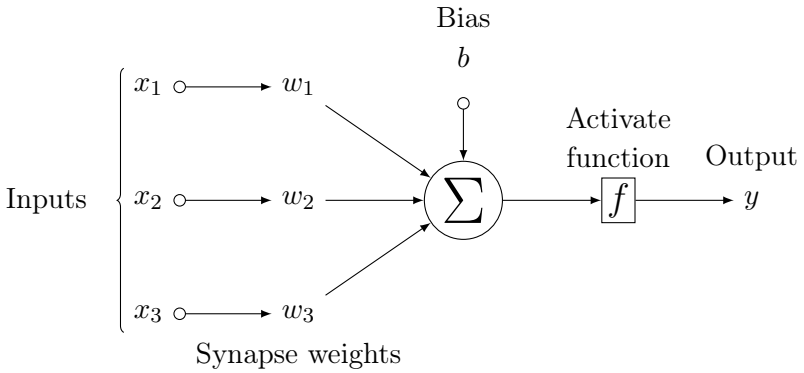


Figure 6.2: Perceptron model — Rosenblatt's model of a neuron, with input vector x, weights w, bias b, activation function f and output y.

Specifically, the perceptron receives N inputs $(x_1, \ldots, x_n)$ from a set of incoming edges, processes them and transmits an output signal (inputs can be either excitatory or inhibitory). Each synapse i has an associated weight w_i, being the weights the tunable parameters of the model.
Therefore, the neuron receives the signals $[x_1 w_1, \ldots, x_n w_n]$ and sums them to produce an output:

$$S = \sum_{i=1}^{n} x_i w_i \tag{6.1}$$

where $W = (w_1, \ldots, w_n)$ collects the synaptic weights and $X = (x_1, \ldots, x_n)$ is the input vector. Plausibly, the bias can be thought of as a w_0 weighing a constant input of -1 and can be included in Eq. (6.1). Afterwards, a non-linear function of the weighted sum, known as the *activation function* f, is calculated to produce the output y:

$$y = f(S)$$

The activation function establishes how the neuron should react to the input signals. It also defines the numerical value emitted as the unit output. If the sum of the inputs exceeds a certain threshold (the bias), the neuron is turned on and can send an impulse through its axon.
Then, the neuron output will be 0 or 1 if S is less or greater than the *activation threshold*:

$$y = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} w_i x_i \leq w_0 \\ 1 & \text{if } \sum_{i=1}^{n} w_i x_i \geq w_0 \end{cases}$$

According to the task to perform, different activation functions can be used.
The activation functions of interest for this dissertation (Figure 6.3) are as follows.

- The logistic sigmoid function:

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

  It maps real numbers to the range $[0, 1]$, making it suitable for binary classification. A generalization of the logistic function is the Softmax function, which is used to normalize the output of a network with multiple outputs to a probability distribution over the predicted output classes.

- The hyperbolic tangent function:

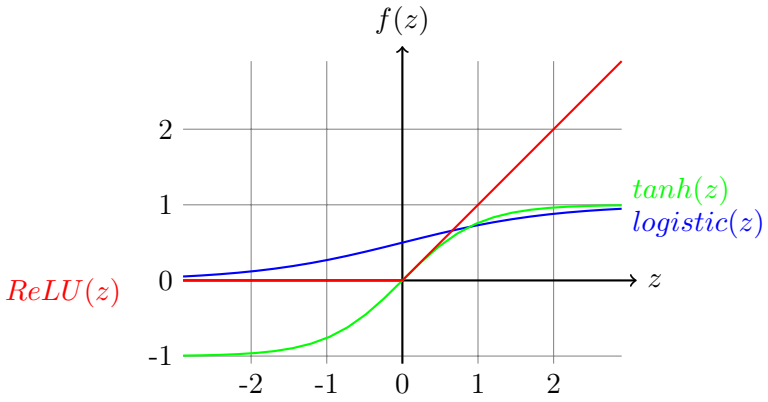


Figure 6.3: Commonly used activation functions: sigmoid, tanh and ReLU.

$$\sigma(x) = \tanh(x)$$

It can take any real value as input and produces an output with values ranging between [-1,1]. It has a sigmoidal shape with a scaling factor and a vertical shift.

- The rectified linear unit (ReLU) [101]:

$$\sigma(x) = max(x, 0)$$

It is currently one of the most popular activations for deep neural networks (since it partially overcomes the vanishing gradient problem and performs better w.r.t. classical sigmoids). The main advantage of its use is that it does not activate all the neurons simultaneously. It returns 0 if it receives any negative input, while for any positive value, it returns that value. The ReLU is widely applied in convolutional neural networks.

Neurons can be organized and connected in different ways depending on the type of activity the network carries out. In the biological field, the most common network architectures present neurons collected in fully connected layers: Each neuron of one layer is connected to all the neurons of the next layer. The input signal propagates through the network in a forward direction and it flows layer by layer from the input to the output (hence the term *feedforward neural network*).

6.2.1 Multi-Layer Perceptron

Architecturally, feedforward neural networks are represented by acyclic graphs. There is a partial ordering between the vertices (neurons) of the graph, while each neuron can be connected to both other neurons and inputs. A Multi-Layer Perceptron (**MLP**) is a particular type of feedforward neural network in which neurons are organized in fully-connected layers, with no intralayer or shortcut connections. The MLP represents the natural extension of the perceptron architecture, combining neurons in multiple layers.
MLPs have been used extensively for classification and prediction tasks because of their simple structure and fast learning process.
Specifically, MLPs — which can classify also non-linearly separable data — are composed of at least three layers, containing one or more neurons.

- Input Layer: level designed to receive information from the outside. The number of neurons (which are buffers, simply passing the information coming though them) depends on the size of the input vector.

- Hidden Layers: intermediate layers located between the input and output layers. There is often more than one hidden layer. Each hidden neuron receives a numerical value in input, which corresponds to a weighted sum of the signals coming from the previous layer (input or hidden).

- Output Layer: the final level that receives input from hidden units and transmit signals outside the system.

The MLP model, shown in Figure 6.4, represents a feedforward neural network with three hidden layers. This architecture allows to learn more complicated features from the input data and, if used as a classifier, to realize complicated separation surfaces. Indeed, a two-layer network with sigmoids (or similar) on all units can represent any Boolean function. Moreover, according to the *universal approximation theorem*, even an MLP with a single sigmoidal hidden layer — composed by a sufficient number of hidden neurons —, trained with enough data, can approximate any bounded continuous real function to any desired accuracy [102].

6.2.2 Network training algorithm

In order to perform a particular task, the network should be appropriately set up and then trained. *Learning* is defined as the process through which the network adapts its weights to external stimuli in order to be able to produce the desired output. Learning in MLP neural networks can happen in supervised

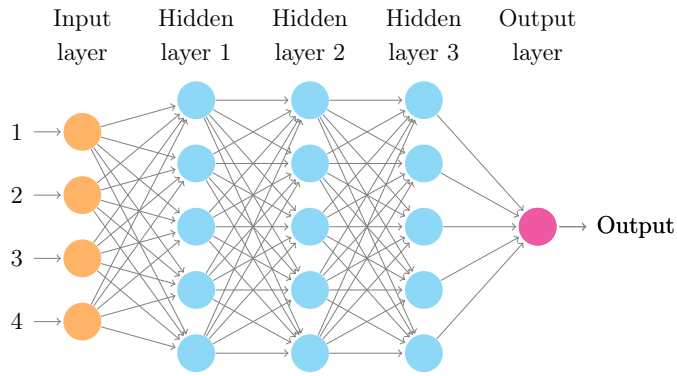


Figure 6.4: Multi-layer perceptron with four neurons in the input layer, five neurons for each hidden layer, and one output neuron.

fashion: the patterns in a given set of samples are predicted using a known labeled subset. Supervised learning aims to build a system that can classify new patterns based on the available knowledge, codified into the connection weights and acquired during the network training. The most used learning algorithm for MLPs is *Backpropagation*: it consists in the iterated modification of the connection weights to minimize the *loss function* [103]. Indeed, the loss function or error function gives a measure of the distance between the network outputs and the expected results at any iteration. The total error that the network commits in the learning phase can be described as:

$$E = \Sigma_p E_p = \Sigma_p (t_p - o_p)^2 \tag{6.2}$$

where p is the number of data supplied to the network during training, E_p is the error for the input pattern p, while t_p and o_p are the expected and observed results for p , respectively. Then, the network parameters are iteratively adjusted to reduce the error, starting from to the top and propagating the error signal backward to the lower layers. The first step in training an MLP consists in choosing the network architecture and setting its initial synaptic weights. Subsequently, the model fits a *training set*, which is representative of the dataset as a whole. A cycle of presentation to the network of all the examples belonging to the training set is called an *epoch*. After processing the whole training set, the next step is to calculate the *loss function*. During the *backward computation*, the ANN modifies its parameters to minimize the error function through a learning algorithm. Finally, to evaluate the final model performance, the network behaviour is verified on a different dataset, called *test set*. For example, when the network represents a classifier, the accuracy of the model is evaluated, i.e. the percentage of test examples for

which the network produces the correct output, over the test set dimension. In fact, the main objective of a supervised learning strategy is to get the *generalization ability*, which consists of producing a correct classification for unknown examples. An optimal generalization requires a trained model able to recognize the difference between signal and noise. However, especially when learning has been done for too long or based on too few examples, the model may work well on the training set but be unable to generalize well, resulting in *overfitting*. In contrast to overfitting, *underfitting* typically occurs when a model is unable to capture the relationship between the input x and the target output y. Both underfitting and overfitting yield poor performance of the machine learning algorithm. Figure 6.5 explains these phenomenons. Generally, the more training data given to the model, the less likely it is to fit too much. When more data are added, the model becomes unable to overuse all the samples and is forced to generalize to make progress.

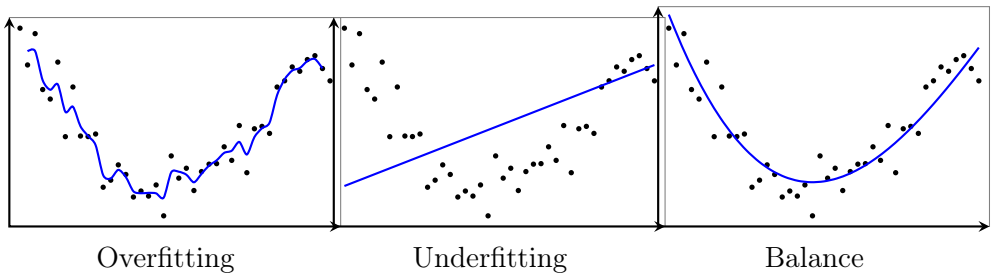


Figure 6.5: .
Supervised learning over/under/good-fitting. By looking the graph on the left side, we can observe that the blue line covers all the points which are present (also noise and outliers). This model tends to cause data overfitting. The middle graph shows that the blue line is not able to capture the point distribution. Such model tends to cause data underfitting. The blue line fits the majority of the points in the graph on the right side, representing a balanced model.

6.2.3 The k-fold cross-validation technique

The k-fold cross-validation is a model validation technique used to guarantee the generalization ability of a network and to avoid the overfitting phenomenon. The available dataset is split randomly into k subsets or folds. The first fold is kept as the test set while the other $k-1$ form the training set. The process is repeated k times, and each time a different fold is used for the validation, as depicted in Figure 6.6. At each of the k steps, the metrics to evaluate the network performance are calculated.
Another strategy to reduce overfitting and improve generalization consists in applying the *early stopping* method. During training, a validation set is used to evaluate the network generalization after each epoch. Once the model performance drops on the validation set, the training process is stopped, possibly before convergence [104].

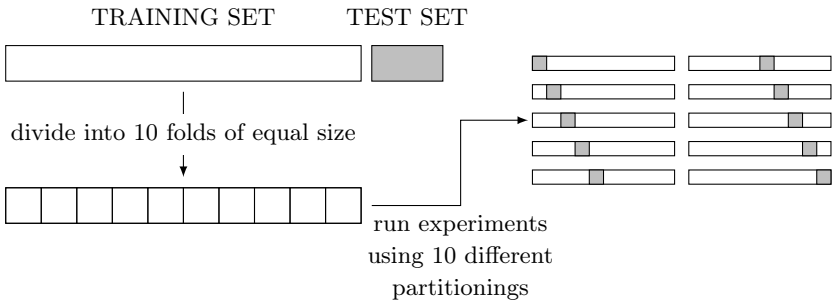


Figure 6.6: Depiction of the k-fold cross validation for 10 folds.

6.2.4 Convolutional Neural Network

Convolutional Neural Networks (ConvNets) [105] are specialized deep neural networks de-signed to process data with a *grid-like structure*. In machine learning applications, the CNN input is usually a multi-dimensional array of data (e.g. 1D for signals and biological sequences; 2D for images or audio spectrograms; 3D for video or volumetric images) and the *kernel* is a multidimensional array of learnable parameters [106]. Convolutional networks are characterized by the use of convolutions in some of their layers. The term *convolution* refers to the mathematical operation

$$s(t) = \int x(a)w(t-a)da \tag{6.3}$$

evaluated on two real-valued functions, namely the input function x and the weighting function w, called the *kernel*. Typically, when processing image-like data, they are repre-sented as tensor of shape $(h \times w \times d)$, with d image channels of height h and width w. Then, a convolution between a two dimensional image I and a kernel K can be defined as

$$S(i,j) = \sum\sum I(m,n)K(i-m, j-n) \tag{6.4}$$

where m and n are the image dimensions in pixels.
In the MLP architecture, each neuron is fully connected to all neurons in the previous layer and is completely independent from the other neurons belonging to the same layer. This means that O($h \times w \times d$) parameters are needed for each neuron in the first hidden layer. Specifically, when the input consists of an image of dimesion $256 \times 256 \times 3$, a single neuron presents two million of parameters, approximately. Therefore, the fully-connected architecture is too expensive in this context, involving an enormous amount of parameters that would lead quickly to overfitting. Differently, CNNs are more effective in this case, since they are more parsimonious in terms of parameters. The general idea is that the convolutional operations can extract valuable information from an image, using a very small amount of parameters.
A typical convolutional neural network architecture consists of a series of interleaved convolutional layers, pooling layers and fully-connected layers, as illustrated in Figure 6.7. Convolutional and pooling layers, perform feature extraction from the image, while a fully connected layer uses the extracted features, for instance, to classify the image.

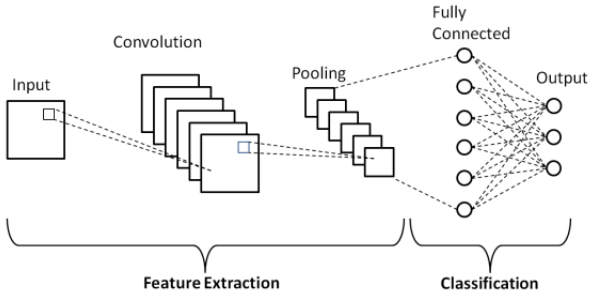


Figure 6.7: An overview of the propagation model of a typical CNN architecture for image classification.

Convolutional layer

In the convolutional layer, a convolutional kernel is applied to the input. During the forward pass, the kernel slides along the image (both in width and in height). Then, the scalar product between the filter and the corresponding matrix in the input grid is calculated. Once N multiple convolutional kernels are applied within a convolutional layer, N so-called *feature maps* of size $F_{fm} \times W_{fm}$ are created, one from each convolutional kernel. The computation of the final volume not only depends on the size of the input ($h \times w \times d$) but also on two hyperparameters, i.e. the stride (S), which represents the sliding size of the kernel, and the spatial extension of the filter (F). For example, given a 3×3 kernel, the input grid can be slid with a stride equal to 1, 2 or 3. If S=1, given an input of dimension $h \times w \times d$ and a filter of dimension $f_h \times f_w \times d$, the output of the convolution operation has a dimension $(h - f_h + 1) \times (w - f_w + 1) \times 1$. The convolution operation performed by the first convolutional layer is depicted in Figure 6.8.

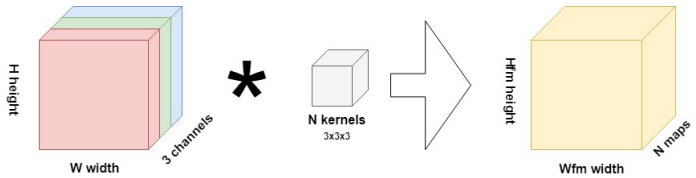


Figure 6.8: Depiction of a convolutional layer composed by 3x3 kernels. The N output feature maps are obtained sliding N kernels along the input.

Furthermore, the size of the feature map is also controlled by the zero-padding procedure. Indeed, pixels on the corners and the edges "are touched" much less than those in the middle and, consequently, the information on the borders of images are not preserved as well as the information in the middle.

Zero-padding allows to solve this problem, "including" the image into a frame of 0s. The convolution computation is usually followed by a ReLU ($= max(0, x)$) layer, which replaces all negative pixel values in the feature map by 0.

Pooling Layer

Another relevant component of CNNs is the pooling layer that aggregates (summing up, averaging or extracting the maximum) the values computed by the convolution filters in small sub-regions (e.g. 2×2) of the input feature map. Its function is to reduce the spatial dimension of the input (width and height) progressively, to diminish the number of parameters and, consequently, the computational requirements. One type of pooling layer is the average pooling layer, which will be used in this dissertation. The average pooling is an operation that computes the average value for each patch across each channel, and uses it to create a downsampled feature map, as depicted in Figure 6.9. In contrast to *max pooling*, the average pooling retains much information about the block.

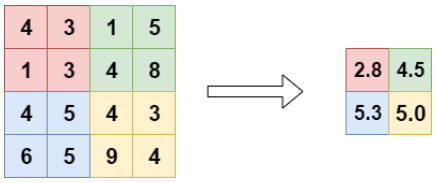


Figure 6.9: The 2×2 average-pool operator, applied on a single-channel image

Fully-Connected Layer

The last level of a CNN architecture is represented by (one or more) Fully-Connected Layer(s), similar to those composing MLPs. The output function used is the *softmax*: It maps the output value of each neuron in the range $[0, 1]$, representing the probability that the input grid belongs to one specific class.

The CNN architecture is based on the formalization of three fundamentals properties: sparse interaction, weight sharing and equivariance to translation.

- Sparse interaction: accomplished by making the kernel smaller than the input; differently from MLP architectures, CNN neurons are connected only locally to neurons in their neighborhood. This means that fewer

parameters must be stored, reducing the memory requirement of the model and improving its statistical efficiency.

- Weight sharing: In a traditional neural network, each element of the weight matrix is used once while, in the CNN model, weights applied to one input are the same as the weight applied elsewhere or, in other words, each member of the kernel is used in all parts of the input.

- Equivariance to translation: The architecture of the CNN, characterized by the weight sharing property, ensures that the convolutional layer can produce the same output when detecting the same pattern in different locations.

6.3 Machine learning applications

Due to their characteristics, neural networks lend themselves well to the analysis of information encoded along a protein sequence. In our specific case, the models we have implemented can perceive the presence of sequence signals made up of particular nucleotide arrangements. Depending on the type of signals that the networks can recognize, they will classify the functional characteristics associated with the signal itself. Once we obtained the reproducible Ribo-seq profiles, we investigated their consensus sequences through machine learning approaches to classify their translation speed.

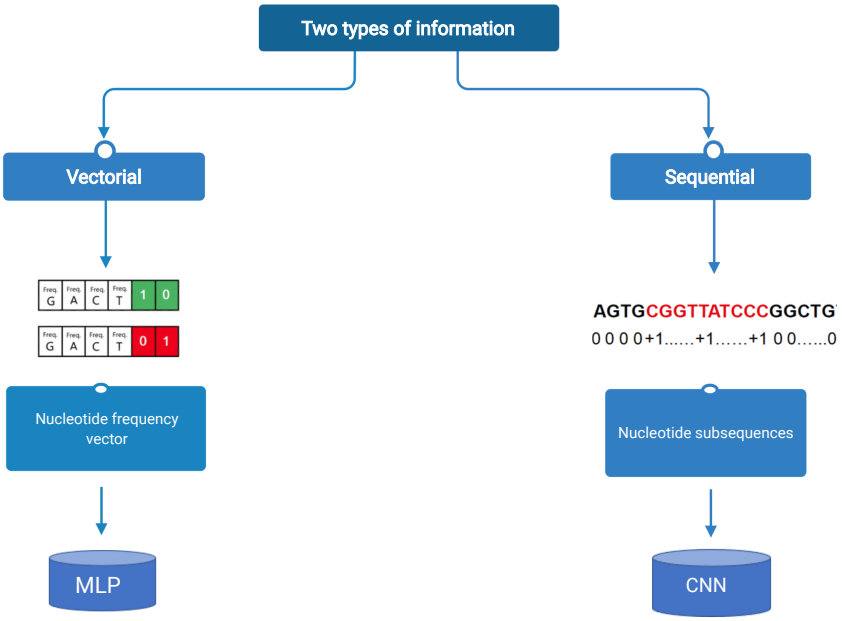


Figure 6.10: Schematic machine learning experimental workflow. Left: The nucleotide frequency vector is analyzed by an MLP. Right: A nucleotide sequence is processed by a 1D CNN.

The consensus sequences labeled with -1 and $+1$ represent the input for the following analyses with the neural network models. In our specific problem, exploiting a network architecture can reveal to us whether there is enough information in the data to classify the subsequences into slow and fast with high accuracy. To accomplish this task, we used two different types of information: vector and sequential. In fact we first considered a four-dimensional array that collects the frequencies of occurrence of the four nucleotides and then

we took into consideration the whole sequences, to understand if the order in which the nucletides are arranged helps to capture the translation speed signal. Consequently, the experiments were carried out by applying two different neural network architectures: MLPs and CNNs, previously described. A sort of comparison between the two models is possible since they have been tested using the same dataset. In particular, as explained in Chapter 6, only subsequences of length from 6 to 18 nucleotides have been selected.

6.3.1 Classification based on nucleotide frequencies by MLPs

In the first experiment, the main purpose is to show the ability of the Multi-layer Perceptron to predict the translation speed of sequences based on their nucleotide frequency. Therefore, once we have identified the 485 subsequences, of which 265 are slow and 220 are fast, we calculated the vector of the relative frequencies of the four nucleotides within them. The target of each example has been coded according to the one-hot encoding with two bits: [0,1] for fast sequences and [1,0] for slow sequences, as shown in Figure 6.11.

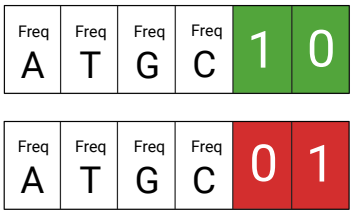


Figure 6.11: One-hot encoding: fast (green) and slow (red) sequences

The examples were split into two sets: 436 made up the training set (about 92% of the entire dataset) and 49 the test set. The first set is used for training the network, while the second to evaluate its generalization. The MLP architecture is shown in Figure 6.12.
Once the dataset has been divided into training and test sets, the hyperparameters have been defined:

- learning rate: 0.05
- epoch: 800
- hidden units: one hidden layer with 10 neurons
- Activaction function: sigmoidal

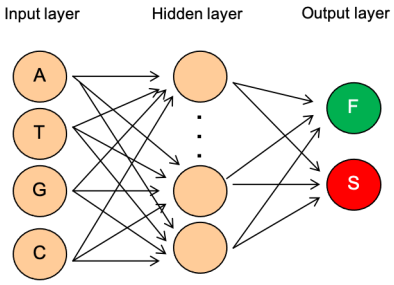


Figure 6.12: An MLP architecture having the nucleotides frequency as input, and predicting the sequence class probability distribution (slow or fast).

In agreement with the size of the input vector and the target, the MLP has four neurons in the input layer, one for each relative frequency value (A, T, G, C). The neurons in the input layer receive the features and propagate them to the neurons in the unique hidden layer. The choice of using a small number of neurons within the hidden layer is closely related to the small size of our dataset: The idea is to get a simpler model that is easier to interpret. When relatively few data are available, the use of too many parameters leads to the overfitting phenomenon.
Our model can be easily implemented using modern machine learning libraries since it is wholly based on a training function that updates weight and bias values according to the resilient backpropagation algorithm [107]. Then, we applied the *Softmax* function at the output layer, which gives a probability distribution for each class label (slow or fast). Finally, we evaluated the accuracy on the test set and then summarized the model performance across five runs. Average and standard deviation of the classification accuracy without cross-validation on five runs are reported in Table 6.1. In order to avoid the overfitting phenomenon and improve the generalization capability of the model, we additionally carried on an experiment using the k-fold cross validation.
Average and standard deviation of the classification accuracy with a 5-fold cross-validation (described in details in Section 6.2.3) are reported in Table 6.2. It can be noted that the network reaches 77.17% and 75.67% of average training accuracy and average test accuracy, respectively. In both experiments, the standard deviation of the accuracy-test does not exceed 1%.
The accuracy percentages are surprising, considering that they are obtained using only nucleotide frequencies as input. The next step is to verify how much the sequential order of the nucleotides can impact the prediction of the speed of translation.

Run	**Training Accuracy (%)**	**Test Accuracy (%)**
1	81.70	78.35
2	79.90	78.35
3	82.73	79.38
4	82.99	77.32
5	82.99	79.38
Average	82.06	78.56
Standard Dev.	1.32	0.86

Table 6.1: Multilayer Perceptron accuracy over 5 runs. We report the obtained Training and Test Accuracy (second and third columns, respectively) obtained over 5 runs. The last two rows report the average and standard deviation of the computed metric.

Run	**Training Accuracy (%)**	**Test Accuracy (%)**
1	76.29	75.26
2	76.03	75.26
3	77.32	77.32
4	77.84	75.26
5	78.35	75.26
Average	77.17	75.67
Standard Dev.	0.99	0.92

Table 6.2: Multilayer Perceptron accuracy with 5-fold cross-validation. We report the obtained Training and Test Accuracy (second and third columns, respectively) obtained over 5 fold with the cross-validation mechanism described in Section 6.2.3. The last two rows report the average and standard deviation of the computed metric over the folds.

6.3.2 Classification based on nucleotide sequences by 1-D CNNs

The good performance of the MLP model leads to the conclusion that much of the information to understand the levels of protein synthesis could lie in the relative frequencies of the four nucleotides. Once the frequencies of the four nucleotides have been analyzed, we decided to use a more complex architecture able to process the sequential data. The main goal of our analysis is to investigate how strongly the order of nucleotides affects the accuracy of the prediction. In our specific case, a one-dimensional convolutional neural network was implemented, considering biological sequences of fixed length. Unlike 2-D images treated with three color channels (R, G, B), our subsequences have been processed considering four channels (A, T, G, C). Firstly, the 1-D CNN recognizes the local patterns in each subsequence through its convolutional layers. Just as with 2-D CNNs, the 1-D pooling operation leads to extract

the 1-D patches from the input. In particular, the average pooling is used for reducing the length of the input.
The 40 reproducible sequences identified through our method are randomly shuffled and divided into the standard training, validation and test sets. Among these, 31 constitute the training set, 6 the validation and the remaining 3 the test set. The validation set is built automatically, selecting 15% of sequences from the training set. See Figure 6.13 for a graphical representation.

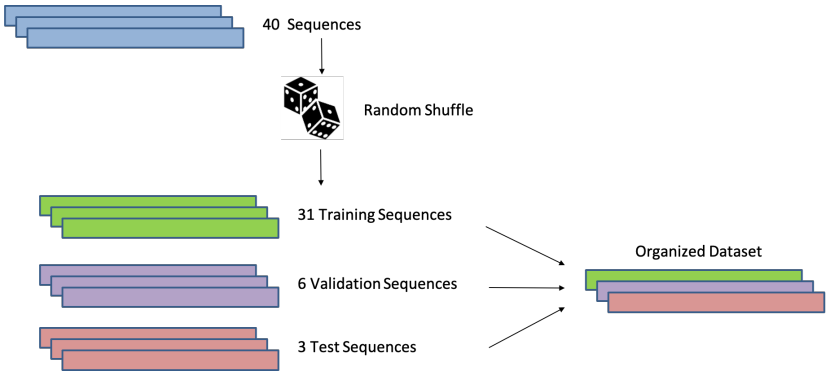


Figure 6.13: The dataset composition. The available 40 Sequences have been randomly split into training (31), validation (6) and test (3) Sequences in order to compose the final dataset.

Our model elaborates some examples constituted by a single nucleotide, of which we want to predict the translation speed, plus a context, that is a set of nucleotides that precede or follow the position under consideration. Each nucleic sequence of fixed length corresponds to an example, which is obtained by "cutting out" from the gene a fragment of 36 nucleotides containing the sequence. Nucleotides with target equal to 0 are ignored. Padding is applied, at the beginning and at the end of the sequences, to reach the fixed context length of 36 nucleotides (18 nucleotides to the left and $18 - X$ to the right of the sequence, if X is the sequence length). The one-hot encoding is used for the targets based on two bits — [0,1] for slow sequences and [1,0] for fast sequences, respectively (see Figure 6.14).
Initially, we defined the CNN model using the Keras deep learning library [108]. Input sequences are processed via the CNN depicted in Figure 6.15, which consists of 2 convolutional layers, equipped with average-pooling, followed by 2 fully connected layers. The architecture is based on 1-D kernels having size 3 and leverages a ReLU activation function. In particular, each hidden convolutional layer is composed by 16 filters. The last fully-connected layer output is projected onto a two-class probability distribution via a softmax function.

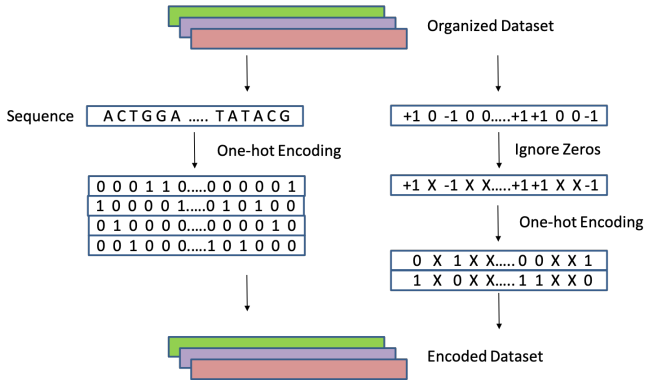


Figure 6.14: Pipeline for the dataset encoding. See the main text for further details.

We trained our model for five thousand epochs, measuring the accuracy on the validation data.

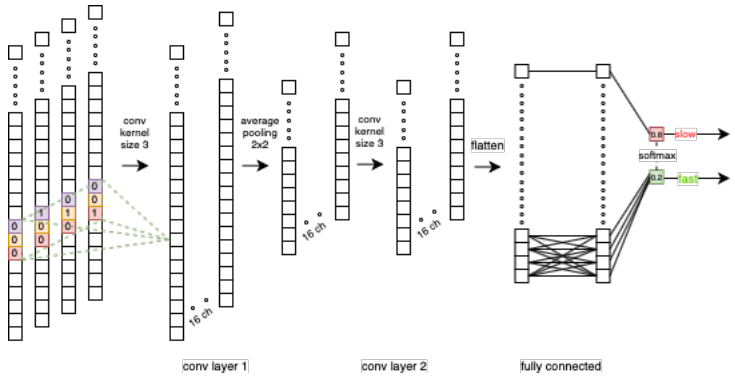


Figure 6.15: The 1-D CNN model exploited for sequence classification. The implemented CNN is capable to process 1-D sequential data employing two convolutional layers each composed by 16 filters of 3x3 kernels. The obtained representation is projected with a fully-connected output layer onto a two-class probability distribution via a softmax function.

We summarize the model performance across five runs in Table 6.3. The metrics used to evaluate a classification model are precision, recall, F1-score and accuracy. We can see that the model performed well achieving a classification accuracy of about 89.39%. It is important to note that, by training a significantly more complex network than the MLP, and by providing a sequential data input, the accuracy increases by 9 percentage points. To further improve performance, we have set up a more complex architecture consisting of seven CNNs: each of

CNN							
Run	**Precision**		**Recall**		**F1- score**		**Accuracy**
	slow	fast	slow	fast	slow	fast	
1	96.00	86.00	90.00	95.00	93.00	90.00	91.84
2	96.00	82.00	87.00	95.00	91.00	88.00	89.80
3	93.00	85.00	90.00	89.00	92.00	87.00	89.80
4	100.00	73.00	77.00	100.00	87.00	84.00	85.71
5	93.00	85.00	90.00	89.00	92.00	87.00	89.80
Average	95.60	95.60	90.00	93.60	91.00	87.20	**89.39**
Standard Dev.	2.88	5.36	5.20	4.67	2.35	2.17	**2.24**

Table 6.3: Summary of the results obtained with the 1-D CNN model. We report the obtained test set metrics computed over 5 different runs. The last two rows report the average over the runs and the corresponding standard deviations. In the case of Precision, Recall, F1-score, we report the results for slow and fast class.

them provides a different prediction, i.e. a pair of probabilities describing the membership of a sequence in a given class (slow or fast).
Each of the seven CNN uses a random 15% of data as the validation set. The models are trained for 5000 epochs using the Adam optimiser. After, we calculate the average of all the predictions. All the experiments are repeated 5 times, reporting the test accuracy corresponding to the best result on the validation data.
The results of the experiments performed by the CNN ensemble are summarised in Table 6.4. Our model reaches a 91% accuracy.

ENSEMBLE: 7_CNN							
Run	**Precision**		**Recall**		**F1- score**		**Accuracy**
	slow	fast	slow	fast	slow	fast	
1	96.00	86.00	90.00	95.00	93.00	90.00	91.84
2	96.00	82.00	87.00	95.00	91.00	88.00	89.80
3	96.00	86.00	90.00	95.00	93.00	90.00	91.84
4	96.00	86.00	90.00	95.00	93.00	90.00	91.84
5	93.00	85.00	90.00	89.00	92.00	87.00	89.80
Average	95.40	95.40	90.00	93.80	92.40	89.00	**91.02**
Standard Dev.	1.34	1.73	1.22	2.68	0.89	1.41	**1.12**

Table 6.4: Summary of the results obtained with the CNN-ensemble model. We report the obtained test set metrics computed over 5 different runs. The last two rows report the average over the runs and the corresponding standard deviations. In the case of Precision, Recall, F1-score, we report the results for slow and fast class.

6.3.3 Conclusions and future work

The usage of complex architectures is a challenge because of the limited number of data. We have chosen the convolutional neural network model since weight sharing allows us to employ a limited number of parameters though using the sequential nature of the data. Our 1-D CNN is a typical architecture built of two convolutional layers, followed by one fully-connected layer.
The obtained results clearly show that this model can extract useful information from a limited amount of data in a better way than MLPs. The high accuracy (89,39 %) and low calculated variance (2.24 %) demonstrate that the training process is stable and, therefore, the results are steady and not influenced by the parameter initialization.
Furthemore, we have proposed an ensemble CNN model to further improve performance. As expected, the accuracy of the ensemble CNN increases while the variance decreases, with results which assess a good ability to discern between fast and slow subsequences (see Table 6.4).

Work in progress

Guided by the results obtained in Chapter 5, we hypothesise that the different codon usage may affect the translation initiation and then, in general, the kinetics of translation. However, even today, the debate on how translation is

regulated under different conditions and what factors greatly affect translation speed remains open. Therefore, to obtain a comprehensive overview, we repeated the machine learning experiments through different types of encoding. Indeed, a viable direction for future research is to build other representations of the same dataset, by enlarging the sliding window or by considering not only a sequence encoding based on nucleotides but also encodings based on codons and amino acids. In the context of the codon dataset, the sequence is processed based on a sliding window on the nucleotide of size 3, where each nucleotide is one-hot encoded. Therefore, the representation of each codon corresponds to the concatenation of the one-hot vectors of its three-nucleotides. Differently, for

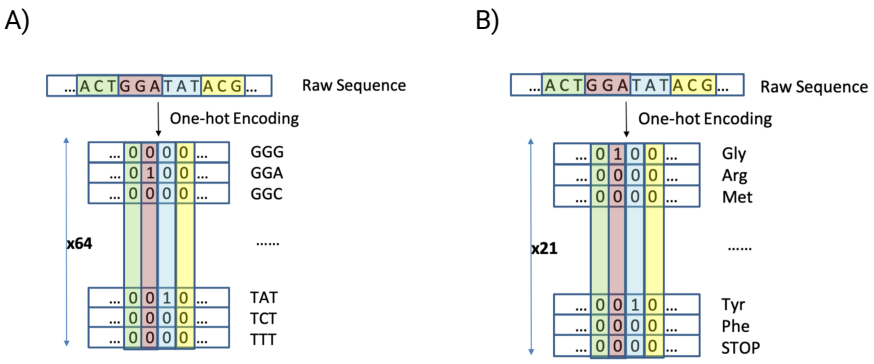


Figure 6.16: Alternative encodings: A) Codon dataset; B) Amino acid dataset.

the amino acid dataset, the sequence is read from codon to codon. The triplets which encode the same amino acid are labelled in the same way. Therefore each codon is encoded by a vector of length 21. Figure 6.16 shows the two other types of encoding we could use in our future experiments. In this way, we can obtain a deeper understanding of the consensus sequence features, in order to define better which is the most useful information to classify nucleic sequences as fast or slow.

This method could represents a starting point to investigate the possible causes of the homogeneous behaviour that characterises the 40 reproducible Ribo-seq profiles identified.

Prompted by good results obtained until now, we think that the classification of all *E.coli* transcriptome through machine learning approaches could be very interesting. Starting from the 40 *E.coli* ORFs, whose consensus sequences are known, we can analyze unlabelled sequences and classify them as fast or slow sequences. In order to do so, we decided to implement different types

of architectures suitable for processing sequential data, namely CNNs, Long-Short Term Memories (LSTMs) and Graph Neural Networks (GNNs). Therefore, our main purpose is to classify all E.coli ORFs, using different data representations (encodings) to choose the best one. In addition, four neural network models — which have been trained to predict the translational speed of codons — will be evaluated to determine which one is the most suitable for this task. The input data to this task consists of the codons for which we want to predict the translational speed and a context. Further analysis allows to find the context size that leads to the most accurate predictions.
In the following, we report the preliminary results, obtained using LSTMs on amino acid sequences. Briefly, LSTMs are a type of recurrent neural network (RNN). LSTMs (and in general RNNs) typically contain some neurons whose activation depends directly or indirectly on their output and can handle input sequences of varying length. In particular, unlike feedforward neural networks, RNNs can use their internal state or *memory* to process sequences of inputs. Through its memory, the neural network gains the ability to integrate information from past inputs [109]. Our architecture processes the sequences one element at a time plus the biological context, which consists of the subsequence of elements following and preceding the element under analysis. The length of both sides of the context is set as a parameter.
For the LSTM model, the penultimate fully-connected top layer has a *Scaled Exponential Linear Unit* (SELU) activaction. The output layer consists of 2 neurons and is equipped with a softmax activation, generating the probability distribution over the predicted output classes (fast and slow). The models are trained for 800 epochs using the Adam SGD optimiser (see Table 6.5 for the parameters).

Model	**Type**	**Dataset**	**Epochs**	**LR**	**Parameters**	**HU LSTM**	**HU Dense**
LSTM-A	Bi-Lstm	Aminoacids	800	10^{-4}	3826	12	30

Table 6.5: LSTM model and training parameters

In the context of unbalanced datasets, accuracy is not an adequate measure, as it does not distinguish between the number of correctly classified examples of different classes. Balanced accuracy is a popular metric used to evaluate a classifier's prediction performance in such scenarios that account for the imbalance by normalizing true positive and true negative predictions based on the number of positive and negative samples. Table 6.6 and Figure 6.17 show the performance obtained with the LSTM model, by varying the length of the context. It is interesting to note that our model can reach 79.20% balanced accuracy using the entire ORF as input and a context equal to 8 and 9. We

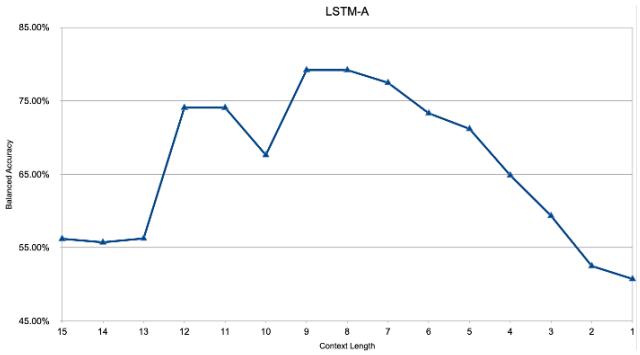


Figure 6.17: Accuracy obtained varying the length (x-axis) of the context sequence.

decided to repeat the experiment with the context equal to 8 because it is computationally less expensive, still maintaining the same accuracy. A context length equal to 8 means that the network will analyse 8 aa upstream and 8 aa downstream plus the central element. Both the average and standard deviation of the accuracy across 5 runs are reported in Table 6.7. As our results show, the context appears to play a very important role in determining improvements in accuracy. In our case, too long or too short contexts seem not to be particularly informative. Therefore according to the literature [43], in addition to codons also the context seems to have an impact to predict the speed of the translation initiation process. Furthermore, as evidenced by the results obtained with the MLP and CNN models, the composition and order of the nucleotides within the sequence is important in predicting the speed of translation. Different runs of the same experiment often show variable results, with a standard deviation of up to 2%. This is likely due to the unbalanced distribution of the two classes in the dataset: 80% of the examples belong to the fast class, while only 20% belong to the slow class. This often leads the network to learn the prior probability distribution, instead of generalizing properly. Even though this issue is usually managed by introducing class weights that can restore the prior probability balance, our network showed to be capable of reaching good performances even without class weights. Advanced optimization schemes, inspired by gradient descent [110] [111], should be implemented to adjust the network parameters to values that produce a performance improvement. As we know, the presence of more data results in better and more accurate models. Therefore, we could increase the number of reproducible Ribo-seq profiles by choosing a less conservative false discovery rate.

LSTM-A	Precision	Recall	Accuracy	F1 Score	Balanced
15	82.92	99.05	82.42	90.15	56.21
14	82.71	99.11	82.27	90.07	55.73
13	83.00	98.70	82.22	90.01	56.26
12	89.57	96.48	87.79	92.79	74.08
11	89.58	96.59	87.85	92.82	74.08
10	87.20	96.94	85.30	91.63	67.63
9	91.48	95.37	89.09	93.37	79.20
8	91.54	94.80	88.75	93.14	79.20
7	90.80	95.05	88.22	92.87	77.47
6	89.07	95.16	86.67	92.01	73.31
5	88.21	95.18	85.87	91.56	71.19
4	85.78	95.56	83.64	90.40	64.86
3	83.76	95.99	81.77	89.46	59.35
2	81.40	98.38	80.57	89.08	52.50
1	80.83	98.67	80.07	88.86	50.74

Table 6.6: Summary results of LSTM context variation. The first column indicates the length of the context. In the case of Precision, Recall, F1-score, we report the results for each length of the context examined.The last column reports the balanced accuracy obtained varying the length of the context sequence.

LSTM-A						
Length	**Run**	**Precision**	**Recall**	**Accuracy**	**F1 Score**	**Balanced**
8(17)	1	90.81	93.64	87.24	92.21	0.771571
	2	90.91	94.72	88.11	92.78	0.776984
	3	92.75	95.26	90.19	93.99	0.821825
	4	91.92	95.40	89.53	93.63	0.802839
	5	91.29	94.99	88.66	93.10	0.786764
	Average	91.54	94.80	88.75	93.14	**0.792023**
	Standard Dev.	0.81	0.70	1.16	0.70	**2.05**

Table 6.7: Results averaged over five runs of LSTM experiments with context equal to 8 (8 aminoacids upstream and 8 aminoacids downstream plus the central element). The last two rows report the average over the runs and the corresponding standard deviations.

Chapter 7

A comparative Ribo-seq profiles analysis: normal vs stress conditions

This chapter describes how our approach evaluates whether performing Ribo-seq experiments performed in conditions different from those characterising the control group might affect experimental reproducibility.

7.1 Impact of heat-shock

Firstly, we examined the samples belonging to the GSE90056 Series listed in Table 7.1, exposed after 10-20 minutes of heat shock at 42°C.
We started this comparative analysis considering the sample GSM2396724, which refers to Ribo-seq data collected from E.coli k-12 MG1655 cultured and exposed to heat-shock (42°C) for 10 min.
Specifically, we investigated whether the experimental variable (in this case control condition vs heat-shock condition) might influence the reproducibility of the Ribo-seq experiment. To achieve this, we performed two experiments,

Dataset	Genotype	Culture's medium	Stress(s)	GEO Series ID	GEO Sample ID
1	E.coli k-12 MG1655	MOPS, 0.2 % glucose	None	GSE64488	GSM1572266
2				GSE90056	GSM2396722
3				GSE72899	GSM1874188
4				GSE53767	GSM1300279
5				GSE51052	GSM1399615
6				GSE77617	GSM2055244
7				GSE35641	GSM872393
8				GSE88725	GSM2344796
Shock 10	E.coli k-12 MG1655	MOPS, 0.2 % glucose	42°C x10'	GSE90056	GSM2396724
Shock 20	E.coli k-12 MG1655	MOPS, 0.2 % glucose	42°C x20'	GSE90056	GSM2396726
Leu stress	E.coli k-12 MG1655	MOPS, 0.2 % glucose	leu starvation	GSE51052	GSM1399610

Table 7.1: Control samples chosen for comparative analysis belonging to different GEO Series (Dataset 1-8). GEO Series ID/GEO Sample ID underlained represent the control samples that will be replaced by the "stressed" samples belonging to the same GEO Series. Column 1: ID Dataset. Column 2: Genotype. Column 3: Culture media. Columns 4 and 5: Samples coordinates (GEO Series ID and GEO Sample ID.

namely "control" and "stressed", testing the bed files with only the 40 reproducible genes identified previously. To balance the analysis between control and stress samples, we create a new control group: it consists of the eight samples analyzed (our benchmark constitutes of 40 genes) and a duplicate of the sample not stressed (GSM2396722) belonging to the same GEO Series (GSE90056) of dataset stressed that will be examined. It turned out that only 30 genes obtained from the new control group have reproducible Ribo-seq profile.
Then, we performed the "stressed" experiment testing the dataset GSE90056-GSM2396724 against our benchmark of 30 genes belonging to the new control group (excluding the duplicate sample). Due to using the Benjamini-Hochberg method, we are aware that adding a duplicate dataset reduces the number of reproducible genes. This is a point we are working on. However, even with this limit, we obtained promising results. Once the sample GSM2396724 are challenged against the benchmark, we found that out of 30 reproducible genes that are in common to all eight control datasets, the 24 genes listed in Table 7.2 have significantly reproducible Ribo-seq profiles. Following the same strategy described above, we compared the sample GSM2396726, subjected to heat-shock (42°C) for 20 min, against our control group. In this comparative analysis, only 30 ORFs obtained from the previous analysis are considered. The table summarises the results of this investigation. According to these results, the different stress condition is a variable to consider in terms of experimental reproducibility. Indeed, up to our results, when the cells are exposed to heat-shock for 20 minutes, the number of reproducible Ribo-seq profiles falls to 19 (see Table 7.3). We observe a significant drop in the number of reproducible Ribo-seq profiles changing from control condition to stressed condition. Furthermore, the increase of exposure time to heat-shock, from 10 min to 20 min, significantly reduced the reproducible Ribo-seq profiles. Thus, our results indicate that translation under heat-shock condition is significantly differentially regulated. As can be seen from Tables 7.2 and 7.3, ompC gene is no longer reproducible when E.coli is exposed to heat stress condition (both for 10 and 20 minutes). Figure 7.2 shows the Riboseq profiles of ompC gene, relating to control condition (top plot), heat shock for 10 minutes (middle plot) and heat shock for 20 minutes (bottom plot).
We report the pvalue matrix obtained for ompC gene. The table 7.4 shows the pairwise comparison between control samples (new control benchmark), while the table 7.5 and table 7.6 report the pairwise comparison between control samples ad heat-shock samples for 10 minutes and 20 minutes, respectively. As evidenced by pvalue observed, ompC is not longer reproducible in stressed condition.

Genes	Annotation
rodZ	Cytoskeleton protein
dnaX	DNA polymerase III subunit tau
glnA	Glutamine synthetase
gltB	Glutamate synthase NADPH large chain
infB	Translation initiation factor IF-2
katG	Catalase-peroxidase
metG	Methionine–tRNA ligase
rne	Ribonuclease E
sucA	2-oxoglutarate dehydrogenase
tufA	Elongation factor Tu 1
tufB	Elongation factor Tu 2
hokB	Toxic component of a type I toxin-antitoxin (TA) system
ubiJ	Ubiquinone biosynthesis protein
lptD	LPS-assembly protein
rpnC	Recombination-promoting nuclease
rpnA	Recombination-promoting nuclease
fdoG	Formate dehydrogenase-O major subunit
wbbH	O-antigen polymerase
wbbI	Beta-1,6-galactofuranosyltransferase
rpnE	Inactive recombination-promoting nuclease-like protein
lpoA	Penicillin-binding protein activator
rsxC	Electron transport complex subunit
yfcI	Recombination-promoting nuclease
gtrS	Uncharacterized protein YfdI

Table 7.2: Set of Ribo-seq profiles that resulted to be reproducible independently of heat-shock temperature of 42°C, 10 minutes

7.2 Impact of amino-acid starvation:

We reiterated our comparative analysis strategy to investigate the effect of leucine starvation on translational control.
More in detail, we considered the sample GSM1399610 belonging to GSE51052 Series that refers to Riboseq data obtained from E.coli k-12 MG1655, characterized by 30 min of leucine starvation. Generally, amino acid starvation can decrease the elongation rate of ribosomes, influencing the expression levels of the protein as evidenced in [74]. In addition, this stress condition may impact the expression levels of the protein due to the reduction of aminoacyl-tRNA concentration [112]. Once again, to balance the analysis, we generated the control group consisting of the eight datasets not stressed and a duplicate of dataset GSM1399615 belonging to the same Series of dataset GSM1399610. For all of them, we generated the Ribo-seq profiles corresponding to the 40 ORFs identified previously

Genes	Annotation
rodZ	Cytoskeleton protein
dnaX	DNA polymerase III subunit tau
gltB	Glutamate synthase NADPH large chain
infB	Translation initiation factor IF-2
metG	Methionine–tRNA ligase
secY	Protein translocase subunit SecY
rne	Ribonuclease E
sucA	2-oxoglutarate dehydrogenase
tufA	Elongation factor Tu 1
tufB	Elongation factor Tu 2
hokB	Toxic component of a type I toxin-antitoxin (TA) system
ubiJ	Ubiquinone biosynthesis protein
lptD	LPS-assembly protein
rpnC	Recombination-promoting nuclease
fdoG	Formate dehydrogenase-O major subunit
wbbH	O-antigen polymerase
yfjI	Uncharacterized protein YfjI
rsxC	Electron transport complex subunit
yfcI	Recombination-promoting nuclease

Table 7.3: Set of Ribo-seq profiles that resulted to be reproducible independently of heat-shock temperature of 42°C, 20 minutes

Control vs Dataset 1	Control vs Dataset 3	Control vs Dataset 4	Control vs Dataset 5	Control vs Dataset 6	Control vs Dataset 7	Control vs Dataset 8
5.10E-05	1.22E-07	8.56E-06	0.002080224	0.000757534	0.002002657	0.000824481

Table 7.4: p-value matrix referring to ompC gene of *E.coli* (Control, GSE90056-GSM2396722, Dataset 2). The columns contain p-values associated to each pairwise comparison.

in chapter 4. According to our method, we generated the corresponding digitalised profiles and then we compared them pairwise. Once we mapped the similarity scores on the corresponding null distributions, we obtained 27 reproducible Ribo-seq profiles which are used as a benchmark for the subsequent analysis. To perform the "stressed" experiment, we substituted the duplicate sample with the corresponding "stressed" dataset GSM1399610 and we compared it against our benchmark.
Finally, we turned that out of 27 possible Ribo-seq profiles only 13 found out to be reproducible, using a FDR threshold of 0.01. We interpreted these results, summarized in Table 7.7, highlighting that the genes that are no longer reproducible are those that, due to stress, have been influenced by the

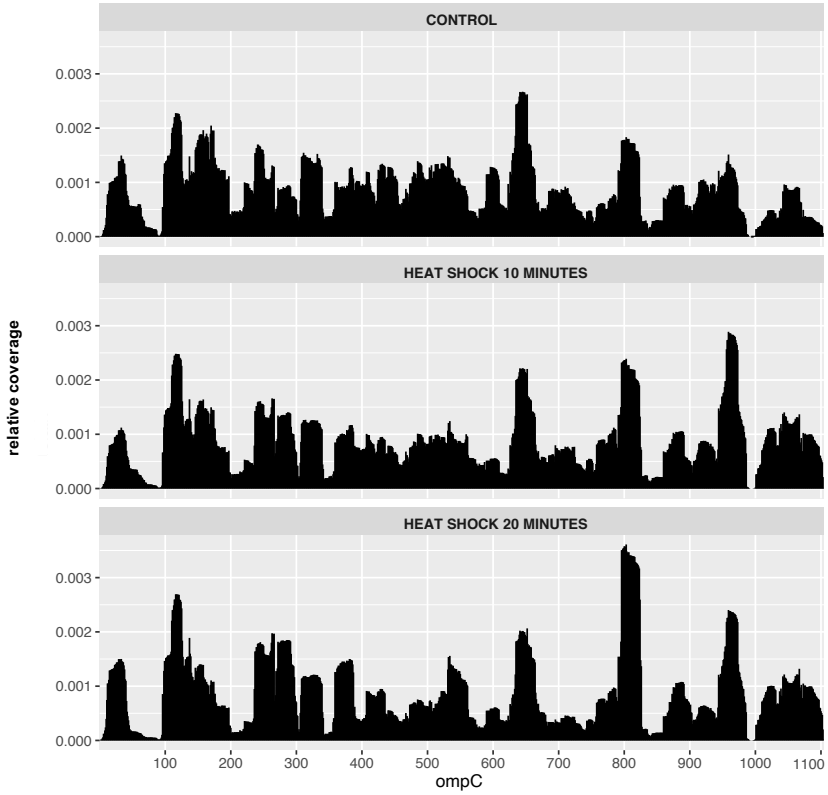


Figure 7.1: Ribo-seq profiles of ompC gene. Top plot depicts the control condition, middle plot the heat shock for 10 minutes, while the bottom plot represents the heat shock for 20 minutes.

Shock 10 vs Dataset 1	Shock 10 vs Dataset 3	Shock 10 vs Dataset 4	Shock 10 vs Dataset 5	Shock 10 vs Dataset 6	Shock 10 vs Dataset 7	Shock 10 vs Dataset 8
0.059723569	0.008329774	0.024878088	0.413734806	0.285821703	0.791321552	0.631009377

Table 7.5: p-value matrix referring to ompC gene of *E.coli* (Shock 10, GSE90056-GSM2396724,42°C for 10 minutes). The columns contain p-values associated to each pairwise comparison.

translational control. Starvation of leucine caused a pronounced change in the distribution of RPFs along the ORFs, as illustrated in Figure 7.2.

Following the same procedure of heat shock comparison, based on the pvalue matrix of ompC gene (see Table 7.8 and Table 7.9), we can state that this gene is not longer reproducible whether *E.coli* grows under leucine starvation conditions for 30 minutes.

Shock 20 vs Dataset 1	Shock 20 vs Dataset 3	Shock 20 vs Dataset 4	Shock 20 vs Dataset 5	Shock 20 vs Dataset 6	Shock 20 vs Dataset 7	Shock 20 vs Dataset 8
0.018967262	0.004085189	0.069977386	0.654819705	0.642425128	0.943959924	0.914184849

Table 7.6: p-value matrix referring to ompC gene of *E.coli* (Shock 20, GSE90056-GSM2396726, 42°C for 20 minutes). The columns contain p-values associated to each pairwise comparison

.

Gene	Annotation
rodZ	Cytoskeleton protein
dnaX	DNA polymerase III subunit tau
gltB	Glutamate synthase, large chain
metG	Methionine–tRNA ligase
tufA	Elongation factor Tu 1
tufB	Elongation factor Tu 2
hokB	Small toxin membrane polypetide
ubiJ	Ubiquinone biosynthesis protein
lptD	LPS-assembly protein
rpnC	Recombination-promoting nuclease
wbbH	O-antigen polymerase
rpnE	Inactive recombination-promoting nuclease-like protein
rsxC	Electron transport complex subunit

Table 7.7: Genes with significantly reproducible Ribo-seq profiles under leucine starvation.

Control vs Dataset 1	Control vs Dataset 2	Control vs Dataset 3	Control vs Dataset 4	Control vs Dataset 6	Control vs Dataset 7	Control vs Dataset 8
0.00321645	3.62E-05	0.00175841	0.00010679	0.00039528	0.00691943	0.00410076

Table 7.8: p-value matrix referring to ompC gene of *E.coli* (Control, GSE51052-GSM1399615, Dataset 5). The columns contain p-values associated to each pairwise comparison.

Leu stress vs Dataset 1	Leu stress vs Dataset 2	Leu stress vs Dataset 3	Leu stress vs Dataset 4	Leu stress vs Dataset 6	Leu stress vs Dataset 7	Leu stress vs Dataset 8
0.51536943	0.0334051	0.28700529	0.36798143	0.48971448	0.57416258	0.16804795

Table 7.9: p-value matrix referring to ompC gene of *E.coli* (Leu stress, GSE51052-GSM1399610). The columns contain p-values associated to each pairwise comparison.

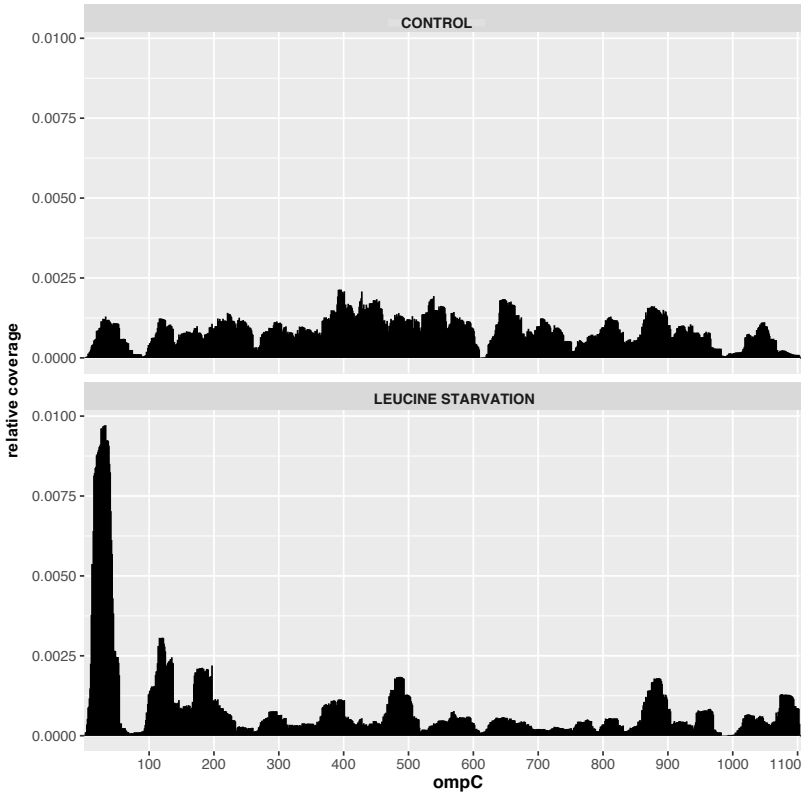


Figure 7.2: ompC Ribo-seq profiles control vs leucine starvation

7.3 Discussion

Once we have substituted the control dataset with the stressed dataset belonging to the same GEO Series, we applied our method to perform a comparative analysis between two different conditions. The results confirm that our method can be used for comparative analyses, allowing us to identify biological differences in the comparative Ribo-seq profiles due to the translation being regulated differently. In the table 7.5, and 7.6, we can observe that the p-values obtained comparing pairwise the control samples with the stressed samples are not significant.

Outmembran porines (ompC and ompF) are expressed when *E.coli* is grown at 37°C. The upregulation of ompC and ompF can occur through responses elicited by osmolarity, pH, ionic strength, and temperature [113]. In particular, their expression levels are controlled by the ompB regulon, which is comprised of positive transcriptional regulator OmpR and an inner membrane sensor histidine kinase EnvZ [114]. The role of kinase EnvZ in response to osmotic stress is known [115]. At high osmolarity, EnvZ autophosphorylates and transfers

the phosphoryl group to the regulator OmpR (aspartate residue), leading to conformational changes. OmpR-P then binds to the promoter regions of the porin genes ompF and ompC and activate their transcription.
The outer membrane is the first barrier between E.coli and the surrounding environment. When the cells are exposed to different stress condition like starvation or temperature changes, the expression of outmembran porines represents a crucial factor in determining the survival of bacterial cells. It is known that the rate of omps protein in the outer membrane is correlated to several factors, including the growth temperature, as evidenced in [116]. In addition, porine genes are subject to complex post-transcriptional regulation by a variety of siRNA molecules, including MicC, RseX and RybB. [117]. It has been demonstrated that MicC inhibits the binding of the small subunit of the ribosome to ompC mRNA, suggesting that MicC is able to prevents the translation initiation [118].
Considering that we assessed the comparative analysis starting from 40 Ribo-seq profiles as a benchmark and only three "stressed" samples, more extensive data sets are required to exploit the power of the analyses, in future studies. Further analysis would be needed to achieve a comprehensive overview of translation regulation in response to a stress condition.

Chapter 8

The human case-study: liver tumours vs their adjacent non-cancerous liver tissues

In this section, we report a preliminary analysis of Riboseq profiles referring to liver tumours and their adjacent noncancerous normal liver tissues from 10 patients with hepatocellular carcinoma (HCC) [119].
The data are stored in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJNA448763. The coordinates for these datasets are reported in Table 8.1.

SRA ID	Dataset
SRR6939924	Dataset 1 normal
SRR6939926	Dataset 1 tumor
SRR6939928	Dataset 2 normal
SRR6939930	Dataset 2 tumor
SRR6939932	Dataset 3 normal
SRR6939934	Dataset 3 tumor
SRR6939936	Dataset 4 normal
SRR6939938	Dataset 4 tumor
SRR6939940	Dataset 5 normal
SRR6939942	Dataset 5 tumor
SRR6939944	Dataset 6 normal
SRR6939946	Dataset 6 tumor
SRR6939948	Dataset 7 normal
SRR6939950	Dataset 7 tumor
SRR6939952	Dataset 8 normal
SRR6939955	Dataset 8 tumor
SRR6939957	Dataset 9 normal
SRR6939959	Dataset 9 tumor
SRR6939961	Dataset 10 normal
SRR6939963	Dataset 10 tumor

Table 8.1: Summary of the ribosome profiling data from HCC patients

The pre-processing procedure of the ribosome profiling data has been described in Chapter 2. The only difference concerning the pre-processing analysis performed on *E.coli* data is related to the read quality score. In this case, low-quality scores lower than 20 are removed, using trimgalore tool. This choice is attributed since the datasets under investigation are tissues, and they result more degraded than the *E.coli* cell cultures.

The goal of this analysis upon the thesis is centered is to provide an effective method that can be applied to each kind of Ribo-seq dataset including those related to complex organisms such humans. In this chapter, our method delivers a unique set of genes for each condition under investigation (i.e. normal and cancer tissues) that have a robust and reproducible ribosome profiles. The obtained high resolution Ribo-seq profiles libraries serve as reliable benchmarks to explore conditions potentially affecting translation control and to detect possible differential translation events which occur during tumorigenesis.

8.0.1 Preliminary results

Firstly, our novel data analysis approach is applied to the human control datasets and then to the cancer datasets, in order to identify the reproducible Ribo-seq profiles for each condition and then to compare them.
According to our method described in Chapter 4, we counted in each row how many p-values resulted significant according to the BH method. In this case, we defined reproducible those Ribo-seq profiles referring to the rows with all the p-values are below a chosen significance threshold ($p < 0.01$).
All the data analysis results are included in the appendix, chapter 11. Following the strategy described in Chapter 4, we found that, out of 1045 genes that are in common to all ten control datasets, the 49 genes listed in Table 11.2 are reproducible. Interestingly, from 3306 genes in common to all ten cancer dataset, 138 Ribo-seq profile are defined as significantly reproducible (Table 11.3). We report here an illustrative example of a reproducible gene identified in the control group, Ornithine carbamoyltransferase (OTC) (See Figure 8.1). OTC is reproducible in the control group but no longer in the tumour group. The Figure 8.2 shows the comparison of OTC Ribo-seq profiles referred to Dataset 1 (control and adjacent cancer sample). In the control sample, the reads seem uniformly distributed across the entire length of the ORF while, in their adjacent cancer sample the distribution of ribosome occupancy results in two isolated footprint peaks along the ORF. OTC is found exclusively specific to the liver mitochondria [120]. It is an enzyme that participates in the urea cycle to detoxify the ammonia produced from amino acid catabolism. Several studies have suggested that accumulated ammonia resulting from OTC deficiency causes chronic liver damage, a potential risk factor of HCC.[121]. It also has been observed that OTC expression is significantly downregulated in HCC [122]. Further analyses are necessary to investigate the role of OTC in the pathogenesis of HCC.
As mentioned above, there is a significant difference between the genes in common between control and cancer datasets that may be related to an

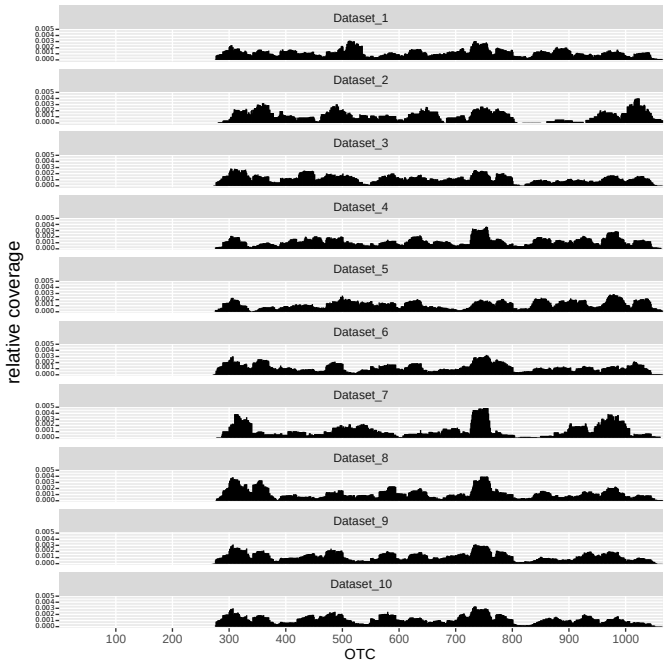


Figure 8.1: OTC Ribo-seq profile across all ten control dataset

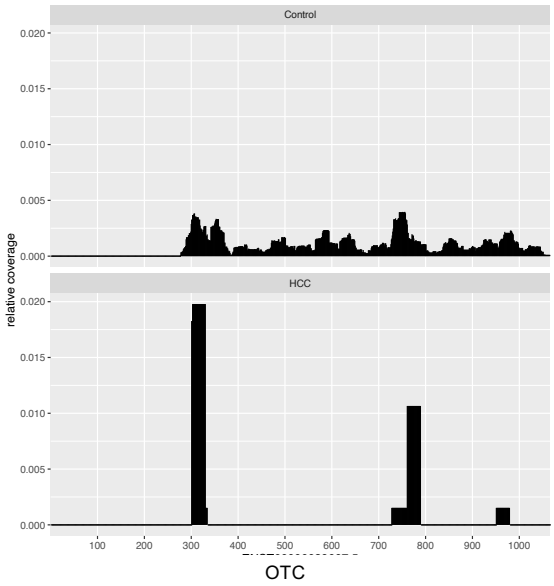


Figure 8.2: OTC Ribo-seq profile between controls vs adjacent cancer tissue (Dataset 1)

experimental issue or a specific pattern expressed by cancer-related genes. As evidenced in [119], it has been observed that the most abundant reads (peaks)

are shifted by 1 or 2 nucleotides between different patients. This occurs because the tissue samples from different patients are collected at different times and then subjected to different digestion efficiency by RNase I treatment. Another explanation can be traced back to varying amounts of starting material of RNA. However, to deepen the reason for these results, we performed an over-representation statistics analysis of reproducible cancer genes using Panther tool [80].
Over-representation analysis determines whether genes from predefined sets (human genes) are found more than would be expected in a subset of our data. In our case, the main objective is to observe whether exist a correlation between reproducible genes and pathway associate with liver cancer. Then, we compared our reproducible cancer gene list to a reference gene list represented by human genes, to determine whether a particular class (e.g. Reactome pathway) of genes is over-represented or under-represented. We observed that 66 pathways are significant, with a FDR of 0.05. For the results of the over-representation analysis, see appendix Table 11.4. If we consider Reactome pathways as a category under investigation and more genes are observed in the test list than expected, we have an over-representation of genes involved in specifics pathways. Otherwise, if fewer genes are observed than expected, we have an under-representation. Interestingly, genes involved in *Signaling by NOTCH4* (R-HSA-9013694) are over-represented (p= 2.87E-04). NOTCH4 is prevalently expressed in endothelial cells [123]. In the liver, this pathway is involved in biliary tree development and tubulogenesis and its dysregulation has been observed as a determinant in the development of HCC [124]. A recent study has highlighted the role of Notch pathway in liver cells transition to the mesenchymal phenotype [125]. We next investigated the disease association of the reproducible cancer genes using the gene–disease association network (DisGeNET) [126] in EnrichR [127] plugin, in order to demonstrate that these genes express a specific liver cancer pattern. Disease enrichment analysis showed liver carcinoma is the most high-ranked (p = 2.16E-13), as illustrated in Figure 8.3.
This result indicates the existence of a specific expression fingerprint of these genes that are shared in patients with HCC cancer.

8.0.2 Future works

For what concerns machine learning analysis, once we have obtained the consensus sequences of reproducible Ribo-seq profiles, we carried on a preliminary analysis exploiting the same neural architectures optimized for the *E.coli* datasets. Unfortunately, this analysis did not result in remarkable performances. Given

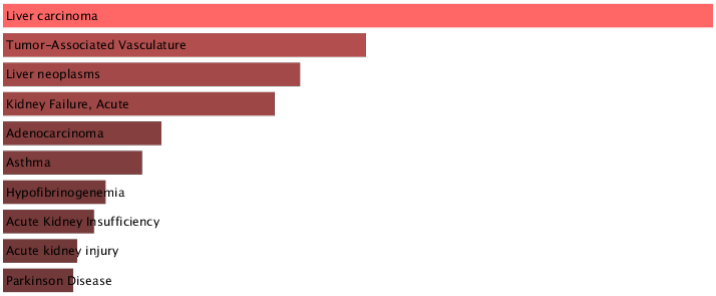


Figure 8.3: Bar graph of top ten enriched disease terms across input reproducible cancer genes, sorted by p-value ranking.

the increased complexity of the human dataset task, we believe that through further analysis, in particular defining ad-hoc neural architectures and with a specialized hyperparameter search, the performances could be greatly improved. Considering that we analyzed Ribo-seq data from 10 liver tumours and their adjacent non-cancerous normal liver tissues, future works will be focused on the assessment of the reproducibility of Ribo-seq experiments from different dataset. Our aim is to analyze samples coming from human cell cultures because they are more homogeneous than tissue: when tissue is taken during a surgical biopsy procedures, it is difficult to have a uniform sample (e.g., it is unavoidable to take parts of the stroma and vessels with the parenchyma of an organ). In addition, during a biopsy or surgery, it is not easy to properly store the tissue before analyzing it.
Future works will be centered to perform a comparative transcriptome and translatome analysis in cancerous vs normal cells, in order to reveals different gene regulatory mechanisms at the transcriptional and translational level between the two conditions. Indeed, several studies in mammalians cells have shown that although most stress responsive genes are regulated at the transcriptional level, mRNA abundance is not always a good proxy of protein concentration [128] [129]. This suggest that the transcriptome needs to be studied in conjunction with translatome and proteome, in order to unravel molecular insights into gene regulatory mechanisms involved in development of cancer.

Chapter 9

Discussion

The Ribosome Profiling technique represents the most advanced tool able to exploits deep sequencing to study the translation of gene expression. Ribo-seq provides a measurement for how the translation is regulated, what is being translated and where a specifics protein is translated. In this dissertation, we have described a new data analysis approach that allows to address the limitations that affect the reproducibility of Ribo-seq experiments. Moreover, we have shown that our method can identify a set of significantly reproducible ribo-seq profiles coming from the comparison of independent Ribo-seq experiments. Although the low number of profile (40 genes), this library represents a comprehensive workbench for comparative experiments aimed to study the factors that influence the translation process.

The consensus sequences built from the 40 reproducible Ribo-seq profiles are labelled with +1 and -1 and correspond to fast and slow regions, respectively. The purpose of this thesis is to verify the existence, within the sequence of the gene, of signals or nucleotide patterns capable to influence the efficiency of translation. Based on several hypotheses present in the literature, to justify the variations of the ribosome speed, we have proposed an investigation on the nucleotide composition in slow and fast sequences. The nucleotide frequencies observed in the fast and slow subsequences are statistically significant and it is highly unlikely to find them by chance.

Our results suggest that the nucleotide composition of the subsequences constitute useful information to discriminate the fast from the slow subsequences. Leveraging such intuition, we exploited statistical information on the nucleotide frequencies to train simple artificial neural networks to predict the nature of the subsequences. In particular, a MLP network is capable to leverage such information in order to predict the nature of the subsequences (i.e., with 75.67% average test accuracy) to distinguish slow and fast subsequences.

Moreover, training a 1-D Convolutional neural network directly on the nucleotide subsequences allowed us to achieve an accuracy of about 89.39% in the same task, confirming the significance of our hypothesis. An increase of performance was expected compared to the frequency-based prediction, because in this

latter approach the information on nucleotides is enriched thanks to sequential and contextual information. The usage of more complex architectures (i.e., 7-CNN ensemble model) improves the prediction of two percentage point (91%). We believe that this represents a good prospect for further analysis, opening the road to an improvement in performances leveraging ad-hoc neural architectures and a specialized hyperparameter search. Considering that we analyzed a relatively small-sized dataset which could hinder our machine learning based method, future works will be devoted to studies on more extensive datasets, in order to fully leverage the representational capability of machine learning approaches. Understanding the factors that affect translation speed, what is the origin of peaks and valleys that we can observe by looking the ribo-seq profiles obtained, remains still unclear. This kind of question could be addressed by considering multiple options such as the presence of optimal or non-optimal codon in the ORF, mRNA secondary structure downstream, mRNA stability, translation pause sites, ribosomal binding sites (e.g IRES), and codon context. Preliminary results on *E.coli* genome revealed that codon-context is a variable to take into account to predict the translation speed. The performance of the experiments changes depending on the length of context chosen. Based on the results obtained using LSTM architecture, too long or too short contexts seems not to be informative.
Furthermore, in Chapter 7, we have shown that our data analysis approach can be used for comparative analysis. The same strategy can be easily applied to other experimental variables in order to detect possible differential mechanisms of translation regulation.
In Chapter 8, our approach has been proved to be very efficient because it allows us to compare large datasets, such as the human one, in a short time. Although the results of data analysis approach are promising, further analyses are required to optimize the neural architectures due to the complexity of the human dataset task. Finally, it is worth mentioning that our method represents an effective approach applied to each kind of Ribo-seq dataset, to investigate an extremely relevant open questions in biology, such the features which could influence the velocity of the ribosome during translation. The application of machine learning-based methodologies has provide an unprecedented point of view in this context, focusing on the role of the context in determining the translation rate. We proposed an innovative methodology bridging two different scientific areas (i.e. biology and data science) and a clear applicative perspective of the obtained results is found in the biotechnology field. It is known that the levels of expression of a target protein can be optimized by several approaches [130] [131] and our methodologies could represent an

important resource for the improvement of translational efficiency. When we don't have information from the Ribo-seq data analysis approach, we can use machine learning methods to predict fast and slow unlabelled translated region. Starting from a small number of reproducible genes, the implementation of efficient neural network architectures allow us to classify all *E.coli* transcriptome. Genetic engineering can exploit these findings in order to understand which region of a not identified ORF is fast or slow, with the purpose to optimise or slow down translation, for instance by changing synonymous codons based on an organism's codon bias.

Chapter 10

Summary of additional research topics

This chapter reviews the research conducted during the PhD period, but not directly covered by this thesis.

10.1 Graph Neural Networks for the Prediction of Protein–Protein Interfaces

More details on this work are available in [132].
Binding site identification allows to determine the function-ality and the quaternary structure of protein–protein complexes. Various approaches to this problem have been proposed without reaching a viable solution. Representing the interacting peptides as graphs, a correspondence graph describing their interaction can be built. Finding the maxi-mum clique in the correspondence graph allows to identify the secondary structure elements belonging to the interaction site. Although the maximum clique problem is NP-complete, Graph Neural Networks make for an approximation tool that can solve the problem in affordable time. Our experimental results are promising and suggest that this direction should be explored further.

10.2 Deep Learning Techniques for Dragonfly Action Recognition

More details on this work are available in [133].
Anisoptera are a suborder of insects belonging to the order of Odonata, commonly identified with the generic term dragonflies. They are characterized by a long and thin abdomen, two large eyes, and two pairs of transparent wings. Their ability to move the four wings independently allows dragonflies to fly forwards, backwards,to stop suddenly and to hover in mid–air, as well as to achieve high flight performance, with speed up to 50km per hour. Thanks to these particular skills, many studies have been conducted on dragonflies, also using machine learning techniques. Some analyze the muscular movements of the flight to simulate dragonflies as accurately as possible, while others try

to reproduce the neuronal mechanisms of hunting dragonflies. The lack of a consistent database and the difficulties in creating valid tools for such complex tasks have significantly limited the progress in the study of dragonflies. We provide two valuable results in this context: first, a dataset of carefully selected, pre–processed and labeled images, extracted from videos, has been released; then some deep neural network models, namely CNNs and LSTMs, have been trained to accurately distinguish the different phases of dragonfly flight, with very promising result.

10.3 AKUImg: A database of cartilage images of Alkaptonuria patients

More details on this work are available in [134]. ApreciseKUre is a multi-purpose digital platform facilitating data collection, integration and analysis for patients affected by Alkaptonuria (AKU), an ultra-rare autosomal recessive genetic disease. We present an ApreciseKUre plugin, called AKUImg, dedicated to the storage and analysis of AKU histopathological slides, in order to create a Precision Medicine Ecosystem (PME), where images can be shared among registered researchers and clinicians to extend the AKU knowledge network (See figure 10.1).
AKUImg includes a new set of AKU images taken from cartilage tissues acquired by means of a microscopic technique. The repository, in accordance to ethical policies, is publicly available after a registration request, to give to scientists the opportunity to study, investigate and compare such precious resources. AKUImg is also integrated with a preliminary but accurate predictive system able to discriminate the presence/absence of AKU by comparing histopatological affected/control images. The algorithm is based on a standard image processing approach, namely histogram comparison, resulting to be particularly effective in performing image classification, and constitutes a useful guide for non-AKU researchers and clinicians.

10.4 A Transcriptional Study of Oncogenes and Tumor Suppressors Altered by Copy Number Variations in Ovarian Cancer

More details on this work are available in [135] The most popular approach to explain cancer is based on the discovery of oncogenes and tumor suppressor genes as a preliminary step in estimating their impact on altered pathways.

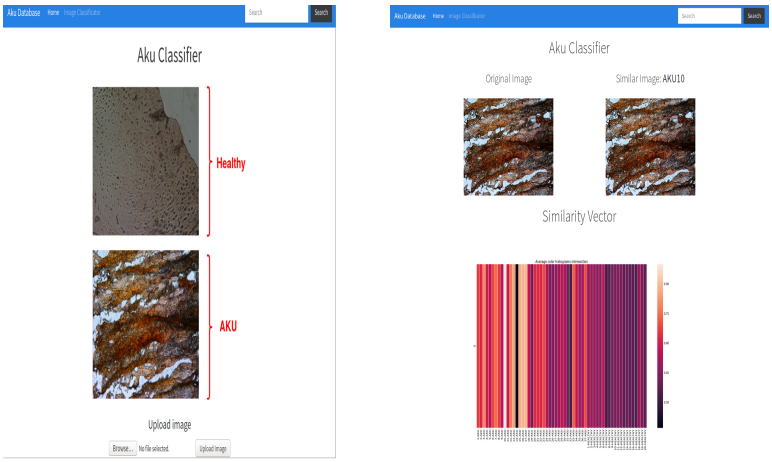


Figure 10.1: A snapshot of the home page of the ApreciseKUre database (left) and a prediction example (right). In particular, an AKU image was fed into the prediction model, which reports similarity values for all AKU images in the dataset (lighter colors indicate a higher similarity).

The present paper proposes a pipeline which aims at detecting "weak" or "indirect" functions impacted by Copy Number Variations (CNVs) of cancer-related genes, integrating such signals over all known oncogenes/tumor suppressor genes of a cancer type. We applied the pipeline to the task of detecting the aberrant functional effects of these alterations across ovarian cancer patients from The Cancer Genome Atlas (TCGA) data.

10.5 Analysis of brain NMR images for age estimation with deep learning

More details on this work are available in [136] During the last decade, deep learning and Convolutional Neural Networks (CNNs) have produced a devastating impact on computer vision, yielding exceptional results on a variety of problems, including analysis of medical images. Recently, these techniques have been extended to 3D images with the downside of a large increase in the computational load. In particular, state-of-the-art CNNs have been used for brain Nuclear Magnetic Resonance (NMR) imaging, with the aim of estimating the patients' age. In fact, a large discrepancy between the real and the estimated age is a clear alarm for the onset of neurodegenerative diseases, such as some types of early dementia and Alzheimer's disease. In this paper, we propose an effective alternative to 3D convolutions that guarantees a significant reduction

of the computational requirements for this kind of analysis. The proposed architectures achieve comparable results with the competitor 3D methods, requiring only a fraction of the training time and GPU memory (See figure 10.2).

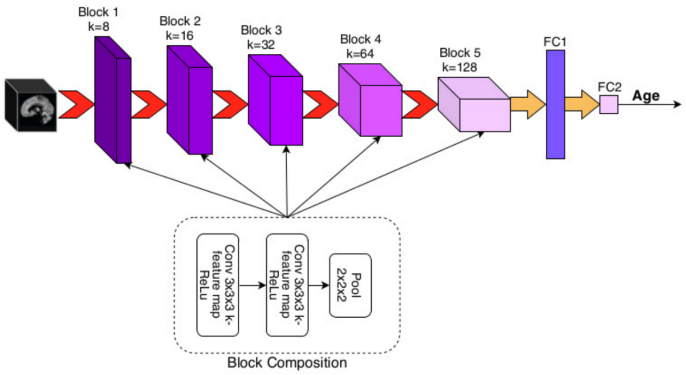


Figure 10.2: 3D–CNN architecture proposed for predicting age from NMR brain images

10.6 Fusion of Visual and Anamnestic Data for the Classification of Skin Lesions with Deep Learning

More details on this work are available in [137] Early diagnosis of skin lesions is essential for the positive outcome of the disease, which can only be resolved with surgical treatment. In this manuscript, a deep learning method is proposed for the classification of cutaneous lesions based on their visual appearance and on the patient's anamnestic data. These include age and gender of the patient and position of the lesion. The classifier discriminates between benign and malignant lesions, mimicking a typical procedure in dermatological diagnostics. Good preliminary results on the ISIC Dataset demonstrate the importance of the information fusion process, which significantly improves the classification accuracy.

Chapter 11

Appendix

	1 vs 2	**1vs3**	**1 vs 4**	**1 vs 5**	**1 vs 6**	**1 vs 7**	**1 vs 8**	**2 vs 3**	**2 vs 4**
rodZ	0,000302	9,9E-10	4,65E-05	0,001141	1,63E-05	0,000622	0,001052	7,05E-05	5,59E-05
arcB	0,016648	2,35E-08	0,001357	0,003275	1,15E-08	0,363792	0,013135	0,000223	9,99E-06
dld	0,002944	1,11E-16	4,91E-09	0,000046	7,71E-08	0,000353	0,007627	0,000185	0,002696
dnaX	4,96E-05	0	9,16E-09	7,61E-09	2,6E-08	0,039177	1,01E-06	0,000139	0,002955
fhuA	0,000314	0	8,3E-12	0,051982	0,001026	0,005756	0,000436	0,004054	5,78E-05
glnA	0,004247	0	6,11E-06	0,0004	4,43E-08	0,000251	0,000209	0,007783	0,000486
gltB	1,58E-09	0	3,5E-09	6,52E-08	2,05E-09	1,51E-09	5,49E-08	2,23E-12	8,92E-13
hisS	0,164359	1,11E-15	7,59E-07	1,16E-10	0,000225	9,39E-07	0,000166	0,024632	7,95E-05
infB	0,00027	0	1,61E-11	0,000115	1,38E-06	1,27E-07	4,31E-09	1,09E-05	7,27E-06
katG	0,000128	0	2,93E-09	3,82E-05	3,94E-11	1,05E-05	0,000158	2,75E-06	1,14E-07
malF	0,000209	4,45E-14	0,000284	0,000199	0	2,62E-14	0	0,011484	0
metG	0,00466	0	7,16E-10	7,81E-05	5,04E-05	3,21E-06	3,55E-06	0,008396	0,001194
mukB	0,000225	0	7,1E-11	5,36E-06	2,28E-06	0,000513	2,4E-06	5,28E-05	1,18E-05
ompC	0,000122	6,22E-15	0,000281	0,00038	4,32E-07	8,82E-05	0,000152	9,01E-08	8,63E-06
parC	0,005318	0	6,62E-08	0,003695	2,28E-05	3,47E-06	3,04E-08	4,95E-06	3,44E-05
secY	0,000179	0	3,72E-05	9,91E-05	6,58E-06	0,013403	0,009769	2,93E-05	8,08E-08
purL	0,008598	0	5,82E-09	0,006168	4,41E-08	6,5E-07	2,92E-07	0,000523	0,016132
rne	0,000197	0	1,1E-10	4,16E-05	6,37E-08	5,06E-05	7,72E-05	8,99E-07	2,25E-10
sucA	0,0074	0	2,77E-07	4,64E-05	1,59E-07	0,000833	0,000491	0,000583	6,99E-12
tufA	0	0	0	0	0	0	0	0	0
tufB	0	0	0	0	0	0	0	0	0
leuA	0,000361	3,08E-10	0,000703	0,000444	0,000483	0,000188	0,000177	1,07E-05	6,28E-06
hokB	0,001322	0,000256	0,001418	0,001901	0,003894	0,002409	0,001682	0,001385	0,001209
acnA	2,77E-05	0	0,035974	0,000556	6,99E-08	5,6E-09	3,14E-07	0,000378	0,004064
ubiJ	2,78E-05	1,05E-10	4,73E-08	6,05E-05	0,000016	2,41E-08	4,73E-09	2,57E-06	5,06E-05
lptD	0,004531	0	1,09E-09	0,000161	1,26E-07	4,62E-05	4,08E-05	0,00011	1,27E-06
rpnC	1,22E-15	0	0	0	0	1,35E-10	0	5,22E-14	2,22E-16
rpnA	1,73E-14	1,8E-12	1,34E-13	0	2,02E-14	2,22E-16	6,41E-14	2,57E-10	1,83E-10
fdoG	3,64E-05	0	0,000645	0,567612	3,63E-10	2,9E-08	1,68E-08	6,28E-06	8,08E-09
wbbH	1,9E-06	4,96E-14	7,03E-05	7,42E-07	6,72E-05	0,000273	0,000249	5,99E-09	5,93E-07
wbbI	0,000019	9,44E-15	1,92E-06	0,008666	0,003151	0,004661	0,000234	2,28E-07	3,59E-09
wbbK	1,57E-05	0	2,96E-07	0,008659	1,94E-10	0,001045	0,002247	0,000143	0,000477
rpnE	0	1,22E-15	1,55E-15	0	0	0	0	5,55E-16	1,5E-14
lpoA	0,00174	0	4,78E-05	0,006497	0,000328	0,03418	0,018665	1,97E-05	0,000817
gspD	0,116793	1,95E-13	0,128488	0,118626	8,39E-08	9,62E-07	1,68E-06	0	0
yfjI	0	5,27E-13	0	0	0	0	0	0,081787	0
rlmL	1,04E-05	0	4,38E-08	8,34E-05	7,42E-05	2,28E-05	4,12E-06	9,43E-08	6,29E-07
rsxC	2,91E-06	0	1,75E-06	1,89E-05	4,86E-07	0,000131	0,002812	1,02E-06	0,000686
yfcI	6,35E-14	6,99E-15	3,97E-14	0,000219	2,51E-14	1,11E-16	0	0	1,4E-12
gtrS	8,01E-06	1,73E-14	1,28E-09	4,19E-06	5,75E-14	8,93E-05	0,00637	7,49E-06	3,55E-05

	2 vs 5	2 vs 6	2 vs 7	2 vs 8	3 vs 4	3 vs 5	3 vs 6	3 vs 7	3 vs 8
rodZ	0,000523	3,35E-07	9,25E-07	0,000177	4,47E-07	0,005145	4,99E-07	0,00026	0,001198
arcB	0,000533	0,000207	8,24E-06	0,000226	0,053975	3,56E-05	1,29E-08	0,023564	0,006714
dld	5,09E-05	0,000418	0,00553	0,011974	1,3E-08	1,85E-05	5,59E-09	0,020211	0,023751
dnaX	3,28E-07	3,95E-07	0,004477	0,001574	2,62E-06	5,98E-08	1,02E-08	0,003527	7,01E-06
fhuA	0,829726	0,001497	7,21E-06	0,000044	1,42E-14	0,005407	2,29E-05	0,000915	6,57E-05
glnA	0,017268	0,002062	0,001453	0,001327	6,75E-08	0,000834	1,26E-09	1,56E-05	3,35E-05
gltB	1,16E-11	9,72E-10	5,79E-12	1,69E-11	1,05E-12	1,57E-09	1,04E-10	2,21E-10	3,28E-12
hisS	0,195895	0,050289	0,003687	0,028351	2,11E-05	3,9E-09	9,31E-05	3,54E-06	0,000813
infB	0,000964	0,000492	7,76E-05	6,73E-05	2,29E-13	6,56E-06	1,06E-06	8,59E-06	2,69E-07
katG	0,039356	1,83E-05	4,54E-06	3,62E-06	1,69E-08	0,001595	2,16E-07	0,000133	0,001866
malF	0	0	5,08E-06	7,34E-12	0,008166	0,011794	0	1,88E-10	0
metG	0,073793	7,33E-05	0,002085	0,001178	1,58E-09	0,000112	2,55E-06	2,03E-09	5,74E-09
mukB	0,000296	1,03E-08	0,002244	0,082593	3,33E-15	3,64E-06	8,06E-10	0,002479	7,76E-05
ompC	0,001204	0,00081	0,00214	0,002844	2,82E-06	0,002592	3,11E-06	0,006417	0,000994
parC	0,001178	0,002148	0,505097	0,208047	6,27E-08	9,66E-05	4,33E-08	6,23E-06	6,7E-06
secY	0,00039	0,000195	0,002245	0,00439	6,39E-07	4,42E-05	5,59E-08	0,001641	0,000189
purL	0,05188	0,026687	0,382599	0,521956	3,99E-14	0,000422	3,21E-08	1,78E-09	7,65E-10
rne	6,74E-05	6,75E-13	2,33E-08	2,48E-10	4,37E-11	7,2E-06	1,13E-05	2,06E-06	2,63E-06
sucA	0,000453	7,45E-07	0,004906	0,002338	3,09E-08	1,56E-05	1,84E-09	3,41E-06	6,49E-06
tufA	0	0	0	0	0	0	0	0	0
tufB	0	0	0	0	0	0	0	0	0
leuA	0,115529	0,000055	0,000337	8,55E-05	0,001154	0,000646	0,005596	0,000141	7,52E-05
hokB	0,006072	0,003628	0,003169	0,000979	0,001706	0,001261	0,001044	0,001904	0,002157
acnA	0,00399	0,007659	0,004207	0,097685	0,00085	0,000318	6,64E-08	3,22E-08	1,73E-06
ubiJ	0,000155	7,06E-07	2,31E-07	4,11E-05	1,23E-07	2,62E-06	3,59E-07	9,94E-10	1,85E-10
lptD	0,001707	3,09E-05	0,015301	0,041123	3,56E-12	1,11E-09	1,64E-07	4,76E-06	0,000016
rpnC	2,22E-16	2,66E-15	4,22E-15	2,55E-15	4,44E-16	1,94E-14	7,66E-13	4,55E-15	4,33E-15
rpnA	8,48E-09	1,33E-15	1,11E-16	2,18E-11	3,33E-16	3,16E-07	3,33E-16	0	0
fdoG	0,006616	2,83E-09	4,82E-10	3,36E-05	0,001655	0,550734	7,13E-10	7,26E-10	5,41E-09
wbbH	6,06E-05	7,79E-08	0,006254	0,000122	3,77E-09	1,29E-06	3E-07	0,000108	1,06E-06
wbbI	0,000277	2,46E-05	0,005735	0,006206	2,65E-09	0,005022	0,005448	0,000288	0,000371
wbbK	0,059825	0,000365	0,040509	0,059113	9,7E-08	0,000726	1,38E-08	9,74E-07	2,99E-06
rpnE	2,25E-13	4,44E-16	6,66E-16	0	8,06E-14	9,14E-13	3,26E-14	0	3,11E-15
lpoA	0,010732	8,74E-06	0,000865	0,000153	0,000227	0,001268	4,3E-07	0,006436	0,000174
gspD	0	0	0	0	0	0	0	0	0
yfjI	0	0	0	0	0,078013	0,07842	5,3E-08	1,72E-09	1,13E-11
rlmL	0,143537	0,087541	8,84E-06	4,01E-05	1,8E-07	5,26E-05	0,001759	1,56E-05	0,000288
rsxC	6,87E-06	3,76E-08	0,003688	0,002035	1,53E-05	0,00038	5,44E-07	2,06E-06	0,000151
yfcI	3,73E-11	0	1,11E-16	0	0	2,71E-09	9,77E-15	2E-15	1,11E-16
gtrS	0,000517	1,11E-08	7,61E-05	0,000232	1,76E-10	1,4E-06	8,22E-12	0,000173	0,001364

	4 vs 5	4 vs 6	4 vs 7	4 vs 8	5 vs 6	5 vs 7	5 vs 8	6 vs 7	6 vs 8	7 vs 8
rodZ	9,11E-05	7,42E-05	0,041224	0,004966	1,88E-09	0,010968	0,000362	1,81E-06	1,61E-07	1,11E-16
arcB	1,78E-05	0,000126	0,004021	0,000721	4,57E-06	0,02243	0,000374	0,006743	0,001304	0
dld	4,77E-10	3,63E-08	0,000173	0,002606	9,12E-07	0,001296	0,00257	0,002538	0,002293	0
dnaX	3,55E-05	3,29E-08	0,025987	0,000335	1,14E-12	0,001396	0,00069	0,006071	8,7E-06	3,89E-15
fhuA	0,249669	0,001028	0,001031	1,02E-05	0,181241	0,512055	0,080773	0,007108	0,004644	0
glnA	0,006572	2,84E-05	0,006554	0,004529	8,23E-06	0,295516	0,140418	0,116553	0,030526	0
gltB	4,24E-08	1,56E-08	3,58E-06	1,97E-06	2,22E-16	1,12E-05	0,000447	1,06E-06	1,24E-06	0
hisS	0,000124	3,81E-05	0,002516	0,007929	2,32E-06	2,71E-05	0,002123	0,002231	0,004787	0
infB	7,88E-15	7,42E-10	0,000533	3,88E-06	3,01E-08	0,000169	1,22E-05	0,003769	8,61E-05	0
katG	1,77E-08	1,44E-13	0,004521	0,00539	1,97E-08	0,569418	0,676552	0,035674	0,030665	0
malF	0	0	5,51E-05	5,19E-09	0	0,000257	4,85E-07	0	0	0
metG	9,14E-05	1,98E-05	2,25E-06	1,34E-06	2,52E-06	0,025737	0,004816	0,000677	0,000147	0
mukB	8,79E-11	1,6E-07	0,043822	0,073723	2,48E-06	0,074124	0,088349	0,001461	0,001418	0
ompC	0,002423	6,9E-06	0,01693	0,003002	0,005893	2,95E-05	8,86E-06	0,000807	0,001234	0
parC	6,33E-05	2,25E-06	0,029424	0,016917	0,000109	0,299582	0,322058	0,006052	0,000414	0
secY	4,88E-06	6,97E-08	0,000384	0,000706	8,33E-09	0,000276	3,62E-05	4,24E-07	0,000012	0
purL	5,01E-08	1,1E-08	9,38E-05	8,86E-06	6,08E-11	0,003062	9,27E-05	0,000475	0,000187	0
rne	4,4E-10	0,000079	0,072103	0,007815	1,39E-06	0,13152	0,298955	0,041551	0,021358	0
sucA	4,68E-07	8,98E-10	0,001823	6,47E-05	2,13E-08	0,204503	0,098146	0,012078	0,000164	0
tufA	0	0	0	0	0	0	0	0	0	0
tufB	0	0	0	0	0	0	0	0	0	0
leuA	0,005073	6,27E-08	0,014206	0,015292	0,002805	0,014118	0,049682	0,003048	0,002536	0
hokB	0,010383	0,00123	0,001018	0,002551	0,000539	0,002035	0,001309	0,001509	0,00139	0,00239
acnA	0,000937	1,52E-05	0,302614	0,346469	0,004689	0,00218	0,001876	0,05163	0,013584	0
ubiJ	4,44E-06	1,26E-05	2,19E-07	1,16E-05	1,6E-06	7,6E-06	0,000104	2,37E-07	2,43E-05	3,99E-11
lptD	2,86E-10	6,06E-09	0,000226	0,000531	3,72E-06	0,001133	0,002534	0,023203	0,000771	0
rpnC	0	3,8E-12	2,85E-07	0	2,21E-12	4,53E-13	0	5,55E-16	1,11E-16	1,11E-15
rpnA	1,57E-13	2,22E-15	1,1E-12	1,11E-16	0	0	1,15E-12	2,22E-16	0	0
fdoG	0,0003	3,72E-07	3,02E-06	2,03E-05	0,571772	0,173064	0,099907	1,67E-10	4,53E-09	0
wbbH	0,000213	1,58E-11	0,01157	0,000328	3,3E-06	0,002542	0,001271	0,016094	0,003545	2,1E-07
wbbI	5,28E-05	0,000098	0,000165	0,00028	1,31E-10	0,004092	0,006792	0,079885	0,230906	1,26E-12
wbbK	5,47E-06	3,77E-09	0,00068	0,006263	2,73E-05	0,001708	0,004416	0,005547	0,037366	0
rpnE	7,35E-14	0	0	3,93E-13	1,78E-13	8,63E-14	2,15E-14	0	0	0
lpoA	1,45E-05	6,38E-05	0,021731	0,001311	1,63E-06	0,056306	0,029831	0,00055	0,000697	0
gspD	0	0	0	0	0	0	0	0	0	0
yfjI	0	0	0	0	0	0	0	0	0	0
rlmL	0,000192	2,17E-08	0,000902	0,001819	4,23E-07	0,301685	0,331031	0,236554	0,251152	0
rsxC	1,14E-05	0,000114	0,058399	0,044739	2,75E-07	0,000647	0,013977	0,008001	0,015512	0
yfcI	1,72E-07	6,53E-14	5,61E-13	3,33E-15	8,44E-12	1,48E-10	3E-15	1,64E-14	4,44E-16	2,22E-16
gtrS	4,35E-06	1,77E-07	3,35E-05	0,047478	1,31E-07	0,083571	0,252352	3,99E-06	8,87E-06	2,78E-10

Table 11.1: List of the ORFs corresponding to the reproducible E.coli Ribo-seq profiles and the p-values associated to each pairwise comparison. The columns contain p-values referring to each performed pairwise comparison

Gene name	Transcript ID	Gene description
POMP	ENST00000380842.5	proteasome maturation protein
LAMP1	ENST00000332556.5	lysosomal associated membrane protein 1
CBR1	ENST00000290349.11	carbonyl reductase 1
ST13	ENST00000216218.8	ST13 Hsp70 interacting protein
TGM2	ENST00000361475.7	transglutaminase 2
C11orf58	ENST00000228136.9	chromosome 11 open reading frame 58
FKBP4	ENST00000001008.6	FKBP prolyl isomerase 4
MYH9	ENST00000216181.11	myosin heavy chain 9
NARS1	ENST00000256854.10	asparaginyl-tRNA synthetase 1
EEF2	ENST00000309311.7	eukaryotic translation elongation factor 2
WBP11	ENST00000261167.7	WW domain binding protein 11
ALDH1A1	ENST00000297785.8	aldehyde dehydrogenase 1 family member A1
OSBP	ENST00000263847.6	oxysterol binding protein
ACTA2	ENST00000224784.10	actin alpha 2 smooth muscle
ATP6V1B2	ENST00000276390.7	ATPase H+ transporting V1 subunit B2

THBS1	ENST00000260356.6	thrombospondin 1
AHNAK	ENST00000378024.9	AHNAK nucleoprotein
MAT1A	ENST00000372213.8	methionine adenosyltransferase 1A
PLVAP	ENST00000252590.9	plasmalemma vesicle associated protein
LRPAP1	ENST00000650182.1	LDL receptor related protein associated protein 1
NANS	ENST00000210444.6	N-acetylneuraminate synthase
PRPF19	ENST00000227524.9	pre-mRNA processing factor 19
DNAJA2	ENST00000317089.10	DnaJ heat shock protein family (Hsp40) member A2
SEL1L	ENST00000336735.9	SEL1L adaptor subunit of ERAD E3 ubiquitin ligase
ZFR	ENST00000265069.13	zinc finger RNA binding protein
PGAM1	ENST00000334828.6	phosphoglycerate mutase 1
PSMD3	ENST00000264639.9	proteasome 26S subunit non-ATPase 3
NUDC	ENST00000321265.10	nuclear distribution C dynein complex regulator
APOB	ENST00000233242.5	apolipoprotein B
ITGA1	ENST00000282588.7	integrin subunit alpha 1
RPL35	ENST00000348462.6	ribosomal protein L35
HEXB	ENST00000261416.12	hexosaminidase subunit beta
SEC63	ENST00000369002.9	SEC63 homolog protein translocation regulator
PLIN2	ENST00000276914.7	perilipin 2
PSMB7	ENST00000259457.8	proteasome 20S subunit beta 7
TM9SF3	ENST00000371142.9	transmembrane 9 superfamily member 3
G6PC1	ENST00000253801.7	glucose-6-phosphatase catalytic subunit 1
SF3A1	ENST00000215793.13	splicing factor 3a subunit 1
PSMB2	ENST00000373237.4	proteasome 20S subunit beta 2
A1BG	ENST00000263100.8	alpha-1-B glycoprotein
SLC25A1	ENST00000215882.10	solute carrier family 25 member 1
TMEM70	ENST00000312184.6	transmembrane protein 70
TMED10	ENST00000303575.9	transmembrane p24 trafficking protein 10
SND1	ENST00000354725.8	staphylococcal nuclease and tudor domain containing 1
PCYOX1	ENST00000433351.7	prenylcysteine oxidase 1
PLPP3	ENST00000371250.4	phospholipid phosphatase 3
UGT2B4	ENST00000305107.7	UDP glucuronosyltransferase family 2 member B4
SPTLC1	ENST00000262554.7	serine palmitoyltransferase long chain base subunit 1
EPAS1	ENST00000263734.5	endothelial PAS domain protein 1
ATP6V1A	ENST00000273398.8	ATPase H+ transporting V1 subunit A
HACD2	ENST00000383657.10	3-hydroxyacyl-CoA dehydratase 2
CNOT11	ENST00000289382.8	CCR4-NOT transcription complex subunit 11
BHMT	ENST00000274353.10	betaine–homocysteine S-methyltransferase
AADAC	ENST00000232892.12	arylacetamide deacetylase
RHEB	ENST00000262187.10	Ras homolog mTORC1 binding
C8G	ENST00000371634.7	complement C8 gamma chain
TUFM	ENST00000313511.8	Tu translation elongation factor mitochondrial
ADAM10	ENST00000260408.8	ADAM metallopeptidase domain 10
PPP2CA	ENST00000481195.6	protein phosphatase 2 catalytic subunit alpha
SLC2A2	ENST00000314251.8	solute carrier family 2 member 2
LRP1	ENST00000243077.8	LDL receptor related protein 1
SUCLG1	ENST00000393868.7	succinate-CoA ligase GDP/ADP-forming subunit alpha
ORM2	ENST00000431067.4	orosomucoid 2
DDX18	ENST00000263239.7	DEAD-box helicase 18
PDIA3	ENST00000300289.10	protein disulfide isomerase family A member 3
B4GALT1	ENST00000379731.5	beta-1 4-galactosyltransferase 1
C3	ENST00000245907.11	complement C3
AARS1	ENST00000261772.13	alanyl-tRNA synthetase 1

SNRNP200	ENST00000323853.10	small nuclear ribonucleoprotein U5 subunit 200
TM4SF4	ENST00000305354.5	transmembrane 4 L six family member 4
DTX3L	ENST00000296161.9	deltex E3 ubiquitin ligase 3L
HGD	ENST00000283871.10	homogentisate 1 2-dioxygenase
COX8A	ENST00000314133.4	cytochrome c oxidase subunit 8A
ABCA1	ENST00000374736.8	ATP binding cassette subfamily A member 1
CAPZA2	ENST00000361183.8	capping actin protein of muscle Z-line subunit alpha 2
IGFBP2	ENST00000233809.9	insulin like growth factor binding protein 2
PPA1	ENST00000373232.8	inorganic pyrophosphatase 1
A2M	ENST00000318602.12	alpha-2-macroglobulin
PLG	ENST00000308192.14	plasminogen
GSTA1	ENST00000334575.6	glutathione S-transferase alpha 1
IFI30	ENST00000407280.4	IFI30 lysosomal thiol reductase
AOX1	ENST00000374700.7	aldehyde oxidase 1
LPCAT3	ENST00000261407.9	lysophosphatidylcholine acyltransferase 3
VCP	ENST00000358901.11	valosin containing protein
LGALS3BP	ENST00000262776.8	galectin 3 binding protein
TPR	ENST00000367478.9	translocated promoter region nuclear basket protein
COX6A1	ENST00000229379.3	cytochrome c oxidase subunit 6A1
DYNC1H1	ENST00000360184.10	dynein cytoplasmic 1 heavy chain 1
ANPEP	ENST00000300060.7	alanyl aminopeptidase membrane
C4BPA	ENST00000367070.8	complement component 4 binding protein alpha
CFH	ENST00000367429.9	complement factor H
VTN	ENST00000226218.9	vitronectin
FGL2	ENST00000248598.6	fibrinogen like 2
HINT1	ENST00000304043.10	histidine triad nucleotide binding protein 1
NDUFB7	ENST00000215565.3	NADH:ubiquinone oxidoreductase subunit B7
SERPINF1	ENST00000254722.9	serpin family F member 1
ORM1	ENST00000259396.9	orosomucoid 1
KDELR2	ENST00000258739.9	KDEL endoplasmic reticulum protein retention receptor 2
GLUD1	ENST00000277865.5	glutamate dehydrogenase 1
PRDX2	ENST00000301522.3	peroxiredoxin 2
FGB	ENST00000302068.9	fibrinogen beta chain
MRPS35	ENST00000081029.8	mitochondrial ribosomal protein S35
UGGT1	ENST00000259253.11	UDP-glucose glycoprotein glucosyltransferase 1
SRRM2	ENST00000301740.13	serine/arginine repetitive matrix 2
CTSC	ENST00000227266.10	cathepsin C
MST1	ENST00000449682.2	macrophage stimulating 1
EMC3	ENST00000245046.6	ER membrane protein complex subunit 3
HSPA9	ENST00000297185.9	heat shock protein family A (Hsp70) member 9
KPNA4	ENST00000334256.9	karyopherin subunit alpha 4
EPRS1	ENST00000366923.8	glutamyl-prolyl-tRNA synthetase 1
PAK2	ENST00000327134.7	p21 (RAC1) activated kinase 2
CALR	ENST00000316448.10	calreticulin
PTCD3	ENST00000254630.12	pentatricopeptide repeat domain 3
TF	ENST00000402696.9	transferrin
PECAM1	ENST00000563924.6	platelet and endothelial cell adhesion molecule 1
SLC40A1	ENST00000261024.7	solute carrier family 40 member 1
HSP90B1	ENST00000299767.10	heat shock protein 90 beta family member 1
ADH5	ENST00000296412.14	alcohol dehydrogenase 5 (class III) chi polypeptide
ARPC3	ENST00000228825.12	actin related protein 2/3 complex subunit 3
PA2G4	ENST00000303305.11	proliferation-associated 2G4
UQCRC1	ENST00000203407.6	ubiquinol-cytochrome c reductase core protein 1

PRPF40A	ENST00000410080.8	pre-mRNA processing factor 40 homolog A
HPX	ENST00000265983.8	hemopexin
KNG1	ENST00000644859.2	kininogen 1
PLOD1	ENST00000196061.5	procollagen-lysine 2-oxoglutarate 5-dioxygenase 1
NDUFA12	ENST00000327772.7	NADH:ubiquinone oxidoreductase subunit A12
SF3B1	ENST00000335508.11	splicing factor 3b subunit 1
NCL	ENST00000322723.9	nucleolin
COPG1	ENST00000314797.10	COPI coat complex subunit gamma 1
CYP27A1	ENST00000258415.9	cytochrome P450 family 27 subfamily A member 1
EIF3I	ENST00000373586.2	eukaryotic translation initiation factor 3 subunit I
PARP1	ENST00000366794.10	poly(ADP-ribose) polymerase 1
SERPINC1	ENST00000367698.4	serpin family C member 1
LRPPRC	ENST00000260665.12	leucine rich pentatricopeptide repeat containing
SF3B4	ENST00000271628.9	splicing factor 3b subunit 4
SRP9	ENST00000304786.12	signal recognition particle 9
SFPQ	ENST00000357214.6	splicing factor proline and glutamine rich
HSPG2	ENST00000374695.8	heparan sulfate proteoglycan 2

Table 11.3: Reproducible genes across all ten cancer dataset

Gene name	Transcript ID	Gene description
OTC	ENST00000039007.5	ornithine transcarbamylase
NAMPT	ENST00000222553.8	nicotinamide phosphoribosyltransferase
CYP3A5	ENST00000222982.8	cytochrome P450 family 3 subfamily A member 5
C5	ENST00000223642.3	complement C5
ACTA2	ENST00000224784.10	actin alpha 2 smooth muscle
ARPC3	ENST00000228825.12	actin related protein 2/3 complex subunit 3
COX6A1	ENST00000229379.3	cytochrome c oxidase subunit 6A1
AADAC	ENST00000232892.12	arylacetamide deacetylase
APOB	ENST00000233242.5	apolipoprotein B
LRP1	ENST00000243077.8	LDL receptor related protein 1
MT2A	ENST00000245185.6	metallothionein 2A
LYVE1	ENST00000256178.8	lymphatic vessel endothelial hyaluronan receptor 1
SDS	ENST00000257549.9	serine dehydratase
CYP27A1	ENST00000258415.9	cytochrome P450 family 27 subfamily A member 1
LRPPRC	ENST00000260665.12	leucine rich pentatricopeptide repeat containing
AARS1	ENST00000261772.13	alanyl-tRNA synthetase 1
LGALS3BP	ENST00000262776.8	galectin 3 binding protein
PECR	ENST00000265322.8	peroxisomal trans-2-enoyl-CoA reductase
PLIN2	ENST00000276914.7	perilipin 2
HGD	ENST00000283871.10	homogentisate 1 2-dioxygenase
SLC51A	ENST00000296327.10	solute carrier family 51 subunit alpha
DEPP1	ENST00000298295.4	DEPP1 autophagy regulator
PRDX2	ENST00000301522.3	peroxiredoxin 2
FGB	ENST00000302068.9	fibrinogen beta chain
TM4SF4	ENST00000305354.5	transmembrane 4 L six family member 4
TUFM	ENST00000313511.8	Tu translation elongation factor
COPG1	ENST00000314797.10	COPI coat complex subunit gamma 1
BGN	ENST00000331595.9	biglycan
LAMP1	ENST00000332556.5	lysosomal associated membrane protein 1
OIT3	ENST00000334011.10	oncoprotein induced transcript 3
SF3B1	ENST00000335508.11	splicing factor 3b subunit 1
SEL1L	ENST00000336735.9	SEL1L adaptor subunit of ERAD E3 ubiquitin ligase
RPL35	ENST00000348462.6	ribosomal protein L35
SND1	ENST00000354725.8	staphylococcal nuclease and tudor domain containing 1
C8A	ENST00000361249.4	complement C8 alpha chain
MT-CO1	ENST00000361624.2	mitochondrially encoded cytochrome c oxidase I
C4BPA	ENST00000367070.8	complement component 4 binding protein alpha
F13B	ENST00000367412.2	coagulation factor XIII B chain
CFH	ENST00000367429.9	complement factor H
FMO3	ENST00000367755.9	flavin containing dimethylaniline monoxygenase 3
OAT	ENST00000368845.6	ornithine aminotransferase
AOX1	ENST00000374700.7	aldehyde oxidase 1
ALDH1B1	ENST00000377698.4	aldehyde dehydrogenase 1 family member B1
MASP2	ENST00000400897.8	mannan binding lectin serine peptidase 2
TF	ENST00000402696.9	transferrin
ADH1C	ENST00000515683.6	alcohol dehydrogenase 1C (class I)
SLC38A3	ENST00000614032.5	solute carrier family 38 member 3
KNG1	ENST00000644859.2	kininogen 1
FGA	ENST00000651975.1	fibrinogen alpha chain

Table 11.2: Reproducible genes across all ten control human dataset

Reactome pathways	Ref List (20595)	Input (137)	Input (expected)	fold Enrichment	raw P-value	FDR
Scavenging by Class F Receptors (R-HSA-3000484)	6	2	0.04	48.69	1.25E-03	4.33E-02
Platelet sensitization by LDL (R-HSA-432142)	17	3	0.12	25.78	3.22E-04	1.84E-02
Cytosolic tRNA aminoacylation (R-HSA-379716)	24	4	0.16	24.34	3.69E-05	3.38E-03
Plasma lipoprotein assembly (R-HSA-8963898)	18	3	0.12	24.34	3.74E-04	1.94E-02
Scavenging by Class A Receptors (R-HSA-3000480)	19	3	0.13	23.06	4.31E-04	2.14E-02
Calnexin/calreticulin cycle (R-HSA-901042)	26	4	0.18	22.47	4.89E-05	4.14E-03
N-glycan trimming in the ER and Calnexin/Calreticulin cycle (R-HSA-532668)	35	5	0.24	20.87	7.36E-06	8.40E-04
Intrinsic Pathway of Fibrin Clot Formation (R-HSA-140837)	22	3	0.15	19.92	6.34E-04	2.73E-02
Formation of Fibrin Clot (Clotting Cascade) (R-HSA-140877)	39	5	0.27	18.73	1.19E-05	1.29E-03
tRNA Aminoacylation (R-HSA-379724)	42	4	0.29	13.91	2.68E-04	1.61E-02
Hh mutants (R-HSA-5362768)	55	5	0.38	13.28	5.47E-05	4.17E-03
Platelet degranulation (R-HSA-114608)	127	11	0.87	12.65	2.67E-09	8.71E-07
Hh mutants abrogate ligand secretion (R-HSA-5387390)	58	5	0.4	12.59	6.93E-05	4.95E-03
Defective CFTR causes cystic fibrosis (R-HSA-5678895)	60	5	0.41	12.17	8.06E-05	5.58E-03
Response to elevated platelet cytosolic Ca2+ (R-HSA-76005)	132	11	0.9	12.17	3.89E-09	1.11E-06
Regulation of activated PAK-2p34 by proteasome mediated degradation (R-HSA-211733)	49	4	0.34	11.92	4.63E-04	2.16E-02
ABC transporter disorders (R-HSA-5619084)	76	6	0.52	11.53	2.03E-05	2.02E-03
Hedgehog ligand biogenesis (R-HSA-5358346)	64	5	0.44	11.41	1.07E-04	7.00E-03
Regulation of Apoptosis (R-HSA-169911)	52	4	0.36	11.24	5.72E-04	2.51E-02
HSP90 chaperone cycle for steroid hormone receptors (SHR) (R-HSA-3371497)	55	4	0.38	10.62	6.98E-04	2.95E-02
Regulation of Insulin-like Growth Factor (IGF) (R-HSA-381426)	124	9	0.85	10.6	3.20E-07	5.63E-05
Integrin cell surface interactions (R-HSA-216083)	84	6	0.58	10.43	3.46E-05	3.29E-03
Iron uptake and transport (R-HSA-917937)	57	4	0.39	10.25	7.92E-04	3.12E-02
Post-translational protein phosphorylation (R-HSA-8957275)	107	7	0.73	9.56	1.30E-05	1.35E-03
Signaling by NOTCH4 (R-HSA-9013694)	80	5	0.55	9.13	2.87E-04	1.68E-02
Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor (R-HSA-1234176)	65	4	0.45	8.99	1.26E-03	4.29E-02
ER-Phagosome pathway (R-HSA-1236974)	83	5	0.57	8.8	3.37E-04	1.83E-02
TP53 Regulates Metabolic Genes (R-HSA-5628897)	83	5	0.57	8.8	3.37E-04	1.79E-02
Respiratory electron transport (R-HSA-611105)	100	6	0.68	8.76	8.68E-05	5.83E-03
Cellular response to heat stress (R-HSA-3371556)	88	5	0.6	8.3	4.35E-04	2.12E-02
Disorders of transmembrane transporters (R-HSA-5619115)	172	9	1.18	7.64	4.20E-06	5.32E-04
Antigen processing-Cross presentation (R-HSA-1236975)	99	5	0.68	7.38	7.26E-04	3.01E-02
COPI-mediated anterograde transport (R-HSA-6807878)	100	5	0.68	7.3	7.58E-04	3.09E-02
ABC-family proteins mediated transport (R-HSA-382556)	102	5	0.7	7.16	8.26E-04	3.20E-02
Respiratory electron transport (R-HSA-163200)	123	6	0.84	7.13	2.55E-04	1.57E-02
Binding and Uptake of Ligands by Scavenger Receptors (R-HSA-2173782)	104	5	0.71	7.02	8.98E-04	3.42E-02
Regulation of Complement cascade (R-HSA-977606)	112	5	0.77	6.52	1.23E-03	4.33E-02
mRNA Splicing - Major Pathway (R-HSA-72163)	180	8	1.23	6.49	4.46E-05	3.92E-03
Asparagine N-linked glycosylation (R-HSA-446203)	303	13	2.07	6.27	2.58E-07	5.88E-05
mRNA Splicing (R-HSA-72172)	188	8	1.29	6.22	5.98E-05	4.41E-03
Platelet activation, signaling and aggregation (R-HSA-76002)	259	11	1.77	6.2	2.46E-06	3.51E-04
Translation (R-HSA-72766)	293	12	2.01	5.98	1.21E-06	1.85E-04
The citric acid (TCA) cycle and respiratory electron transport (R-HSA-1428517)	173	7	1.18	5.91	2.35E-04	1.49E-02
Neutrophil degranulation (R-HSA-6798695)	478	19	3.27	5.81	1.23E-09	4.69E-07
ER to Golgi Anterograde Transport (R-HSA-199977)	153	6	1.05	5.73	7.73E-04	3.10E-02
Transport to the Golgi and subsequent modification (R-HSA-948021)	184	7	1.26	5.56	3.37E-04	1.88E-02
Processing of Capped Intron-Containing Pre-mRNA (R-HSA-72203)	238	9	1.63	5.52	4.99E-05	4.07E-03
Signaling by NOTCH (R-HSA-157118)	200	7	1.37	5.11	5.46E-04	2.44E-02
Extracellular matrix organization (R-HSA-1474244)	299	10	2.05	4.89	5.16E-05	4.06E-03
Biological oxidations (R-HSA-211859)	219	7	1.5	4.67	9.15E-04	3.43E-02
Cellular responses to stress (R-HSA-2262752)	547	17	3.74	4.54	3.00E-07	5.72E-05
Innate Immune System (R-HSA-168249)	1105	34	7.57	4.49	1.57E-13	1.79E-10
Cellular responses to external stimuli (R-HSA-8953897)	561	17	3.84	4.43	4.23E-07	6.90E-05
Metabolism of carbohydrates (R-HSA-71387)	286	8	1.96	4.09	9.24E-04	3.40E-02
Metabolism of RNA (R-HSA-8953854)	661	17	4.53	3.76	3.72E-06	4.99E-04
Hemostasis (R-HSA-109582)	669	17	4.58	3.71	4.34E-06	5.22E-04
Diseases of signal transduction (R-HSA-5663202)	366	9	2.51	3.59	1.08E-03	3.93E-02
Metabolism of amino acids and derivatives (R-HSA-71291)	367	9	2.51	3.58	1.10E-03	3.94E-02
Disease (R-HSA-1643685)	1126	27	7.71	3.5	1.47E-08	3.73E-06
Infectious disease (R-HSA-5663205)	464	11	3.18	3.46	4.12E-04	2.09E-02
Metabolism of proteins (R-HSA-392499)	1977	43	13.54	3.18	4.88E-12	3.72E-09
Metabolism (R-HSA-1430728)	2079	42	14.23	2.95	9.51E-11	5.43E-08
Post-translational protein modification (R-HSA-597592)	1388	28	9.5	2.95	2.66E-07	5.52E-05
Vesicle-mediated transport (R-HSA-5653656)	725	14	4.96	2.82	5.23E-04	2.39E-02
Immune System (R-HSA-168256)	2158	41	14.77	2.78	1.07E-09	4.88E-07
Transport of small molecules (R-HSA-382551)	719	13	4.92	2.64	1.49E-03	5.00E-02
Signal Transduction (R-HSA-162582)	2728	34	18.68	1.82	4.39E-04	2.09E-02

Table 11.4: Summary results of over-representation test of Reproducible cancer genes. The first column contains the name of the annotation data category (Reactome pathway). The second column contains the number of genes in the reference list (human genes) that map to reactome pathway category. The third column contains the number of reproducible cancer genes that map to the reactome pathway data category. The fourth column contains the expected value (the number of genes we would expect in our list for this category, based on the reference list). The fifth column shows the Fold Enrichment of the reproducible cancer genes observed over the expected value. If it is greater than 1, it means that the category is over-represented. Otherwise, the category is under-represented if it is less than 1. The sixth column is the raw p-value as determined by Fisher's exact test. The seventh column is the False Discovery Rate as calculated by the Benjamini-Hochberg procedure (FDR < 0.05

Bibliography

[1] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes," *Cell*, vol. 147, no. 4, pp. 789–802, 2011.

[2] N. T. Ingolia, "Ribosome profiling: new views of translation, from single codons to genome scale," *Nature Reviews Genetics*, vol. 15, no. 3, pp. 205–213, 2014.

[3] A. Valleriani and D. Chiarugi, "A workbench for the translational control of gene expression," *bioRxiv*, 2020.

[4] S. Ahnert, T. Fink, and A. Zinovyev, "How much non-coding dna do eukaryotes require?" *Journal of theoretical biology*, vol. 252 4, pp. 587–92, 2008.

[5] G. Storz, "An expanding universe of noncoding rnas," *Science*, vol. 296, no. 5571, pp. 1260–1263, 2002.

[6] L. Srijyothi, S. Ponne, T. Prathama, C. Ashok, and S. Baluchamy, *Roles of non-coding RNAs in transcriptional regulation.* IntechOpen London, UK, 2018, vol. 55.

[7] P. L. DeHaseth, M. L. Zupancic, and M. T. Record, "Rna polymerase-promoter interactions: the comings and goings of rna polymerase," *Journal of bacteriology*, vol. 180, no. 12, pp. 3019–3025, 1998.

[8] M. Raffaelle, E. I. Kanin, J. Vogt, R. R. Burgess, and A. Z. Ansari, "Holoenzyme switching and stochastic release of sigma factors from rna polymerase in vivo," *Molecular Cell*, vol. 20, no. 3, pp. 357–366, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1097276505016813

[9] B. Z. Ring, W. S. Yarnell, and J. W. Roberts, "Function of e. coli rna polymerase σ factor-σ70 in promoter-proximal pausing," *Cell*, vol. 86, no. 3, pp. 485–493, 1996.

[10] S. Banerjee, J. Chalissery, I. Bandey, and R. Sen, "Rho-dependent transcription termination: more questions than answers," *Journal of microbiology (Seoul, Korea)*, vol. 44, no. 1, p. 11, 2006.

[11] A. Kremling, J. Geiselmann, D. Ropers, and H. de Jong, "Understanding carbon catabolite repression in escherichia coli using quantitative models," *Trends in microbiology*, vol. 23, no. 2, pp. 99–109, 2015.

[12] C. L. Turnbough, "Regulation of bacterial gene expression by transcription attenuation," *Microbiology and Molecular Biology Reviews*, vol. 83, no. 3, 2019.

[13] L. A. Simmons, J. J. Foti, S. E. Cohen, and G. C. Walker, "The sos regulatory network," *EcoSal Plus*, vol. 2008, 2008.

[14] D. Panne, "The enhanceosome," *Current opinion in structural biology*, vol. 18, no. 2, pp. 236–242, 2008.

[15] A. G. Arimbasseri, K. Rijal, and R. J. Maraia, "Comparative overview of rna polymerase ii and iii transcription cycles, with focus on rna polymerase iii termination and reinitiation," *Transcription*, vol. 5, no. 1, p. e27369, 2014.

[16] G. Orphanides and D. Reinberg, "A unified theory of gene expression," *Cell*, vol. 108, no. 4, pp. 439–451, 2002.

[17] M. L. Grace, M. B. Chandrasekharan, T. C. Hall, and A. J. Crowe, "Sequence and spacing of tata box elements are critical for accurate initiation from the β-phaseolin promoter," *Journal of Biological Chemistry*, vol. 279, no. 9, pp. 8102–8110, 2004.

[18] J. K. Rimel and D. J. Taatjes, "The essential and multifunctional tfiih complex," *Protein Science*, vol. 27, no. 6, pp. 1018–1037, 2018.

[19] K. M. André, E. H. Sipos, and J. Soutourina, "Mediator roles going beyond transcription," *Trends in Genetics*, 2020.

[20] I. Jonkers and J. T. Lis, "Getting up to speed with transcription elongation by rna polymerase ii," *Nature reviews Molecular cell biology*, vol. 16, no. 3, pp. 167–177, 2015.

[21] K. Adelman and J. T. Lis, "Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans," *Nature Reviews Genetics*, vol. 13, no. 10, pp. 720–731, 2012.

[22] J. T. Lis, P. Mason, J. Peng, D. H. Price, and J. Werner, "P-tefb kinase recruitment and function at heat shock loci," *Genes & development*, vol. 14, no. 7, pp. 792–803, 2000.

[23] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[24] G. M. Cooper, R. E. Hausman, and R. E. Hausman, *The cell: a molecular approach*. ASM press Washington, DC, 2007, vol. 4.

[25] F. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, "General nature of the genetic code for proteins," 1961.

[26] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew *et al.*, "The complete genome sequence of escherichia coli k-12," *science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[27] J. Shine and L. Dalgarno, "The 3'-terminal sequence of escherichia coli 16s ribosomal rna: complementarity to nonsense triplets and ribosome binding sites," *Proceedings of the National Academy of Sciences*, vol. 71, no. 4, pp. 1342–1346, 1974.

[28] D. Dottavio-Martin, D. P. Suttle, and J. M. Ravel, "The effects of initiation factors if-1 and if-3 on the dissociation of escherichia coli 70 s ribosomes," *FEBS letters*, vol. 97, no. 1, pp. 105–110, 1979.

[29] S. Gottesman, E. Roche, Y. Zhou, and R. T. Sauer, "The clpxp and clpap proteases degrade proteins with carboxy-terminal peptide tails added by the ssra-tagging system," *Genes & development*, vol. 12, no. 9, pp. 1338–1347, 1998.

[30] D. A. Mangus, M. C. Evans, and A. Jacobson, "Poly (a)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression," *Genome biology*, vol. 4, no. 7, pp. 1–14, 2003.

[31] A. G. Johnson, R. Grosely, A. N. Petrov, and J. D. Puglisi, "Dynamics of ires-mediated translation," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1716, p. 20160177, 2017.

[32] M. Verma, J. Choi, K. A. Cottrell, Z. Lavagnino, E. N. Thomas, S. Pavlovic-Djuranovic, P. Szczesny, D. W. Piston, H. S. Zaher, J. D. Puglisi *et al.*, "A short translational ramp determines the efficiency of protein synthesis," *Nature communications*, vol. 10, no. 1, pp. 1–15, 2019.

[33] J. Frank, H. Gao, J. Sengupta, N. Gao, and D. J. Taylor, "The process of mrna–trna translocation," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19 671–19 678, 2007.

[34] M. Kozak, "Regulation of translation via mrna structure in prokaryotes and eukaryotes," *Gene*, vol. 361, pp. 13–37, 2005.

[35] B. Hetrick, K. Lee, and S. Joseph, "Kinetics of stop codon recognition by release factor 1," *Biochemistry*, vol. 48, no. 47, pp. 11 178–11 184, 2009.

[36] S. Altuvia, A. Zhang, L. Argaman, A. Tiwari, and G. Storz, "The escherichia coli oxys regulatory rna represses fhla translation by blocking ribosome binding," *The EMBO journal*, vol. 17, no. 20, pp. 6069–6075, 1998.

[37] E. Nudler and A. S. Mironov, "The riboswitch control of bacterial metabolism," *Trends in biochemical sciences*, vol. 29, no. 1, pp. 11–17, 2004.

[38] J. D. Ho, N. C. Balukoff, P. R. Theodoridis, M. Wang, J. R. Krieger, J. H. Schatz, and S. Lee, "A network of rna-binding proteins controls translation efficiency to activate anaerobic metabolism," *Nature communications*, vol. 11, no. 1, pp. 1–16, 2020.

[39] S. M. García-Mauriño, F. Rivero-Rodríguez, A. Velázquez-Cruz, M. Hernández-Vellisca, A. Díaz-Quintana, M. A. De la Rosa, and I. Díaz-Moreno, "Rna binding protein regulation and cross-talk in the control of au-rich mrna fate," *Frontiers in molecular biosciences*, vol. 4, p. 71, 2017.

[40] K. Masaki, Y. Sonobe, G. Ghadge, P. Pytel, P. Lépine, F. Pernin, Q.-L. Cui, J. P. Antel, S. Zandee, A. Prat *et al.*, "Rna-binding protein altered expression and mislocalization in ms," *Neurology-Neuroimmunology Neuroinflammation*, vol. 7, no. 3, 2020.

[41] B. Pereira, M. Billaud, and R. Almeida, "Rna-binding proteins in cancer: old players and new actors," *Trends in cancer*, vol. 3, no. 7, pp. 506–528, 2017.

[42] K. Pakos-Zebrucka, I. Koryga, K. Mnich, M. Ljujic, A. Samali, and A. M. Gorman, "The integrated stress response," *EMBO reports*, vol. 17, no. 10, pp. 1374–1395, 2016.

[43] F. F. Chevance, S. Le Guyon, and K. T. Hughes, "The effects of codon context on in vivo translation speed," *PLoS Genet*, vol. 10, no. 6, p. e1004392, 2014.

[44] Q. Yang, C.-H. Yu, F. Zhao, Y. Dang, C. Wu, P. Xie, M. S. Sachs, and Y. Liu, "erf1 mediates codon usage effects on mrna translation efficiency through premature termination at rare codons," *Nucleic acids research*, vol. 47, no. 17, pp. 9243–9258, 2019.

[45] J. B. Plotkin and G. Kudla, "Synonymous but not the same: the causes and consequences of codon bias," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 32–42, 2011.

[46] C. A. Waudby, C. M. Dobson, and J. Christodoulou, "Nature and regulation of protein folding on the ribosome," *Trends in biochemical sciences*, vol. 44, no. 11, pp. 914–926, 2019.

[47] G. A. Brar and J. S. Weissman, "Ribosome profiling reveals the what, when, where and how of protein synthesis," *Nature reviews Molecular cell biology*, vol. 16, no. 11, pp. 651–664, 2015.

[48] N. Sonenberg and A. G. Hinnebusch, "Regulation of translation initiation in eukaryotes: mechanisms and biological targets," *Cell*, vol. 136, no. 4, pp. 731–745, 2009.

[49] S. Wang, I. B. Rosenwald, M. J. Hutzler, G. A. Pihan, L. Savas, J.-J. Chen, and B. A. Woda, "Expression of the eukaryotic translation initiation factors 4e and 2α in non-hodgkin's lymphomas," *The American journal of pathology*, vol. 155, no. 1, pp. 247–255, 1999.

[50] N. Burwick and B. H. Aktas, "The eif2-alpha kinase hri: a potential target beyond the red blood cell," *Expert opinion on therapeutic targets*, vol. 21, no. 12, pp. 1171–1177, 2017.

[51] F. Pettersson, C. Yau, M. C. Dobocan, B. Culjkovic-Kraljacic, H. Retrouvay, R. Puckett, L. M. Flores, I. E. Krop, C. Rousseau, E. Cocolakis *et al.*, "Ribavirin treatment effects on breast cancers overexpressing eif4e, a biomarker with prognostic specificity for luminal b-type breast cancer," *Clinical Cancer Research*, vol. 17, no. 9, pp. 2874–2884, 2011.

[52] N. Robichaud, S. V. del Rincon, B. Huor, T. Alain, L. A. Petruccelli, J. Hearnden, C. Goncalves, S. Grotegut, C. H. Spruck, L. Furic *et al.*, "Phosphorylation of eif4e promotes emt and metastasis via translational control of snail and mmp-3," *Oncogene*, vol. 34, no. 16, pp. 2032–2042, 2015.

[53] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments," *Nature protocols*, vol. 7, no. 8, pp. 1534–1550, 2012.

[54] Y. Liu, A. Beyer, and R. Aebersold, "On the dependency of cellular protein levels on mrna abundance," *Cell*, vol. 165, no. 3, pp. 535–550, 2016.

[55] V. Benes, J. Blake, and K. Doyle, "Ribo-zero gold kit: improved rna-seq results after removal of cytoplasmic and mitochondrial ribosomal rna," *Nature Methods*, vol. 8, no. 11, pp. iii–iv, 2011.

[56] R. C. Scarpulla, "Transcriptional paradigms in mammalian mitochondrial biogenesis and function," *Physiological reviews*, vol. 88, no. 2, pp. 611–638, 2008.

[57] N. Hornstein, D. Torres, S. D. Sharma, G. Tang, P. Canoll, and P. A. Sims, "Ligation-free ribosome profiling of cell type-specific translation in the brain," *Genome biology*, vol. 17, no. 1, pp. 1–15, 2016.

[58] S. Schafer, E. Adami, M. Heinig, K. E. C. Rodrigues, F. Kreuchwig, J. Silhavy, S. Van Heesch, D. Simaite, N. Rajewsky, E. Cuppen *et al.*, "Translational regulation shapes the molecular landscape of complex disease phenotypes," *Nature communications*, vol. 6, no. 1, pp. 1–9, 2015.

[59] S. Kuersten, A. Radek, C. Vogel, and L. O. Penalva, "Translation regulation gets its 'omics' moment," *Wiley Interdisciplinary Reviews: RNA*, vol. 4, no. 6, pp. 617–630, 2013.

[60] N. R. Guydosh and R. Green, "Dom34 rescues ribosomes in 3' untranslated regions," *Cell*, vol. 156, no. 5, pp. 950–962, 2014.

[61] K. E. Baker and R. Parker, "Nonsense-mediated mrna decay: terminating erroneous gene expression," *Current opinion in cell biology*, vol. 16, no. 3, pp. 293–299, 2004.

[62] A. Dana and T. Tuller, "Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells," *PLoS Comput Biol*, vol. 8, no. 11, p. e1002755, 2012.

[63] C. Sin, D. Chiarugi, and A. Valleriani, "Quantitative assessment of ribosome drop-off in e. coli," *Nucleic acids research*, vol. 44, no. 6, pp. 2528–2537, 2016.

[64] M. V. Gerashchenko and V. N. Gladyshev, "Ribonuclease selection for ribosome profiling," *Nucleic acids research*, vol. 45, no. 2, pp. e6–e6, 2017.

[65] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[66] K. L. Howe, B. Contreras-Moreira, N. De Silva, G. Maslen, W. Akanni, J. Allen, J. Alvarez-Jarreta, M. Barba, D. M. Bolser, L. Cambell *et al.*, "Ensembl genomes 2020—enabling non-vertebrate genomic research," *Nucleic acids research*, vol. 48, no. D1, pp. D689–D695, 2020.

[67] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett, "FastQC," Babraham Institute, Babraham, UK, Jan. 2012.

[68] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[69] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[70] C. J. Woolstenhulme, N. R. Guydosh, R. Green, and A. R. Buskirk, "High-precision analysis of translational pausing by ribosome profiling in bacteria lacking efp," *Cell reports*, vol. 11, no. 1, pp. 13–21, 2015.

[71] G. J. Morgan, D. H. Burkhardt, J. W. Kelly, and E. T. Powers, "Translation efficiency is maintained at elevated temperature in escherichia coli," *Journal of Biological Chemistry*, vol. 293, no. 3, pp. 777–793, 2018.

[72] F. Mohammad, C. J. Woolstenhulme, R. Green, and A. R. Buskirk, "Clarifying the translational pausing landscape in bacteria by ribosome profiling," *Cell reports*, vol. 14, no. 4, pp. 686–694, 2016.

[73] G.-W. Li, D. Burkhardt, C. Gross, and J. S. Weissman, "Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources," *Cell*, vol. 157, no. 3, pp. 624–635, 2014.

[74] A. R. Subramaniam, B. M. Zid, and E. K. O'Shea, "An integrated approach reveals regulatory controls on bacterial translation elongation," *Cell*, vol. 159, no. 5, pp. 1200–1211, 2014.

[75] D. H. Burkhardt, S. Rouskin, Y. Zhang, G.-W. Li, J. S. Weissman, and C. A. Gross, "Operon mrnas are organized into orf-centric structures that predict translation efficiency," *Elife*, vol. 6, p. e22037, 2017.

[76] G.-W. Li, E. Oh, and J. S. Weissman, "The anti-shine–dalgarno sequence drives translational pausing and codon choice in bacteria," *Nature*, vol. 484, no. 7395, pp. 538–541, 2012.

[77] N. E. Baggett, Y. Zhang, and C. A. Gross, "Global analysis of translation termination in e. coli," *PLoS genetics*, vol. 13, no. 3, p. e1006676, 2017.

[78] M. S. Guo, T. B. Updegrove, E. B. Gogol, S. A. Shabalina, C. A. Gross, and G. Storz, "MicI, a new σe-dependent srna, combats envelope stress by repressing synthesis of lpp, the major outer membrane lipoprotein," *Genes & development*, vol. 28, no. 14, pp. 1620–1634, 2014.

[79] H. Nikaido, "Porins and specific diffusion channels in bacterial outer membranes," *The Journal of biological chemistry (Print)*, vol. 269, no. 6, pp. 3905–3908, 1994.

[80] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "Large-scale gene function analysis with the panther classification system," *Nature protocols*, vol. 8, no. 8, pp. 1551–1566, 2013.

[81] D. B. Goodman, G. M. Church, and S. Kosuri, "Causes and effects of n-terminal codon bias in bacterial genes," *Science*, vol. 342, no. 6157, pp. 475–479, 2013.

[82] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, "Coding-sequence determinants of gene expression in escherichia coli," *science*, vol. 324, no. 5924, pp. 255–258, 2009.

[83] R. C. Hunt, V. L. Simhadri, M. Iandoli, Z. E. Sauna, and C. Kimchi-Sarfaty, "Exposing synonymous mutations," *Trends in Genetics*, vol. 30, no. 7, pp. 308–321, 2014.

[84] E. M. Novoa and L. R. de Pouplana, "Speeding with control: codon usage, trnas, and ribosomes," *Trends in Genetics*, vol. 28, no. 11, pp. 574–581, 2012.

[85] A. Tats, M. Remm, and T. Tenson, "Highly expressed proteins have an increased frequency of alanine in the second amino acid position," *BMC genomics*, vol. 7, no. 1, pp. 1–13, 2006.

[86] K. Saito, R. Green, and A. R. Buskirk, "Translational initiation in e. coli occurs at the correct sites genome-wide in the absence of mrna-rrna base-pairing," *Elife*, vol. 9, p. e55002, 2020.

[87] J. F. Kane, "Effects of rare codon clusters on high-level expression of heterologous proteins in escherichia coli," *Current opinion in biotechnology*, vol. 6, no. 5, pp. 494–500, 1995.

[88] R. Spanjaard and J. Van Duin, "Translation of the sequence agg-agg yields 50% ribosomal frameshift," *Proceedings of the National Academy of Sciences*, vol. 85, no. 21, pp. 7967–7971, 1988.

[89] S. R. Maloy, V. J. Stewart, R. Taylor, and S. I. Miller, "Genetic analysis of pathogenic bacteria," *Trends in Microbiology*, vol. 4, no. 12, p. 504, 1996.

[90] G. Hanson and J. Coller, "Codon optimality, bias and usage in translation and mrna decay," *Nature reviews Molecular cell biology*, vol. 19, no. 1, p. 20, 2018.

[91] R. Saunders and C. M. Deane, "Synonymous codon usage influences the local protein structure observed," *Nucleic acids research*, vol. 38, no. 19, pp. 6719–6728, 2010.

[92] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318.

[93] D. Grapov, J. Fahrmann, K. Wanichthanarak, and S. Khoomrung, "Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine," *Omics: a journal of integrative biology*, vol. 22, no. 10, pp. 630–636, 2018.

[94] S. Bonechi, M. Bianchini, P. Bongini, G. Ciano, G. Giacomini, R. Rosai, L. Tognetti, A. Rossi, and P. Andreini, "Fusion of visual and anamnestic data for the classification of skin lesions with deep learning," in *New Trends in Image Analysis and Processing – ICIAP 2019*, M. Cristani, A. Prati, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham: Springer International Publishing, 2019, pp. 211–219.

[95] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar *et al.*, "A universal snp and small-indel variant caller using deep neural networks," *Nature biotechnology*, vol. 36, no. 10, pp. 983–987, 2018.

[96] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019.

[97] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[98] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[99] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[100] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli," *Nucleic acids research*, vol. 10, no. 9, pp. 2997–3011, 1982.

[101] A. Coates, A. Ng, H. Lee, G. Gordon, D. Dunson, and M. Dudík, "Proceedings of the fourteenth international conference on artificial intelligence and statistics," 2011.

[102] M. R. Baker and R. B. Patil, "Universal approximation theorem for interval neural networks," *Reliable Computing*, vol. 4, no. 3, pp. 235–239, 1998.

[103] T. Rauber and K. Berns, "Kernel multilayer perceptron," in *2011 24th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2011, pp. 337–343.

[104] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[105] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[106] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[107] N. Prasad, R. Singh, and S. P. Lal, "Comparison of back propagation and resilient propagation algorithm for spam classification," in *2013 Fifth international conference on computational intelligence, modelling and simulation*. IEEE, 2013, pp. 29–34.

[108] N. Ketkar, "Introduction to keras," in *Deep learning with Python*. Springer, 2017, pp. 97–111.

[109] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[110] D. Maclaurin, D. Duvenaud, and R. Adams, "Gradient-based hyperparameter optimization through reversible learning," in *International conference on machine learning*. PMLR, 2015, pp. 2113–2122.

[111] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, "Forward and reverse gradient-based hyperparameter optimization," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1165–1173.

[112] S. L. Svenningsen, M. Kongstad, T. S. Stenum, A. J. Muñoz-Gómez, and M. A. Sørensen, "Transfer rna is highly unstable during early amino acid starvation in escherichia coli," *Nucleic acids research*, vol. 45, no. 2, pp. 793–804, 2017.

[113] X. Lin, C. Wang, C. Guo, Y. Tian, H. Li, and X. Peng, "Differential regulation of ompc and ompf by atpb in escherichia coli exposed to nalidixic acid and chlortetracycline," *Journal of proteomics*, vol. 75, no. 18, pp. 5898–5910, 2012.

[114] S. Forst, J. Delgado, G. Ramakrishnan, and M. Inouye, "Regulation of ompc and ompf expression in escherichia coli in the absence of envz." *Journal of bacteriology*, vol. 170, no. 11, pp. 5080–5085, 1988.

[115] S. Norioka, G. Ramakrishnan, K. Ikenaka, and M. Inouye, "Interaction of a transcriptional activator, ompr, with reciprocally osmoregulated genes, ompf and ompc, of escherichia coli." *Journal of Biological Chemistry*, vol. 261, no. 36, pp. 17 113–17 119, 1986.

[116] W. Alphen and B. Lugtenberg, "Influence of osmolarity of the growth medium on the outer membrane protein pattern of escherichia coli." *Journal of bacteriology*, vol. 131, no. sc2, pp. 623–630, 1977.

[117] M. Á. De la Cruz and E. Calva, "The complexities of porin genetic regulation," *Journal of molecular microbiology and biotechnology*, vol. 18, no. 1, pp. 24–36, 2010.

[118] J. Vogel and K. Papenfort, "Small non-coding rnas and the bacterial outer membrane," *Current opinion in microbiology*, vol. 9, no. 6, pp. 605–611, 2006.

[119] Q. Zou, Z. Xiao, R. Huang, X. Wang, X. Wang, H. Zhao, and X. Yang, "Survey of the translation shifts in hepatocellular carcinoma with ribosome profiling," *Theranostics*, vol. 9, no. 14, p. 4141, 2019.

[120] B. C. Tennant and S. A. Center, "Chapter 13 - hepatic function," in *Clinical Biochemistry of Domestic Animals (Sixth Edition)*, sixth edition ed., J. J. Kaneko, J. W. Harvey, and M. L. Bruss, Eds. San Diego: Academic Press, 2008, pp. 379–412. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123704917000131

[121] A. Laemmle, R. C. Gallagher, A. Keogh, T. Stricker, M. Gautschi, J.-M. Nuoffer, M. R. Baumgartner, and J. Häberle, "Frequency and pathophysiology of acute liver failure in ornithine transcarbamylase deficiency (otcd)," *PLoS One*, vol. 11, no. 4, p. e0153358, 2016.

[122] L. He, X. Cai, S. Cheng, H. Zhou, Z. Zhang, J. Ren, F. Ren, Q. Yang, N. Tao, and J. Chen, "Ornithine transcarbamylase downregulation is associated with poor prognosis in hepatocellular carcinoma," *Oncology letters*, vol. 17, no. 6, pp. 5030–5038, 2019.

[123] H. Uyttendaele, G. Marazzi, G. Wu, Q. Yan, D. Sassoon, and J. Kitajewski, "Notch4/int-3, a mammary proto-oncogene, is an endothelial cell-specific mammalian notch gene," *Development*, vol. 122, no. 7, pp. 2251–2259, 1996.

[124] K. M. Sokolowski, M. Balamurugan, S. Kunnimalaiyaan, T. C. Gamblin, and M. Kunnimalaiyaan, "Notch signaling in hepatocellular carcinoma: molecular targeting in an advanced disease," *Hepatoma Research*, vol. 1, pp. 11–18, 2015.

[125] J. Lu, Y. Xia, K. Chen, Y. Zheng, J. Wang, W. Lu, Q. Yin, F. Wang, Y. Zhou, and C. Guo, "Oncogenic role of the notch pathway in primary liver cancer (review) corrigendum in/10.3892/ol. 2016.5145," *Oncology letters*, vol. 12, no. 1, pp. 3–10, 2016.

[126] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, p. gkw943, 2016.

[127] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W90–W97, 2016.

[128] Y. Xu, M. Poggio, H. Y. Jin, Z. Shi, C. M. Forester, Y. Wang, C. R. Stumpf, L. Xue, E. Devericks, L. So *et al.*, "Translation control of the immune checkpoint in cancer and its therapeutic targeting," *Nature medicine*, vol. 25, no. 2, pp. 301–311, 2019.

[129] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature reviews genetics*, vol. 13, no. 4, pp. 227–232, 2012.

[130] S. Schlegel, E. Rujas, A. J. Ytterberg, R. A. Zubarev, J. Luirink, and J.-W. De Gier, "Optimizing heterologous protein production in the periplasm of e. coli by regulating gene expression levels," *Microbial cell factories*, vol. 12, no. 1, pp. 1–12, 2013.

[131] D. M. Francis and R. Page, "Strategies to optimize protein expression in e. coli," *Current protocols in protein science*, vol. 61, no. 1, pp. 5–24, 2010.

[132] N. Pancino, A. Rossi, G. Ciano, G. Giacomini, S. Bonechi, P. Andreini, F. Scarselli, M. Bianchini, and P. Bongini, "Graph neural networks for the prediction of protein–protein interfaces."

[133] M. Monaci, N. Pancino, P. Andreini, S. Bonechi, P. Bongini, A. Rossi, G. Ciano, G. Giacomini, F. Scarselli, and M. Bianchini, "Deep learning techniques for dragonfly action recognition." in *ICPRAM*, 2020, pp. 562–569.

[134] A. Rossi, G. Giacomini, V. Cicaloni, S. Galderisi, M. S. Milella, A. Bernini, L. Millucci, O. Spiga, M. Bianchini, and A. Santucci, "Akuimg: A database of cartilage images of alkaptonuria patients," *Computers in Biology and Medicine*, vol. 122, p. 103863, 2020.

[135] G. Giacomini, G. Ciravegna, M. Pellegrini, R. D'Aurizio, and M. Bianchini, "A transcriptional study of oncogenes and tumor suppressors altered by copy number variations in ovarian cancer," in *Innovation in Medicine and Healthcare*. Springer, 2020, pp. 159–169.

[136] A. Rossi, G. Vannuccini, P. Andreini, S. Bonechi, G. Giacomini, F. Scarselli, and M. Bianchini, "Analysis of brain nmr images for age estimation with deep learning," *Procedia Computer Science*, vol. 159, pp. 981–989, 2019.

[137] S. Bonechi, M. Bianchini, P. Bongini, G. Ciano, G. Giacomini, R. Rosai, L. Tognetti, A. Rossi, and P. Andreini, "Fusion of visual and anamnestic data for the classification of skin lesions with deep learning," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 211–219.