TECHNISCHE UNIVERSITÄT MÜNCHEN

FACHGEBIET FÜR BIOINFORMATIK

SEQUENCE AND STRUCTURE DETERMINANTS OF TRANSLATION

Fei Qi

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Aphrodite Kapurniotu
Prüfer der Dissertation:	
	1. Prof. Dr. Dmitrij Frischmann
	2. Prof. Dr. John Parsch (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 07.11.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 18.12.2017 angenommen.

Abstract

Translation is a fundamental process in gene expression. The regulation of translation process is crucial for both of the quantity and quality control of protein synthesis, and has long been in the spotlight of biological research. Although recent technical advances, such as high-throughput RNA structure probing and ribosome profiling, have greatly promoted the research of translation regulation, some problems still remain in this area. This thesis presents our attempts to solve several longstanding problems in the research of translation regulation.

We first investigated the dependence of sequence-structure relationships in mRNA coding regions on temperature, aiming to find a distinguishing feature for identification of functionally important RNA structural elements. Our result shows that high and low thermostability is indicative of functional RNA structural and sequence elements, respectively. Therefore, melting temperature is a crucial parameter, which highlights functional RNA structures and sequence segments. This work is present in chapter 2.

In chapter 3, we conducted an evolutionary analysis of polyproline motifs, which induce translational pauses in translation elongation. Our analysis reveals that polyproline motifs are disfavored during evolution because of their translational burdens. Against the overall trend of polyproline motif depletion, we observed their enrichment in the vicinity of translational start sites, in the inter-domain regions of multi-domain proteins, and downstream of transmembrane helices, which indicate the regulatory role of polyproline motifs in translation.

Chapter 4 explores the cooperation of codon usage, RNA structure and polyproline motif in the regulation of translation elongation. We found a cooperation of slow-translating codons and polyproline motifs, prolonging the translational pauses at these motifs. We also discovered a region containing fast-translating codons immediately downstream of the translational pause sites induced by RNA structures, which maintains the functionally important RNA structures during translation by ensuring an enough space between ribosomes at and downstream of the translational pause sites.

Collectively, our work will contribute to the study of the sequence and structure determinants of translation, and improve the overall understanding of the translation process.

Zusammenfassung

Translation ist ein fundamentaler Prozess bei der Expression von Genen. Regulierung des Translationsprozesses beeinflusst die Proteinsynthese sowohl quantitativ, als auch qualitativ und ist bereits lange Zeit im Mittelpunkt biologischer Forschung. Trotz neuster technischer Fortschritte, wie dem sogenannten High-troughput RNA structure probing und Ribosome profiling, bleiben weiterhin einige Schwierigkeiten in diesem Forschungsfeld bestehen. In der folgenden Arbeit versuchen wir einige langjährige Problemstellungen im Bereich der Translationsregulation zu adressieren und neue Lösungsansätze darzubieten.

Zunächst wurden mRNA-kodierende Regionen unter verschiedenen Temperaturzuständen untersucht. Hierbei lag der Fokus auf der Beziehung zwischen Sequenz und Struktur mit dem Ziel der Identifikation neuer RNA Strukturelemente, die von funktioneller Wichtigkeit sind. Die Ergebnisse in Kapitel 2 zeigen, dass Unterschiede in der Thermostabilität bezeichnend für die funktionelle RNA Struktur und das zugehörige Sequenzelement sind. Die Schmelztemperatur erwies sich als ein zentraler Parameter im Hinblick auf die Sequenz und Struktur Beziehung.

Kapitel 3 umfasst die evolutionäre Analyse von Polyprolin-Motifen, welche Unterbrechungen in der Elongationsphase des Translationsprozesses induzieren können. Es konnte gezeigt werden, dass Polyprolin-Motife auf Grund ihres Einflusses auf die Translation evolutionär benachteiligt sind. Eine Anreicherung von Polyprolin-Motifen konnte im Bereich des Translationsstarts, in inter-domain Regionen von multi-domain Proteinen und downstream der Transmembranhelices gefunden werden. Daraus lässt sich auf eine regulatorische Rolle der Polyprolin-Motife im Translationsvorgang schließen.

Das Zusammenspiel von codon usage, RNA Struktur und Polyprolin-Motifen im Bezug auf die Regulierung der Elongation wird in Kapitel 4 erläutert. Eine Verbindung aus langsam-translatierenden Codons und Polyprolin-Motifen zeigte eine Verlängerung der Translationspausen. Desweiteren konnten wir schnell-translatierende Codons identifizieren, welche direkt downstream der RNA Strukturen mit translations-verlangsamenden Codons lokalisiert sind. Schnell-translatierende Codons sorgen während der Translation für eine stabile Distanz zwischen Ribosomen und den Regionen der Translationspausen. Daher wird angenommen, dass sie für die Aufrechterhaltung einer funktionellen RNA Struktur zuständig sind.

Diese Arbeit trägt zu einem besseren Verständnis von Sequenz- und Struktureigenschaften im Translationsprozess bei und gibt neue Einblicke in den Ablauf der Translation.

Publications

1. **Qi**, **F.** and Frishman, D. (2017) Melting temperature highlights functionally important RNA structure and sequence elements in yeast mRNA coding regions. *Nucleic Acids Research*, 45(10), 6109–6118. doi: 10.1093/nar/gkx161.

2. **Qi**, **F.**, Motz, M., Jung, K., Lassak, J. and Frishman, D. (2018) Evolutionary analysis of polyproline motifs in *Escherichia coli* reveals their regulatory role in translation. *PLOS Computational Biology*, 14(2), e1005987. doi: 10.1371/journal.pcbi.1005987.

Acknowledgements

It is almost 5 years since my arrival in German to pursue a PhD. In this long journey, I received helps from a lot of people, and I owe a debt of gratitude to all of them.

First, I would like to thank my supervisor, Prof. Dmitrij Frishman, for giving me the opportunity to work in his group. I am really appreciate all his guidances in the scientific work and his helps in my life.

I would also like to thank all secretaries in our Department for all their administrative helps. Special thanks to Mrs. Leonie Corry for all her helps in my life.

I am grateful to my collaborators in Ludwig-Maximilians-Universität München, Prof. Kirsten Jung, Dr. Jürgen Lassak and Magdalena Motz, for the fruitful cooperation and their helps in preparing the manuscript. I would also like to thank Prof. John Parsch for his insightful discussion.

I wish to express my appreciation to all my colleagues in our group. Special thanks to Hongen Xu, Yanping Zhang, Hengyuan Liu, Jinlong Ru, Peter Hönigschmid, Evans Kataka, Martina Weigl, Usman Saeed, Andrey Chursov, Michael Kiening, Anja Mösch and Drazen Jalsovec for all their helps and the conversations during my doctoral study.

I also wish to thank all my friends, Haitao Liu, Saiqi Yang, Tingting Chen, Guo Chen, Shuai Zhao, Hui Wang, Di Xu, Baopeng Ma, Fang Yang, Xiangdong Zhao, Chendong Zheng, Chenguang Zhou, Sufu Gan, Yu Zhuang and Wenbo Yu. I enjoy all the leisure time we spent together.

I am thankful for the financial support from the China Scholarship Council, and also grateful to Prof. Dmitrij Frishman and Prof. Kirsten Jung for their kindly financial support.

Finally, I would like to give special thanks to all my families, especially my mother Xiaofang Zhang and my wife Yi Yang. I cannot make it without your support.

Contents

List of Figures

List of Tables

Abbreviations

a.u.	Arbitrary unit
CAMT	Codon-anticodon matching time
EF-P	Elongation factor P
Frag-seq	Fragmentation sequencing
IF-5A	Initiation factor 5A
IRES	Internal ribosome entry site
ncRNA	Non-coding RNA
nTR	Non-transmembrane region
PARS	Parallel analysis of RNA structure
PARTE	Parallel analysis of RNA structures with temperature elevation
PCPF	Proline codon pair frequency
PGL	Propensity of gene loss
PPI	Poly proline helix I
PPII	Poly proline helix II
pS	Synonymous polymorphism rate
PSEC	Propensity of stalling effect change
RMSD	Root-mean-square deviation
SEM	Standard error of the mean
SHAPE-seq	Selective 2′-hydroxyl acylation analyzed by primer extension sequencing
SRP	Signal recognition particle
T_m	Melting temperature
TMH	Transmembrane helix
UTR	Untranslated region
ValS	Val-tRNA synthetase

Chapter 1

Background

1.1 Translation elongation

The process by which a functional gene product is synthesized based on the genetic information from a gene is called gene expression [1]. Although the productions could be functional RNAs [2], they are often proteins. In such case, translation is a fundamental stage of the gene expression. Translation is the process that ribosomes synthesize proteins using an mRNA as template. Translation consists of 3 phases: initiation, elongation and termination. Initiation involves the formation of the initiation complex, which contains the ribosome associated with the target mRNA and the first aminoacyl-tRNA [1]. Elongation encompasses all reactions of the synthesis of the polypeptide [1]. Termination includes all steps of the release of the completed polypeptide and the dissociation of the ribosome from the mRNA [1]. In this work, we focused on the translation elongation process.

Translation elongation proceeds in a cycle of 3 steps. The first step is the matching of mRNA codon to specific tRNA anticodon [3]. In this step, an aminoacyl-tRNA is placed in the ribosome A site by an elongation factor [3]. Any aminoacyl-tRNA can be carried into the A site by the elongation factor [1]. However, only the one whose anticodon correctly pairs to the mRNA codon can make stabilizing contacts with ribosomal RNAs, which hold the aminoacyl-tRNA in the A site and promote the next step of elongation [1]. Incorrectly paired aminoacyl-tRNAs cannot make these stabilizing contacts and thus diffuse out of the A site [1]. The second step is the formation of peptide bond, i.e. the ribosomal peptidyl transfer reaction [3]. This reaction transfers the polypeptide from the P-site peptidyl-tRNA to the aminoacyl-tRNA in the A site, and thus adds an amino acid residue to the polypeptide [3]. The last step is the so-called translocation [3]. In this step, the ribosome moves one codon along the mRNA, the peptidyl-tRNA is placed in the P site and the deacylated tRNA gets out of the ribosome through ribosomal E site. Thus, the A site becomes empty for the next aminoacyl-tRNA [4]. These steps repeat to synthesis the polypeptide until the translation termination.

1.2 Regulation of translation elongation rate

Translation elongation is a non-uniform process and subjected to strict regulation [5, 6], both in term of the overall and the intra-molecular variation of the elongation rate [7–10]. The overall elongation rate controls the quantity of the translation products while the intra-molecular variation of the elongation rate coordinates multiple co-translational process and thus ensures the quality the synthesized proteins [7–10]. Several factors have been found to regulate the translation elongation rate, including codon usage, RNA secondary structure, and amino acid sequence [5, 11–14].

1.2.1 Codon usage

Codon degeneracy—multiple codons specify a same amino acid—is an important characteristic of the genetic code, which improves the flexibility of the mRNA sequences. The existence of synonymous codons makes it possible for the mRNAs to carry additional information while keeping their primary function—encoding the amino acid sequences of the proteins [10, 15, 16]. Many of the synonymous codons could be recognized by different spices of tRNAs. Since species of tRNAs usually differ in abundance, synonymous codons are therefore translated at different rate. Codon pairing to low-abundance isoaccepting tRNAs requires a longer time for the cognate aminoacyl-tRNA to get into the ribosome A site, and is thus translated at a lower speed than codon pairing to high-abundance isoaccepting tRNAs [17]. Codon usage is exploited as a kinetic control of translation elongation [10], and subjected to strict evolutionary selection [6]. A typical example is the unequal usage of synonymous codons—slow-translating codons are used much more rarely than fast-translating codons, which reflects an adaption of the codon usage to the available tRNA pool and optimizes the efficiency of protein synthesis [15, 18]. The evolutionary constraint is also indicated by the fact that many synonymous mutations in human genes are associated with disease [10].

The regulatory role of codon usage on translation elongation is relatively well studied. First, the usage of genetic codons regulates the overall translation elongation rate. Ikemura discovered a strong positive correlation between tRNA abundance and codon usage in both of *Escherichia coli* and yeast, and found that the extent of this correlation depends on the expression levels of individual genes [18]. As a consequence, codon bias exists between genes with high and low expression level [19]. Highly expressed gene preferentially uses fast-translating codons to maximize the translation efficiency, and thus ensures a high expression level of the encoded protein [15, 18]. Codon usage also plays an important role in the regulation of the intra-molecular variation of elongation rate. Sharp et al. introduced a measurement named relative synonymous codon usage [19]. They suggested that changes of relative synonymous codon usage are associated with changes of the local translation elongation rate [10, 19]. Several experimental works proved this concept and demonstrated that even a single codon change can have pronounced effect on location elongation rate [20–22].

In recent years, genome-wide studies were applied to identify conserved patterns of the distribution of fast- and slow-translating codons across individual mRNAs and in the transcriptome [10]. In 2010, Tuller et al. found an enrichment of slow-translating codons in the first 50 codons of mRNAs, which slows down the translation elongation immediately after initiation [23]. This so-called "codon ramp" maintains a suitable distance between adjacent ribosomes and thus prevents ribosome congestions, which may result in misfolded or truncated proteins [10, 23–25]. After the ramp, translation elongation goes into a rapid phase [10, 26], which is accomplished by using the same codon for amino acid recurring in the protein sequence [10, 27]. This reuse of codons enables a rapid recycle of tRNAs [10, 27], since the aminoacyl tRNA synthetases can form complexes associated with ribosomes [28, 29].

The usage of synonymous codons also coordinates multiple co-translational processes. This is achieved by clusters of fast- and slow-translating codons, which occur at strategic locations of the co-translational processes [10]. First, the non-uniform distribution of synonymous codons with different translation rate fine-tunes the co-translational folding of proteins [9, 30, 31]. In co-translational protein folding, structural elements of a protein may influence each other [32]. Due to the cooperativity between different parts of the structure, the timing of translation is crucial for proper folding [11]. It has been known for a long time that the clusters of slow-translating codons tend to occur in the inter-domain linkers of multi-domain proteins [15]. These clusters induce translational pauses between structural domains, which provide time delays in translation elongation and thus facilitate independent folding of domains to minimize the chance of misfolding [15, 16, 33]. In 2014, O'Brien et al. found that fast translation of mRNA stretches coding for structural domains decreases their chances of misfolding, indicating that the fast-translating codons also play an important role in coordinating co-translational protein folding [34].

Another typical co-translational process is the targeting of α-helical transmembrane proteins to the translocons, mediated by the signal recognition particle (SRP), and their insertion into the membrane [35, 36]. This process has been found to be facilitated by translational pauses [8, 37–41]. Pechmann et al. found clusters of slow-translating codons located 35–40 codons after the SRP-binding site [41]. Considering that about 28 amino acids are required to span the ribosome exit tunnel [42], these clusters of slow-translating codons induce a translational pause just after the SRP-binding elements protruded from the ribosome exit tunnel, which promotes the SRP-mediated co-translational recognition and targeting of transmembrane proteins [41]. In the fungus *Emericella nidulans*, Dessen and Képès identified two translational pauses occurring approximately 45 and 70 codons downstream of transmembrane helices (TMHs), caused by clusters of slow-translating codons [38]. Given that about 28 amino acids can be accommodated in the ribosome exit tunnel [42], the first pause may occur after the TMH has emerged from the ribosome exit tunnel and is being inserted into the membrane by translocon, and the second pause may occur when two TMHs are forming a hairpin [38]. Therefore, the time delays caused by the two translational pauses may facilitate the efficient insertion of TMH [38].

Collectively, the codon usage profile of genes influences the overall and intra-molecular variation of the translation elongation rate, and is thus crucial for both of the quantity and quality control of protein synthesis.

1.2.2 RNA secondary structure

An RNA molecular usually contains only one nucleic acid polymer strand, but every RNA chain tends to fold back on itself for thermodynamic reasons [43]. RNA secondary structure refers to the base pairing interaction within a single RNA molecular. Stems, loops, bulges and junctions are the 4 basic elements of the RNA secondary structure, and stem-loop is the most common RNA structural element [44].

These structural elements, as well as the global folding patterns, play a fundamental role in the function and regulation of RNAs [45–50]. For some non-coding RNAs (ncRNAs), such as rRNAs and tRNAs, their structures are crucial, and disruptions on theirs structures may have lethal consequences. For a number of years, most of the attention was given to functional secondary structures of ncRNAs, but more recently mRNA structure also moved to the spotlight of genomics and bioinformatics research. Secondary structures of the three functional mRNA domains — 5′ UTR, the coding region and 3′ UTR — are largely independent, since base pairs across domain borders are rare [51]. While the primary function of the coding regions is to encode amino acid sequences of proteins, they are presumed to contain even more RNA secondary structures than UTRs [45]. mRNA structure has been found influencing multiple stages of gene expression and protein synthesis [52], and regulating the translation elongation is one of the important regulatory roles it plays [12].

How do RNA structures influence translation elongation? First, stable local RNA secondary structures induce translational pauses [53]. Translating ribosomes must unwind the RNA structures they encounter to move along the mRNA, and thus stable local structural elements could act as energetic hurdles, which arrest the translating ribosomes and induce translational pauses [53]. These translational pauses are involved in the regulation of co-translational protein folding [54]. Bartoszewski et al. investigated the ΔF508 mutation in the human cystic fibrosis transmembrane conductance regulator, which is the most frequent cause of cystic fibrosis [55]. The ΔF508 mutation results in the loss of phenylalanine at the 508 position and a synonymous codon change of the isoleucine (ATC to ATT) at the 507 position of the protein, and eventually leads to protein misfolding [55]. Bartoszewski et al. identified that the synonymous codon change alters the RNA secondary structure, prolongs the translational pause at its position, and thus causes misfolding [55]. Recently, Del Campo et al. identified 71 translational pauses induced by mRNA secondary structures in *E. coli* mRNAs, by combining high-throughput RNA secondary structure probing method and ribosome profiling data [56]. Several studies found stable local RNA structures at the beginning of mRNAs, which may have a similar function as the aforementioned codon ramp [10, 57, 58].

Although stable local RNA structures slow down the translation elongation, a strong positive correlation was found between the global mRNA structuredness of and the protein abundance [59, 60]. Genes with higher expression level are under stronger selection pressure for their mRNA to fold [59]. This is surprising, since one would expect that highly folded mRNA has a decreased translation elongation rate and thus a low expression level. Mao et al. investigated this counterintuitive phenomenon, taking into account the dynamics of RNA secondary structure during translation [61]. They found that RNA secondary structures shorten the distance between ribosomes during translation elongation [61]. When two adjacent ribosomes are close enough, RNA structures between them disappear [61]. Therefore, in highly structured regions of mRNAs, the RNA secondary structures result in a shorter ribosomal distance, which in turn eliminates the RNA structures and leads to a higher translation elongation rate [61]. This finding explains the counterintuitive correlation between mRNA structuredness and protein abundance [61]. However, another question arises in this context: how do the functionally important mRNA structures, for example the structures inducing functional translational pauses, get maintained during translation to affect every passing ribosome? Although Mao et al. hypothesized that it may be achieved by cooperation between codon usage and RNA secondary structure [61], this question remains unstudied.

In recent years, several experimental approaches have been developed for genome-wide probing of RNA structures. These approaches greatly promoted the study of mRNA structures, which were for a long time blocked by the low-throughput of traditional RNA structure probing methods and the inaccuracy of computational RNA structure prediction programs [52]. These methods, including fragmentation sequencing (Frag-seq) [62], parallel analysis of RNA structure (PARS) [45] and selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq) [63] combine RNA structure probing by structure-specific enzymes and chemical modifications at double- or single-stranded bases with high-throughput next generation sequencing. The experiment can probe millions of molecules at single nucleotide resolution within one experiment and therefore enables comprehensive study of RNA structures [64]. Although local structural elements and global folding patterns of mRNAs can be gleaned from these experiments, challenges remain in study of mRNA structures. An important challenge is to distinguish functionally important structural elements from non-functional ones, since every RNA chain tends to fold back on itself for thermodynamic reasons [43]. This challenge, to some extent, is limiting the progress of mRNA structure studies.

1.2.3 Amino acid sequence

Translation elongation rate is also influenced by the amino acid sequences of the encoded proteins. A variety of amino acid sequence segments were found inducing translational pauses either by interacting with the ribosome exit tunnel or interfering with the peptidyl transfer reaction [5, 13, 14].

During translation elongation, the growing nascent polypeptide chain goes through the ribosomal exit tunnel before emerging from the ribosome. The exit tunnel is negatively charged and requires about 28 amino acids to span [42, 65]. It is found that some amino acid sequence segments chemically or electrostatically interact with the ribosome exit tunnel and thus arrest the translation elongation [65–68]. This phenomenon is also captured by several genome-wide ribosomal profiling studies [69–71]. Additionally, Ramu et al. found that specific nascent polypeptide chains in the exit tunnel affects functional properties of the A site of the peptidyl transferase center, which prevents a particular subset of amino acids from being incorporated into the nascent polypeptide chain and thus leads to ribosome stalling [14].

In addition to the nascent-polypeptide-mediated translational pause, the substrates of the ribosomal peptidyl transfer reaction also affect translation elongation [5]. The elongation rate strongly depends on the amino acids to be incorporated into the growing polypeptide chain [6], and the incorporation of proline is especially slow [72–74]. The pyrrolidine ring of proline gives it an exceptional conformational rigidity compared to all other amino acids, which makes it not only a poor A-site peptidyl acceptor [74], but also a poor P-site peptidyl donor [72, 73]. Translation of two and more consecutive prolines dramatically impairs the peptidyl transfer reaction and leads to ribosome stalling [72, 75–79].

Although decreasing translation efficiency, consecutive proline sequences are important for protein structure and function. Proline is unique in terms of being the sole amino acid to adopt *cis* and *trans* conformations, both of which are nearly energetically equal and naturally occur in proteins [80–82]. Notably, a sequence of consecutive prolines results in the formation of either the right-handed poly proline helix I (PPI) or the left-handed poly proline helix II (PPII). Beside α-helix and β-sheet, PPII helix is considered to be the third major secondary structure element in proteins, and plays an important role in mediating protein-protein and protein-nucleic acid interactions [83–86]. Three consecutive prolines are also an integral part of the active center in the universally conserved Val-tRNA synthetase (ValS) [87]. The proline triplet in ValS is essential for efficient charging of the tRNA with valine and for preventing mischarging by threonine. These two examples illustrate why nature has evolved a specialized translation elongation factor, referred to as EF-P in bacteria or a/eIF-5A in archaea and eukaryotes, to alleviate ribosome stalling at consecutive prolines [5, 72, 75–79, 88]. However, EF-P and a/eIF-5A cannot fully prevent translational pauses imposed by the translation of consecutive prolines [5].

In this context, two important questions arise. Given that the evolutionary selection for high efficiency of protein synthesis shapes mRNA sequences, for example the adaption of codon usage to the available tRNA pool [89, 90], the first question is that whether translation-efficiency-based evolutionary pressure also shapes protein sequences. This question could be answered by investigating that whether the amino acid sequence segments causing translational pause are under selection due to their impairment on translation efficiency. Recently, Sabi and Tuller found that short peptides which induce ribosome stalling in yeast by interacting with the ribosomal exit tunnel, tend to be either over- or underrepresented in the proteome [91]. They hypothesized that those short peptide sequences were under evolutionary selection based on their synthetic efficiency [91]. However, still little is known about the evolution of consecutive proline sequences. The second question is that whether those amino acid sequence segments are purely translational burden and genes carry them only because of their protein structure/function, or the translational pauses they induced are widely exploited as a regulatory mechanism of translation. Although the effect of amino acid sequence segments on translation has recently attracted growing attention, this question remains largely explored.

1.2.4 Cooperation of factors to regulate translation elongation

Most studies of the regulation of translation elongation focus on the individual effect of each aforementioned factor. However, the fact that codon usage and RNA structure both contribute to the slow ramp at the beginning of mRNAs indicates that multiple factors may be operational in the same regulatory mechanism [10]. Yet few studies have been conducted on the cooperation of these factors. In 2014, Mao et al. found a combined effect of codon usage and RNA secondary structure increasing translation efficiency [61]. In the next year, Gorochowski et al. identified a trade-off between tRNA-abundance based codon usage and RNA secondary structure, which results in a smoothed translation elongation rate [7]. Despite these 2 studies, the whole picture and the detail about the cooperative manner in which those factors work together remain largely unclear.

1.3 The objective and outline of this work

In this work, our general objective is to improve the understanding of the sequence and structure determinants of translation, especially the elongation process. Aiming this goal, we made comprehensive attempts to answer some of the aforementioned questions: (i) how to distinguish functionally important RNA structural elements from non-functional ones; (ii) are sequences of consecutive prolines subjected to evolutionary selection due to their impairment on translation efficiency; (iii) do the amino acid sequence segments inducing ribosome stallings play a regulatory role in translation; and (iv) how do multiple factors cooperate to regulate translation elongation.

The results of our contributing efforts are presented in the following chapters. In chapter 2, we mainly tested the feasibility of discriminating between functionally important and unimportant RNA structural elements by their thermostability. We found that high thermostability highlights functionally important RNA structural elements while low thermostability indicates functional RNA sequence segments. Chapter 3 presents our evolutionary analysis of consecutive proline sequences in *E. coli* proteomes. Our analysis reveals an evolutionary selection pressure against the sequences of consecutive prolines as a consequence of the reduced translation efficiency. We also observed an unequal distribution of these sequences in proteins, indicating their involvement in the regulation of translation. Chapter 4 describes our investigation on the cooperation of amino acid sequence, codon usage and RNA structure to regulate translation elongation. Our result shows a correlation between slow-translating codons and consecutive proline sequences which may results in longer translational pauses. We also discovered a mechanism utilizing fast-translating codons to maintain functional RNA structures during translation elongation.

Finally, the last chapter summarizes our conclusions and briefly discusses the possible applications of our findings.

Chapter 2

Melting Temperature Highlights Functionally Important RNA Structure and Sequence Elements in Yeast mRNA Coding Regions

Secondary structure elements in the coding regions of mRNAs play an important role in gene expression and regulation, but distinguishing functional from non-functional structures remains challenging. Here we investigate the dependence of sequence-structure relationships in the coding regions on temperature based on the recent PARTE data by Wan et al. Our main finding is that the regions with high and low thermostability (high T_{m} and low T_{m} regions) are under evolutionary pressure to preserve RNA secondary structure and primary sequence, respectively. Sequences of low T_{m} regions display a higher degree of evolutionary conservation compared to high T_{m} regions. Low T_{m} regions are under strong synonymous constraint, while high T_{m} regions are not. These findings imply that high T_{m} regions contain thermostable functionally important RNA structures, which impose relaxed evolutionary constraint on sequence as long as the base-pairing patterns remain intact. By contrast, low thermostability regions contain single-stranded functionally important conserved RNA sequence elements accessible for binding by other molecules. We also find that theoretically predicted structures of paralogous mRNA pairs become more similar with growing temperature, while experimentally measured structures tend to diverge, which implies that the melting pathways of RNA structures cannot be fully captured by current computational approaches.

This chapter has been published as: Qi, F. and Frishman, D. (2017) Melting temperature highlights functionally important RNA structure and sequence elements in yeast mRNA coding regions. *Nucleic Acids Research*, 45(10), 6109–6118. doi: 10.1093/nar/gkx161.

2.1 Introduction

Secondary structure elements and global folding patterns play a fundamental role in the function and regulation of RNAs [45–50]. For a number of years, most of the attention was given to functional secondary structures of ncRNAs, but more recently mRNA structure also moved to the spotlight of genomics and bioinformatics research. mRNA structures have been found to influence multiple stages of gene expression and protein synthesis, including transcription, splicing, RNA transport, translation initiation, elongation and termination, as well as RNA degradation [12, 43, 52, 92–95]. Secondary structures of the three functional mRNA domains — 5′ UTR, the coding region and 3′ UTR — are largely independent, since base pairs across domain borders are rare [51]. A broad variety of functional structural elements were described in UTRs [96–99]. While the primary function of the coding regions is to encode amino acid sequences of proteins, they are presumed to contain even more RNA secondary structures than UTRs [45], with some of them already proven to be functional [57, 100–104]. The redundancy of the genetic code makes it possible for the coding regions to carry overlapping functions, which manifest themselves at the level of protein and RNA sequences and structures [51, 105, 106].

In recent years, several experimental approaches have been developed for genome-wide measurement of RNA structures. These methods, including Frag-seq [62], PARS [45] and SHAPE-seq [63] combine RNA structure probing by structure-specific enzymes and chemical modifications at double- or single-stranded bases with high-throughput next generation sequencing. These approaches can probe millions of molecules at single nucleotide resolution within one experiment and therefore enable comprehensive studies of RNA structures [64]. An important question, which arises in this context, is to which extent the results of high-throughput structure probing experiments are reproducible and compatible with each other.

Owing to the availability of RNA structure probing data many of the classical problems in molecular evolution, which have been extensively addressed for protein molecules, can now be examined for mRNAs as well. In particular, it is of great interest to investigate to which extent secondary and/or tertiary structure of mRNAs constrains sequence variation and how strongly mRNA structures are conserved in evolution, as was done for protein 3D structures long ago [107]. Recently, Wan et al. published parallel analysis of RNA structures with temperature elevation (PARTE) experiment, in which secondary structures of yeast RNAs were probed and melting temperatures (T_m) were derived at single nucleotide resolution at five temperatures [43]. Using these data, we demonstrate that high and low thermostability regions in the mRNA coding regions highlight functionally important RNA structures and sequence segments, respectively. We report a surprising pattern of structural divergence between sequence-similar mRNAs along the temperature ladder, which cannot be captured by the currently available computational approaches. There is a considerable reproducibility between the high-throughput RNA structure probing experiments, PARS and PARTE.

2.2 Materials and Methods

2.2.1 Experimental data on secondary structures of yeast mRNAs

Secondary structure profiles of 3002 yeast mRNAs determined at room temperature by PARS experiment were downloaded from `http://genie.weizmann.ac.il/pubs/PARS10` [45]. For each individual nucleotide position of mRNAs, a PARS score reflects its likelihood to be in a double-stranded conformation based on the number of sequencing reads upon treatment by two structure-specific enzymes, RNase V1 and nuclease S1, which cleave at double-stranded and single-stranded regions, respectively. A total of 4 405 020 bases in the 3002 mRNAs are covered by PARS scores. For a given mRNA sequence, the vector of its PARS scores is referred to as its PARS structure.

Another mRNA structure dataset used in this work was obtained by a PARTE experiment, in which 4562 yeast mRNAs were structure-probed by RNase V1 at five temperatures (23, 30, 37, 55 and 75℃; two biological replicates were performed for each temperature) [43]. PARTE reveals T_m of each base, with double-stranded regions being progressively eliminated as temperature increases. V1 reads resulting from this experiment were downloaded from the GEO database [108] (GSE39680), and their counts were normalized exactly as described by Wan et al. [43]: (i) for each library, peaks were defined as those bases that are covered by more reads than the bases on their left and their right and whose read coverage is greater than the average coverage of bases on the same gene and the average coverage of all bases; (ii) using the PoissonSeq algorithm [109], the library size of each sequencing lane was estimated based on the high confidence peaks observed in both duplicate libraries at at least one of the five temperatures; and (iii) V1 read numbers were normalized by dividing the counts in each sample by the corresponding library size. For each RNA nucleotide position the $\log_2$ value of the mean of the two normalized V1 read numbers from two duplicate samples was treated as its PARTE score (each mean V1 read number was augmented by 0.001 to avoid the undefined logarithm values for those bases where the read count is zero). A total of 7 497 468 bases in the 4562 mRNAs are covered by PARTE scores. For a given mRNA sequence, vectors of its PARTE scores are referred to as its PARTE structure.

2.2.2 Predicted secondary structures of yeast mRNAs

Sequences of 6686 yeast mRNAs were downloaded from SGD (release 57-1-1) [110]. Base pairing probabilities for each yeast mRNA were calculated using the RNAfold algorithm from the ViennaRNA package [111]. In order to simulate the PARTE experiment, which was carried out at five different temperatures, RNAfold was run five times for each yeast mRNA using five different values of the `-T` parameter (23, 30, 37, 55 and 75). For a given mRNA, vectors of its theoretically predicted base pairing probabilities are referred to as its predicted structures.

2.2.3 Yeast paralogous mRNAs

We considered 246 pairs of aligned mRNA sequences of yeast paralogs, as well as the percent identity between aligned coding regions of each pair, as described previously [112]. Each of these pairs of paralogous proteins shares over 50% amino acid sequence identity and < 10% difference in sequence length.

2.2.4 Distances between secondary structures of yeast paralogs

For a given pair of aligned mRNA sequences, we employed root-mean-square deviation (RMSD) as the measure of the distance between their experimental structures. RMSD values were calculated between the vectors of PARS or PARTE scores for all aligned positions without gaps. Sequence positions not probed in the experiments (i.e. those with read number 0) were also taken into account in this calculation — their PARS score equals 0 and their PARTE score is $\log_2(0 + 0.001)$. Similarly, distances between predicted structures for a given pair of aligned mRNAs were calculated as RMSD between vectors of predicted base pairing probabilities.

2.2.5 Paired and unpaired bases of yeast mRNAs

We subdivided the bases of yeast mRNAs into two classes according to their PARS scores: (i) paired bases (PARS score ≥ 0) and (ii) unpaired bases (PARS score < 0). In all the mRNAs covered by both PARS and PARTE experiments we identified 3 514 124 paired bases and 884 030 unpaired bases.

2.2.6 Melting temperatures of RNA structures in yeast mRNAs

Data on T_m of RNA structures covering over 320 000 bases in yeast mRNAs were kindly provided by Yue Wan and Howard Y. Chang from the Howard Hughes Medical Institute and the Program in Epithelial Biology at Stanford University School of Medicine. For each RNA sequence position, the T_m value was calculated as the mean of the two temperatures between which the position transitioned from the double-stranded to the single-stranded state in the PARTE experiment, i.e. 26.5, 33.5, 46 and 65℃ for the transitions between 23 and 30℃, 30 and 37℃, 37 and 55℃, and 55 and 75℃, respectively. Positions that remained double-stranded at 75℃ were assigned the T_m value of 80℃ [43]. We only considered 1262 mRNAs that have at least 5% of bases with probed melting temperatures.

2.2.7 Regions with high or low T_m in pairs of paralogous yeast mRNAs

We applied a sliding window of 100 nt with a step size of 10 bases to the alignments of paralogous yeast mRNAs and calculated average T_m values for each sequence in a window. Only those windows in which the alignment had < 10% of gaps and the two aligned sequences both had at least 10% of bases with probed T_m were included in the analysis. We defined two classes of windows — high T_m windows and low T_m windows — dependent on whether the two aligned sequences in a window both had T_m values among the top or bottom 25% of all T_m values. Windows with the middle range of T_m values were excluded from further analysis. For each window, the T_m values of the two aligned sequences were calculated as the arithmetic mean of all the T_m values of their individual bases. Overlapping windows belonging to the same class (high or low T_m) were merged, yielding the total of 167 regions with high T_m and 96 regions with low T_m among the 246 pairs of paralogous mRNAs.

2.2.8 Thermostable and meltable positions in pairs of yeast paralogous mRNAs

For each pair of yeast paralogous mRNAs, a position in the alignment was defined as thermostable or meltable dependent on whether both aligned bases in this position had $T_\mathrm{m} \geq 65$°C or $T_\mathrm{m} \leq 46$°C, respectively. This procedure yielded 1323 thermostable and 256 413 meltable positions, out of which 734 thermostable and 20 790 meltable positions were located in high T_m regions while low T_m regions contained 22 765 meltable positions and no thermostable position.

2.2.9 Synonymous base substitutions between yeast paralogs

The synonymous polymorphism rate (pS) of yeast paralogs was estimated by the equation $pS = Sd/S$, where Sd is the number of observed synonymous substitutions and S is the number of potential synonymous substitutions. In this work, we employed the SNAP software [113] to calculate the pS value. We compared the pS value of each high/low T_m region (pS_{region}) with the pS value of the entire alignment of yeast paralogous mRNAs ($pS_{alignment}$) by calculating $\Delta pS = pS_{region} - pS_{alignment}$.

2.2.10 Zipcodes in yeast mRNAs

Positions of 12 functional motifs in yeast mRNAs responsible for binding with mRNA transport proteins — the so-called zipcodes — were obtained from the study of

Jambhekar et al. [114]. Five of these zipcodes (ASH1-E1min, TPO1N, ERG2N, WSC2C and SRL1C), which are located in the coding regions and covered by the PARTE melting temperature data, were included in further analysis.

2.3 Results

2.3.1 Correlation between PARS and PARTE scores

Reproducibility of results is a crucial aspect in the evaluation of experiment strategies. To assess the correlation between the PARS and PARTE data, we computed Spearman's rank correlation coefficients between PARS and PARTE scores over all 4 398 154 bases in those 2995 mRNA sequences that are contained both in the PARS and in the PARTE datasets. The correlation coefficients were relatively low (0.325, 0.323, 0.321, 0.312 and 0.250 at the PARTE temperatures of 23, 30, 37, 55 and 75℃, respectively) but highly significant (all P-values < $2.2e^{-16}$). As expected, the highest correlation was detected at 23℃ as it is closest to the room temperature at which the PARS experiment was carried out. The lower correlation at high temperatures can be explained by progressive unfolding of RNA structures.

Correlation between PARS and PARTE scores is not surprising given the similarity of these two experimental strategies [43, 45]. PARS and PARTE scores both reflect the likelihood of individual bases to be in a double-stranded conformation [43, 45]. The relatively low correlation coefficients are due to a key difference between the PARS and the PARTE experiments—the enzymes used to detect RNA structures. While PARS relies both on RNase V1 and nuclease S1 to probe the bases in a double- and single-stranded conformation, respectively, PARTE only probes bases in a double-stranded conformation by RNase V1 [43, 45]. Therefore, while PARS can capture the likelihood of bases to be in a single-stranded conformation based on reads stemming from the nuclease S1, the PARTE experiment does not deliver this information. Indeed, the correlation between PARS and PARTE scores was much stronger (correlation coefficient 0.587, P-value < $2.2e^{-16}$) when only bases in double-stranded conformation were considered (data not shown).

2.3.2 Dependence of structure divergence on sequence identity

In our previous work, we explored sequence-structure relationships in yeast mRNAs based on PARS data [112]. Upon comparing secondary structures between sequence-similar paralogous yeast mRNAs, we found that coding regions of mRNAs are not under strong evolutionary pressure to preserve a particular global shape, which implies that global secondary structure of the coding regions does not play a major role in gene regulation. The recent availability of PARTE data for yeast mRNAs [43] has made it possible to investigate sequence-structure divergence at different temperature levels. As seen in Figure S2.1, at all five temperatures the similarity of PARTE structures shows no correlation with the sequence similarity in the range of sequence identity between 50% (the lowest level considered) and roughly 85–90%. In this range, the distance between experimental structures of paralogous mRNAs does not differ from the median distance between randomly selected mRNA pairs (dashed horizontal lines in Figure S2.1). By contrast, at sequence identity levels over 85–90%, the distance between experimental structures of paralogous mRNAs displays a near linear dependence on sequence identity (Figure S2.2 and Table S2.1).

This finding is in line with our previous analysis of structure probing data obtained by the PARS method [45], in which we found that the global structural conformation of the coding regions is not crucial for gene expression and regulation. This result is compatible with the notion that mRNA conformation depends on interactions with the solvent as well as with proteins and other ligands and that mRNAs adopt a highly dynamic ensemble of conformations instead of a single global structure [52]. An important insight provided by our analysis is that interrogation of mRNA structures by PARS and PARTE leads to qualitatively similar evolutionary conclusions, indicating the reproducibility of the high-throughput RNA structure probing experiments.

2.3.3 Variation of the distance between paralogous mRNA structures along the temperature ladder

The availability of the PARTE structure-probing data opens up the possibility to investigate how the distance between secondary structures of similar RNA molecules varies with temperature and to obtain clues about the RNA structure unfolding pathways during the melting process. We therefore calculated structural distances between paralogous yeast mRNAs along the temperature ladder. Intuitively, one would expect the structural distance to be inversely proportional to temperature: as the temperature grows, more and more base pairs melt, and an ever-increasing portion of both molecules becomes single-stranded and, thus, more similar to each other. However, the experimentally determined PARTE structures show a strikingly different behavior. The distances between randomly selected and all paralogous mRNA pairs do not appear to vary with temperature at all, while the distances between highly similar paralogous mRNA pairs actually become larger at higher temperatures (Figure 2.1A). We speculate that this surprising pattern may, to some extent, be due to the limitations of the experimental approach. First, the PARTE experiment probes the *in vitro* re-folded RNA structures rather than *in vivo* structures [43]. Second, as noted by Wan et al., in the PARTE data 20% of the bases show a transition for increased V1 reads at higher temperatures, which may indicate that a considerable proportion of thermostable RNA secondary structures became accessible to RNase V1 only upon dissolution of tertiary structures [43]. This implies that the differences between these structures were only detected at higher temperatures. We also cannot rule out the possibility that the RNA unfolding pathways during the melting process are actually quite different even between similar molecules, presumably due to complex tertiary interactions and dynamic effects.

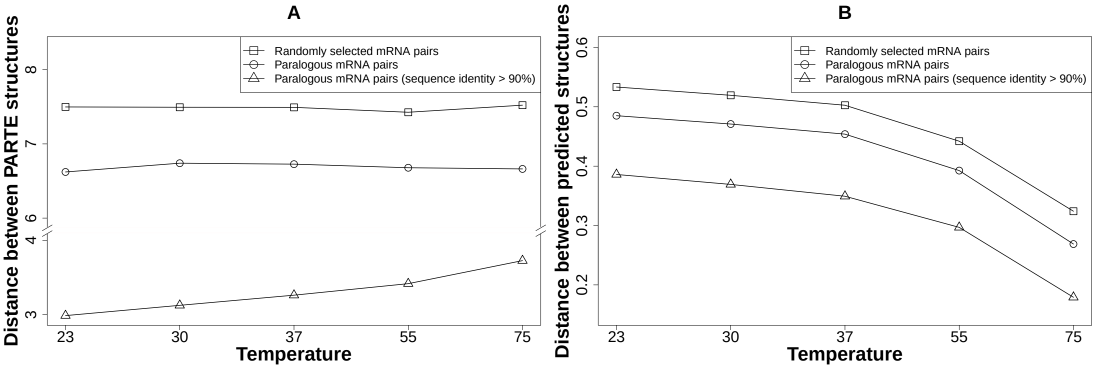


Figure 2.1 Variation of the distance between secondary structures of paralogous mRNA pairs along the temperature ladder. Points are the median levels of the distance at each temperature. **(A)** Distance between PARTE structures. **(B)** Distance between predicted structures.

This unexpected trend could not be captured by RNAfold predictions. As seen in Figure 2.1B, the distances between the predicted structures behave exactly as intuitively expected. The distances between the predicted structures of randomly selected, all paralogous and highly similar paralogous mRNA pairs all become smaller as the temperature grows from 23 to 75℃. The same pattern was also obtained with the RNAplfold program, which computes local base pair probabilities (Figure S2.3). This may be due to a number of inherent limitations of the current computational structure prediction approaches, especially when applied to long RNA sequences and at large deviations in temperature from standard conditions. Exponential growth of the number of possible secondary structures with the sequence length necessitates the introduction of approximations into the folding algorithms [115]. Modeling pseudoknots and prediction of long-range interactions continue to be an unsolved problem [116]. As well, energy calculations are parametrized at 37℃ and become less reliable at other temperatures [117, 118].

2.3.4 Melting temperature highlights functionally important structure and sequence elements in the coding regions of mRNAs

As discussed above, the global secondary structure of the mRNA coding regions is poorly conserved in evolution and probably does not play a role in gene regulation. Instead, RNA structure is more likely to be functional at the level of local structural elements situated in the coding regions. However, it is very hard to distinguish functionally important structural elements from non-functional ones, since every RNA chain tends to fold back on itself for thermodynamic reasons [43]. One important and experimentally measurable feature that may be indicative of functionality is the thermostability of RNA structures. It has been demonstrated that in ncRNAs functionally important structures have more stable structures than random RNAs of the same length and dinucleotide frequency [119, 120]. Many known functional structured RNA regulatory elements were identified in yeast mRNA 3′ UTRs by locating thermostable base pairs [43]. It is therefore conceivable that functionally important structural elements in the coding regions of mRNAs could also be discriminated by their thermostability.

Locally stable structures can be gleaned from the genome-wide PARTE experiment, in which secondary structures of yeast RNAs were probed and T_m were derived at single nucleotide resolution at five temperatures [43]. However, proving that such local structures actually fulfill a biological function is a challenging task. One approach to this problem could be based on assessing RNA-level selective constraints acting on protein-coding regions, including synonymous constraint and compensatory mutations. These unique patterns of sequence-structure relationships are a hallmark of the functionally important RNA elements in the coding regions.

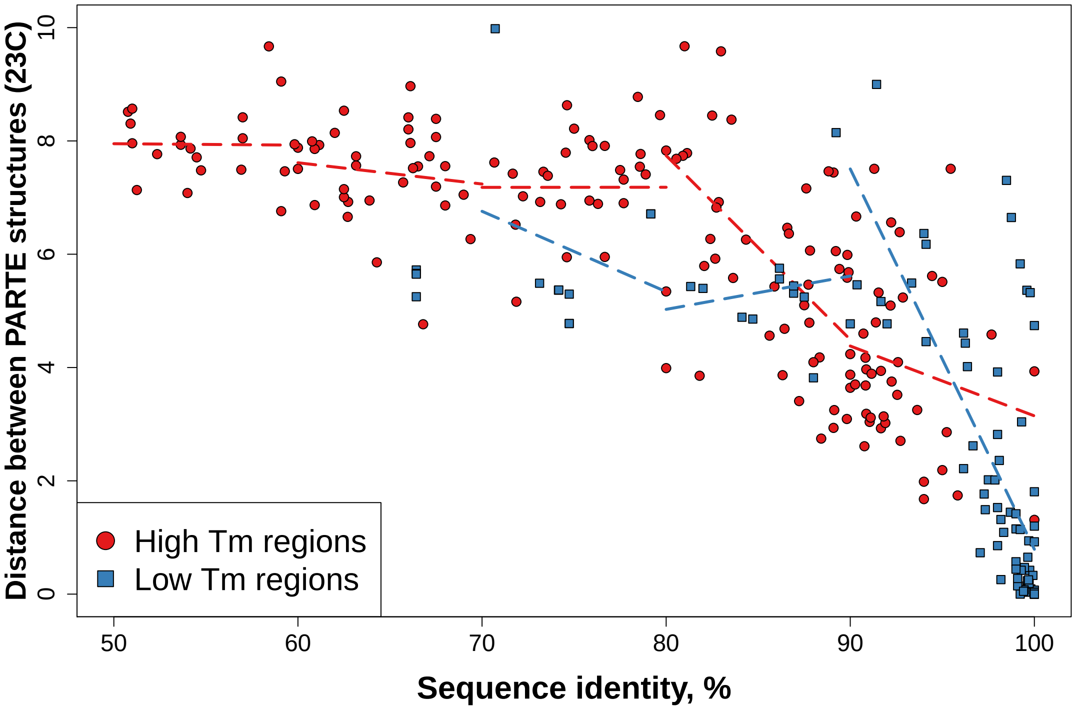


Figure 2.2 Sequence–structure relationships in the high/low T_m regions of paralogous mRNA pairs. For high T_m regions, the distance between structures shows a linear dependence from sequence identity for sequence identity values over 80% (correlation coefficient −0.54, *P*-value = $2.0e^{-7}$). For low T_m regions, the distance between structures shows a linear dependence from sequence identity for the sequence identity values over 90% (correlation coefficient −0.69, *P*-value = $2.1e^{-11}$). Linear regression for each 10% range of sequence identity is shown by a dashed line with the corresponding color. PARTE structures at 23℃ were used.

We identified 167 high T_m regions and 96 low T_m regions in 246 pairs of paralogous mRNAs. As seen in Figure 2.2, high T_m regions show a much stronger sequence-structure relationship than the low T_m regions in paralogous mRNA pairs. The distance between the structures of high T_m regions depends linearly on their sequence similarity for the sequence identity levels over 80%, while in low T_m regions this dependence only becomes apparent for sequences that share more than 90% identity. Low T_m regions show a higher sequence identity than high T_m regions (Figure 2.3A), while high T_m regions display a smaller structural distance upon controlling for sequence identity level (Figure 2.3B). Thus, high T_m and low T_m regions are under evolutionary pressure to preserve secondary RNA structure and primary sequence, respectively, and would therefore be expected to contain functionally important RNA structure elements and sequence segments, respectively. Indeed, high thermostability is a prerequisite for functionally important RNA structure elements [43, 119–121] while low thermostability ensures sufficient accessibility of functionally important RNA sequence elements [122, 123]. Melting temperature is thus a crucial parameter, which correlates with the distribution of functionally important structure and sequence elements along the coding regions of mRNAs.

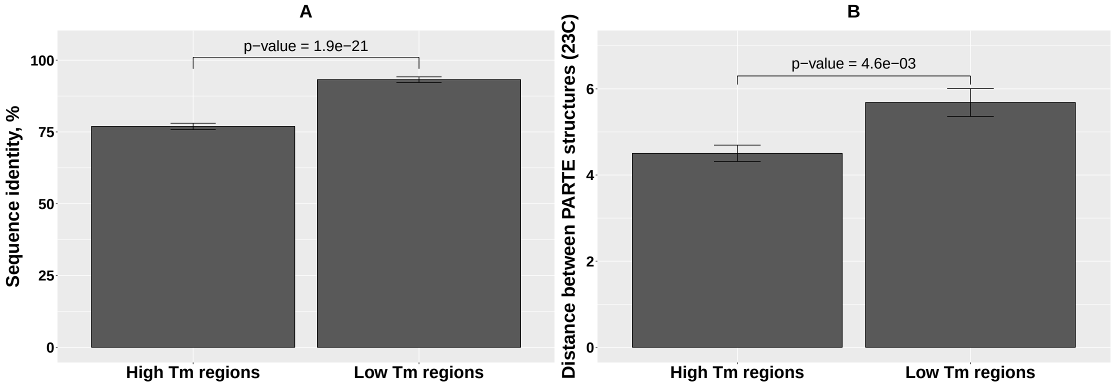


Figure 2.3 Low T_{m} regions in paralogous mRNA pairs are more conserved in sequence while high T_{m} regions are more conserved in RNA secondary structure. **(A)** Sequence identity between mRNAs (Mann-Whitney-Wilcoxon test, P-value = $1.9e^{-21}$). **(B)** Distance between RNA secondary structures (PARTE structures at 23℃; Mann-Whitney-Wilcoxon test, P-value = $4.6e^{-3}$). The differences are significant according to Mann-Whitney-Wilcoxon test. Error bars indicate standard error. The investigation of structural distances was effected upon controlling for sequence identity. Only regions with sequence identity 85–95% were considered in this analysis, while the regions with sequence identity < 85%, for which the distance between structures does not differ from randomly selected mRNA pairs as well as the regions with sequence identity > 95%, among which almost no high T_{m} regions exists, were excluded from consideration.

Our finding that low T_{m} regions are more conserved in the nucleotide sequence than high T_{m} regions does not contradict to the conclusion of Wan et al. that thermostable bases in yeast mRNAs are significantly more conserved than meltable bases [43]. In contrast to the analysis of Wan et al., which was performed at single-nucleotide resolution in full-length mRNA sequences, our results are solely based on low and high T_{m} regions in the coding portions of mRNAs. We were able to reproduce the results of Wan et al. and confirm that at single-base resolution, thermostable positions are more conserved than meltable positions when all individual positions of the entire coding regions in yeast paralogous mRNA alignments were examined together (Figure S2.4). However, when only positions located in high T_m and low T_m regions were examined separately, in high T_m regions the thermostable positions exhibited higher sequence conservation than meltable positions, while in low T_m regions meltable positions displayed a very high conservation level and thermostable position were completely absent (Figures 2.4 and 2.5). When considering the conservation of coding mRNA regions both at the structure and sequence level, it becomes apparent that the relatively high conservation level of thermostable positions in high T_m regions reflects evolutionary pressure to preserve RNA structure. The highest sequence conservation level observed in meltable positions of the low T_m regions is a reflection of the relatively high evolutionary pressure to preserve primary RNA sequence experienced by low T_m regions.

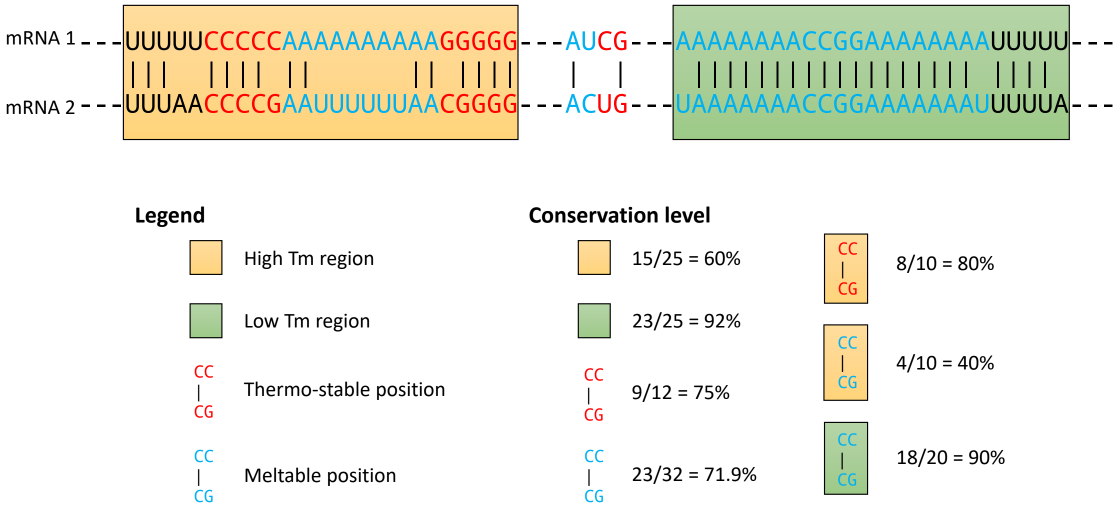

mRNA 1 ---UUUUUCCCCCAAAAAAAAAAGGGGG---AUCG---AAAAAAAACCGGAAAAAAAAUUUUU---
mRNA 2 ---UUUAACCCCGAAUUUUUUAACGGGG---ACUG---UAAAAAAACCGGAAAAAAAUUUUUA---
Legend
High Tm region
Low Tm region
CC | CG Thermo-stable position
CC | CG Meltable position
Conservation level
15/25 = 60%
23/25 = 92%
9/12 = 75%
23/32 = 71.9%
8/10 = 80%
4/10 = 40%
18/20 = 90%

Figure 2.4 Schematic illustration of the conservation levels of high/low T_m regions and thermostable/meltable positions. The alignment of two mRNAs is shown on the top. The low T_m region displays a higher sequence identity than the high T_m region (92% *versus* 60%). When all thermostable and meltable positions are considered together, the thermostable positions show a higher conservation level than the meltable positions (75% *versus* 71.9%). When the thermostable and the meltable positions located in high T_m and low T_m regions are considered separately, the meltable positions in the low T_m region are most conserved (90%), followed by the thermostable positions in the high T_m region (80%), while the meltable positions in the high T_m region are least conserved (40%).

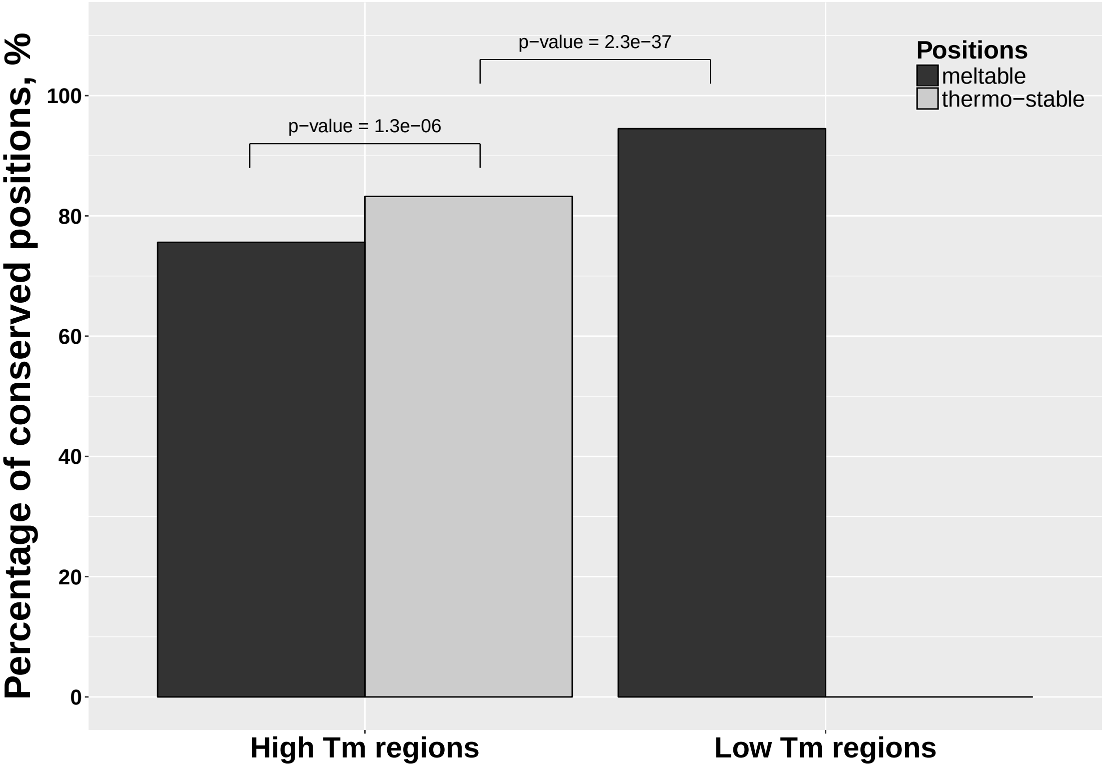


Figure 2.5 Conservation levels of thermostable and meltable positions in high/low T_m regions. Positions in high T_m and low T_m regions are examined separately. In high T_m regions the thermostable positions exhibit higher sequence conservation than meltable positions (Z-test for two proportions, P-value = $1.3e^{-6}$), while in low T_m regions meltable positions display a very high conservation level (Z-test for two proportions, P-value = $2.3e^{-37}$) and thermostable position is completely absent.

2.3.5 Low T_m regions are under synonymous constraint while high T_m regions exhibit relaxed sequence constraint

It is currently believed that RNA-level functions in coding regions manifest themselves by synonymous constraint [105, 106]. We therefore compared the synonymous polymorphism rate (pS) in the high and low T_m regions with the pS values calculated over the entire alignment of yeast paralogous mRNAs. Most low T_m regions exhibit negative ΔpS values while most high T_m regions exhibit positive ΔpS values (chi-squared test, P-values < 0.01) (Figure 2.6), which indicates that the low T_m regions are under synonymous constraint and may harbor functionally important nucleotide sequence motifs, such as ncRNA and protein binding sites [105, 106, 124]. This notion is compatible with the complete absence of thermostable nucleotide base pairs in low T_m regions, ensuring good accessibility of binding sites. High concentration of synonymous substitutions in high T_m regions may points to relaxed RNA sequence constraint, which may provide an evolutionary advantage for these regions in terms of accommodating functionally important RNA secondary structure elements [50, 105].

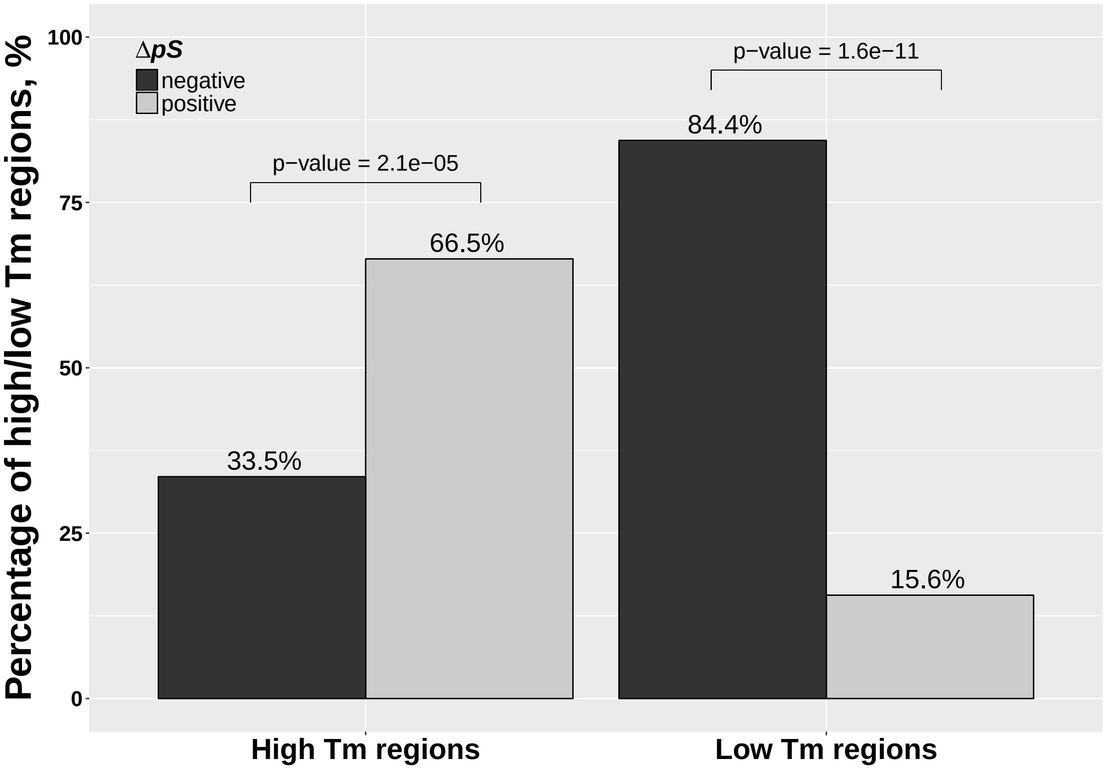


Figure 2.6 Most low T_m regions exhibit negative ΔpS values while most high T_m regions exhibit positive ΔpS values. The differences are significant according to chi-squared test.

2.3.6 Functionally important structure elements in the coding regions of yeast mRNAs tend to be thermostable

A typical class of functionally important structure elements in yeast mRNAs is constituted by the so-called zipcodes—regions of mRNAs recognized by the RNA-binding protein She2p [49, 114, 125]. Localized mRNAs are transported to the bud tip of the daughter cell by the She protein complex depending on the interaction between She2p and the loop-stem-loop structure of the zipcode [49, 114, 125]. Out of the 12 functional zipcodes in yeast mRNAs identified in a previous study [114], 5 zipcodes

(ASH1-E1min, TPO1N, ERG2N, WSC2C and SRL1C) are located in the coding regions and covered by the PARTE melting temperature data. These 5 zipcodes range from 49 (ASH1-E1min) to 178 (SRL1C) nucleotides in length. Another functionally important structure element in yeast mRNA coding regions is the *URE2* internal ribosome entry site (IRES) element, which locates between nucleotides 205 and 309 in the *URE2* coding region and folds into a stem-loop structure [126]. This IRES element mediates the cap-independent internal initiation of translation resulting in the expression of an N-terminal truncated form of the Ure2p protein [127]. We calculated the T_m of each structure element by averaging the T_m values of every PARTE-probed base within the element. All six structure elements show high T_m values (ASH1-E1min: 46℃, TPO1N: 51.9℃, ERG2N: 63.8℃, WSC2C: 80℃, SRL1C: 54.7℃ and *URE2* IRES: 53.6℃), which fall into the typical range of high T_m regions and thus exhibit high thermostability (Figure 2.7). This finding supports the hypothesis that high thermostability is indicative of functionally important RNA structure elements in mRNA coding regions.

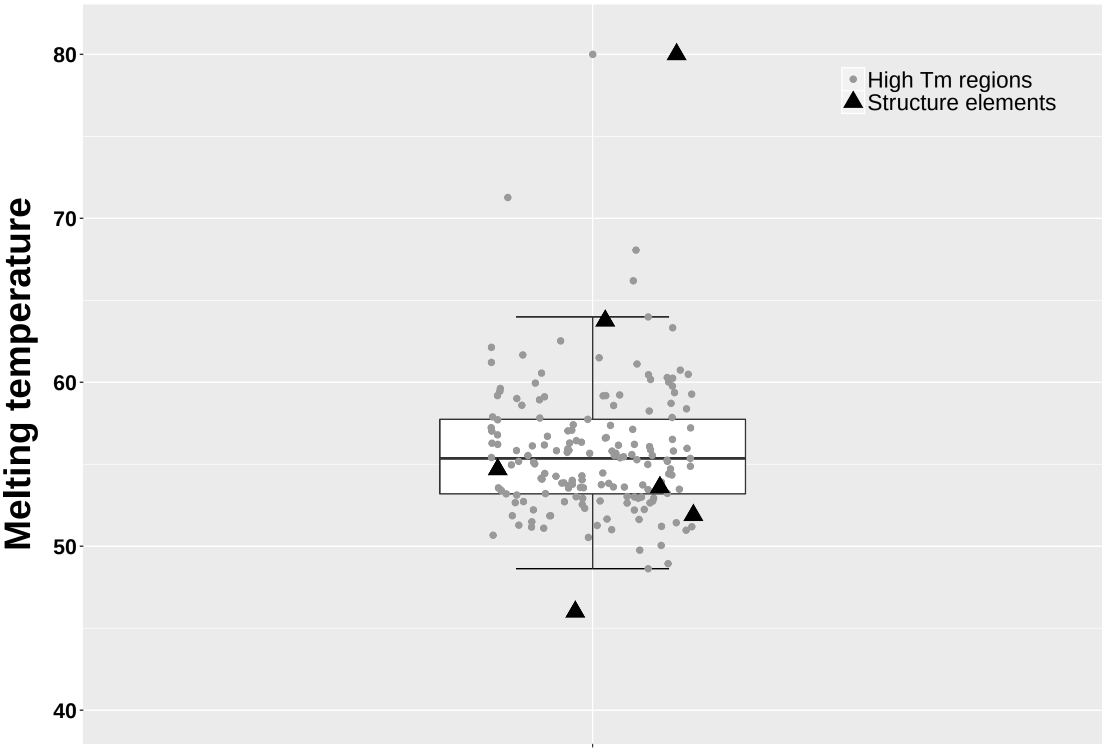


Figure 2.7 Melting temperatures of high T_m regions (boxplot and grey dots) and experimentally validated structure elements (black triangles).

2.4 Discussion

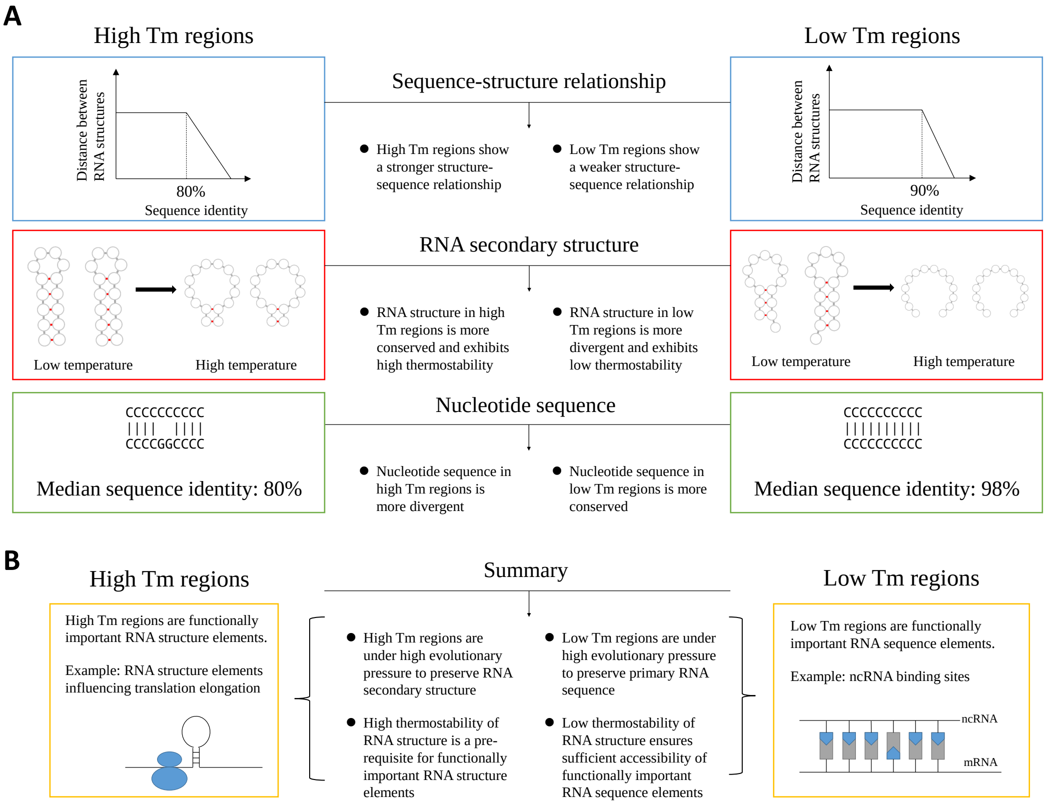

A
High Tm regions
Low Tm regions
Distance between RNA structures
80%
Sequence identity
90%
Sequence identity
Sequence-structure relationship
● High Tm regions show a stronger structure-sequence relationship
● Low Tm regions show a weaker structure-sequence relationship
RNA secondary structure
Low temperature
High temperature
● RNA structure in high Tm regions is more conserved and exhibits high thermostability
● RNA structure in low Tm regions is more divergent and exhibits low thermostability
Nucleotide sequence
CCCCCCCCCC
CCCCGGCCCC
Median sequence identity: 80%
CCCCCCCCCC
CCCCCCCCCC
Median sequence identity: 98%
● Nucleotide sequence in high Tm regions is more divergent
● Nucleotide sequence in low Tm regions is more conserved
B
High Tm regions
Summary
Low Tm regions
High Tm regions are functionally important RNA structure elements.
Example: RNA structure elements influencing translation elongation
● High Tm regions are under high evolutionary pressure to preserve RNA secondary structure
● High thermostability of RNA structure is a pre-requisite for functionally important RNA structure elements
● Low Tm regions are under high evolutionary pressure to preserve primary RNA sequence
● Low thermostability of RNA structure ensures sufficient accessibility of functionally important RNA sequence elements
Low Tm regions are functionally important RNA sequence elements.
Example: ncRNA binding sites
ncRNA
mRNA

Figure 2.8 **(A)** Summary of the findings about high T_m and low T_m regions. **(B)** Inferences based on these findings.

Figure 2.8 summarizes our findings about high/low T_m regions as well as our inference based on these findings. High T_m regions exhibit a stronger sequence-structure relationship, conserved and thermostable RNA secondary structures and relatively divergent nucleotide sequences, while low T_m regions display a weaker sequence-structure relationship, divergent and less thermostable RNA secondary structures and highly conserved nucleotide sequences. These findings suggest that high T_m regions are under high evolutionary pressure to preserve RNA secondary structure, whereas low T_m regions are under high evolutionary pressure to preserve primary RNA sequence. We therefore hypothesize that high T_m regions may contain thermostable functionally important RNA structure elements [128–133] and thus experience relatively high evolutionary pressure to preserve the RNA structure and a relaxed evolutionary constraint on the nucleotide sequence, as long as the thermostable nucleotide base pairs which are crucial for the RNA structure remain intact. Considering the highly conserved nucleotide sequence and low thermostability of low T_{m} regions, we hypothesize that low T_{m} regions may contain functionally important RNA sequence elements, for example, binding sites which are conserved in sequence and require a good accessibility to interact with ligands. High and low thermostability is, respectively, indicative of functionally important RNA structures and sequence segments in mRNA coding regions. We therefore speculate that the melting temperature is a crucial parameter for the identification of functionally important RNA structure and sequence elements. We have been able to verify the association of high thermostability with functional importance for two types of RNA structure elements — zipcodes in the yeast mRNA coding regions and *URE2* IRES. The lack of experimentally determined and precisely characterized sequence motifs in the coding regions of yeast mRNAs prevented us from directly assessing the functional implications of low thermostability. While in previous research coding regions carrying RNA-level functions were associated with synonymous constraint elements and relaxed protein structure constraints [105], we find that synonymous constraint is only apparent in functionally important sequence regions (e.g. binding sites) and that functionally important RNA structures are not under synonymous constraint. This finding may prove useful in future investigations of functionally important elements in mRNA coding regions, as the overall attention to mRNAs structure grows. A typical example is constituted by RNA thermometers — temperature-sensitive RNA structural elements that are typically located in the 5′ UTR of mRNAs and form a secondary structure that traps the ribosome binding site and/or the translation initiation codon [134]. In response to temperature changes, an RNA thermometer undergoes a conformational transition, which impacts translation efficiency and eventually regulates gene expression [135]. RNA thermometers have recently attracted growing attention [136], and efforts have been made to discover new elements of this type [117]. Secondary structures of RNA thermometers are more conserved than their primary sequences [137], which is analogous to the high T_{m} regions described in this work. We therefore speculate that our findings may facilitate the search for new RNA thermometers in mRNA coding regions. The characteristics of high T_{m} regions, including strong sequence-structure relationships, conservation patterns of thermostable and meltable positions, and relaxed sequence constraint could serve as features to narrow down the search space in RNA thermometer discovery.

2.5 Supplementary Materials

Table S2.1 Correlation between PARTE structure distances and sequence identity for different sequence identity ranges.

Sequence identity range (%)	23℃		30℃		37℃		55℃		75℃	
	ρ	P-value	ρ	P-value	ρ	P-value	ρ	P-value	ρ	P-value
50–60	0.04	0.75	0.06	0.67	0.05	0.73	0.03	0.81	0.02	0.88
60–70	0.08	0.49	0.01	0.93	0.09	0.48	0.08	0.50	0.08	0.53
70–80	0.05	0.82	0.07	0.74	0.15	0.48	0.21	0.32	0.22	0.31
80–90	−0.48	$3.90e^{-03}$	−0.56	$5.93e^{-4}$	−0.57	$4.88e^{-04}$	−0.55	$8.54e^{-4}$	−0.48	$3.77e^{-03}$
90–100	−0.84	$1.13e^{-17}$	−0.82	$1.30e^{-15}$	−0.83	$1.48e^{-16}$	−0.86	$1.57e^{-18}$	−0.82	$3.14e^{-16}$

† ρ: rank correlation coefficient

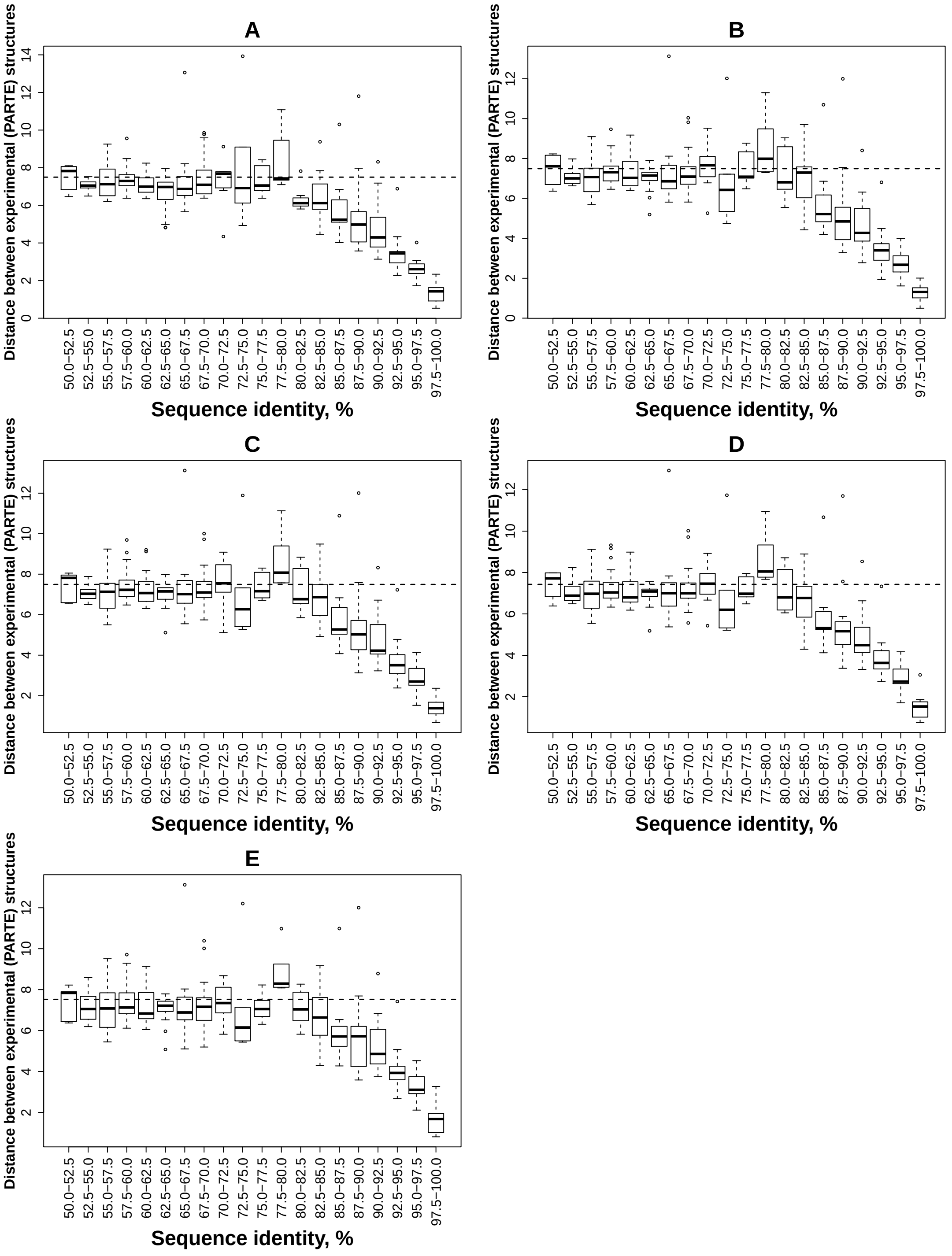


Figure S2.1 Boxplots of distances between experimental (PARTE) structures of paralogous mRNA pairs at 23℃ **(A)**, 30℃ **(B)**, 37℃ **(C)**, 55℃ **(D)** and 75℃ **(E)**. Every box corresponds to the range of identity 2.5%. The dashed lines are median levels of distances between PARTE structures of randomly selected mRNA pairs.

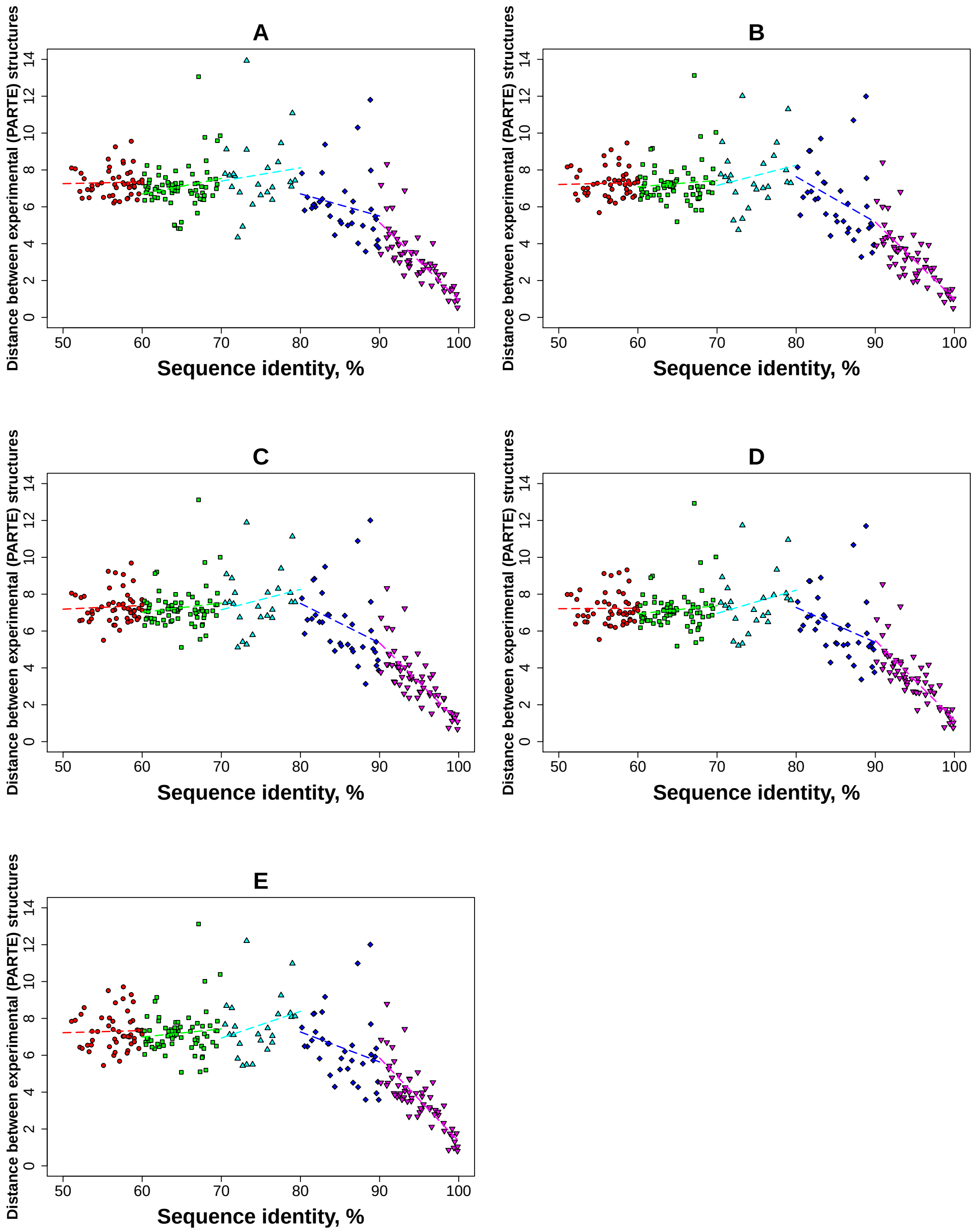


Figure S2.2 Sequence identity *versus* distance between experimental (PARTE) structures of paralogous mRNA pairs at 23℃ **(A)**, 30℃ **(B)**, 37℃ **(C)**, 55℃ **(D)** and 75℃ **(E)**. Dots are colored according to sequence identity level (50–60%: red, 60–70%: green, 70–80%: cyan, 80–90%: blue, 90–100%: purple). Linear regression for each range of sequence identity is shown by a dashed line with the corresponding color.

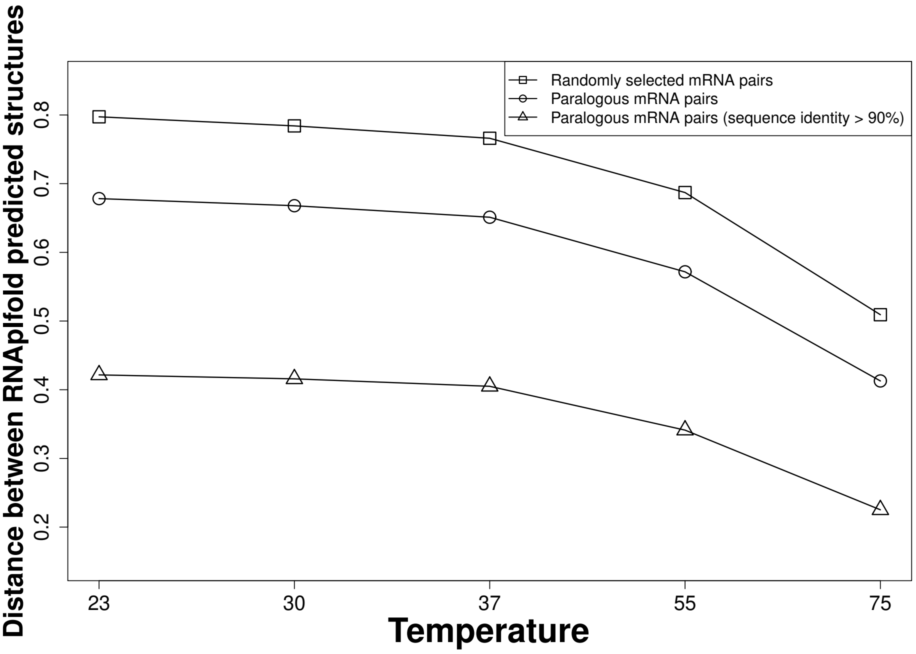


Figure S2.3 Variation of the distance between RNAplfold predicted structures of paralogous mRNA pairs along the temperature ladder. Points are the median levels of the distance at each temperature.

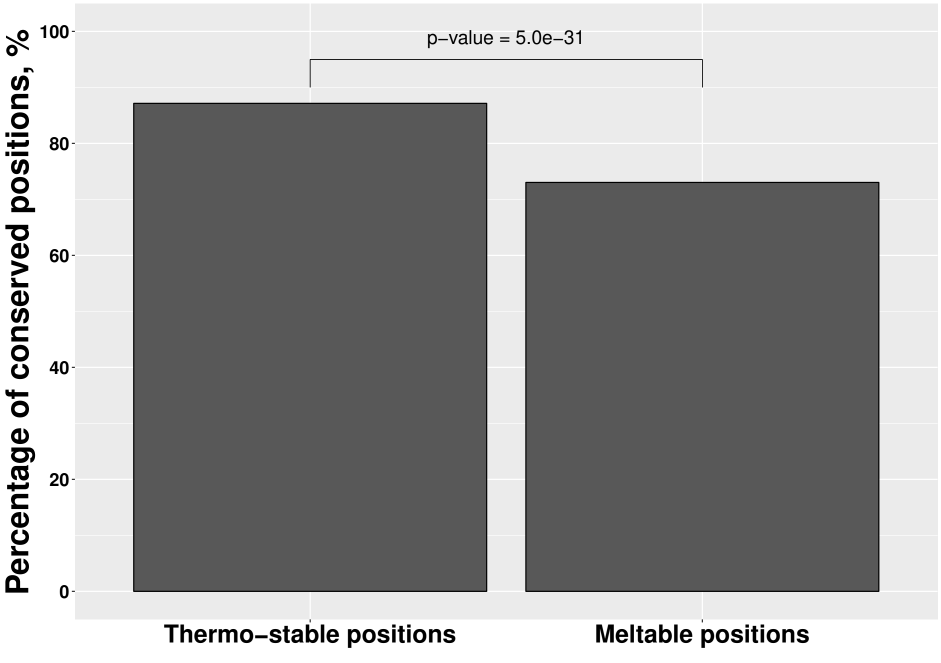


Figure S2.4 Conservation levels of thermostable and meltable positions in all pairs of yeast paralogous mRNAs. Thermostable positions are significantly more conserved (Z-test for 2 proportions, P-value = $5.0e^{-31}$).

Chapter 3

Evolutionary Analysis of Polyproline Motifs in *Escherichia coli* Reveals Their Regulatory Role in Translation

Translation of consecutive prolines causes ribosome stalling, which is alleviated but cannot be fully compensated by the elongation factor P. However, the presence of polyproline motifs in about one third of the *E. coli* proteins underlines their potential functional importance, which remains largely unexplored. We conducted an evolutionary analysis of polyproline motifs in the proteomes of 43 *E. coli* strains and found evidence of evolutionary selection against translational stalling, which is especially pronounced in proteins with high translational efficiency. Against the overall trend of polyproline motif loss in evolution, we observed their enrichment in the vicinity of translational start sites, in the inter-domain regions of multi-domain proteins, and downstream of transmembrane helices. Our analysis demonstrates that the time gain caused by ribosome pausing at polyproline motifs might be advantageous in protein regions bracketing domains and transmembrane helices. Polyproline motifs might therefore be crucial for co-translational folding and membrane insertion.

This chapter has been published as: Qi, F., Motz, M., Jung, K., Lassak, J. and Frishman, D. (2018) Evolutionary analysis of polyproline motifs in *Escherichia coli* reveals their regulatory role in translation. *PLOS Computational Biology*, 14(2), e1005987. doi: 10.1371/journal.pcbi.1005987. This study was designed, directed and coordinated by Prof. Kirsten Jung, Prof. Dmitrij Frishman and Dr. Jürgen Lassak. Magdalena Motz, Prof. Kirsten Jung and Dr. Jürgen Lassak established the rules for the classification of polyproline motifs. Prof. Dmitrij Frishman and I conceived the bioinformatics part, and I performed the bioinformatics analyses.

3.1 Introduction

Ribosomes facilitate the synthesis of proteins by translating the nucleotide sequence from an mRNA template. The speed of mRNA translation significantly varies and strongly depends on the amino acids to be incorporated into the growing polypeptide chain [6]. Especially slow is the incorporation of proline [72–74]. The pyrrolidine ring gives proline an exceptional conformational rigidity compared to all other amino acids and makes it not only a poor A-site peptidyl acceptor [74], but also a poor P-site peptidyl donor [72, 73]. Translation of two and more consecutive prolines dramatically impairs the peptidyl transfer reaction and eventually causes ribosomes to stall [72, 75–79]. Although basically all diproline comprising motifs cause translational stalling [77, 138], the arrest strength is influenced by physical and chemical properties of the adjacent amino acids that affect the conformation of the nascent polypeptide chain. Based on proteomic approaches combined with systematic *in vivo* and *in vitro* analyses, a hierarchy of arrest peptides was described [76, 77, 138, 139]. Thereby triplets such as PPP, D/PP/D, PPW, APP, G/PP/G and PPN cause strong ribosome stalling whereas e.g. L/PP/L, CPP or HPP result in a rather weak translational pause. Moreover, the stalling strength is modulated by amino acids located — up to position −5 — upstream of the arrest motif [138, 140, 141]. In this respect, H, K, Q, R or W further pronounce the arrest whereas C, G, L, S or T attenuate it. We therefore define a "polyproline motif" as a consecutive stretch of prolines with flanking residues: $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$, $n \geq 2$; where $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ can be any amino acid.

Regardless of the difficulties to translate consecutive proline coding sequences, they occur frequently within prokaryotic and eukaryotic proteomes [87, 142]. This in turn implies that the benefits of retaining polyproline motifs significantly outweigh their costs to incorporate them into the nascent polypeptide chain [5]. Proline is unique in terms of being the sole amino acid to adopt *cis* and *trans* conformations, both of which are nearly energetically equal and naturally occur in proteins [80–82]. Notably, a sequence of consecutive prolines results in the formation of either the right-handed poly proline helix I (PPI) or the left-handed poly proline helix II (PPII). Beside α-helix and β-sheet, PPII helix is considered to be the third major secondary structure element in proteins and plays an important role in mediating protein-protein and protein-nucleic acid interactions [83–86]. Three consecutive prolines are also an integral part of the active center in the universally conserved Val-tRNA synthetase (ValS) [87]. The proline triplet in ValS is essential for efficient charging of the tRNA with valine and for preventing mischarging by threonine. These two examples illustrate why nature has evolved a specialized translation elongation factor, referred to as EF-P in bacteria or a/eIF-5A in archaea and eukaryotes, to alleviate ribosome stalling at polyproline motifs [5, 72, 75–79, 88]. The importance of polyproline motifs in proteins is further underlined by the fact that *efp* mutants are characterized by pleiotropic defects. Reportedly, the absence of EF-P impairs bacterial fitness [143, 144], membrane integrity [145], motility [146], antibiotic sensitivity [147] and is ultimately lethal for certain bacteria such as *Mycobacterium tuberculosis* [148] and *Neisseria meningitides* [149]. Similarly, IF-5A is an essential protein in archaea [150] as well as in eukaryotes [151] where eIF-5A is associated e.g. with cancer [152] and HIV infection [153].

EF-P alleviates polyproline motif-dependent translational arrest, but cannot fully prevent ribosome pausing at these sequences [78, 138, 154]. The fact that polyproline motifs form a functionally important structural element — the PPII helix — and at the same time interfere with translation poses a major evolutionary conundrum. Are polyproline motifs disfavored during evolution due to their translational burden? Does ribosome stalling caused by polyproline motifs regulate the speed of translation at the protein level in the same way as rare genetic codons and secondary structures cause translational pause at the RNA level [11, 155]? To address these questions, we conducted an evolutionary analysis of polyproline motifs in the proteomes of 43 *E. coli* strains. Our analysis revealed evolutionary selection against polyproline motifs as a consequence of the reduced translation efficiency. Against the overall background of polyproline motif depletion, we observed their frequent occurrence in the vicinity of translational start sites, in the inter-domain regions of multi-domain proteins, and downstream of transmembrane helices, where slow-translating codons are also enriched. This indicates the potential involvement of polyproline motifs in co-translational protein folding and transmembrane helix insertion.

3.2 Materials and Methods

3.2.1 Proteomes and orthologous groups of *E. coli*

We obtained *E. coli* proteomes and orthology assignments from the OMA database [156]. The total of 206 360 protein sequences from six out of seven *E. coli* phylotypes [157] were downloaded (Table S3.1). We also obtained 11 356 orthologous groups covering 195 056 proteins.

3.2.2 The core- and accessory proteomes of *E. coli*

The core- and accessory proteomes were defined based on the occurrence of orthologous groups. An orthologous group was classified as belonging to the core proteome if it was present in all the 43 *E. coli* proteomes, otherwise it was considered belonging to the accessory proteome. All proteins not assigned to any orthologous group were classified as belonging to the accessory proteome. This procedure yielded a core proteome of *E. coli* covering 73 745 proteins and an accessory proteome covering 132 615 proteins.

3.2.3 Identification of polyproline motifs in real and random sequences

Using the program fuzzpro from the EMBOSS package [158], we identified polyproline motifs in the *E. coli* proteins. The same procedure was applied to randomly generated sequences. Each amino acid sequence in our dataset was shuffled 1000 times while maintaining its composition using the program shuffleseq from the EMBOSS package [158], yielding 1000 sets of random *E. coli* protein sequences.

3.2.4 Enrichment and depletion of polyproline motifs

We used the SPatt algorithm [159] to assess the enrichment and depletion of polyproline motifs, taking into account occurrence patterns of proline in various parts of protein structure. SPatt determines the expected occurrence of a sequence motif based on a Markov chain model of order m (model Mm), compares the observed occurrence with expected one, and calculates the P-value for the significance of a motif's enrichment or depletion. Choosing a model Mm means taking into account the m-mer and $(m + 1)$-mer compositions while determining the expected occurrence. For example, the model M0 solely takes into account the amino acid composition, while choosing the model M1 takes into account the compositions of amino acid monomers and dimers. For a motif of length l, the maximum m is $(l - 2)$. In our case, although a polyproline motif can have more than 2 residues, the essential part of a polyproline motif is the proline stretch with at least two consecutive proline residues. Therefore, we chose model M0 in our tests.

3.2.5 Normalization of polyproline motif occurrence

The occurrence of polyproline motifs in proteins was normalized by the polyproline motif occurrence in randomly generated sequences. Each amino acid sequence (either full protein sequences or specific sequence segments of interest) was shuffled 1000 times while maintaining its composition using the program shuffleseq from the EMBOSS package [158], yielding 1000 sets of random sequences. The number of times the polyproline motif occurred in a real sequence was then divided by the number of times the same motif occurred in each of the 1000 random sequences, yielding a vector of 1000 ratios between the observed and the expected polyproline motif occurrence. The Mann-Whitney-Wilcoxon test was employed to assess the significance of the difference between two such vectors corresponding to two different sequences or sequence segments. This procedure was carried out for each strain of *E. coli* separately.

3.2.6 Classification of polyproline motifs

The rules for classification of polyproline motifs were established by our collaborators, Magdalena Motz, Kirsten Jung and Jürgen Lassak (unpublished data). Polyproline motifs were classified into three groups (strong, medium and week) according to their predicted ribosomal translation arrest strength. The prediction is based on experimental data both from systematic *in vitro* and *in vivo* analyses [5, 87, 138, 141, 142] (Tables S3.2 and S3.3, from Motz et al., unpublished data).

As described in Section 3.1, the ribosome stalling strength of a $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motif is dependent on the number of consecutive prolines and on the flanking amino acids. First, the flanking residues $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ were classified according to their influence on the ribosome stalling strength (motifs involving ambiguous amino acids were excluded from consideration). If a flanking residue of the polyproline motif in an *E. coli* strain lacking *efp* (Δ*efp*) was responsible for a decrease of the translational output by ≥ 70% compared to a wildtype control, the residue was defined as strong [77, 138, 141]. In cases where the protein synthesis was reduced by 30–60%, the stalling strength was classified as medium. In all other cases, the polyproline sequence context was assumed to cause only a weak arrest. All possible $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motifs and their respective arrest strength are listed in Table S3.3 (Motz et al., unpublished data). Based on the classification, the predicted motif strength was correlated to available ribosome profiling data [138]. Woolstenhulme et al. compared the ribosome occupancy at a diprolyl motif with the occupancy downstream of the motif in an Δ*efp* strain [138]. Stalling was ranked according to the observed asymmetry (ratio) between these two values. When an asymmetry quotient of 2.00 was set as a threshold for proteins subject to strong translation arrest, more than 75% of these proteins possess at least one medium or strong polyproline motif. This number further increases to ~80% and ~90% when applying more stringent cutoffs of 3.00 and 5.00 to the asymmetry score, respectively (Table S3.4, from Motz et al., unpublished data).

3.2.7 Word frequency in protein sequences

Frequencies of each single and dimer amino acid in protein sequences were calculated using the compseq program from the EMBOSS package [158]. For each amino acid dimer, an expected frequency was additionally calculated based on the observed frequency of single amino acid.

3.2.8 Multiple alignment of protein sequences

Multiple alignment of protein sequences in each orthologous group were computed using the Clustal Omega software [160] with all default parameters.

3.2.9 Construction of phylogenetic trees

Phylogenetic tree for each orthologous group with at least three proteins containing at least one polyproline motif were reconstructed using the PhyML software [161]. These trees were then rooted at midpoint.

3.2.10 Reconstruction of evolutionary events

In order to reconstruct the gain and loss of the ribosome stalling effect in the evolutionary history of *E. coli* protein families, we first assigned one of the four possible ribosome stalling states [S (strong), M (medium), W (weak) and N (none)] to all the exterior nodes (leaves) of the phylogenetic trees. Subsequently, the Maximum Likelihood algorithm [162] was employed to reconstruct the states of ancestral nodes (internal nodes). The change of state between a given node and its ancestral node from a stronger stalling effect state to a weaker or no stalling effect state was defined as a loss of the stalling effect, while the change from a weaker or no stalling effect state to a stronger state was defined as a gain event.

3.2.11 Propensity of stalling effect change

We defined propensity of stalling effect change (PSEC) similar to propensity of gene loss (PGL) frequently used in evolutionary studies [163]. PGL captures the idea that the longer the time during which a gene could have been lost but was not, the lower the propensity of this gene to be lost. PGL is thus defined as the ratio between the total length of branches in which the gene is lost and the total length of branches in which the gene could have been lost [164, 165]. Similarly, PSEC captures the idea that the longer the time during which the stalling effect of a motif could have been gained/lost but was not, the lower the propensity of the stalling effect to be gained/lost. However, our model is somewhat more complex than the PGL model, since the PGL only considers gene loss and we have to consider both gain and loss of the stalling effect. Therefore, the PSEC is calculated as the difference between the propensities of gain and loss of the stalling effect:

$$PSEC = \frac{\sum B_g}{\sum B_{cg}} - \frac{\sum B_l}{\sum B_{cl}} \tag{3.1}$$

where B_g and B_l are the lengths of the branches in which the stalling effect was gained and lost, respectively, and B_{cg} and B_{cl} are the lengths of branches in which the stalling effect could have been gained and lost, respectively. Thus, a positive PSEC indicates that the stalling effect of a sequence motif tends to be gained, while a negative PSEC indicates that it tends to be lost during evolution.

3.2.12 Protein abundance, gene expression, and translation efficiency

Protein abundance data used in this study was from [166, 167], covering 2163 proteins. Microarray data on transcription levels of 2710 genes from *E. coli* K-12 MG1655 under standard growth conditions was downloaded from the ASAP database [168]. Translation efficiency for each of the 1743 genes present in both datasets was calculated as:

$$Translation\_efficiency_i = \frac{Protein\_abundance_i}{Transcription\_level_i} \tag{3.2}$$

3.2.13 Domain composition of the *E. coli* proteins

Sequence positions of 7398 structural domains in 4080 *E. coli* K-12 MG1655 proteins were obtained from the Gene3D database [169].

3.2.14 Transmembrane segments

We obtained the sequence positions of 5672 transmembrane segments within 912 α-helical transmembrane proteins from the UniProt database [170]. Since reviewed data on transmembrane proteins of *E. coli* K-12 MG1655 (taxonomy ID 511145) are not available in the UniProt database, we used the reviewed data of *E. coli* K-12 (taxonomy ID 83333) instead.

3.3 Results

3.3.1 Polyproline motifs are underrepresented in *E. coli* proteomes

We first investigated the overall frequency of polyproline motifs in *E. coli* strains and found 99 386 polyproline motifs within 68 710 (33.3%) proteins from the 43 proteomes considered in this study. Out of these 68 710 proteins, 47 056 proteins (68.5%) harbor only one polyproline motif, 15 027 proteins (21.9%) have two polyproline motifs, and 6627 proteins (9.6%) have more than 2 polyproline motifs (Figure S3.1). We identified 22 253 (22.4%), 21 953 (22.1%) and 55 149 (55.5%) polyproline motifs with strong, medium and weak ribosome stalling effect, respectively. We found that polyproline motifs are significantly underrepresented in all the 43 *E. coli* proteomes compared with randomly generated protein sequences (Figures 3.1A and S3.2). Pairs of consecutive prolines show the lowest ratio between the observed and expected frequency (0.84) compared to all other pairs of identical amino acids in *E. coli* K-12 MG1655. Moreover, normalized by the random level, the numbers of polyproline motifs negatively depend on the strength of the ribosome pausing effect in all *E. coli* proteomes: in *E. coli* K-12 MG1655, for example, polyproline motifs with strong, medium and weak ribosome stalling effect constitute 55.5%, 70.9% and 104.4% of the random level, respectively (Figures 3.1B and S3.3). Collectively, these findings suggest the existence of evolutionary pressure against ribosome pausing.

To investigate this hypothesis, we grouped the proteins into the core proteome, which encompasses conserved, evolutionary older sequences, and the accessory proteome, which mainly contains proteins of younger origin. Assuming that evolution disfavors polyproline motifs, one would expect them to occur less frequently in the core proteome. Indeed, significantly fewer polyproline motifs were found in proteins belonging to the *E. coli* core set, independent of the arrest strength (Figures 3.1C and S3.4).

3.3.2 Variation of ribosome stalling strength in *E. coli* evolution

We next investigated changes in ribosome stalling strength caused by polyproline motifs in the *E. coli* proteins by considering 3280 orthologous groups with at least 3 proteins and at least one polyproline motif. Within these orthologous groups, we identified 4980 aligned regions containing polyproline motifs, of which 1568 showed changes of the ribosome stalling effect states. Out of the 1923 evolutionary events 955 were gain events (change from a weaker or no stalling effect state to a stronger state) and 968 were loss events (change from a stronger stalling effect state to a weaker or no stalling effect state). The propensity of stalling effect change (PSEC) was calculated for each of these aligned regions as described in Section 3.2. In the core proteome, substantially more aligned regions displayed a negative PSEC (Figure 3.1D), indicating that the ribosome stalling effect tends to be lost in evolution. In line with this finding, in the phylogenetically younger accessory proteome, PSEC still displayed no strong preference with 51.5% and 48.5% aligned regions possessing positive and negative PSEC, respectively. These results are also in line with the notion that evolution generally disfavors polyproline motifs in *E. coli*.

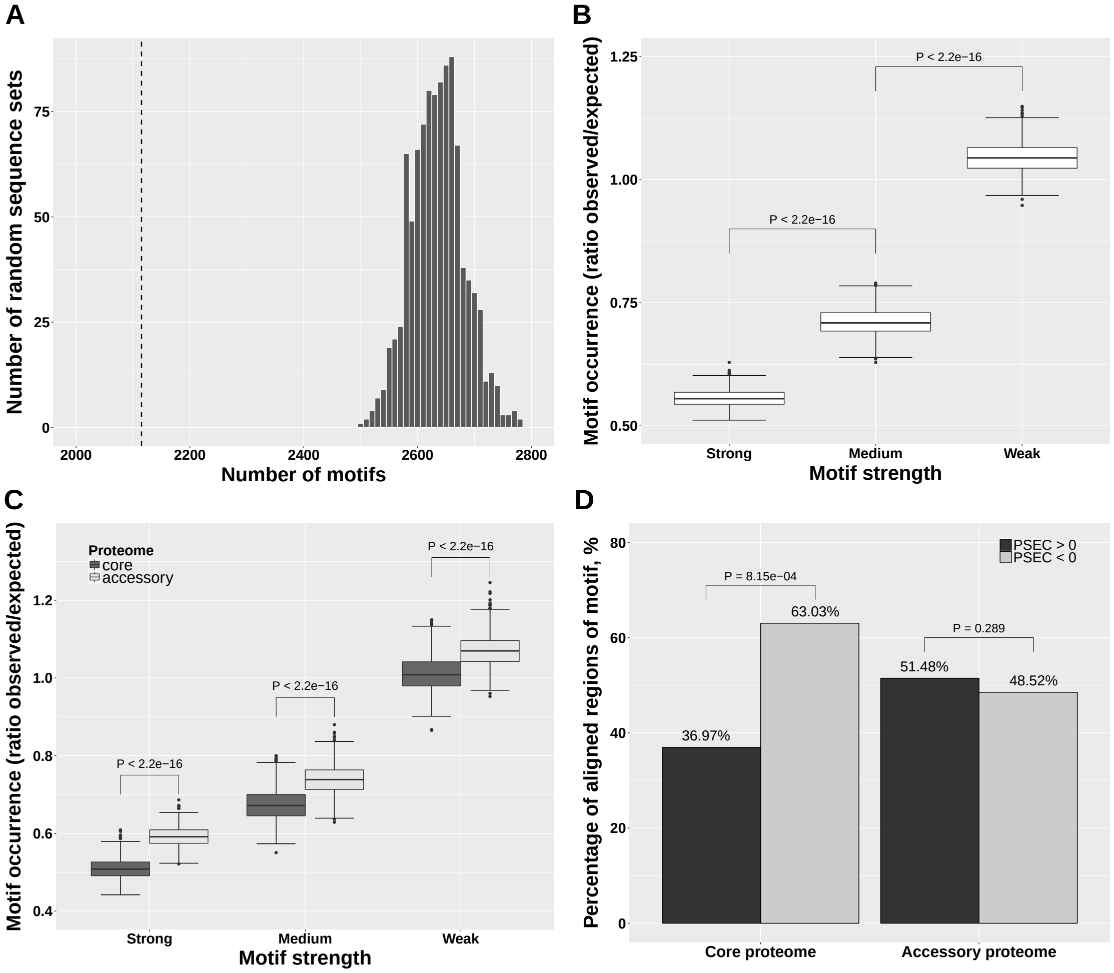


Figure 3.1 Distribution and conservation of polyproline motifs. **(A)** Occurrence of polyproline motifs in *E. coli* K-12 MG1655 is lower than the random level. The histogram shows the numbers of motifs found in 1000 sets of random sequences, and the dashed line shows the number of motifs found in real sequences. **(B)** Numbers of polyproline motifs negatively correlate with the strength of the ribosome stalling effect in *E. coli* K-12 MG1655. The differences are significant according to Mann-Whitney-Wilcoxon test. **(C)** Occurrence of polyproline motifs in the core proteome of *E. coli* K-12 MG1655 is lower than that in the accessory proteome. The differences are significant according to Mann-Whitney-Wilcoxon test. **(D)** In the core proteome more aligned regions have a negative PSEC (chi-squared test) while in the accessory proteome PSEC values display no strong preference.

3.3.3 Translational efficiency is the evolutionary driving force for selecting against polyproline motifs

The efficiency of translation and consequently biosynthesis correlates with both translation initiation and elongation rates [171]. Translation elongation rate in turn depends on multiple factors, such as codon bias [154], tRNA levels [17] and the amino acid to be incorporated [71, 74], but can also be influenced by an amino acid sequence such as consecutive prolines [77, 79, 138]. Accordingly, we investigated whether there is a connection between the relative frequency of polyproline motifs and translational efficiency in *E. coli* K-12 MG1655, and found that they are negatively correlated (Figure 3.2A), which is especially evident in the top 25% of most efficiently translated proteins and for polyproline motifs known to cause a strong translational pause. Occurrence of polyproline motifs also anti-correlates with relative protein abundance (Figure 3.2B). Thus, in the course of evolution, polyproline motifs are more disfavored in those proteins that have a high copy number per cell and need to be efficiently translated, implying a translation-efficiency-driven selection pressure against polyproline motifs.

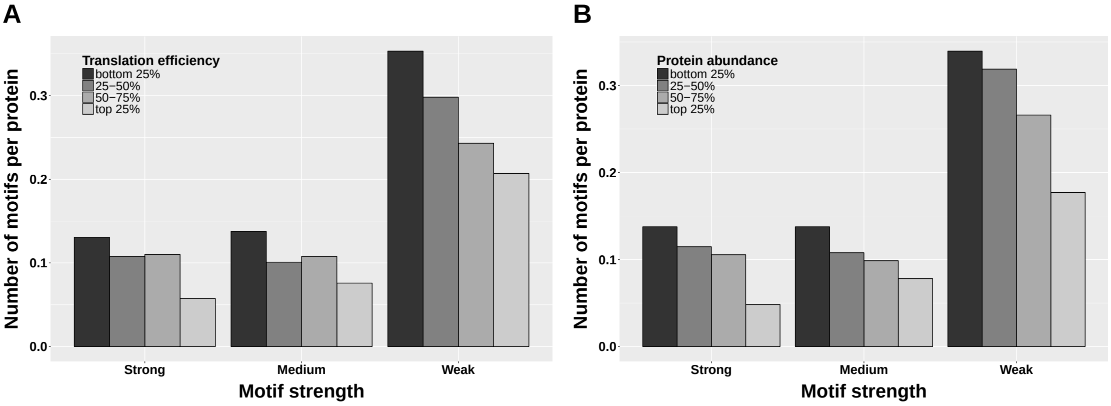


Figure 3.2 Correlation between translation efficiency, protein abundance and frequency of polyproline motifs. **(A)** Proteins with high translation efficiency tend to have fewer polyproline motifs (Spearman's rho = −0.114, *P*-value = $3.18e^{-10}$). **(B)** High abundance proteins tend to have fewer motifs (Spearman's rho = −0.147, *P*-value = $9.36e^{-7}$).

3.3.4 Polyproline motifs as regulatory elements in protein synthesis

We next investigated whether polyproline-mediated ribosome pausing is exploited in the regulation of translation, focusing on the reference strain *E. coli* K-12 MG1655 comprising 2115 polyproline motifs in 1477 proteins (33.9% of the whole proteome). In 2010, Tuller et al. discovered reduced translation efficiency within the first 50 codons of the coding regions [23]. The authors suggested that a slow ramp at the beginning of the ORF might serve as a late stage of translation initiation, being a probate means to reduce ribosomal traffic jams in order to minimize the cost of protein biosynthesis [23]. We were therefore curious whether there exists an enrichment of polyproline motifs around the start sites of *E. coli* K-12 MG1655 proteins. In the 2115 polyproline motifs of *E. coli* K-12 MG1655, 325 were found in the first 50 amino acids, and 1771 located elsewhere in the protein sequence. After normalization by random level, we found a clear enrichment of polyproline motifs in the N-terminal 50 residues (Figure 3.3A). Thus, similar to the specific codon bias in this region, an accumulation of polyproline motifs might allow adjustment of translational speed in order to minimize the cost of protein production.

3.3.5 Polyproline motifs coordinate co-translational folding of proteins

Protein folding is a co-translational process, and it is generally believed that structural elements of a protein may influence each other during the folding process [32]. Due to the cooperativity between different parts of the structure, the timing of translation is crucial for proper folding [11]. The non-uniform distribution of synonymous codons with different translation rates fine-tunes the co-translational folding of proteins [9, 30, 31]. Fast translation of the mRNA stretches coding for structural domains helps to avoid misfolded intermediates [34], while translational pauses induced by clusters of slow-translating codons in the inter-domain linkers of multi-domain proteins facilitate independent folding of domains to minimize the chance of misfolding [15, 16, 30, 33]. By analogy, we hypothesized that polyproline motifs may coordinate co-translational folding by slowing down translation of inter-domain linkers, and as a consequence, would be expected to occur more frequently between rather than within structural domains.

We therefore investigated the positional preference of polyproline motifs in globular multi-domain proteins. Sequence positions of 7398 structural domains within 4080 *E. coli* K-12 MG1655 proteins were obtained from Gene3D database [169]. Out of these proteins, 1868 (45.8%) are multi-domain proteins possessing the total of 5186 domains. An inter-domain linker was defined as the sequence span between the boundaries of two consecutive domains (if such a span was shorter than 5 amino acids, it was expanded downstream to achieve the length of 5 amino acids). This procedure yielded 3318 inter-domain linkers between 5186 domains.

Indeed, we found that polyproline motifs are significantly depleted in structural domains (P-value = $7.86e^{-80}$), but not in inter-domain linkers (P-value = 0.912). We then investigated the relative location of polyproline motifs with respect to domain boundaries. As seen in Figure 3.3B, polyproline motifs frequently occur in two regions: (i) −12 to −2 residues relative to the domain start; and (ii) −2 to +9 residues relative to the domain end. Polyproline motifs are significantly enriched in these two regions (P-values < 0.05). Thus, there is a strong correlation between the location of polyproline motifs and the structural domain boundaries, which was also observed for clusters of slow-translating codons [15, 16, 172, 173]. These findings imply that the ribosome stalling effect caused by the polyproline motifs within structural domains may interfere with their folding, while stalling at domain boundaries may facilitate it.

3.3.6 Polyproline motifs facilitate co-translational insertion of TMHs

Another typical co-translational process is the targeting of α-helical transmembrane proteins to the translocons, mediated by the signal recognition particle (SRP), and their insertion into the membrane [35, 36]. This process has been found to be facilitated by translational pause [8, 37–41]. A recent study by Fluman et al. identified two translational pauses, triggered by Shine-Dalgarno-like elements in *E. coli* mRNAs, that contribute to the SRP-mediated targeting of transmembrane proteins [8]. The first pause occurs before the nascent peptide emerges from the exit tunnel of the ribosome (16 to 30 codons of the protein) and the second one occurs after the emergence of the first transmembrane helix (−5 to +1 codons relative to the start of the second TMH). In the fungus *Emericella nidulans*, Dessen and Képès identified two translational pauses occurring at the distance of approximately 45 and 70 codons from TMHs, caused by clusters of slow-translating codons and presumed to facilitate translocon-mediated co-translational insertion of TMH [38].

We investigated the occurrence and location of polyproline motifs in transmembrane proteins. Based on the UniProt [170] annotation, we identified 912 α-helical transmembrane proteins from *E. coli* K-12 containing the total of 5672 TMHs. We found that 39.3% (358) of these transmembrane proteins harbor polyproline motifs, which is even higher than the percentage of soluble proteins (32.6%; chi-squared test, P-value = $1.6e^{-4}$). No enrichment of polyproline motifs around the pause sites identified by Fluman et al. was observed (data not shown). However, as seen in Figure 3.3C, we found that (i) polyproline motifs rarely occur within TMH; and (ii) polyproline motifs display a relatively high occurrence in four positions (positions −17 to −1, 23 to 32, 49 to 59 and 77 to 87 relative to TMH start; termed here site I, II, III and IV, respectively). The depletion of polyproline motifs in TMH is significant (P-value = $1.65e^{-27}$) implying that the ribosome stalling effect caused by the polyproline motifs may interfere with the folding of transmembrane proteins. It should be noted that the site positions are shown relative to the start of a TMH, and thus in some cases the given region can actually be located in another TMH (see Figure 3.3D for illustration). We therefore tested the enrichment/depletion of the polyproline motifs in each of the four sites described above separately in TMH and in non-transmembrane regions. For example, out of the 4439 site IV regions, 3013 and 1426 regions are located in TMH and non-transmembrane regions, respectively. For all four sites, significant depletion of polyproline motifs was evident in TMH regions (P-values for sites I, II, III and IV are $2.63e^{-6}$, $1.86e^{-3}$, $7.84e^{-3}$ and $1.98e^{-3}$, respectively), while in non-transmembrane regions a significant enrichment of polyproline motifs was observed for site III (P-value = 0.035). The location of this site is similar to the location of one of the translational pauses (approximately 45 codons from TMHs) identified by Dessen and Képès. Considering that most of the TMHs are 21 residues in length (Figure S3.5) and that about 28 amino acids can be accommodated in the ribosome exit tunnel [42], ribosome stalling at site III may occur after the TMH has emerged from the ribosome exit tunnel and is being inserted into the membrane by translocon [35, 174]. We therefore speculate that the translational pause at site III could provide a time delay for the efficient insertion of TMH.

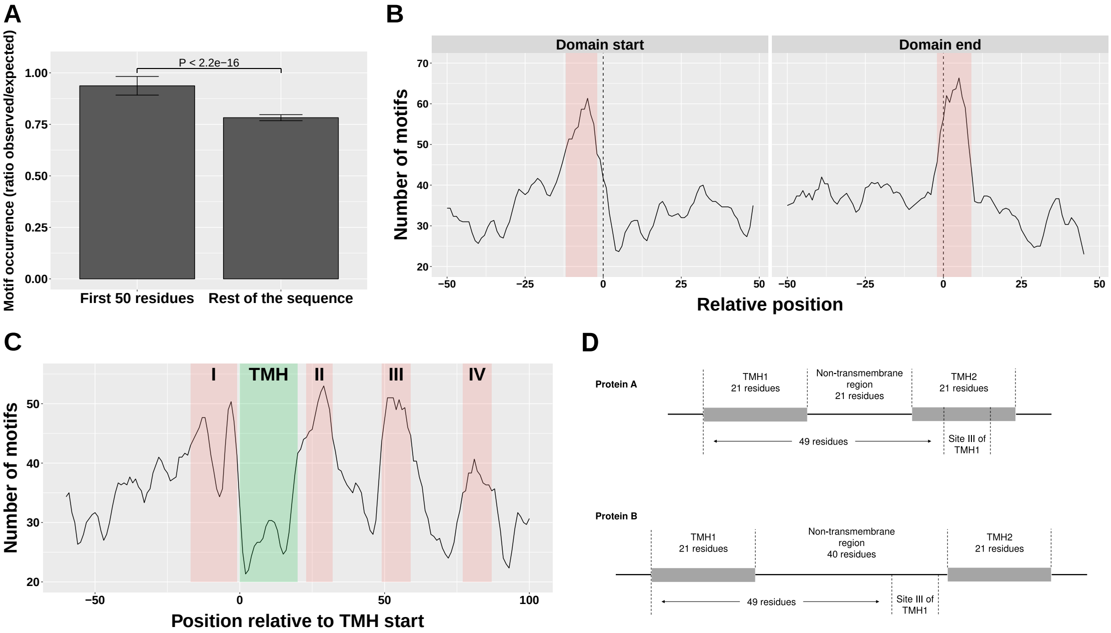
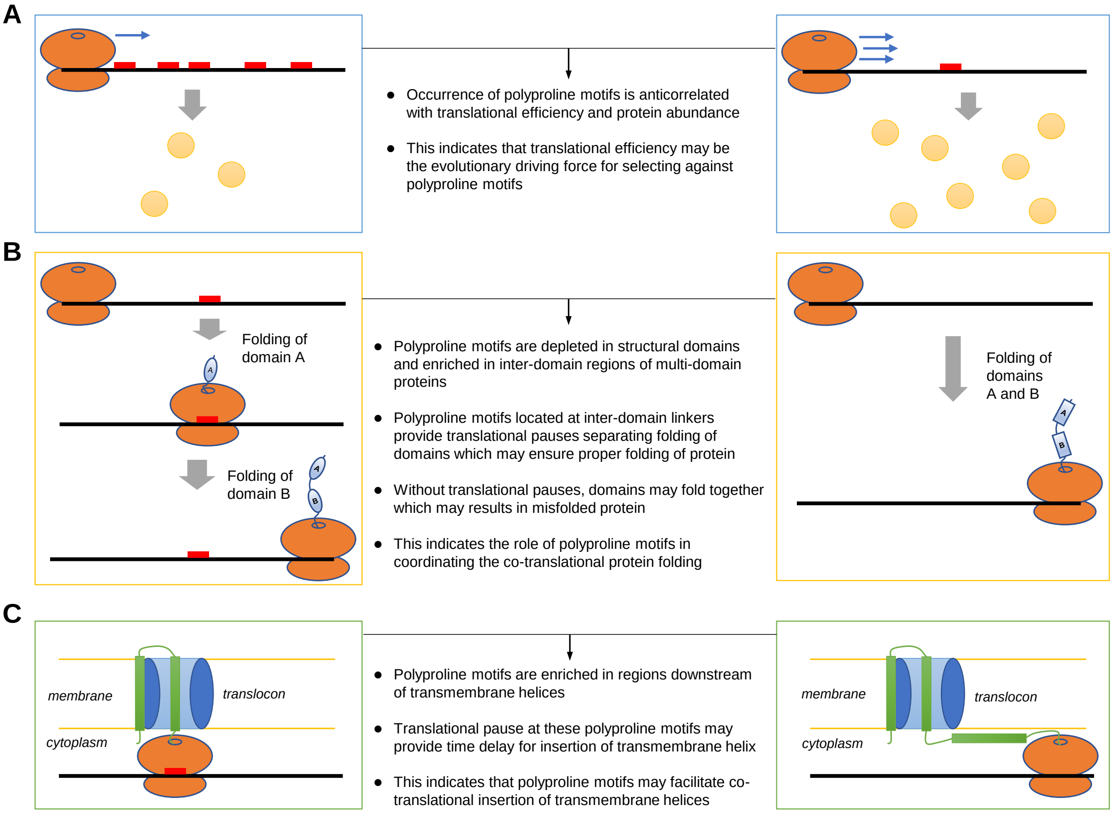


Figure 3.3 Functional role of polyproline motifs. **(A)** Occurrence of polyproline motifs in the first 50 residues is higher than elsewhere in the protein sequence (Mann-Whitney-Wilcoxon test). Error bars indicate the standard deviation. **(B)** Occurrence of polyproline motifs is associated with domain boundaries. Regions with relatively high motif occurrence are marked red. Data are smoothed over a three-residue window. Left: frequency of motifs relative to domain start (dashed line). Right: frequency of motifs relative to domain end (dashed line). The enrichment of motifs in these two regions is significant (P-values < 0.05). **(C)** Frequency of polyproline motifs relative to the start position of TMH. TMH is marked green (assuming the typical length of 21 residues). Regions with high motif frequency are marked red. Data are smoothed over a three-residue window. **(D)** Schematic illustration of the site III location relative to TMH and the non-transmembrane region. In protein A site III of TMH1 locates in the TMH2 while in protein B site III of TMH1 is in the non-transmembrane region.

3.4 Discussion

Proline is a poor substrate for the ribosomal peptidyl transfer reaction [72–74], and consecutive prolines cause ribosome stalling [77]. The bacterial elongation factor P (EF-P) and its archaeal and eukaryotic orthologs a/eIF-5A alleviate this stalling to some degree, but cannot fully compensate the translational burden imposed by polyproline motifs [75, 78, 79, 138]. The presence of a large number of such motifs in bacterial proteomes might imply their biological significance, yet their precise functional role remains poorly understood [87, 142].

In this study, we made a comprehensive attempt to shed light on the functional role of polyproline motifs by investigating their distribution and evolution in the proteomes of 43 *E. coli* strains. Figure 3.4 summarizes our findings about polyproline motifs. We found evidence of evolutionary selection pressure against translational stalling caused by polyproline motifs. Translational efficiency and protein abundance negatively correlate with the frequency of polyproline motifs and thus might be the driving force for their loss. Against the general trend of losing polyproline motifs during the course of evolution, we observed accumulation of polyproline motifs close to protein N-terminus, in inter-domain regions of multi-domain proteins as well as downstream of transmembrane helices. We therefore speculate that the time gain caused by translational pause at polyproline motifs might be crucial for translational regulation, domain folding, and proper membrane insertion.



- Occurrence of polyproline motifs is anticorrelated with translational efficiency and protein abundance
- This indicates that translational efficiency may be the evolutionary driving force for selecting against polyproline motifs

- Polyproline motifs are depleted in structural domains and enriched in inter-domain regions of multi-domain proteins
- Polyproline motifs located at inter-domain linkers provide translational pauses separating folding of domains which may ensure proper folding of protein
- Without translational pauses, domains may fold together which may results in misfolded protein
- This indicates the role of polyproline motifs in coordinating the co-translational protein folding

- Polyproline motifs are enriched in regions downstream of transmembrane helices
- Translational pause at these polyproline motifs may provide time delay for insertion of transmembrane helix
- This indicates that polyproline motifs may facilitate co-translational insertion of transmembrane helices

Figure 3.4 Summary of the findings about polyproline motifs. Red bars indicate the positions of polyproline motifs mapped to RNA sequences. **(A)** Translation efficiency is presumably the driving force for evolutionary selection against polyproline motifs. **(B)** Polyproline motifs coordinate co-translational folding of proteins. **(C)** Polyproline motifs facilitate co-translational insertion of transmembrane helices.

Evolutionary selection for high efficiency of protein synthesis is one of the forces shaping mRNA sequences. For example, unequal usage of synonymous codons reflects an adaption of the codon usage to the available tRNA pool, with slow-translating codons used much more rarely than fast-translating codons [89, 90]. However, protein sequence elements were also found to influence the translation rate by interacting with the ribosome exit tunnel or impairing the peptidyl transfer reaction [79, 175]. An important question, which arises in this context, is whether there exists protein-level evolutionary selection for high translation efficiency. Recently, Sabi and Tuller found that short peptides, which induce ribosome stalling in yeast by interacting with the ribosomal exit tunnel, tend to be either over- or underrepresented in the proteome [91]. They hypothesized that short peptide sequences were under evolutionary selection based on their synthetic efficiency. Our results show that polyproline motifs, which induce ribosome stalling by slowing down the peptidyl transfer reaction, are significantly underrepresented in *E. coli* proteomes, and that selection is more evident against motifs causing stronger ribosome stalling and in proteins with higher translation efficiency. These findings support the conjecture that translation-efficiency-based evolutionary pressure shapes protein sequences.

Against the overall background of polyproline motif depletion, our investigation of the intra-molecular distribution pattern of polyproline motifs revealed their overrepresentation at several strategic locations, indicating their regulatory role in translation elongation. Translation elongation is a non-uniform process, which is subject to strict regulation [5, 6] both in terms of the quantity of the translation products [7] and the intra-molecular variation of the elongation rate, which ensures the quality of the synthesized proteins by coordinating co-translational processes [8–10]. The role of polyproline motifs in the regulation of the overall translation elongation rate is exemplified by the lysine-dependent acid stress response regulator CadC of *E. coli* [5, 79, 176]. This membrane-integrated pH-sensor and transcriptional activator contains two polyproline motifs, which allow for fine-tuning of its copy number. The amount of the CadC protein is crucial for regulating the expression of the target operon. Analogously, precisely regulated translational output of the polyproline-containing receptor CpxA is required for *Shigella flexneri* virulence [177]. The intra-molecular variation of elongation rate has so far been thought to be regulated by *cis*-acting elements embedded in the translated mRNA, such as clusters of slow-translating codons [11] and Shine-Dalgarno-like RNA sequences [8] (although the latter notion has recently been challenged [178]), as well as by *trans*-acting molecules, such as the signal recognition particle, which arrests translation elongation while targeting proteins to the membrane [179, 180]. Our study highlights the role of polyproline motifs in coordinating the co-translational protein folding and transmembrane helix insertion, implying that they could serve as protein-level *cis*-acting elements, which directly regulate the rate of translation elongation.

3.5 Supplementary Materials

Table S3.1 List of all 43 *E. coli* strains used in this study.

Taxonomy id	OMA id	Name	Phylogenetic group	Pathogenicity
1133853	ECO1E	Escherichia coli O104:H4 (strain 2009EL-2071)	A	yes (EAEC)
511693	ECOBB	Escherichia coli (strain B / BL21)	A	no
469008	ECOBD	Escherichia coli (strain B / BL21-DE3)	A	no
413997	ECOBR	Escherichia coli (strain B / REL606)	A	no
595496	ECOBW	Escherichia coli (strain K12 / MC4100 / BW2952)	A	no
536056	ECOD1	Escherichia coli (strain ATCC 33849 / DSM 4235 / NCIB 12045 / K12 / DH1)	A	no
316385	ECODH	Escherichia coli (strain K12 / DH10B)	A	no
316401	ECOH1	Escherichia coli O78:H11 (strain H10407 / ETEC)	A	yes (ETEC)
331112	ECOHS	Escherichia coli O9:H4 (strain HS)	A	no
481805	ECOLC	Escherichia coli (strain ATCC 8739 / DSM 1576 / Crooks)	A	no
511145	ECOLI	Escherichia coli (strain K12 / MG1655)	A	no
1040638	ECOLX	Escherichia coli O104:H4 LB226692	A	yes (EHEC)
585395	ECO10	Escherichia coli O103:H2 (strain 12009 / EHEC)	B1	yes (EHEC)
585396	ECO1A	Escherichia coli O111:H- (strain 11128 / EHEC)	B1	yes (EHEC)
331111	ECO24	Escherichia coli O139:H28 (strain E24377A / ETEC)	B1	yes (ETEC)
573235	ECO26	Escherichia coli O26:H11 (strain 11368 / EHEC)	B1	yes (EHEC)
585055	ECO55	Escherichia coli (strain 55989 / EAEC)	B1	yes (EAEC)
585034	ECO8A	Escherichia coli O8 (strain IAI1)	B1	no
595495	ECOKO	Escherichia coli (strain ATCC 55124 / KO11)	B1	no

Table S3.1 List of all 43 *E. coli* strains used in this study. *(continued)*

Taxonomy id	OMA id	Name	Phylogenetic group	Pathogenicity
566546	ECOLW	Escherichia coli (strain ATCC 9637 / CCM 2024 / DSM 1116 / NCIMB 8666 / NRRL B-766 / W)	B1	no
409438	ECOSE	Escherichia coli (strain SE11)	B1	no
574521	ECO27	Escherichia coli O127:H6 (strain E2348/69 / EPEC)	B2	yes (EPEC)
585035	ECO45	Escherichia coli O45:K1 (strain S88 / ExPEC)	B2	yes (ExPEC)
585057	ECO7I	Escherichia coli O7:K1 (strain IAI39 / ExPEC)	B2	yes (ExPEC)
585397	ECO81	Escherichia coli O81 (strain ED1a)	B2	no
685038	ECO8N	Escherichia coli O83:H1 (strain NRG 857C / AIEC)	B2	yes (AIEC)
655817	ECOAB	Escherichia coli OR:K5:H- (strain ABU 83972)	B2	yes (ABU)
885275	ECOC1	Escherichia coli (strain clone D i14)	B2	yes (UPEC)
885276	ECOC2	Escherichia coli (strain clone D i2)	B2	yes (UPEC)
405955	ECOK1	Escherichia coli O1:K1 / APEC	B2	yes (APEC)
714962	ECOKI	Escherichia coli O18:K1:H7 (strain IHE3034 / ExPEC)	B2	yes (ExPEC)
362663	ECOL5	Escherichia coli O6:K15:H31 (strain 536 / UPEC)	B2	yes (UPEC)
199310	ECOL6	Escherichia coli O6:H1 (strain CFT073 / ATCC 700928 / UPEC)	B2	yes (UPEC)
431946	ECOS5	Escherichia coli O150:H5 (strain SE15)	B2	no
869729	ECOUM	Escherichia coli (strain UM146)	B2	yes (AIEC)
364106	ECOUT	Escherichia coli (strain UTI89 / UPEC)	B2	yes (UPEC)
216592	ECO44	Escherichia coli O44:H18 (strain 042 / EAEC)	D	yes (EAEC)
585056	ECOLU	Escherichia coli O17:K52:H18 (strain UMN026 / ExPEC)	D	yes (ExPEC)
439855	ECOSM	Escherichia coli (strain SMS-3-5 / SECEC)	D	no
155864	ECO57	Escherichia coli O157:H7	E	yes (EHEC)
444450	ECO5E	Escherichia coli O157:H7 (strain EC4115 / EHEC)	E	yes (EHEC)
544404	ECO5T	Escherichia coli O157:H7 (strain TW14359 / EHEC)	E	yes (EHEC)
701177	ECOCB	Escherichia coli O55:H7 (strain CB9615 / EPEC)	F	yes (EPEC)

Table S3.2 Classification of the effect amino acids at position $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ exert on ribosomal stalling strength (from Motz et al., unpublished data).

Effect on stalling strength	Amino acid at position $X_{(-2)}$	Amino acid at position $X_{(-1)}$	Consecutive proline stretch	Amino acid at position $X_{(+1)}$
From [138]				
Strong	A, E, H, I, K, P, R, V	D, G, P	PP	D, E, K, N, P, S, W
Moderate	D, G, N, Q, Y	A	PP	A, G, Q,
Weak	C, F, L, M, T, S	C, E, F, H, I, K, L, M, N, Q, R, S, T, V, W, Y	PP	C, F, H, I, L, M, R, T, V, Y
From [77, 141]				
Strong	D, E, H, K, P, Q, R, Y, W	A, D, P	PP	D, G, N, P, W
Moderate	F, G, I, M, N, V	E, G, S	PP	E, Q, S, T
Weak	A, C, L, S, T	C, F, H, I, K, L, M, N, Q, R, T, V, W, Y	PP	A, C, F, H, I, K, L, M, R, V, Y
Combined data				
strong	A, D, E, G, H, I, K, P Q, R, V, W, Y	A, D, G, (P)	nP	D, E, G, N, (P), S, W
Moderate	F, M, N	S, E	nP	A, K, Q, T
Weak	C, L, S, T	C, F, H, I, K, L, M, N, Q, R, T, V, W, Y	nP	C, F, H, I, L, M, R, V, Y

Table S3.3 Rules for the prediction of ribosomal stalling strength induced by a $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$ motif (from Motz et al., unpublished data).

$X_{(-2)}$	$X_{(-1)}$	nP	$X_{(+1)}$	Motif strength	Example
Rules for n = 3: put the first proline at $X_{(-1)}$ position					
strong	P	PP	strong	strong	KPPPD
strong	P	PP	medium	strong	KPPPK
strong	P	PP	weak	strong	KPPPC
medium	P	PP	strong	strong	FPPPD
medium	P	PP	medium	medium	FPPPK
medium	P	PP	weak	medium	FPPPC
weak	P	PP	strong	strong	SPPPD
weak	P	PP	medium	medium	SPPPK
weak	P	PP	weak	weak	SPPPC
Rules for n = 2					
If $X_{(+1)}$ = strong, classify $X_{(-1)}$ using the rule for $X_{(-2)}$					
NA	strong	PP	strong	strong	KPPD
NA	medium	PP	strong	medium	FPPD
NA	weak	PP	strong	weak	SPPD
If $X_{(+1)}$ = medium and $X_{(-2)}$ or $X_{(-1)}$ = weak, classify $X_{(-1)}$ using the rule for $X_{(-2)}$					
NA	strong	PP	medium	medium	TIPPK
NA	medium	PP	medium	medium	AFPPK
NA	weak	PP	medium	weak	ACPPK
Other situations					
strong	strong	PP	medium	strong	KGPPK
strong	medium	PP	medium	medium	KEPPK
medium	strong	PP	medium	medium	FGPPK
medium	medium	PP	medium	medium	FEPPK
strong	strong	PP	weak	strong	KGPPC
strong	medium	PP	weak	medium	KEPPC
strong	weak	PP	weak	weak	KCPPC
medium	strong	PP	weak	medium	FGPPC
medium	medium	PP	weak	medium	FEPPC
medium	weak	PP	weak	weak	FCPPC
weak	strong	PP	weak	weak	SGPPC
weak	medium	PP	weak	weak	SEPPC
weak	weak	PP	weak	weak	SCPPC

[1] Rules for n > 3: all motifs are classified as strong

Table S3.4 Matching the predicted stalling strength of polyproline motifs with the ribosome profiling data from [138] (from Motz et al., unpublished data).

Gene	Motif	Asymmetry score	Predicted Stalling Strength
cadC	SPPPI, ATPPE	84.37	W, S
malZ	EDPPQ, RMPPA	68.72	M, M
gntX	KPPPW, YAPPL	47.69	S, S
nanS	GGPPC	41.23	S
yodB	VPPPA	35.18	S
valS	IPPPN	28.87	S
ytfM	RPPPK, KVPPD	22.76	S, S
acnB	KNPPA, RVPPG, VAPPT	17.49	M, S, S
treA	PQPPD, YVPPE, SQPPF	15.68	S, S, W
cysQ	EDPPG, ARPPL	14.85	S, S
gpr	GPPPG	14.85	S
ygdH	INPPE, AEPPN	13.74	M, S
lepA	IPPPE	13.53	S
uvrA	SDPPK	13.41	S
yeiG	LPPPR, TPPPV, TQPPW	13.29	W, W, S
rnb	IPPPQ	13.00	S
poxB	AIPPQ	11.30	S
cpxA	NDPPN	9.42	S
dapE	VVPPG, INPPF	9.15	S, W
csiD	AAPPS	7.66	S
pyrC	LAPPV	7.63	W
eutL	KLPPH, VAPPL, VPPPS	6.97	W, S, S
pfkB	SLPPG, VPPPV	6.94	W, S
yhbW	LPPPI	6.70	W
qorA	YPPPS	6.61	S
yhjG	GNPPD, DTPPF	6.43	M
uvrB	EPPPT	6.43	S
acnA	VVPPG, ASPPL	6.29	S
atpD	NEPPG	6.19	S
hofP	FKPPE, CEPPQ	5.87	S, M
uxuA	DDPPR	5.81	S
ybhB	AAPPK	5.69	S
yqjH	FVPPT, PRPPS	5.65	M, S
ybiU	RRPPG	5.61	S
pdxB	CDPPR	5.56	W
aes	DLPPW, EVPPC	5.52	W, W
recB	APPPD, GCPPL	5.39	S, W
lon	KIPPE, VGPPG	5.04	S, S

Table S3.4 Matching the predicted stalling strength of polyproline motifs with the ribosome profiling data from [138] (from Motz et al., unpublished data). *(continued)*

Gene	Motif	Asymmetry score	Predicted Stalling Strength
pgm	HNPPE, YNPPN	4.98	M, M
fabF	TSPPE, AVPPT	4.81	W, M
gltB	LVPPA, TNPPI, SPPPH, NNPPF, IRPPV	4.80	W, W, W, W
bcsB	APPPG, NLPPD, TMPPV	4.59	S, W, W
yjjK	VVPPK, FIPPG	4.52	S, S
ligT	RQPPR, IPPPG	4.46	W, S
glnL	GIPPH	4.40	S
hslU	LIPPA, MAPPG	4.39	M, S
rbsR	MTPPL	4.35	W
gadW	QSPPM	4.30	M
sufS	EMPPW	4.30	M
ytfE	TPPPE	4.26	S
alaS	GGPPG	4.04	S
tyrB	SSPPN	3.92	W
yeaN	CGPPL	3.82	W
ycfS	PLPPA, YYPPG	3.78	W
ycaN	IGPPV	3.76	S
ampD	SLPPG	3.75	W
sbmA	ATPPT	3.73	W
clpA	GAPPG	3.64	S
ptsG	IWPPI	3.55	S
wzyE	VAPPE	3.49	S
mrp	DMPPG	3.44	M
fkpA	GIPPN	3.40	S
betT	MQPPE, SLPPE	3.34	S, W
proW	ALPPI	3.28	W
pstA	TPPPN	3.14	S
cytR	NLPPM, LPQPPT, CDPPL	3.06	W, S, W
rbbA	VIPPY, EAPPV, EQPPL	3.04	W, S, W
ubiX	IMPPV	3.01	W
tdh	GAPPA, GIPPS	2.97	S, S
yagU	QTPPN, LNPPY, LTPPL	2.92	W, W, W
mqo	GAPPM	2.91	S
bcsG	TAPPT	2.90	M
clpB	CAPPG	2.85	S
puuA	LQPPC	2.82	W
zwf	MSPPS	2.76	W
gsiA	QAPPI	2.76	S

Table S3.4 Matching the predicted stalling strength of polyproline motifs with the ribosome profiling data from [138] (from Motz et al., unpublished data). *(continued)*

Gene	Motif	Asymmetry score	Predicted Stalling Strength
ppc	LPPPE	2.65	S
cbpA	TIPPG	2.53	S
nupG	VMPPK	2.47	M
rhmA	RYPPY	2.46	W
yhgF	DEPPK	2.42	M
mcrB	QGPPG	2.37	S
malQ	GAPPD, GLPPM	2.29	S, W
amiD	AVPPR	2.28	W
yfaY	PQPPV, QLPPG	2.27	W, W
gcd	TSPPI, FTPPS	2.27	W, W
ytfF	QMPPL	2.27	W
gsiB	VVPPS	2.19	S
ampG	YTPPF	2.18	W
rtn	EIPPD	2.18	S
lysU	GLPPT	2.15	W
hrpA	KKPPK, KLPPA	2.15	M, W
tpiA	IAPPE	2.14	S
yrfF	FAPPA, DYPPQ	2.12	M
dhaM	VAPPT, PVPPV	2.06	S, W
yhdN	CLPPE	2.05	W
efeO	AFPPS	2.04	M
codB	AIPPV	1.91	W
yjfP	LFPPL	1.89	W
rbn	GVPPG	1.76	S
ytfN	KMPPS, EIPPA, TVPPM, SGPPD	1.75	M, M, W, S
intA	GFPPD	1.73	M
fruA	MVPPL	1.72	W
phnP	YGPPD, SHPPR	1.68	S, W
dapF	VEPPY	1.65	M
fepB	KLPPQ	1.61	W
yjiY	NTPPA	1.59	W
sad	YYPPT	1.56	M
fhuF	MVPPL	1.49	W
rpoE	RRPPS	1.43	S
cdd	TLPPL	1.40	W
malE	PNPPK	1.33	M
gpmM	DTPPR	1.33	W
dapD	VVPPA	1.26	M

Table S3.4 Matching the predicted stalling strength of polyproline motifs with the ribosome profiling data from [138] (from Motz et al., unpublished data). *(continued)*

Gene	Motif	Asymmetry score	Predicted Stalling Strength
gcvR	PRPPM	1.18	W
carB	VIPPY	1.18	M
otsA	IAPPD, PLPPK	1.13	S, W
fruK	AKPPS	1.06	S
proB	GAPPA	1.02	S
argE	KLPPF, ECPPN	1.00	W
sdhB	QNPPA	1.00	M
edd	LMPPL	0.85	W
aroA	NYPPL	0.78	M
mprA	VLPPQ	0.76	W
ibpA	GYPPY	0.67	M
frmB	YLPPK	0.55	W

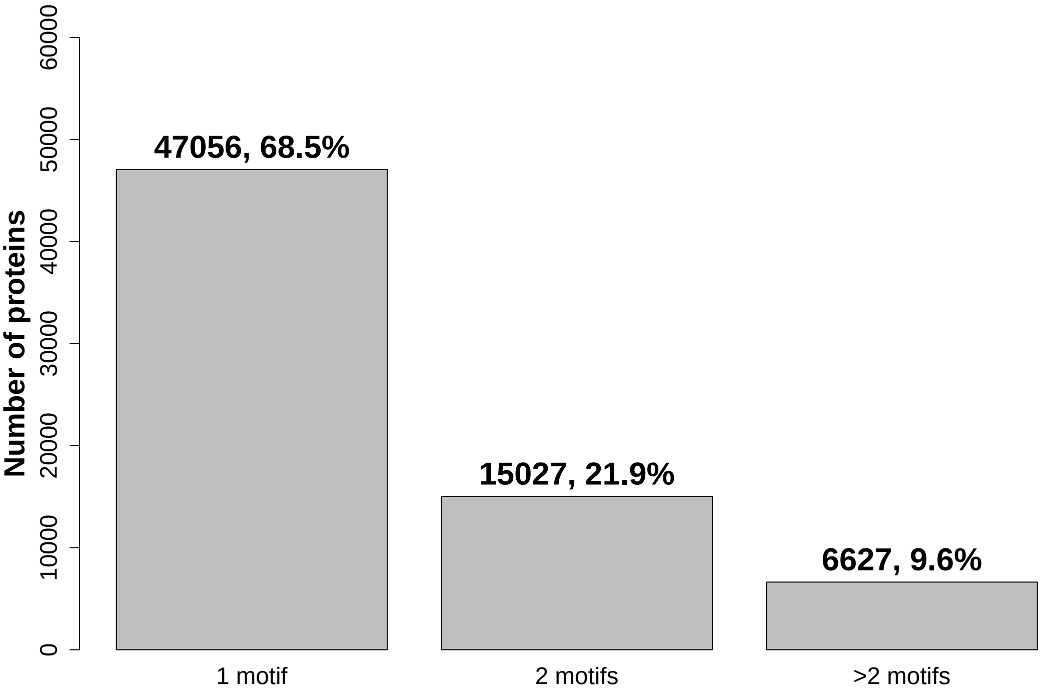


Figure S3.1 Numbers of *E. coli* proteins with 1, 2 and > 2 polyproline motifs.

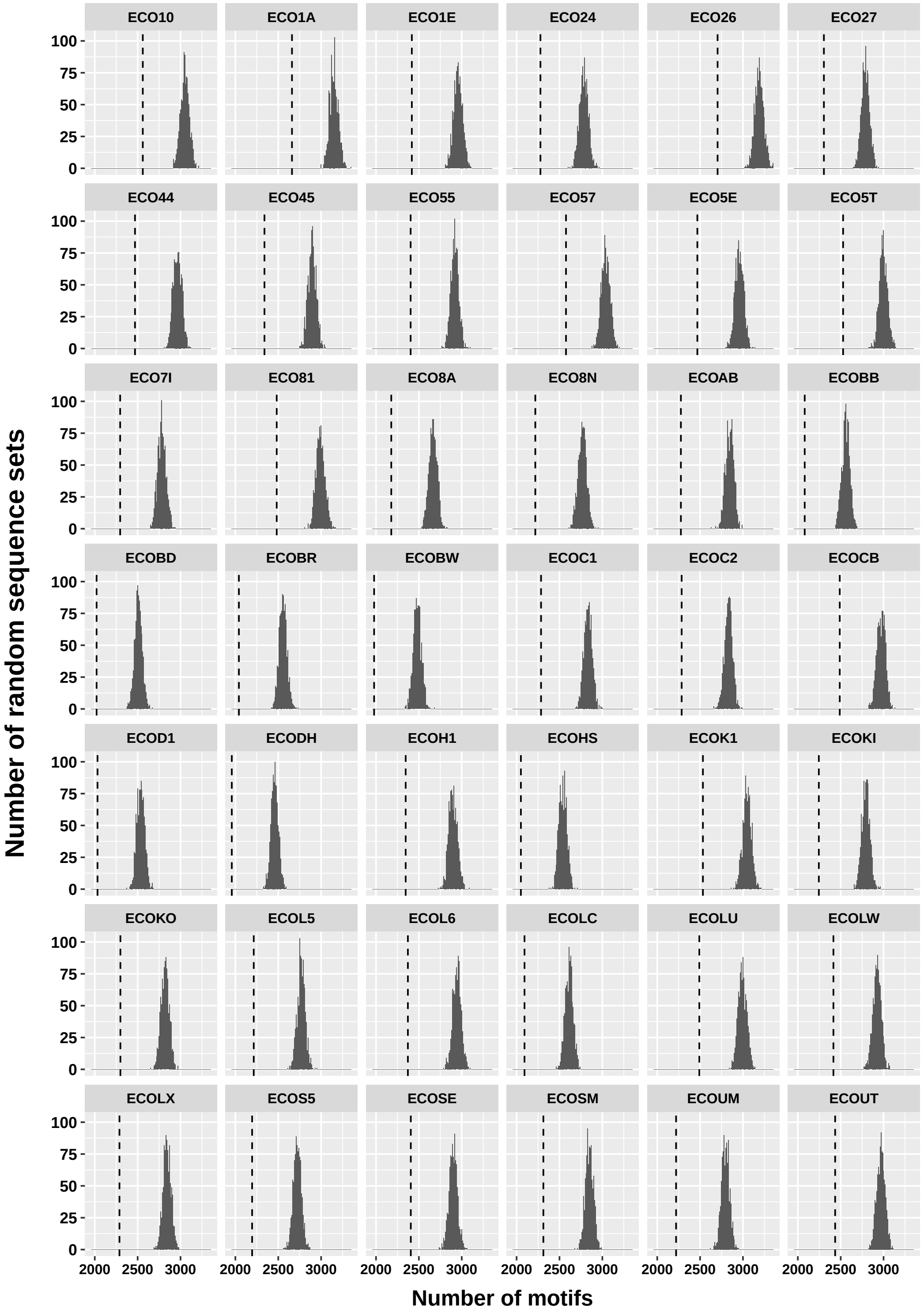


Figure S3.2 Occurrence of polyproline motifs is lower than the random level. The histogram shows the numbers of motifs found in 1000 sets of random sequences, and the dashed line shows the number of motifs found in real sequences. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id of each strain is shown in the panel title. For mapping OMA ids to names of strains, please see Table S3.1.

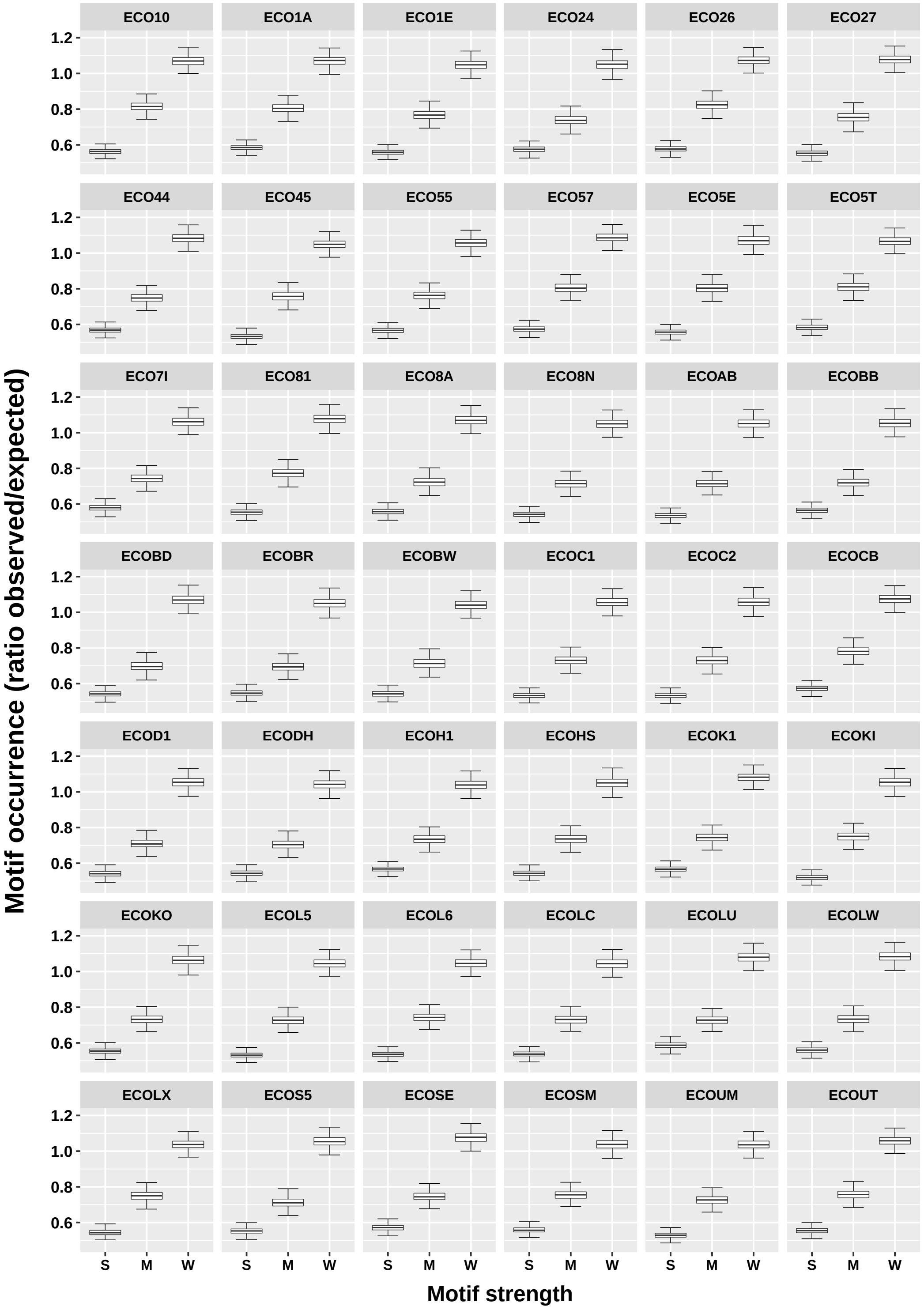


Figure S3.3 Numbers of polyproline motifs negatively correlate with the strength of the ribosome stalling effect. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id of each strain is shown in the panel title. For mapping OMA ids to names of strains, please see Table S3.1. All the differences are significant according to Mann-Whitney-Wilcoxon test, *P*-values < $2.2e^{-16}$.

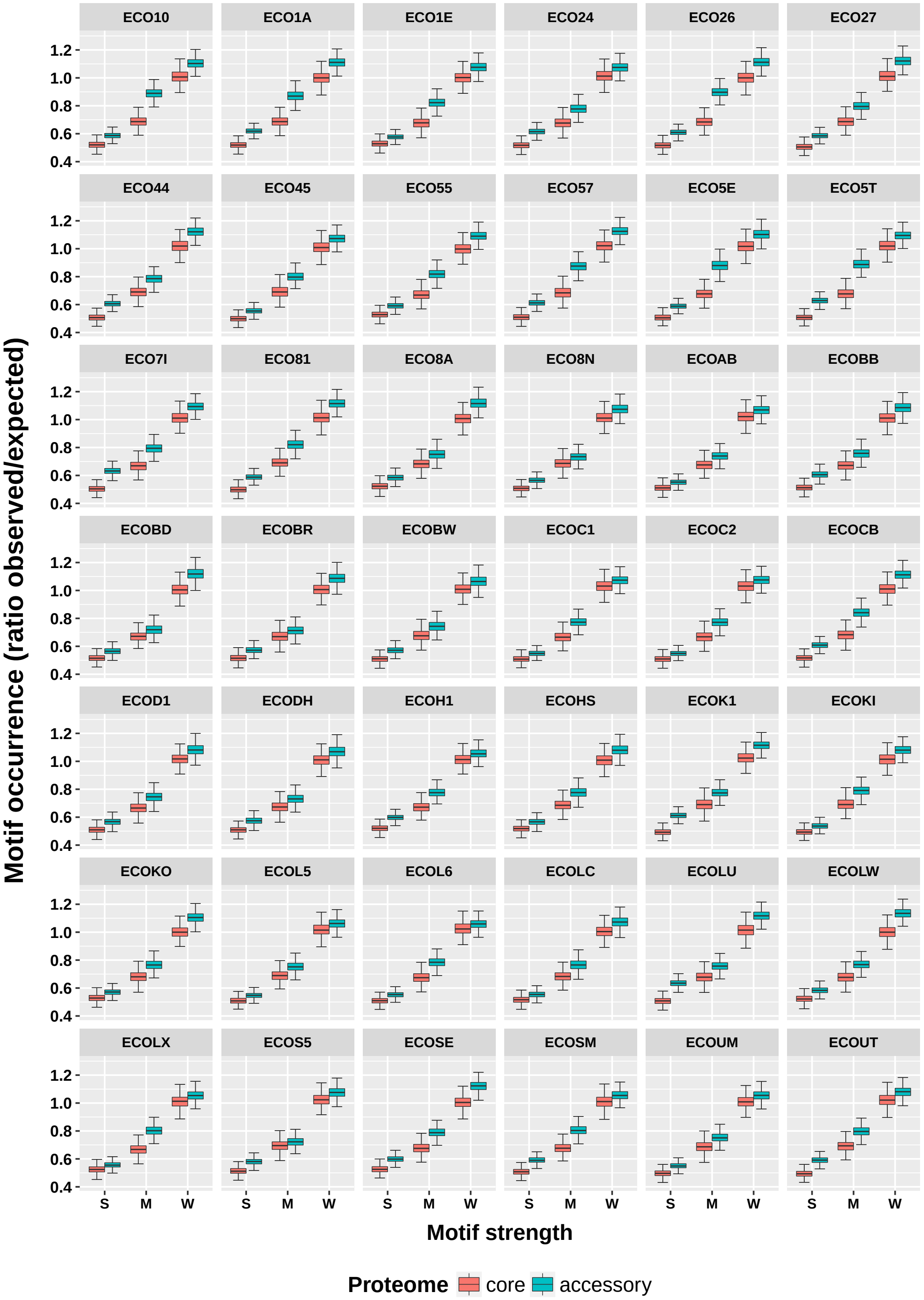


Figure S3.4 Occurrence of polyproline motifs in the core proteome is lower than that in the accessory proteome. The results for 42 *E. coli* strains (except for *E. coli* K-12 MG1655) are shown. The OMA id of each strain is shown in the panel title. For mapping OMA ids to names of strains, please see Table S3.1. All the differences are significant according to Mann-Whitney-Wilcoxon test, *P*-values < $2.2e^{-16}$.

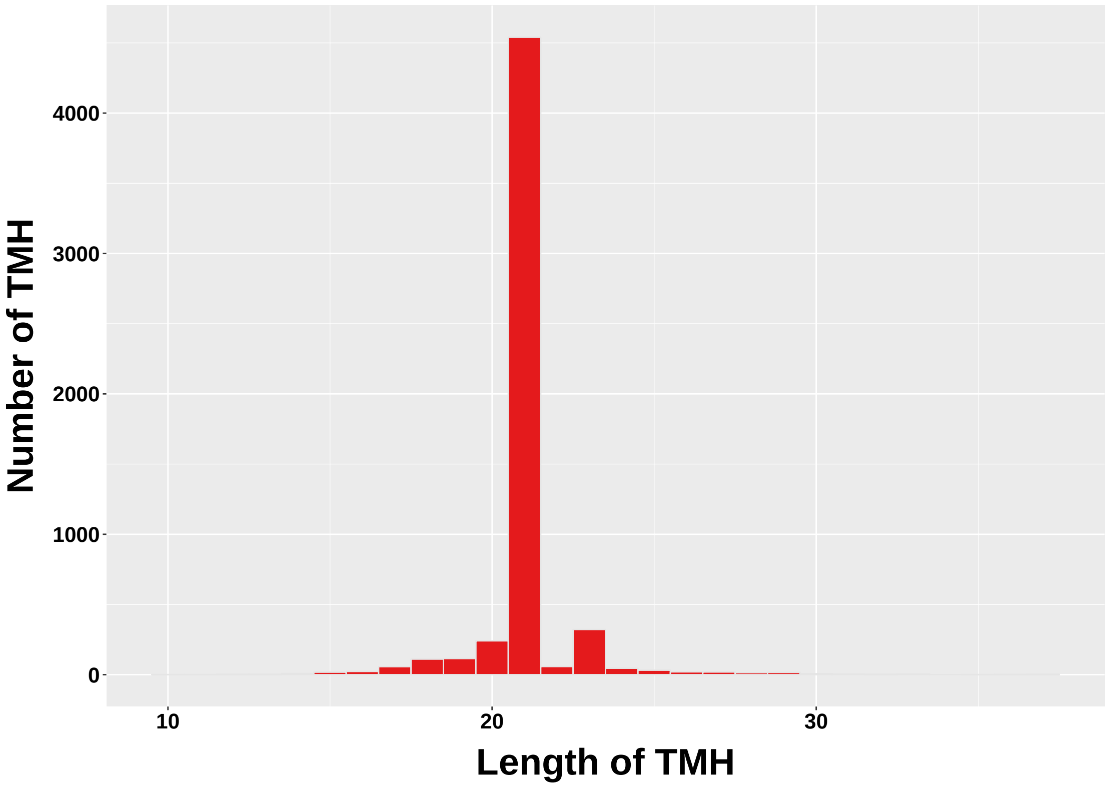


Figure S3.5 Histogram of the TMH length.

Chapter 4

Cooperation of Codon usage, RNA Structure and Polyproline Motif to Regulate Ribosome Kinetics

Ribosome kinetics is important for gene expression and regulation. Translation elongation consists of 3 steps: codon-anticodon matching, peptidyl transfer reaction and ribosomal translocation. The speed of these steps is influenced respectively by the concentration of the cognate tRNA for the A-site codon, polyproline motif which dramatically impairs the peptidyl transfer reaction, and stable RNA structure which slows down the ribosomal translocation. Although the individual function for each of these 3 elements is relatively well studied, the co-effect of these elements remains largely unexplored. Here, we investigated the cooperation of the 3 elements regarding their co-effect on ribosome kinetics. Our analysis reveals that polyproline motifs and slow-translating codons pairing to low-abundance tRNAs tend to occur at the same point, resulting in longer translational pauses. This correlation is especially pronounced in polyproline motifs with stronger ribosome stalling strength and disappears in motifs within transmembrane helices, indicating that it arises from the regulation of ribosome kinetics. We also find a region consisting of fast-translating codons after translational pause sites induced by RNA structures, which may play a functional role in the maintenance of functionally important RNA structures during translation. We observe no cooperation between polyproline motifs and RNA structures regarding ribosome kinetics.

4.1 Introduction

Ribosome kinetics plays an important role in the regulation of gene expression and the quality control of protein synthesis [11]. Fast translation of mRNA stretches coding for structural domains helps avoiding misfolded intermediates [34], and translational pause provides a time delay facilitating multiple co-translational processes [30], such as co-translational protein folding [16] and membrane insertion [37, 38]. Translation elongation consists of 3 steps: matching of tRNA anticodon to mRNA codon in ribosomal A site (codon-anticodon matching), formation of the peptide bond (peptidyl transfer reaction) and moving of ribosome along the mRNA (ribosomal translocation). The speed of translation elongation is influenced by multiple factors, including RNA structure [181, 182], codon usage [31] and amino acid sequence [6, 76, 77]. A stable RNA structure provides an energetic hurdle, which slows down the ribosome moving along the mRNA [54, 181, 182]. A slow-translating codon pairs to low-abundance tRNA species, and thus requires a longer time for a cognate aminoacyl-tRNA to be carried into the ribosomal A site [17]. Proline is a poor A-site peptidyl acceptor [74] and also a poor P-site peptidyl donor [72, 73]. In the translation of consecutive prolines, the peptidyl transfer reaction is dramatically impaired, and thus ribosome stalls [72, 75–79]. Although basically all consecutive-prolines comprising motifs cause translational stalling [77, 138], the arrest strength is influenced by physical and chemical properties of the adjacent amino acids that affect the conformation of the nascent polypeptide chain. Moreover, the stalling strength is modulated by amino acids located — up to position −5 — upstream of the arrest motif [138, 140, 141]. We therefore define a "polyproline motif" as a consecutive stretch of prolines with flanking residues: $X_{(-2)}X_{(-1)}$-nP-$X_{(+1)}$, $n \geq 2$; where $X_{(-2)}$, $X_{(-1)}$ and $X_{(+1)}$ can be any amino acid. This definition is exactly the same as describe in Section 3.1.

These 3 elements — codons usage, RNA structure and polyproline motif — influence different steps of translation elongation and are encoded in different layers of genetic information. Therefore, it is possible to combine these elements at a point or in a region of gene sequences. An important question, which arises in this context, is whether there exits cooperation between these elements regarding the regulation of ribosome kinetics, and if there exists, how do they cooperate. Although a previous study has found a trade-off between tRNA-abundance-based codon usage and RNA secondary structure [7], this question is still far away from being well answered. In this study, we investigated the correlation between the occurrences of polyproline motifs, slow- and fast-translating codons, and RNA structures in *Escherichia coli*, with respect to the regulation of ribosome kinetics. We report a correlation between the occurrences of polyproline motifs and slow-translating codons, which results in longer translational pauses at polyproline motifs. We found that this correlation is more evident in polyproline motifs with stronger ribosome arrest strength and does not exist in motifs within transmembrane helices, indicating the correlation arises from the regulation of ribosome kinetics. We also observed a "boost region" immediately downstream of translational pause sites induced by RNA secondary structures. This boost region consists of fast-translating codons, and maintains the functionally important RNA structures during translation by launching ribosomes to rapidly pass this region. We found no cooperation between polyproline motifs and RNA secondary structures regarding ribosome kinetics.

4.2 Materials and Methods

4.2.1 cDNA and protein sequences

cDNA and protein sequences of 4352 *E. coli* K-12 MG1655 genes were downloaded from the OMA database [156].

4.2.2 Codon-anticodon matching time of genetic codons

The codon-anticodon matching time (CAMT) of genetic codons in *E. coli* was taken from the study of Gorochowski et al. [7] and is shown in Table S4.1 (this data was named "translation time of codon" in [7], and we rename it CAMT to avoid ambiguity). The CAMT was predicted based on tRNA availability and is in arbitrary unit (a.u.).

Difference of the CAMT between consecutive and discrete residues (ΔCAMT) was calculated as:

$$\Delta CAMT = \frac{CAMT_c - CAMT_d}{CAMT_c + CAMT_d} \tag{4.1}$$

where $CAMT_c$ and $CAMT_d$ are the mean CAMT of consecutive and discrete residues, respectively.

4.2.3 Consecutive and discrete residues

Consecutive and discrete residues in *E. coli* proteins were identified using the program fuzzpro from the EMBOSS package [158].

4.2.4 Ribosome stalling strength of polyproline motifs in *E. coli*

The classification of the ribosome stalling strength of polyproline motifs is exactly the same as described in Section 3.2.6.

4.2.5 Transmembrane segments of *E. coli* proteins

This data is exactly the same as described in Section 3.2.14.

4.2.6 Secondary structures of *E. coli* mRNAs

Experimentally probed secondary structures of *E. coli* mRNAs are from the study of Del Campo et al. [56]. These structures were determined by PARS experiment, in which RNase V1 and A/T1 were used to probe bases in double- and single-stranded conformation, respectively. We downloaded the normalized reads counts of V1- and A/T1-treated samples from the GEO database [108] (GSE63817). Exactly the same as described by Del Campo et al. [56], we calculated a PARS score for each nucleotide position as the $\log_2$ value of the ratio between the normalized reads counts from the V1- and A/T1-treated samples. In this PARS experiment, RNase A hydrolyzes at single-stranded C and U nucleotides and RNase T1 cuts at single-stranded G nucleotides [56]. Therefore, adenines were excluded from the analysis. Nucleotide positions which were probed by neither RNase V1 nor A/T1 were also excluded from consideration.

Base pairing probabilities of *E. coli* mRNAs were calculated using RNAfold algorithm from ViennaRNA package [111].

4.3 Results

4.3.1 Consecutive prolines in polyproline motifs contain more slow-translating codons than discrete prolines

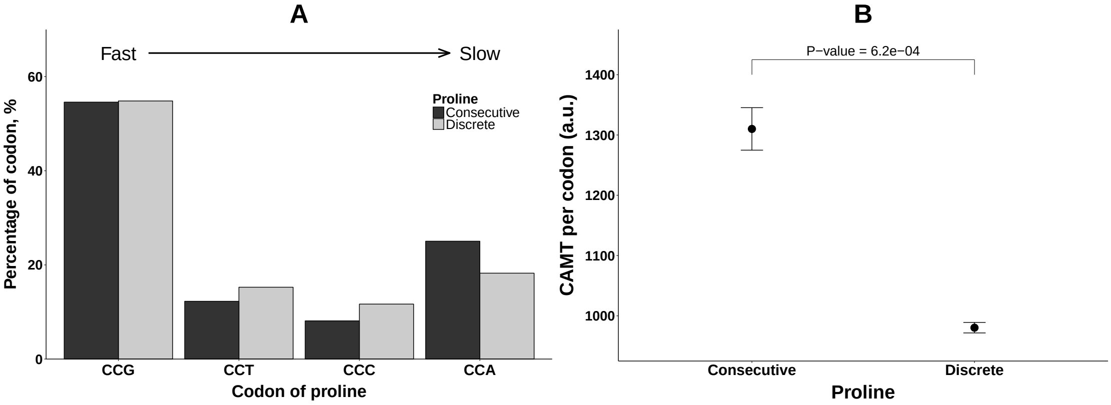


Figure 4.1 Comparison of codon usage and CAMT between consecutive and discrete prolines in *E. coli* proteins. **(A)** Codon usage. The difference is significant according to chi-squared test, P-value = $1.7e^{-30}$. **(B)** CAMT. The mean and standard error of the mean (SEM) are plotted. The P-value of the Mann-Whitney-Wilcoxon test is shown.

First, we investigated the correlation between tRNA-abundance-based codon choice and polyproline motifs. Translation of two or more consecutive prolines results in ribosome stalling [72, 75–79], whereas translation of discrete prolines does not. Consequently, if correlation exists between polyproline motifs and tRNA-abundance-based codon choice with respect to the regulation of ribosome kinetics, consecutive prolines in polyproline motifs and discrete prolines would differ in codon usage. Therefore, we compared the codon usage of consecutive and discrete prolines. In this comparison, we excluded the first 50 residues of proteins in order to avoid potential bias induced by regulation of translation initiation, for example, the selection for codon usage and RNA structure at the beginning of genes [57, 182, 183]. Because a trade-off was found between tRNA abundance and RNA structure supporting a smoothed translation elongation rate [7], one would expect also a trade-off between tRNA abundance and polyproline motifs. However, we found that consecutive prolines in polyproline motifs have more slow-translating codons and a longer codon-anticodon matching time (CAMT) than discrete prolines (Figures 4.1). This result indicates a correlation between slow-translating codons and polyproline motifs, leading to stronger translational pauses at these motifs and an uneven translation elongation rate in *E. coli* proteins.

4.3.2 Correlation between slow-translating codons and polyproline motifs arises from regulation of ribosome kinetics

We then checked whether the correlation between slow-translating codons and polyproline motifs is due to that they are both enriched in proteins with low translation efficiency. We redid the above analysis while restricted to only proteins containing polyproline motifs. In this analysis, we also observed significant differences of codon usage and CAMT between consecutive and discrete prolines (Figure S4.1). This finding indicates the correlation between slow-translating codons and polyproline motifs is not a result of their co-enrichments in proteins with low translation efficiency.

Another possible mechanism explaining the codon bias between consecutive and discrete prolines is replication slippage. DNA replication slippage may lead to a trinucleotide expansion in DNA sequence, and thus produce a consecutive proline stretch of the same codon. Consequently, such stretches may induce codon bias between consecutive and discrete prolines. To investigate the effect of replication slippage on consecutive/discrete proline codon bias, we checked whether the consecutive proline stretches were derived from replication slippage, i.e. whether they tend to have the same codon for each residue.

In *E. coli* proteins, we identified 2008, 95 and 9 consecutive proline stretches of length 2, 3 and > 3, respectively. For stretches of length 2, the proportion of stretches using the same codon for each residue does not differ significantly from expected purely by chance (Figure S4.2A). For stretches of length 3, the proportion of stretches using the same codon for each residue is significantly higher than expected by chance (Figure S4.2B). These results indicate that in *E. coli*, consecutive proline stretches of length 3 are affected by replication slippage, while stretches of length 2 are not. For the 9 consecutive proline stretches of length > 3, none of them uses the same codon for each residue. Therefore, we then investigated the codon usage of consecutive proline stretches of length 2 and 3, separately. As seen in Figure S4.3, for consecutive proline stretches of length 2, their codon usage and CAMT are both significantly different from discrete prolines; whereas for stretches of length 3, their codon usage shows significant difference from discrete prolines while their CAMT does not. These results demonstrate that although replication slippage affects the consecutive proline stretches of length 3, it is not the driving factor for the differences of codon usage and CAMT between consecutive and discrete prolines in *E. coli* proteins.

After ruling out the above 2 possibilities, we hypothesized that the correlation between slow-translating codons and polyproline motifs in *E. coli* proteins arises from the regulation of ribosome kinetics. That is, slow-translating codons and polyproline motifs tend to occur at the same points and thus provide longer time delays in translation elongation.

To verify this hypothesis, we first investigated to what extent the CAMT differs between consecutive and discrete prolines. Significant differences of CAMT were observed between consecutive and discrete residues for more than a half of the 20 amino acids (Figure S4.4). However, most of the amino acids show a slight difference of CAMT between consecutive and discrete residues, and proline is the only one that has a relatively dramatic difference (Figure 4.2A).

We then investigated the codon usage of consecutive prolines in different regions of *E. coli* proteins. If the correlation between polyproline motifs and slow-translating codons arises from selection for stronger translational pauses, the preference for slow-translating codons in consecutive prolines would disappear in protein structures where fast translation is required and translational pause is disfavored. A typical structure of this type is a transmembrane helix. In Chapter 3, we found that polyproline motifs are significantly underrepresented in TMHs, indicating that translational pause impairs the folding of TMH and is therefore depleted. As seen in Figures 4.2B and 4.2C, indeed we found that the codon usage and CAMT of consecutive prolines within TMHs are not significantly different from discrete prolines.

We also investigated the codon usage of consecutive prolines in polyproline motifs with different ribosome stalling strength. The strength of ribosome stalling caused by a polyproline motif is affected by the length of the consecutive proline stretch and the type of flanking residues. Therefore, we classified the polyproline motifs into 3 groups (strong, medium and weak) according to their predicted ribosome stalling strengths, which were assigned based on previously published experimental data (see Section 4.2 for detail).

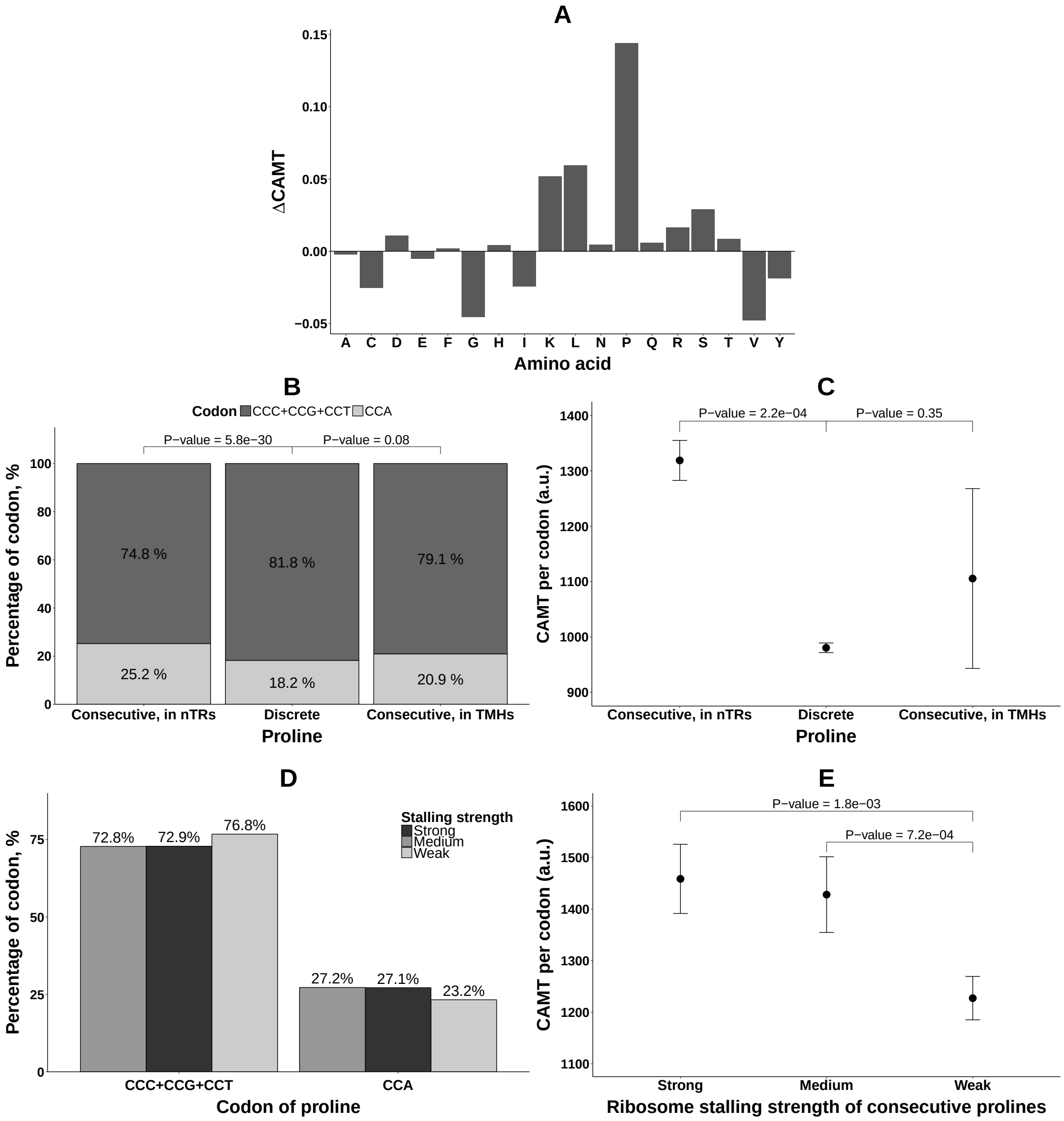


Figure 4.2 **(A)** Difference of CAMT between consecutive and discrete residues for each amino acid in *E. coli*. Methionine and tryptophan were excluded, since they have only one codon. The ΔCAMT of proline is significantly higher than the other amino acids (permutation test, *P*-value = $1.6e^{-3}$). **(B)** and **(C)** Consecutive prolines in non-transmembrane regions (nTRs) show significant differences of codon usage and CAMT to discrete prolines, while consecutive prolines in TMHs do not. **(B)** Codon usage. The *P*-values of chi-squared test are shown. To better display the result, codons of proline are shown in 2 groups: CCA and CCC+CCG+CCT, since CCA has an overwhelmingly longer CAMT [7]. **(C)** CAMT. Error bars indicate SEM. The *P*-values of the Mann–Whitney–Wilcoxon test are shown. **(D)** and **(E)** Polyproline motifs with strong/medium stalling strength have larger proportions of slow-translating codon and a longer CAMT than motifs with weak stalling strength. **(D)** Codon usage. The difference is significant according to chi-squared test, *P*-value = $4.2e^{-3}$. **(E)** CAMT. Error bars indicate SEM. The *P*-values of the Mann–Whitney–Wilcoxon test are shown.

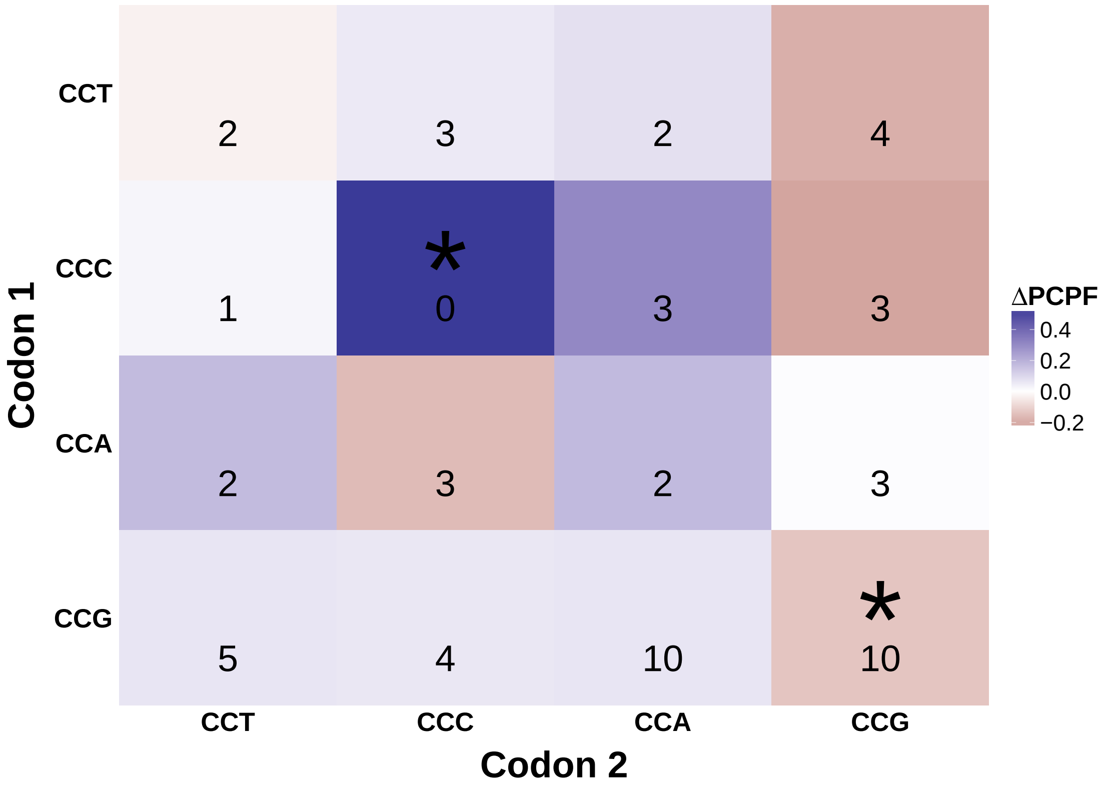


Figure 4.3 Difference of proline codon pair frequency (ΔPCPF) between polyproline motifs with strong/medium and weak strengths. The ΔPCPF was calculated as $(PCPF_{i\_sm} - PCPF_{i\_w})/(PCPF_{i\_sm} + PCPF_{i\_w})$, where $PCPF_{i\_sm}$ is the PCPF of proline codon pair i in polyproline motifs with strong/medium strengths, and $PCPF_{i\_w}$ is that in motifs with weak strength. The number in each cell is the translation speed of the corresponding codon pair, which is taken from [154] and transformed as the fastest-translating codon pair is 10 and the slowest-translating pair is 0. The asterisk indicates a significant difference (P-value of chi-squared test < 0.01).

As seen in Figures 4.2D and 4.2E, consecutive prolines of polyproline motifs with strong and medium stalling strength have larger proportions of slow-translating codons and longer CAMTs than that of motifs with weak stalling strength. This result indicates that the correlation between polyproline motifs and slow-translating codons is stronger for motifs with stronger ribosome arrest strength, and thus demonstrates that the correlation is because of selection for stronger translational pauses.

Another support for this hypothesis is the preference of proline codon pairs in polyproline motifs. Chevance et al. recently investigated the effect of codon context on *in vivo* translation speed in *Salmonella enterica* [154]. They measured the translation speed of all possible codon pairs in a consecutive proline stretch of length 2 within a HHHPPHH context, and found that different codon pairs were translated in different speed. Therefore, we calculated the proline codon pair frequency (PCPF) of polyproline motifs with 2 consecutive prolines in *E. coli* proteins, and compared the PCPF between polyproline motifs with different ribosome stalling strengths. As seen in Figure 4.3, polyproline motifs with strong/medium stalling strength have significantly more CCC-CCC pairs (the slowest-translating pair) and fewer CCG-CCG pairs (one of the two fastest-translating pairs) than motifs with weak stalling strength. This result indicates that the polyproine motifs with stronger stalling strength are more likely to have a codon context which slows down the translation elongation even more and thus enhances the translational pauses.

All the above findings support our hypothesis that the correlation between slow-translating codons and polyproline motifs in *E. coli* proteins arises from the regulation of ribosome kinetics. The slow-translating codons enhance the translational pauses at polyproline motifs and result in longer time delays in translation elongation, which may facilitate co-translational processes, such as co-translational protein folding and membrane insertion.

4.3.3 A boost region exists after the translational pause sites induced by RNA structures

Recently, Del Campo et al. identified 71 ribosome stalling sites induced by RNA secondary structures in *E. coli*, by analyzing data from PARS and ribosome profiling experiments [56]. Using this data, we investigated the correlation between codon usage and RNA secondary structure regarding ribosome kinetics. We aligned the mRNA sequences to these ribosome stalling sites and then compared the CAMT of codons up- and downstream of these sites. We found that ~21 codons downstream of the pause sites are fast-translating codons (Figures 4.4A and S4.5A). This region shows a significantly shorter CAMT than the regions flanking it (Figure 4.4B), and may serve as a "boost region" launching ribosomes from the pause sites.

Why such boost region exists and what is its function? A previous study reported that when 2 adjacent translating ribosomes are close, the nucleotides between them cannot fold to a structure [61]. Then, assuming that these translational pauses are functionally important and therefore necessary to every translating ribosome, we found the function of this boost region becomes obvious: the boost region maintains the functionally important RNA secondary structures during translation elongation.

While a translating ribosome stalls at a pause site induced by RNA secondary structure, following ribosomes accumulate behind it. After the first ribosome unwound the RNA structure and passed the pause site, the boost region gives it a faster translation speed than the following ribosome coming to the pause site. Finally, when the second ribosome arrives at the pause site, an enough space has already appeared between these 2 ribosomes, which enables the re-formation of the functionally important RNA secondary structure. Thus, the functionally important RNA secondary structure gets maintained during translation elongation and arrests every passing ribosome.

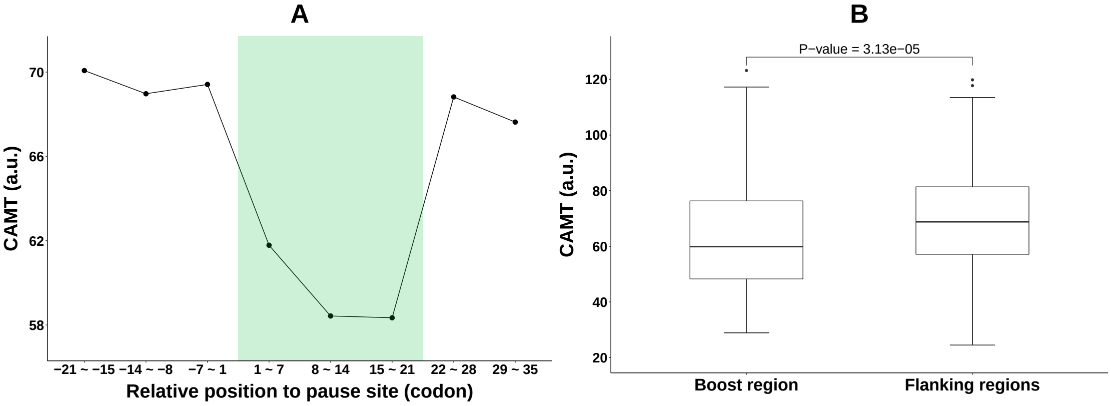


Figure 4.4 **(A)** A boost region (marked green) consisting of fast-translating codons exists downstream of the translational pause sites induced by RNA structures. **(B)** The boost region has a shorter CAMT than regions flanking it. The *P*-value of the Mann–Whitney–Wilcoxon test is shown.

Such boost region does not exist after the translational pause sites induced by polyproline motifs (Figure S4.5B). This is not surprising, because maintenance of sequence elements requires no such mechanism.

4.3.4 No correlation was observed between polyproline motifs and RNA structures regarding ribosome kinetics

We then investigated the correlation between the occurrences of polyproline motifs and RNA secondary structures with respect to their effects on ribosome kinetics. We aligned PARS scores of nucleotides to consecutive/discrete prolines. As seen in Figure 4.5, the nucleotides coding for prolines show higher PARS scores, presumably due to the high GC content of proline codons (CCN). However, RNA structures at consecutive prolines should not affect the translational pauses induced by polyproline motifs. When a ribosome stalls at a polyproline motif, the proline codons are covered by the translating ribosome, and the RNA structures at these codons have already been unwound. Therefore, RNA secondary structures which influence the translational pauses induced by polyproline motifs should locate ~15 nucleotides downstream of these motifs. We observed no significant difference between the downstream structuredness of consecutive and discrete prolines, in either PARS scores or RNAfold predicted base pairing probabilities (Figure 4.5). This result implies that no correlation exits between polyproline motifs and RNA secondary structures regarding ribosome kinetics.

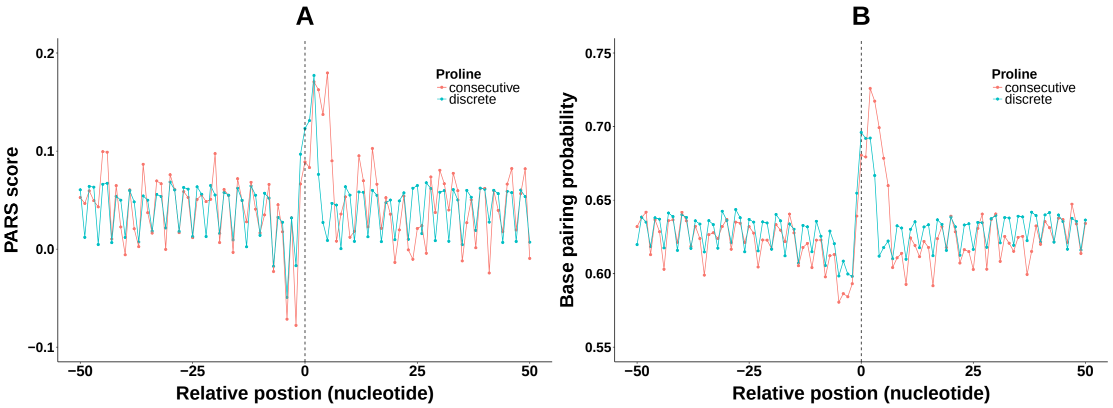


Figure 4.5 No significant difference exists between the structuredness of the nucleotides downstream of consecutive and discrete prolines, in either PARS scores **(A)** or RNAfold predicted base pairing probabilities **(B)**. The nucleotide sequences were aligned to the first nucleotide of consecutive proline stretches or discrete prolines. For each position, the mean value of PARS scores or base pairing probabilities is shown.

4.4 Discussion

Translation is the process that ribosomes synthesize proteins based on genetic information. A variety of processes, such as protein folding and membrane insertion, can happen co-translationally and are essential for producing functional proteins [16, 37, 38]. Intra-molecular variation of translation elongation rate coordinates these co-translational processes [30, 34]. Therefore, the ribosome kinetics is crucial for the quality control of protein synthesis and subject to strict regulation [5, 6]. Multiple elements play regulatory roles in translation elongation, including codon usage [31], RNA structure[181, 182] and polyproline motif [6, 76, 77]. Although the individual effects of these elements are relatively well studied, the cooperation of these elements remains largely unexplored.

In this study, we made a comprehensive attempt to shed light on the cooperation of codon usage, RNA structure and polyproline motif to regulate ribosome kinetics. We investigated the correlation between the occurrences of these elements regarding translational pauses in *E. coli.* We found an enrichment of slow-translating codons in polyproline motifs, which presumably arises from the regulation of ribosome kinetics and may enhance the translational pauses at these motifs. We also observed a boost region, which locates immediately downstream of the translational pause sites induced by RNA structures. This region consists of fast-translating codons, and thus gives a faster translation speed to the ribosome just passed the translational pause site than the ribosome coming to the site. Therefore, this boost region ensures an enough space between these 2 ribosomes, which enables the re-formation of the RNA structures arresting ribosomes. Hence, this boost region serves as a mechanism maintaining the functionally important RNA structures during translation. For polyproline motif and RNA structure, we observed no correlation between them regarding the regulation of ribosome kinetics. Collectively, our study improves the understanding about how multiple elements work as a coherent whole to regulate ribosome kinetics and will be a good base for further research.

4.5 Supplementary Materials

Table S4.1 CAMT of codons in *E. coli* proteins. The CAMT is taken from [7] and is in arbitrary unit (a.u.).

Amino acid	Codon	CAMT
A	GCT	89.3
A	GCC	105.3
A	GCA	67.6
A	GCG	40.8
C	TGT	129.9
C	TGC	59.2
D	GAT	42.9
D	GAC	71.9
E	GAA	16.3
E	GAG	84.0
F	TTT	123.5

Table S4.1 CAMT of codons in *E. coli* proteins. The CAMT is taken from [7] and is in arbitrary unit (a.u.). *(continued)*

Amino acid	Codon	CAMT
F	TTC	126.6
G	GGT	32.5
G	GGC	27.2
G	GGA	71.9
G	GGG	52.1
H	CAT	256.4
H	CAC	166.7
I	ATT	29.9
I	ATC	64.9
I	ATA	204.1
K	AAA	44.2
K	AAG	140.8
L	TTA	185.2
L	TTG	24.8
L	CTT	158.7
L	CTC	120.5
L	CTA	2000.0
L	CTG	12.6
M	ATG	24.5
N	AAT	119.0
N	AAC	99.0
P	CCT	101.0
P	CCC	188.7
P	CCA	5000.0
P	CCG	56.2
Q	CAA	84.7
Q	CAG	73.5
R	CGT	24.8
R	CGC	36.6
R	CGA	163.9
R	CGG	101.0
R	AGA	74.6
R	AGG	153.8
S	TCT	73.5
S	TCC	149.3
S	TCA	222.2
S	TCG	80.6
S	AGT	129.9

Table S4.1 CAMT of codons in *E. coli* proteins. The CAMT is taken from [7] and is in arbitrary unit (a.u.). *(continued)*

Amino acid	Codon	CAMT
S	AGC	70.9
T	ACT	106.4
T	ACC	74.6
T	ACA	303.0
T	ACG	66.2
V	GTT	50.8
V	GTC	112.4
V	GTA	84.7
V	GTG	35.6
W	TGG	68.5
Y	TAT	79.4
Y	TAC	53.2

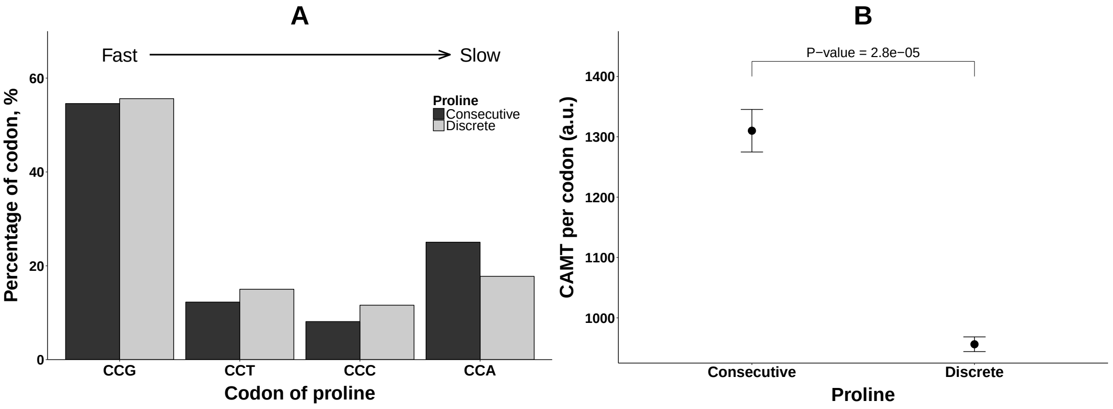


Figure S4.1 Comparison of codon usage and CAMT between consecutive and discrete prolines in *E. coli* proteins containing polyproline motifs. **(A)** Codon usage. The difference is significant according to chi-squared test, *P*-value = $2.1e^{-30}$. **(B)** CAMT. The mean and SEM are plotted. The *P*-value of the Mann-Whitney-Wilcoxon test is shown.

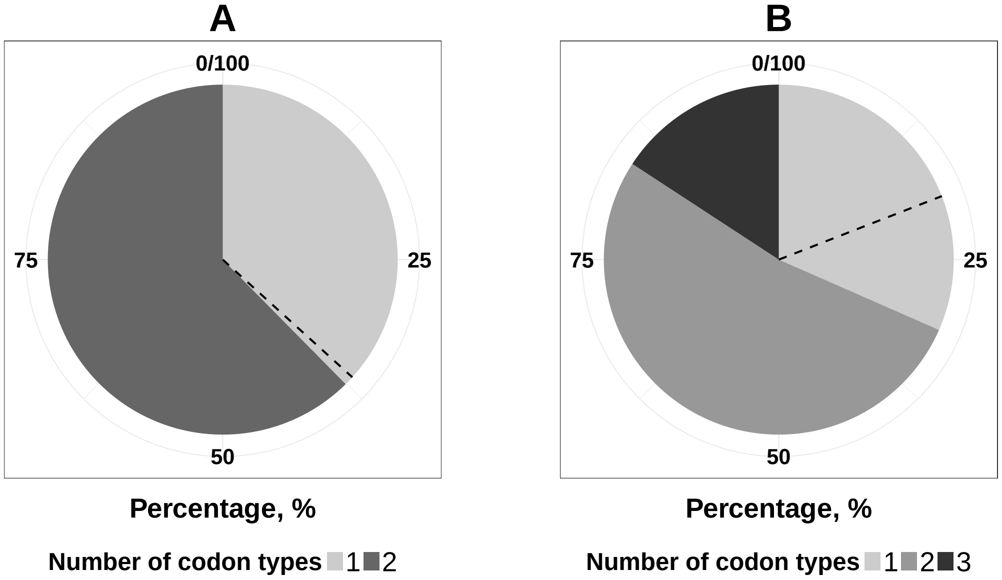


Figure S4.2 Pie charts of the proportion of consecutive proline stretches using the same codon for each residue. Dashed lines indicate the proportions expected by chance. **(A)** Stretches of length 2. The difference is not significant according to chi-squared test, 37.6% *versus* 36.7%, *P*-value = 0.43. **(B)** Stretches of length 3. The difference is significant according to chi-squared test, 31.6% *versus* 19.1%, *P*-value = $3.0e^{-3}$.

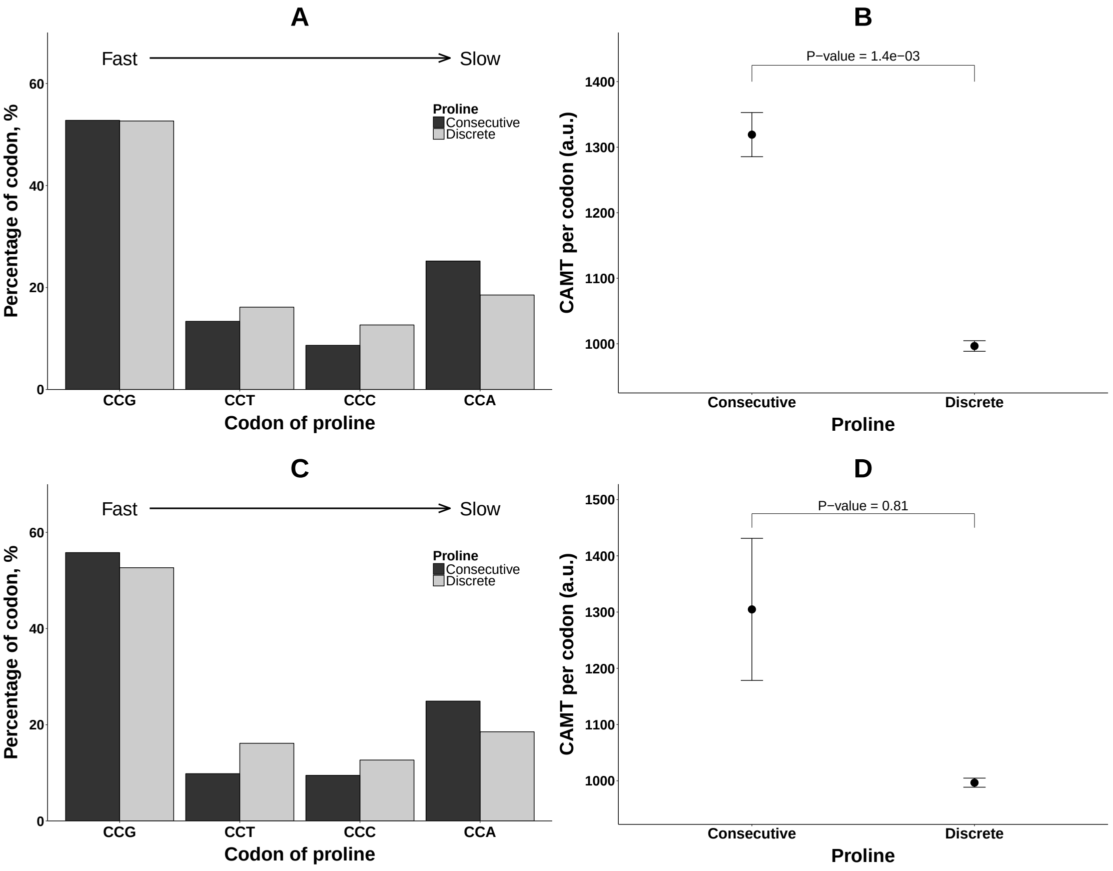


Figure S4.3 For consecutive proline stretches of length 2 in *E. coli*, the codon usage **(A)** and CAMT **(B)** both differ significantly from discrete prolines. **(A)** Codon usage. The difference is significant according to chi-squared test, *P*-value = $3.7e^{-33}$. **(B)** CAMT. The *P*-value of the Mann–Whitney–Wilcoxon test is shown. Error bars indicate SEM. For consecutive proline stretches of length 3 in *E. coli*, the codon usage **(C)** is significantly different from discrete prolines, while the CAMT **(D)** is not. **(C)** Codon usage. The difference is significant according to chi-squared test, *P*-value = $1.1e^{-3}$. **(D)** CAMT. The *P*-value of the Mann–Whitney–Wilcoxon test is shown. Error bars indicate SEM.

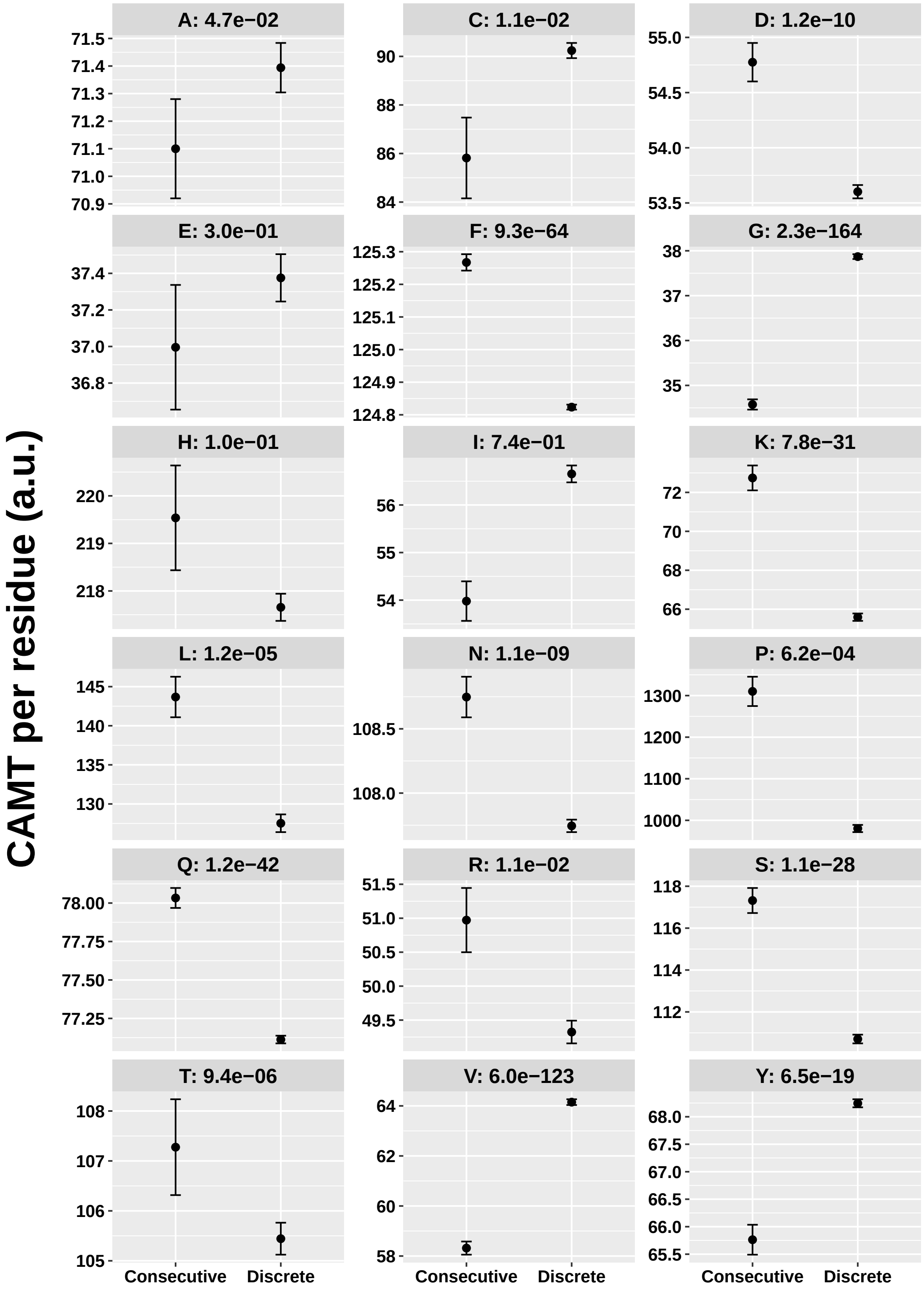


Figure S4.4 CAMT per codon of consecutive and discrete residues of each amino acid in *E. coli*. Methionine and tryptophan were excluded, since they have only one codon. The abbreviation of each amino acid and the *P*-value of Mann–Whitney–Wilcoxon test are shown in the title of each panel. Error bars indicate SEM.

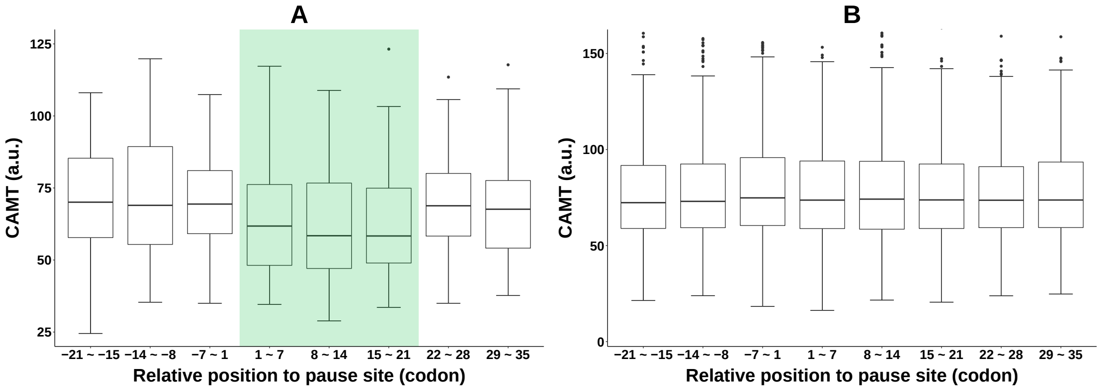


Figure S4.5 The boost region (green) exists after the translational pause sites induced by RNA structures **(A)**, but not after pause sites induced by polyproline motifs **(B)**.

Chapter 5

Summary and Outlook

The regulation of translation elongation is crucial for the quantity and quality control of protein synthesis [7–10]. Several factors are involved in this regulation, including codon usage, RNA structure and amino acid sequence [5, 11–14]. Although the regulation of translation elongation has received growing attentions in recent years, several problems still remain unsolved: (i) there is no practical method to massively distinguish functional RNA structural elements from non-functional ones; (ii) it is still unclear that whether the short amino acid sequences inducing ribosome stallings are under evolutionary selection due to their translational burden and whether they play a regulatory role in translation; and (iii) how these factors work cooperatively remains largely unexplored. In this work, we made comprehensive attempts to shed light on these questions.

In chapter 2, we investigated the sequence-structure relationship in mRNA coding regions and its dependence on temperature, utilizing the recently published PARTE data. We observed a considerable reproducibility of the high-throughput RNA structure probing experiments, and discovered a surprising pattern of how the distance between paralogous mRNA structures varies when temperature grows. Most importantly, we found that high and low thermostability is, respectively, indicative of functionally important RNA structural and sequence elements in mRNA coding regions. Therefore, melting temperature is a crucial parameter, which highlights functionally important RNA structures and sequence segments from non-functional ones, and could serve as a distinguishing feature in the identification of mRNA structural and sequence elements.

In chapter 3, we conducted an evolutionary analysis of polyproline motifs in 43 *E. coli* proteomes. Our result shows that these motifs, which cause ribosome stalling during translation, are disfavored during evolution because of their impairment on translation efficiency. We also observed enrichment of polyproline motifs in the vicinity of protein N-terminus, in inter-domain regions of multi-domain proteins and downstream of transmembrane helices, indicating that translational pause at polyproline motifs is beneficial to translational regulation, co-translational protein folding and proper membrane insertion. This result demonstrates that polyproline motif is exploited as a protein-level *cis*-acting element in the regulation of translation elongation.

In chapter 4, we studied the cooperative manner in which codon usage, RNA structure and polyproline motif work together to regulate translation elongation. We observed that polyproline motifs prefer slow-translating codons, which may result in longer translational pauses at these motifs. We found evidence that this preference arises from regulation of translation elongation. In this study, we also discovered a boost region, which consists of fast-translating codons and locates immediately downstream of translational pause sites induced by RNA structures. This boost region maintains functionally important RNA structures during translation, by ensuring an enough space between ribosomes at and downstream of the pause site.

Recent years, several novel experimental methods have been developed, which greatly promoted the study on the regulation of translation. The high-throughput RNA structure probing methods open a new era of the RNA structure research. These methods enable genome-wide detections of RNA secondary structure profiles. Based on these profiles, Chursov et al. found that the global folding pattern of mRNA coding regions does not play an important role in the regulation of gene expression [112]. These methods also provide numerous candidates for functionally important mRNA structural elements, since the local RNA structures can be easily gleaned from the output of these methods. In this background, our finding that high thermostability is a hallmark of functional mRNA structural elements could be used to narrow down the search space in the identification of these elements, such as RNA thermometer and riboswitch. Ribosome profiling is a technique which produces the translatome — a global "snapshot" of the translation status of all mRNAs at a particular moment [184]. This technique is widely used in the investigation of translation process. It facilitates the identification of translational pause sites, the discovery of novel translational-pause-programming sequences/structures and the study of elongation factors. Recently, Mohammad et al. published a clarification of the bacterial translational pause landscape using ribosome profiling, which challenges the previously reported pauses mediated by Shine-Dalgarno-like sequences [178]. As we identified a regulatory role of polyproline motifs in translation, verifying the predicted functional translational pauses at polyproline motifs by ribosome profiling would be a possible follow-up study. A longstanding problem in the study of translation is that the ribosome translating speed cannot be easily measured. Recently, Chevance et al. developed a bacterial genetic system, which indirectly measures the relative *in vivo* speed of a ribosome translating a short RNA sequence [154]. Using this system, Chevance et al. investigated the effect of codon context on translation [154]. Arising from our finding that polyproline motifs preferentially use slow-translating codons, it would be worth to utilize this system to check whether the use of slow-translating codons in polyproline motifs really prolongs the translational pauses at these motifs. A further goal of the translation study would be to predict the translation profile of a gene based on its sequence [185]. Dana and Tuller recently developed a program which predicts translation elongation rate of a gene by calculating the mean of codon decoding rates derived from ribosome profiling data [17, 185]. Since multiple factors, in addition to codon usage, influence the elongation process, it would be necessary to include the effects of these factors in the prediction. Several experimental methods have been developed to monitor dynamic translation process of a single mRNA molecular and/or by a single ribosome [53, 155, 186]. These experiments would provide valuable parameters to the prediction of gene translation profile. Our research would also contribution to this goal, especially the result about the cooperation of factors.

Bibliography

[1] B. Lewin. *Genes IX*. Jones and Bartlett Publishers, Sudbury, Mass, 2008.

[2] P. P. Amaral, M. E. Dinger, T. R. Mercer, and J. S. Mattick. The eukaryotic genome as an RNA machine. *Science*, 319(5871):1787–9, 2008.

[3] R. Green and H. F. Noller. Ribosomes and translation. *Annu. Rev. Biochem.*, 66:679–716, 1997.

[4] A. Griffiths. *Introduction to genetic analysis*. W.H. Freeman and Co, New York, 2008.

[5] J. Lassak, D. N. Wilson, and K. Jung. Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A. *Mol. Microbiol.*, 99(2):219–35, 2016.

[6] S. Varenne, J. Buc, R. Lloubes, and C. Lazdunski. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.*, 180(3):549–76, 1984.

[7] T. E. Gorochowski, Z. Ignatova, R. A. L. Bovenberg, and J. A. Roubos. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res.*, 43(6):3022–32, 2015.

[8] N. Fluman, S. Navon, E. Bibi, and Y. Pilpel. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. *Elife*, 3:e03440, 2014.

[9] E. P. O'Brien, P. Ciryam, M. Vendruscolo, and C. M. Dobson. Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.*, 47(5):1536–44, 2014.

[10] Z. E. Sauna and C. Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.*, 12(10):683–91, 2011.

[11] M. Marin. Folding at the rhythm of the rare codon beat. *Biotechnol. J.*, 3(8):1047–57, 2008.

[12] M. Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005.

[13] T. Tenson and M. Ehrenberg. Regulatory nascent peptides in the ribosomal tunnel. *Cell*, 108(5):591–4, 2002.

[14] H. Ramu, N. Vázquez-Laslop, D. Klepacki, Q. Dai, J. Piccirilli, R. Micura, and A. S. Mankin. Nascent peptide in the ribosome exit tunnel affects functional properties of the A-site of the peptidyl transferase center. *Mol. Cell*, 41(3):321–30, 2011.

[15] T. A. Thanaraj and P. Argos. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.*, 5(8):1594–612, 1996.

[16] A. A. Komar. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.*, 34(1):16–24, 2009.

[17] A. Dana and T. Tuller. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.*, 42(14):9171–81, 2014.

[18] T. Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2(1):13–34, 1985.

[19] P. M. Sharp, T. M. Tuohy, and K. R. Mosurski. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, 14(13):5125–43, 1986.

[20] A. A. Komar, T. Lesnik, and C. Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation. *FEBS Lett.*, 462(3):387–91, 1999.

[21] F. Bonekamp and K. F. Jensen. The AGG codon is translated slowly in *E. coli* even at very low expression levels. *Nucleic Acids Res.*, 16(7):3013–24, 1988.

[22] L. S. Folley and M. Yarus. Codon contexts from weakly expressed genes reduce expression *in vivo*. *J. Mol. Biol.*, 209(3):359–78, 1989.

[23] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54, 2010.

[24] E. T. Powers and W. E. Balch. Costly mistakes: translational infidelity and protein homeostasis. *Cell*, 134(2):204–6, 2008.

[25] M. Ehrenberg, P. P. Dennis, and H. Bremer. Maximum *rrn* promoter activity in *Escherichia coli* at saturating concentrations of free RNA polymerase. *Biochimie*, 92(1):12–20, 2010.

[26] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23, 2009.

[27] G. Cannarozzi, G. Cannarrozzi, N. N. Schraudolph, M. Faty, P. v. Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral. A role for codon order in translation dynamics. *Cell*, 141(2):355–67, 2010.

[28] M. Kaminska, S. Havrylenko, P. Decottignies, P. Le Maréchal, B. Negrutskii, and M. Mirande. Dynamic organization of aminoacyl-tRNA synthetase complexes in the cytoplasm of human cells. *J. Biol. Chem.*, 284(20):13746–54, 2009.

[29] M. P. Deutscher. The eucaryotic aminoacyl-tRNA synthetase complex: suggestions for its structure and function. *J. Cell Biol.*, 99(2):373–7, 1984.

[30] D. A. Nissley and E. P. O'Brien. Timing is everything: unifying codon translation rates and nascent proteome behavior. *J. Am. Chem. Soc.*, 136(52):17892–8, 2014.

[31] C.-H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol. Cell*, 59(5):744–54, 2015.

[32] B. Hardesty, T. Tsalkova, and G. Kramer. Co-translational folding. *Curr. Opin. Struct. Biol.*, 9(1):111–4, 1999.

[33] G. Zhang, M. Hubalewska, and Z. Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.*, 16(3):274–80, 2009.

[34] E. P. O'Brien, M. Vendruscolo, and C. M. Dobson. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nat. Commun.*, 5:2988, 2014.

[35] F. Cymer, G. v. Heijne, and S. H. White. Mechanisms of integral membrane protein insertion and folding. *J. Mol. Biol.*, 427(5):999–1022, 2015.

[36] T. A. Rapoport. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170):663–9, 2007.

[37] F. Képès. The "+70 pause": hypothesis of a translational control of membrane protein assembly. *J. Mol. Biol.*, 262(2):77–86, 1996.

[38] P. Dessen and F. Képès. The PAUSE software for analysis of translational control over protein targeting: application to *E. nidulans* membrane proteins. *Gene*, 244(1-2):89–96, 2000.

[39] M. H. H. Nørholm, S. Light, M. T. I. Virkki, A. Elofsson, G. v. Heijne, and D. O. Daley. Manipulating the genetic code for membrane protein production: what have we learnt so far? *Biochim. Biophys. Acta*, 1818(4):1091–6, 2012.

[40] A. S. Morgunov and M. M. Babu. Optimizing membrane-protein biogenesis through nonoptimal-codon usage. *Nat. Struct. Mol. Biol.*, 21(12):1023–5, 2014.

[41] S. Pechmann, J. W. Chartron, and J. Frydman. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP *in vivo*. *Nat. Struct. Mol. Biol.*, 21(12):1100–5, 2014.

[42] T. Bornemann, J. Jöckel, M. V. Rodnina, and W. Wintermeyer. Signal sequence-independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. *Nat. Struct. Mol. Biol.*, 15(5):494–9, 2008.

[43] Y. Wan, K. Qu, Z. Ouyang, M. Kertesz, J. Li, R. Tibshirani, D. L. Makino, R. C. Nutter, E. Segal, and H. Y. Chang. Genome-wide measurement of RNA folding energies. *Mol. Cell*, 48(2):169–81, 2012.

[44] I. Tinoco and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293(2):271–81, 1999.

[45] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7, 2010.

[46] Y. Arava, Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.*, 100(7):3889–94, 2003.

[47] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U. S. A.*, 99(9):5860–5, 2002.

[48] P. A. Takizawa, J. L. DeRisi, J. E. Wilhelm, and R. D. Vale. Plasma membrane compartmentalization in yeast by messenger RNA transport and a septin diffusion barrier. *Science*, 290(5490):341–4, 2000.

[49] K. a. Shepard, a. P. Gerber, A. Jambhekar, P. a. Takizawa, P. O. Brown, D. Herschlag, J. L. DeRisi, and R. D. Vale. Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 100(20):11429–34, 2003.

[50] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microRNA target recognition. *Nat. Genet.*, 39(10):1278–84, 2007.

[51] S. A. Shabalina. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.*, 34(8):2428–2437, 2006.

[52] Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, 12(9):641–55, 2011.

[53] J.-D. Wen, L. Lancaster, C. Hodges, A.-C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, and I. Tinoco. Following translation by single ribosomes one codon at a time. *Nature*, 452(7187):598–603, 2008.

[54] G. Faure, A. Y. Ogurtsov, S. A. Shabalina, and E. V. Koonin. Role of mRNA structure in the control of protein folding. *Nucleic Acids Res.*, 44(22):10898–10911, 2016.

[55] R. A. Bartoszewski, M. Jablonsky, S. Bartoszewska, L. Stevenson, Q. Dai, J. Kappes, J. F. Collawn, and Z. Bebok. A synonymous single nucleotide polymorphism in ΔF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J. Biol. Chem.*, 285(37):28741–28748, 2010.

[56] C. Del Campo, A. Bartholomäus, I. Fedyunin, and Z. Ignatova. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet.*, 11(10):1–23, 2015.

[57] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, 324(5924):255–8, 2009.

[58] S. M. Studer and S. Joseph. Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol. Cell*, 22(1):105–15, 2006.

[59] C. Park, X. Chen, J.-r. Yang, and J. Zhang. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.*, 110(8):E678–86, 2013.

[60] H. Zur and T. Tuller. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae. EMBO Rep.*, 13(3):272–7, 2012.

[61] Y. Mao, H. Liu, Y. Liu, and S. Tao. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae. Nucleic Acids Res.*, 42(8):4813–22, 2014.

[62] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, 7(12):995–1001, 2010.

[63] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci.*, 108(27):11063–11068, 2011.

[64] S. a. Mortimer, M. A. Kidwell, and J. a. Doudna. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, 15(7):469–79, 2014.

[65] S. Bhushan, M. Gartmann, M. Halic, J.-P. Armache, A. Jarasch, T. Mielke, O. Berninghausen, D. N. Wilson, and R. Beckmann. α-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. *Nat. Struct. Mol. Biol.*, 17(3):313–7, 2010.

[66] B. Seidelt, C. A. Innis, D. N. Wilson, M. Gartmann, J.-P. Armache, E. Villa, L. G. Trabuco, T. Becker, T. Mielke, K. Schulten, T. A. Steitz, and R. Beckmann. Structural insight into nascent polypeptide chain-mediated translational stalling. *Science*, 326(5958):1412–5, 2009.

[67] H. Nakatogawa and K. Ito. The ribosomal exit tunnel functions as a discriminating gate. *Cell*, 108(5):629–36, 2002.

[68] M. G. Lawrence, L. Lindahl, and J. M. Zengel. Effects on translation pausing of alterations in protein and RNA components of the ribosome exit tunnel. *J. Bacteriol.*, 190(17):5862–9, 2008.

[69] R. Sabi and T. Tuller. A comparative genomics study on the effect of individual amino acids on ribosome stalling. *BMC Genomics*, 16 Suppl 1:S5, 2015.

[70] A. Dana and T. Tuller. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS Comput. Biol.*, 8(11):e1002755, 2012.

[71] C. A. Charneski and L. D. Hurst. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.*, 11(3):e1001508, 2013.

[72] L. K. Doerfel, I. Wohlgemuth, C. Kothe, F. Peske, H. Urlaub, and M. V. Rodnina. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science*, 339(6115):85–8, 2013.

[73] L. K. Doerfel, I. Wohlgemuth, V. Kubyshkin, A. L. Starosta, D. N. Wilson, N. Budisa, and M. V. Rodnina. Entropic contribution of elongation factor P to proline positioning at the catalytic center of the ribosome. *J. Am. Chem. Soc.*, 137(40):12997–3006, 2015.

[74] M. Y. Pavlov, R. E. Watts, Z. Tan, V. W. Cornish, M. Ehrenberg, and A. C. Forster. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc. Natl. Acad. Sci. U. S. A.*, 106(1):50–4, 2009.

[75] E. Gutierrez, B.-S. Shin, C. J. Woolstenhulme, J.-R. Kim, P. Saini, A. R. Buskirk, and T. E. Dever. eIF5A promotes translation of polyproline motifs. *Mol. Cell*, 51(1):35–45, 2013.

[76] S. J. Hersch, M. Wang, S. B. Zou, K.-M. Moon, L. J. Foster, M. Ibba, and W. W. Navarre. Divergent protein motifs direct elongation factor P-mediated translational regulation in *Salmonella enterica* and *Escherichia coli*. *MBio*, 4(2):e00180–13, 2013.

[77] L. Peil, A. L. Starosta, J. Lassak, G. C. Atkinson, K. Virumäe, M. Spitzer, T. Tenson, K. Jung, J. Remme, and D. N. Wilson. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proc. Natl. Acad. Sci. U. S. A.*, 110(38):15265–70, 2013.

[78] D. R. Tanner, D. A. Cariello, C. J. Woolstenhulme, M. A. Broadbent, and A. R. Buskirk. Genetic identification of nascent peptides that induce ribosome stalling. *J. Biol. Chem.*, 284(50):34809–18, 2009.

[79] S. Ude, J. Lassak, A. L. Starosta, T. Kraxenberger, D. N. Wilson, and K. Jung. Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science*, 339(6115):82–5, 2013.

[80] K. P. Lu, G. Finn, T. H. Lee, and L. K. Nicholson. Prolyl *cis-trans* isomerization as a molecular timer. *Nat. Chem. Biol.*, 3(10):619–29, 2007.

[81] D. R. Macinga, M. M. Parojcic, and P. N. Rather. Identification and analysis of aarP, a transcriptional activator of the 2′-N-acetyltransferase in *Providencia stuartii*. *J. Bacteriol.*, 177(12):3407–13, 1995.

[82] R. Thapar. Roles of prolyl isomerases in RNA-mediated gene expression. *Biomolecules*, 5(2):974–99, 2015.

[83] A. A. Adzhubei, F. Eisenmenger, V. G. Tumanyan, M. Zinke, S. Brodzinski, and N. G. Esipova. Third type of secondary structure: noncooperative mobile conformation. Protein Data Bank analysis. *Biochem. Biophys. Res. Commun.*, 146(3):934–8, 1987.

[84] A. A. Adzhubei, F. Eisenmenger, V. G. Tumanyan, M. Zinke, S. Brodzinski, and N. G. Esipova. Approaching a complete classification of protein secondary structure. *J. Biomol. Struct. Dyn.*, 5(3):689–704, 1987.

[85] A. A. Adzhubei, M. J. E. Sternberg, and A. A. Makarov. Polyproline-II helix in proteins: structure and function. *J. Mol. Biol.*, 425(12):2100–32, 2013.

[86] A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28):9691–702, 2005.

[87] A. L. Starosta, J. Lassak, L. Peil, G. C. Atkinson, C. J. Woolstenhulme, K. Virumäe, A. Buskirk, T. Tenson, J. Remme, K. Jung, and D. N. Wilson. A conserved proline triplet in Val-tRNA synthetase and the origin of elongation factor P. *Cell Rep.*, 9(2):476–83, 2014.

[88] L. K. Doerfel and M. V. Rodnina. Elongation factor P: Function and effects on bacterial fitness. *Biopolymers*, 99(11):837–45, 2013.

[89] M. d. Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–44, 2004.

[90] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, 151(3):389–409, 1981.

[91] R. Sabi and T. Tuller. Computational analysis of nascent peptides that induce ribosome stalling and their proteomic distribution in *Saccharomyces cerevisiae. RNA*, 23(7):983–994, 2017.

[92] N. L. Garneau, J. Wilusz, and C. J. Wilusz. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.*, 8(2):113–26, 2007.

[93] M. B. Warf and J. A. Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci.*, 35(3):169–78, 2010.

[94] K. C. Martin and A. Ephrussi. mRNA localization: gene expression in the spatial dimension. *Cell*, 136(4):719–30, 2009.

[95] R. R. Breaker. Riboswitches and the RNA World. *Cold Spring Harb. Perspect. Biol.*, 4(2):a003566, 2012.

[96] P. C. Bevilacqua and J. M. Blose. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.*, 59(1):79–103, 2008.

[97] B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15(3):342–348, 2005.

[98] M. Mandal and R. R. Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.*, 11(1):29–35, 2004.

[99] E. Nudler and A. S. Mironov. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, 29(1):11–7, 2004.

[100] A. G. Nackley, S. A. Shabalina, I. E. Tchivileva, K. Satterfield, O. Korchynskyi, S. S. Makarov, W. Maixner, and L. Diatchenko. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, 314(5807):1930–3, 2006.

[101] D. B. Carlini, Y. Chen, and W. Stephan. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics*, 159(2):623–33, 2001.

[102] P. O. Ilyinskii, T. Schmidt, D. Lukashev, A. B. Meriin, G. Thoidis, D. Frishman, and A. M. Shneider. Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS*, 13(5):421–30, 2009.

[103] J. Duan, M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. Synonymous mutations in the human *dopamine receptor D2 (DRD2)* affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, 12(3):205–216, 2003.

[104] E. Goz and T. Tuller. Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. *BMC Genomics*, 16 Suppl 1(Suppl 10):S4, 2015.

[105] M. Macossay-Castillo, S. Kosol, P. Tompa, and R. Pancsa. Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. *PLoS Comput. Biol.*, 10(5):e1003607, 2014.

[106] M. F. Lin, P. Kheradpour, S. Washietl, B. J. Parker, J. S. Pedersen, and M. Kellis. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res.*, 21(11):1916–28, 2011.

[107] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 5(4):823–6, 1986.

[108] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. a. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(D1):D991–5, 2013.

[109] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–38, 2012.

[110] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, 40(D1):D700–5, 2012.

[111] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6:26, 2011.

[112] A. Chursov, M. C. Walter, T. Schmidt, A. Mironov, A. Shneider, and D. Frishman. Sequence-structure relationships in yeast mRNAs. *Nucleic Acids Res.*, 40(3):956–62, 2012.

[113] B. Korber. HIV signature and sequence variation analysis. In *Comput. Anal. HIV Mol. Seq.* Part 4, pages 55–72. Dordrecht, Netherlands: Kluwer Academic Publishers, 2000.

[114] A. Jambhekar, K. McDermott, K. Sorber, K. a. Shepard, R. D. Vale, P. a. Takizawa, and J. L. DeRisi. Unbiased selection of localization elements reveals *cis*-acting determinants of mRNA bud localization in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A.*, 102(50):18005–18010, 2005.

[115] D. H. Mathews. Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, 359(3):526–32, 2006.

[116] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5(1):104, 2004.

[117] A. Churkin, A. Avihoo, M. Shapira, and D. Barash. RNAthermsw: direct temperature simulations for predicting the location of RNA thermometers. *PLoS One*, 9(4):e94340, 2014.

[118] A. Chursov, S. J. Kopetzky, G. Bocharov, D. Frishman, and A. Shneider. RNAtips: analysis of temperature-induced changes of RNA secondary structure. *Nucleic Acids Res.*, 41(W1):W486–91, 2013.

[119] P. Clote, F. Ferré, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, 2005.

[120] P. G. Higgs. RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.*, 33(3):199–253, 2000.

[121] M. Ringnér and M. Krogh. Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, 1(7):e72, 2005.

[122] X. Li, G. Quon, H. D. Lipshitz, and Q. Morris. Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–107, 2010.

[123] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, 34(17):e117, 2006.

[124] S. Steigele, W. Huber, C. Stocsits, P. F. Stadler, and K. Nieselt. Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biol.*, 5(1):25, 2007.

[125] C. Olivier, G. Poirier, P. Gendron, A. Boisgontier, F. Major, and P. Chartrand. Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.*, 25(11):4752–66, 2005.

[126] L. C. Reineke, A. a. Komar, M. G. Caprara, and W. C. Merrick. A small stem loop element directs internal initiation of the *URE2* internal ribosome entry site in *Saccharomyces cerevisiae. J. Biol. Chem.*, 283(27):19011–25, 2008.

[127] A. A. Komar, T. Lesnik, C. Cullin, W. C. Merrick, H. Trachsel, and M. Altmann. Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [*URE3*] propagation in yeast cells. *EMBO J.*, 22(5):1199–1209, 2003.

[128] D. L. Bentley. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, 15(3):163–75, 2014.

[129] A. Kanhere, K. Viiri, C. C. Araújo, J. Rasaiyaah, R. D. Bouwman, W. a. Whyte, C. F. Pereira, E. Brookes, K. Walker, G. W. Bell, A. Pombo, A. G. Fisher, R. a. Young, and R. G. Jenner. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol. Cell*, 38(5):675–88, 2010.

[130] D. Y. Vargas, K. Shah, M. Batish, M. Levandoski, S. Sinha, S. a. E. Marras, P. Schedl, and S. Tyagi. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, 147(5):1054–65, 2011.

[131] M. Meyer, M. Plass, J. Pérez-Valle, E. Eyras, and J. Vilardell. Deciphering 3′ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol. Cell*, 43(6):1033–9, 2011.

[132] B. Zamft, L. Bintu, T. Ishibashi, and C. Bustamante. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl. Acad. Sci. U. S. A.*, 109(23):8948–53, 2012.

[133] L. P. Eperon, I. R. Graham, a. D. Griffiths, and I. C. Eperon. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell*, 54(3):393–401, 1988.

[134] F. Righetti and F. Narberhaus. How to find RNA thermometers. *Front. Cell. Infect. Microbiol.*, 4:132, 2014.

[135] F. Narberhaus, T. Waldminghaus, and S. Chowdhury. RNA thermometers. *FEMS Microbiol. Rev.*, 30(1):3–16, 2006.

[136] J. Kortmann and F. Narberhaus. Bacterial RNA thermometers: molecular zippers and switches. *Nat. Rev. Microbiol.*, 10(4):255–65, 2012.

[137] T. Waldminghaus, L. C. Gaubig, and F. Narberhaus. Genome-wide bioinformatic prediction and experimental evaluation of potential RNA thermometers. *Mol. Genet. Genomics*, 278(5):555–564, 2007.

[138] 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 Rep.*, 11(1):13–21, 2015.

[139] F. Cymer, R. Hedman, N. Ismail, and G. v. Heijne. Exploration of the arrest peptide sequence space reveals arrest-enhanced variants. *J. Biol. Chem.*, 290(16):10208–15, 2015.

[140] S. Elgamal, A. Katz, S. J. Hersch, D. Newsom, P. White, W. W. Navarre, and M. Ibba. EF-P dependent pauses integrate proximal and distal signals during translation. *PLoS Genet.*, 10(8):e1004553, 2014.

[141] A. L. Starosta, J. Lassak, L. Peil, G. C. Atkinson, K. Virumäe, T. Tenson, J. Remme, K. Jung, and D. N. Wilson. Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site. *Nucleic Acids Res.*, 42(16):10711–9, 2014.

[142] A. Mandal, S. Mandal, and M. H. Park. Genome-wide analyses and functional classification of proline repeat-rich proteins: potential role of eIF5A in eukaryotic evolution. *PLoS One*, 9(11):e111800, 2014.

[143] J. Lassak, E. C. Keilhauer, M. Fürst, K. Wuichet, J. Gödeke, A. L. Starosta, J.-M. Chen, L. Søgaard-Andersen, J. Rohr, D. N. Wilson, S. Häussler, M. Mann, and K. Jung. Arginine-rhamnosylation as new strategy to activate translation elongation factor P. *Nat. Chem. Biol.*, 11(4):266–70, 2015.

[144] T. Yanagisawa, T. Sumida, R. Ishii, C. Takemoto, and S. Yokoyama. A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P. *Nat. Struct. Mol. Biol.*, 17(9):1136–43, 2010.

[145] S. B. Zou, S. J. Hersch, H. Roy, J. B. Wiggers, A. S. Leung, S. Buranyi, J. L. Xie, K. Dare, M. Ibba, and W. W. Navarre. Loss of elongation factor P disrupts bacterial outer membrane integrity. *J. Bacteriol.*, 194(2):413–25, 2012.

[146] D. B. Kearns, F. Chu, R. Rudner, and R. Losick. Genes governing swarming in *Bacillus subtilis* and evidence for a phase variation mechanism controlling surface motility. *Mol. Microbiol.*, 52(2):357–69, 2004.

[147] A. Rajkovic, S. Erickson, A. Witzky, O. E. Branson, J. Seo, P. R. Gafken, M. A. Frietas, J. P. Whitelegge, K. F. Faull, W. Navarre, A. J. Darwin, and M. Ibba. Cyclic rhamnosylated elongation factor P establishes antibiotic resistance in *Pseudomonas aeruginosa*. *MBio*, 6(3):e00823, 2015.

[148] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.*, 48(1):77–84, 2003.

[149] T. Yanagisawa, H. Takahashi, T. Suzuki, A. Masuda, N. Dohmae, and S. Yokoyama. *Neisseria meningitidis* translation elongation factor P and its active-site arginine residue are essential for cell viability. *PLoS One*, 11(2):e0147907, 2016.

[150] K. Gäbel, J. Schmitt, S. Schulz, D. J. Näther, and J. Soppa. A comprehensive analysis of the importance of translation initiation factors for *Haloferax volcanii* applying deletion and conditional depletion mutants. *PLoS One*, 8(11):e77188, 2013.

[151] T. E. Dever, E. Gutierrez, and B.-S. Shin. The hypusine-containing translation factor eIF5A. *Crit. Rev. Biochem. Mol. Biol.*, 49(5):413–25, 2014.

[152] H. Sievert, N. Pällmann, K. K. Miller, I. Hermans-Borgmeyer, S. Venz, A. Sendoel, M. Preukschas, M. Schweizer, S. Boettcher, P. C. Janiesch, T. Streichert, R. Walther, M. O. Hengartner, M. G. Manz, T. H. Brümmendorf, C. Bokemeyer, M. Braig, J. Hauber, K. E. Duncan, and S. Balabanov. A novel mouse model for inhibition of DOHH-mediated hypusine modification reveals a crucial function in embryonic development, proliferation and oncogenic transformation. *Dis. Model. Mech.*, 7(8):963–76, 2014.

[153] I. Hauber, D. Bevec, J. Heukeshoven, F. Krätzer, F. Horn, A. Choidas, T. Harrer, and J. Hauber. Identification of cellular deoxyhypusine synthase as a novel target for antiretroviral therapy. *J. Clin. Invest.*, 115(1):76–85, 2005.

[154] F. F. V. Chevance, S. Le Guyon, and K. T. Hughes. The effects of codon context on *in vivo* translation speed. *PLoS Genet.*, 10(6):e1004392, 2014.

[155] C. Chen, H. Zhang, S. L. Broitman, M. Reiche, I. Farrell, B. S. Cooperman, and Y. E. Goldman. Dynamics of translation by single ribosomes through mRNA secondary structures. *Nat. Struct. Mol. Biol.*, 20(5):582–588, 2013.

[156] A. M. Altenhoff, N. Škunca, N. Glover, C.-M. Train, A. Sueki, I. Piližota, K. Gori, B. Tomiczek, S. Müller, H. Redestig, G. H. Gonnet, and C. Dessimoz. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.*, 43(D1):D240–9, 2015.

[157] O. Clermont, J. K. Christenson, E. Denamur, and D. M. Gordon. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.*, 5(1):58–65, 2013.

[158] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6):276–7, 2000.

[159] G. Nuel. Significance score of motifs in biological sequences. In M. A. Mahdavi, editor, *Bioinforma. - Trends Methodol.* Pages 173–94. InTech, 2011.

[160] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7(1):539, 2011.

[161] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–21, 2010.

[162] E. Paradis, J. Claude, and K. Strimmer. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–90, 2004.

[163] D. M. Krylov, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.*, 13(10):2229–35, 2003.

[164] R. Albalat and C. Cañestro. Evolution by gene loss. *Nat. Rev. Genet.*, 17(7):379–91, 2016.

[165] E. Borenstein, T. Shlomi, E. Ruppin, and R. Sharan. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucleic Acids Res.*, 35(1):e7, 2007.

[166] J. R. Wiśniewski and D. Rakus. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the *Escherichia coli* proteome. *J. Proteomics*, 109:322–31, 2014.

[167] J. R. Wiśniewski and D. Rakus. Quantitative analysis of the *Escherichia coli* proteome. *Data Br.*, 1:7–11, 2014.

[168] J. D. Glasner, P. Liss, G. Plunkett, A. Darling, T. Prasad, M. Rusch, A. Byrnes, M. Gilson, B. Biehl, F. R. Blattner, and N. T. Perna. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, 31(1):147–51, 2003.

[169] S. D. Lam, N. L. Dawson, S. Das, I. Sillitoe, P. Ashford, D. Lee, S. Lehtinen, C. A. Orengo, and J. G. Lees. Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.*, 44(D1):D404–9, 2016.

[170] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.*, 43(D1):D204–12, 2015.

[171] T. E. F. Quax, N. J. Claassens, D. Söll, and J. v. d. Oost. Codon bias as a means to fine-tune gene expression. *Mol. Cell*, 59(2):149–61, 2015.

[172] A. A. Komar and R. Jaenicke. Kinetics of translation of *γ*B crystallin and its circularly permutated variant in an *in vitro* cell-free system: possible relations to codon distribution and protein folding. *FEBS Lett.*, 376(3):195–8, 1995.

[173] I. J. Purvis, A. J. Bettany, T. C. Santiago, J. R. Coggins, K. Duncan, R. Eason, and A. J. Brown. The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. A hypothesis. *J. Mol. Biol.*, 193(2):413–7, 1987.

[174] L. Tu, P. Khanna, and C. Deutsch. Transmembrane segments form tertiary hairpins in the folding vestibule of the ribosome. *J. Mol. Biol.*, 426(1):185–98, 2014.

[175] D. N. Wilson, S. Arenz, and R. Beckmann. Translation regulation via nascent polypeptide-mediated ribosome stalling. *Curr. Opin. Struct. Biol.*, 37:123–33, 2016.

[176] L. Tetsch, C. Koller, I. Haneburger, and K. Jung. The membrane-integrated transcriptional activator CadC of *Escherichia coli* senses lysine indirectly via the interaction with the lysine permease LysP. *Mol. Microbiol.*, 67(3):570–83, 2008.

[177] H. E. Marman, A. R. Mey, and S. M. Payne. Elongation factor P and modifying enzyme PoxA are necessary for virulence of *Shigella flexneri*. *Infect. Immun.*, 82(9):3612–21, 2014.

[178] F. Mohammad, C. J. Woolstenhulme, R. Green, and A. R. Buskirk. Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Rep.*, 14(4):686–94, 2016.

[179] M. Halic, T. Becker, M. R. Pool, C. M. T. Spahn, R. A. Grassucci, J. Frank, and R. Beckmann. Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, 427(6977):808–14, 2004.

[180] N. Mason, L. F. Ciufo, and J. D. Brown. Elongation arrest is a physiologically important function of signal recognition particle. *EMBO J.*, 19(15):4164–74, 2000.

[181] D. M. Mauger, N. A. Siegfried, and K. M. Weeks. The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Lett.*, 587(8):1180–8, 2013.

[182] T. Tuller, Y. Y. Waldman, M. Kupiec, and E. Ruppin. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.*, 107(8):3645–50, 2010.

[183] K. Bentele, P. Saffert, R. Rauscher, Z. Ignatova, and N. Blüthgen. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.*, 9:675, 2013.

[184] N. T. Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, 15(3):205–13, 2014.

[185] A. Dana and T. Tuller. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. *G3 (Bethesda).*, 5(1):73–80, 2014.

[186] X. Yan, T. A. Hoek, R. D. Vale, and M. E. Tanenbaum. Dynamics of translation of single mRNA molecules *in vivo*. *Cell*, 165(4):976–89, 2016.