Inferring the translation speed and determining its relationship to the protein structure produced

Alistair G. Martin

Department of Statistics

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

Green Templeton College

April 2017

Acknowledgements

So here it is, my thesis, presented neatly and in proper formatting, spanning nearly two hundred pages, and ultimately representing four years of blood, sweat, and tears in its production. I am proud of what it contains and the effort it took. The results won't change the world; they likely won't have any lasting impact; however, I smile knowing that in producing this thesis I discovered things about our world that no one before me knew. Science is always so exciting when you think about it that way.

As with nearly all theses, this was not a solitary effort. I need to acknowledge the many family members and friends, whether they be from Scotland, GTBC, the Oxford MTG community, the DTC, or the OPIG. Without their collective friendship, encouragement and support I likely would not be writing this acknowledgement. They have made undertaking a doctorate in Oxford an incredible experience, one that I will always remember fondly.

Personal thanks should be given to Paul Taylor and Beth McMillan, board game buddies and bad influences. Anna Kotova, who, while short in stature, is unmissable in her comradery. Jinwoo Leem, a friend who survived the stress of the Deane group, captaincy, and the DPhil simultaneously to me. Charlotte Deane, you always pushed me just the right amount. Lastly, Sophie, you have been a rock during this final year. Thank you all so very much

Abstract

The central dogma outlines the flow of information within a cell, whereby a DNA sequence is transcribed into RNA, which, in turn, is translated into a protein. This flow is unidirectional, meaning that once constructed, a protein should retain no knowledge of either the RNA or DNA sequence which led to its creation. Due to the degeneracy of the genetic code, multiple synonymous mRNA sequences can result in the same protein being produced. However, an increasing volume of experimental work shows that while these synonymous sequences produce the same amino acid sequence, the encoded proteins may differ in their physical properties. These results suggest that there is information contained in the mRNA sequence pertaining to the structure of the encoded protein above and beyond mere specification of the amino acid sequence. This thesis investigates whether the speed with which a codon is translated biases the protein structure produced. The initial chapters focus on determining a suitable metric for the translation speed, comparing various theoretical estimates to a new experimental measure. Finding that the estimators perform poorly, we construct a transcriptome-wide database relating the experimentally derived translation speeds directly to a large number of experimentally derived protein structures. Using this database to test our hypothesis, we observe various associations between the translation speed and the protein structure produced. Our analysis is the first time that the relationship between translation speed and protein structure has been investigated on a transcriptome-wide scale using purely experimental data. Our findings provide strong support for the cotranslational folding hypothesis which suggests a protein folds while it is being produced.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Introduction

Proteins are the macromolecular machines behind life; they are the mechanics for when a cell is damaged, the trash collector for any cellular waste, and the emergency services that respond when a cell is put into unpleasant environments [1–3]. The number of different proteins that exist within a single cell reaches into the thousands in bacteria and the tens of thousands in more complex multicellular organisms [4, 5]. Most of these proteins have a specific role within the cell and, hence, have a single defined function [6]. These functions are often related to the protein's three-dimensional structure, and the three-dimensional structure is thought to be determined solely by the protein sequence [7, 8]. Likewise, some protein's functionality is enabled by a lack a defined structure, which has also been found to be a property encoded within the protein sequence [9]. In turn, the protein sequence is derived directly from the genome [10]. For the information in the genome to encode the protein sequence, it must pass from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA), via transcription, and then from RNA to protein, via translation [10]. This flow of information is referred to as the central dogma of biology and suggests that the genomic sequence (DNA) uniquely defines the protein sequence, and therefore both the structure and function [11, 12].

In actuality, this process has been shown to be far more complex and intertwined than the original formulation of the central dogma leads one to believe [13]. In this chapter, we outline how proteins are defined and then created from the information contained within the genome. We discuss how the protein structure is commonly described alongside established theories as to how it forms. We then present evidence indicating that proteins take on their structure as they are undergoing translation and, therefore, suggest that the rate of translation along the RNA may be a significant factor in the resultant structure. We give evidence to support this hypothesis, showing that both the degeneracy within the genome and translational kinetics can, in fact, produce different proteins. This evidence leads to our thesis investigation, namely "Inferring the translation speed and determining its relationship to the protein structure produced".

1.2 The central dogma

1.2.1 DNA

The genome contains all the genetic information of an organism encoded over several DNA molecules, referred to as the chromosomes [10]. Each DNA molecule consists of two long polymer chains, or strands, woven around each other to create the signature double helix shape [14]. As shown in Figure 1.1, the chain is constructed from four different fundamental units called nucleotides or bases, namely two pyrimidines, cytosine (C) and thymine (T), and two purines, guanine (G) and adenine (A). The nucleotides of the two chains are not selected independently as only particular pairs of bases are found to align due to base pairing; a term that refers to the strong hydrogen bonds that form between the chains when particular nucleotides align [15, 16]. Specifically, purine A pairs with pyrimidine T to create two hydrogen bonds, and purine G pairs with pyrimidine C to create three hydrogen bonds. Other pairings do not generate as many hydrogen bonds as can be seen via the chemical structures given in Figure 1.1. For example, ATCG in one strand will be mirrored by TAGC in the other strand. As a result,

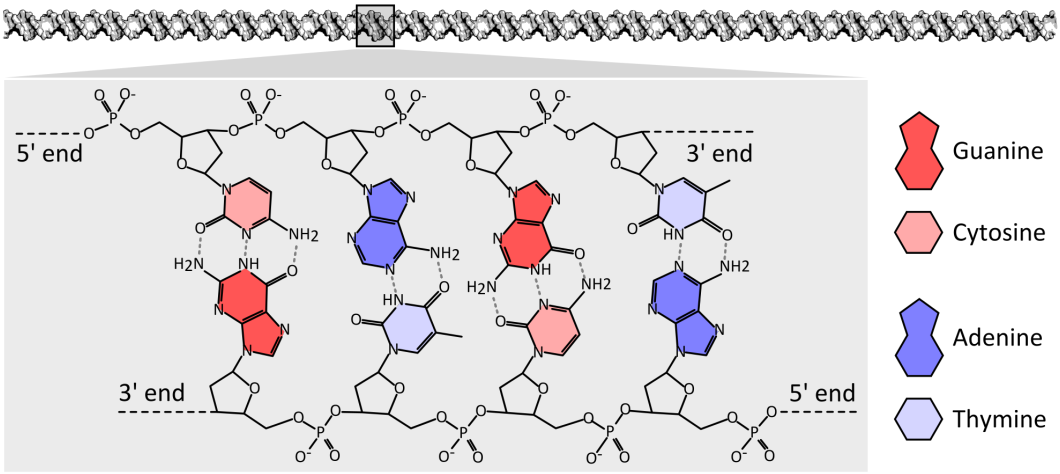


Fig. 1.1 Chemical structure of DNA. DNA is constructed from two long polymer chains wrapped around one another to form the signature double helix shape shown above in grey. Each DNA strand is constructed from four different sugar molecules called bases which are coloured differently in the diagram. The bases are joined via a covalent bond between the hydroxyl group of one and the phosphate of the next, resulting in a long polymer chain with a sugar-phosphate backbone. As such, there exists an unused phosphate at one end of the strand and an unused hydroxyl group at the other. These termini are referred to as the 5' and 3' ends respectively. The two strands of DNA are bound together by hydrogen bonds between their respective bases which only form when particular bases are aligned. Specifically, adenine pairs with thymine to create two hydrogen bonds, and guanine pairs with cytosine to create three hydrogen bonds.
Diagram repurposed from that of Thomas Shafee (Wikicommons, License: CC BY 4.0)

even though the chains are not physically identical to each other, the information encoded within one strand of the DNA can be used to infer the information contained within the other.

In addition to the paired DNA chains being physically distinct in terms of the sequence of nucleotides, the nucleotides are bonded together in a manner which invokes directionally [15, 16]. Shown in Figure 1.1, the covalent bond between neighbouring bases occurs between the phosphate group on the 5' carbon on one base and the hydroxyl group on the 3' carbon of the next. The labels, read as x primed, where x is the number, refer to specific atoms within each nucleotide. As such, at one end of a DNA strand, there exists an unused phosphate group, while at other there is an available hydroxyl group. Given this, the ends of a DNA strand are called 5' and 3' in reference to the available phosphate and hydroxyl groups respectively.

1.2.2 RNA

From the exact sequence of nucleotides in DNA, RNA, which is also a series of nucleotides, is specified [11, 12]. However, while DNA is constructed from A, G, C, and T, RNA is built from A, G, C, and uracil (U) [10]. Notably, RNA uses the ribose versions of these bases over the deoxyribose version used in DNA, as the ribose versions are less stable [17]. A lower stability is required for RNA as it is a transient molecule that the cell needs to be able to decompose readily once it is no longer needed [18]. In contrast, DNA contains the genetic information, so ideally should be as stable as possible. The switch from T to U is relatively small as the methylation of U produces T. As such the two nucleotides are closely related in chemical space, so much so that A now pairs with U for determining base pairing within RNA in an analogous manner to the A-T pairing found in DNA [10]. In fact, the nucleotides U and T are often used interchangeably when stating an RNA sequence due to both simplicity and their similarity. In addition, to the switch of U to T, over 100 other non-standard nucleotides can be found within various RNA molecules, these allowing for greater diversity in the various functions RNA can take on within the cell [19]. These non-standard nucleotides are exclusively the product of naturally occurring modifications made to the canonical nucleotides (A, G, C, and U) after the RNA has been produced.

A further difference to DNA is that RNA is a single stranded molecule, which means that it does not readily form a double helix structure due to lacking a complementary strand. Instead, the bases on the single RNA chain interact with each other using a similar base pairing schema outlined for DNA above. Such interactions cause sections of the RNA chain to pinch and align in various places which produce an overall structure of loops, bulges and double helices to emerge [20]. Lastly, DNA and RNA chains differ significantly in their lengths; while the former can be millions or even billions of bases long, RNA rarely reaches into the thousands of bases [21]. This shortening is primarily caused by the drop in the stability of ribose bases compared to their deoxyribose counterparts mentioned above. As such, only certain fragments

of the genome specify RNA. The regions that do encode for RNA are easily identified via the presence of a promoter region, which is a small sequence that causes RNA polymerase, the large protein-based machine that performs transcription, to bind near the location [22]. Other regions, those that do not contain information pertaining to RNA, can account for up to 90% of the genome [23]. It is still hotly debated as to this region's role within the cell, some referring to it as junk DNA, while others attribute it a plethora of biological functions [24, 25].

1.2.3 Transcription

As mentioned above, regions that encode RNA are most easily identified by nearby promoter regions as shown in Figure 1.2, which gives an overview of translation and transcription. These promoter regions are found upstream of the section to be transcribed, which refers to the fact it is nearer the 3' end of the DNA [22]. The 3' end is considered upstream as the RNA polymerase, a large molecular machine that facilitates transcription, reads the DNA in the 3' to 5' direction [10]. As the polymerase only reads one of the DNA strands, referred to as the coding strand, it splits the helix apart locally to form a bubble which it then moves along the DNA during transcription. The DNA strand not being read by the polymerase is referred to as the template strand. The RNA polymerase creates RNA from the DNA via the same base pairing rules that specify the interaction between the paired DNA strands, though with the T-U substitution mentioned above. Each nucleotide in the DNA specifies a nucleotide in the produced RNA. Note that this means that the RNA created has an equivalent sequence to that of the analogous section of the template strand, bar the T-U substitution. This also means the RNA is produced by the polymerase in the opposite direction to the coding strand, i.e., starting with the 5' end. In a similar manner to the promotor region defining the beginning of a region to be transcribed, the RNA polymerase continues transcribing until it encounters a specific sequence, referred to as the terminator, at which point transcription ceases, and the polymerase detaches [26].

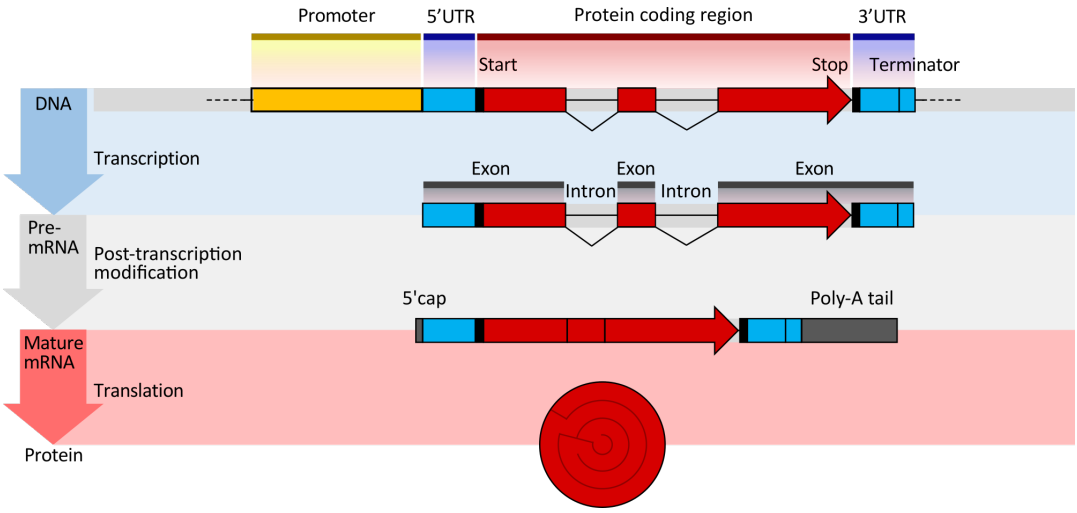


Fig. 1.2 Transcription and translation of protein coding genes within eukaryotes. The information to encode a protein originates within the DNA, shown at the top running in the 5' (left) to 3' direction. Alongside the protein coding region (red), which is broken into exons and introns, there is also the UTR regions (blue) and the signalling region (yellow). The signalling region indicates the section of DNA to be transcribed into mRNA, namely the protein coding region flanked by the two UTR regions. Once transcribed, the mRNA undergoes a series of post-transcriptional modifications, capping, polyadenylation, and splicing. At this point, the mRNA is mature and can undergo translation to produce a protein.
Diagram repurposed from that of Thomas Shafee (Wikicommons, License: CC BY 4.0)

1.2.4 Post-transcriptional modifications

Not all RNA, once transcribed are then used to define a protein; in fact, the vast majority do not [27]. For example, microRNA (miRNA) is used by cells to signal changes gene expression [28]. Other RNA are used as a unit in the construction of various biomolecules, most notably, ribosomes out of ribosomal RNA (rRNA) and the amino acid carriers using transfer RNA (tRNA). Those that are destined to encode proteins are referred to as messenger RNA (mRNA) or transcripts. In prokaryotic systems, a domain of life which features mostly single-celled bacteria, these mRNA can be used to specify proteins immediately [10]. However, in eukaryotic systems, which is a domain of life with more complex organisms, such as *H. sapiens*, before mRNA is useable, its antecedent, precursor messenger RNA (pre-mRNA), must undergo a series of post-transcriptional modifications [10]. These modifications can change the sequence

and, hence, the specific protein encoded may be altered before the mRNA becomes functional and mature.

The first modification is a cap added to the 5' end of the pre-mRNA to stop ribonuclease, a protein that digests RNA, from degrading it [29]. Similarly, at the 3' end, between 200 and 300 adenine bases are added and referred to as the poly(A) tail [29]. This tail also helps prevent digestion from ribonuclease. The most notable and impactful modification, however, is splicing [30]. Splicing involves removing various sections of the pre-mRNA called introns, with the remaining sections referred to as called exons. The combined exons, which connected after splicing removes the intervening introns, define the protein sequence that will be produced. Notable, exons can also be deleted in this process, which means that an individual pre-mRNA can lead to multiple different mRNAs, and hence multiple different protein sequences, depending on the exact series of exons remaining post-splicing. The removal of even a single exon can drastically change the function of the resultant protein [31]. Once all post-transcriptional modifications have occurred, the mRNA consists of the following from the 5' to 3' end: the 5' cap, a section of non-coding RNA upstream of the coding called the 5' untranslated region (5' UTR), the coding region which contains the information specifying the protein to be produced, another section of RNA downstream of the coding region called the 3' untranslated region (3' UTR), and, finally, the poly(A) tail. All these segments are clearly shown in Figure 1.2. The two UTR regions which surround the coding region are used primarily for regulating the expression of the protein encoded by a given mRNA [32].

1.3 Translation

Translation is the biophysical process by which a protein is constructed using the information contained within the mRNA sequence. Specifically, the bases within the mRNA are read in groups of three called codons, and each codon specifies a particular amino acid, which are also known as residues, within the protein produced. For example, the codon GGA

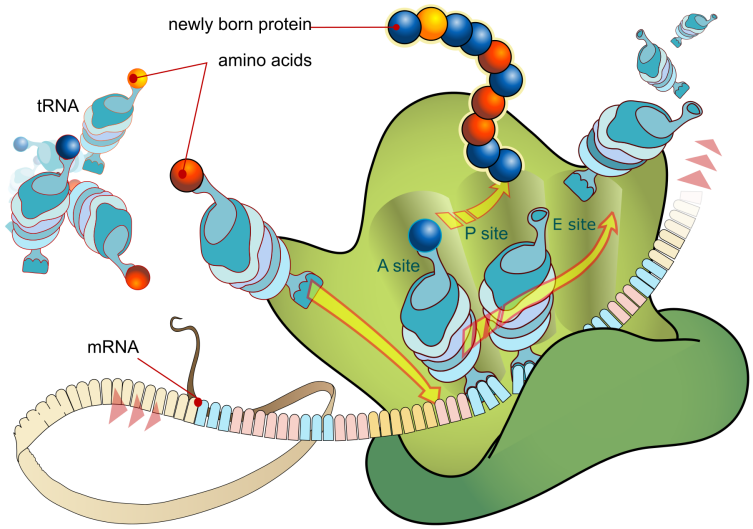


Fig. 1.3 Biophysical process of translation. The ribosome is a large complex molecular machine that reads the mRNA in groups of three called codons. The codon located at the A-site of the ribosome at any given time specifies the next amino acid to be incorporated into the protein undergoing construction. Once the tRNA carrying the specified amino acid associates with the ribosome, the nascent chain is transferred onto the new tRNA to incorporate the residue. At this point, the ribosome moves forward a single codon simultaneously to the tRNAs contained within it being manoeuvred between the ribosome sites. The tRNA with the growing protein attached is now once again located at the P-site, while the spent deacylated tRNA is at the E-site. The E-site is open to the cellular pool, so the deacylated tRNA is now free to exit the ribosome and be recycled.

designates the amino acid glycine. The translation process is facilitated by the ribosome, a large macromolecular machine constructed from both rRNA and various proteins. The structure of the ribosome has been measured by different methods but broadly described the complex consists of two large fragments of different sizes referred to as the large and small subunits [33]. There are slight variations in these subunits between eukaryotes and prokaryotes, the former containing a higher quantity of protein as well as being roughly 25% larger [34]. However, the core structure and mechanism by which the ribosome functions remains consistent across most forms of life. Below we describe in detail the physical process of translation which is also shown in Figure 1.3.

1.3.1 Initiation

To begin translating, but before attaching to the mRNA, the small ribosomal subunit must bind with a host of initiation factors and proteins, as well as a specialised initiation tRNA [35, 36]. This initiation tRNA carries a single methionine, the starting amino acid of nearly all proteins, and differs in both structure and sequence to the tRNA used to supply methionine at other parts of the protein sequence. Once the small subunit complex forms, it binds to the mRNA in the 5' UTR region, where it then scans along the transcript until it reaches the start codon. Once there, it then waits for the association of the larger ribosomal unit to form the complete ribosome. Note that the presence of a start codon alone is not sufficient within bacteria, where nearby specific sequences, such as the Shine-Dalgarno sequence in *E. coli* or transcription factors within the 5' UTR are also required for translation to occur [35, 36]. Once the small and larger ribosomal units have combined, the ribosome's structure contains inside three codon-width pockets side-by-side, referred to as the A, P, and E-sites. The P-site resides above the start codon, with the A and E-sites on the neighbouring downstream and upstream codons respectively. Inside the P-site sits the initiation tRNA with its methionine still attached. This amino acid forms the base for the growing protein chain, which is referred to as the nascent chain until complete.

1.3.2 Elongation

The downstream A-site resides above the next codon in the RNA sequence to be translated [37]. At initiation, this is the codon immediately after the start codon. The ribosome must wait for the tRNA carrying the correct amino acid to associate with it before it can incorporate it into the nascent chain. The various tRNA molecules have attached a specific amino acid at one end and contain the corresponding anticodon sequence at the other. The interaction strength of the anticodon pairing to the codon within the ribosome's A-site determines when the correct amino acid is present on a given tRNA. The strongest and, hence, most readily accepted pairings

are when the anticodon is the codon reverse-transcribed, e.g., GAU pairs best with AUC. A tRNA with the correct anticodon for a given codon is referred to as the cognate tRNA. Once the cognate tRNA is present within the A-site, the ribosome chemically bonds the nascent chain on the P-site tRNA, which at initiation consists solely of the methionine, to the amino acid. Once the nascent chain has transitioned between tRNAs, the ribosome steps one codon downstream while the codons are manoeuvred between the sites in a motion referred to as translocation. The P-site again now contains a tRNA with the nascent chain attached. The chain grows from a point adjacent to the A and P-site to the surface of the ribosome through a gap in the overall structure. This gap is referred to as the ribosome tunnel and can contain between 30 to 40 amino acids, the number varying depending on the compactness of the nascent chain [38]. At this point, the spent deacylated tRNA, which is a tRNA without an amino acid attached, resides in the E-site. The E-site is accessible to the cellular environment and, as such, the deacylated tRNA can exit the ribosome and be used again to carry another amino acid to a ribosome once replenished.

1.3.3 Termination

This process continues along the RNA until the ribosome reaches a stop codon [39]. Stop codons do not encode for an amino acid and, as such, no cognate tRNA exist to associate with the ribosome meaning that the ribosome will pause indefinitely at this codon. However, given sufficient delay at any given codon, the ribosome will cleave the nascent chain, which at the stop codon is complete, from the tRNA within the P-site. Once cleaved, the residues still contained within the ribosome tunnel can defuse out, and the protein can dissociate from the ribosome. Since translation is now complete, the ribosome subunits proceed to break apart and detach from the mRNA, to be used again elsewhere within the cell. Finally, it is important to note that the translation of a given strand of mRNA by a single ribosome does not impede

either the attachment or translation of the strand by others. As such, from a single strand of mRNA, many copies of the protein may be generated simultaneously [40].

1.4 Degeneracy

1.4.1 The genetic code

The protein defined by the coding region of the mRNA is deciphered via the genetic code, which is shown in Figure 1.4 [10]. The nucleotides within the mRNA are grouped into sets of three called codons, with each codon specifying an amino acid within the derived protein sequence. There are 64 possible combinations of three nucleotides (4^3), so consequently there are 64 different possible codons; however, there are only 20 possible amino acids that these codons can specify (though rare organisms may differ from 20). This disparity is the result of two factors. Firstly, three of the codons, namely UAG, UGA, and UAA, define the end of the coding region, and hence protein sequence, and cause translation to stop instead of specifying an amino acid. There is also a start codon, AUG, which defines where a protein begins within an mRNA chain, but, in contrast to the stop codons, this does encode for an amino acid, methionine in multi-cellular organisms and formylmethionine in bacteria [35, 36]. Secondly, and more notably, the genetic code is degenerate, which means that more than one codon may encode the same amino acid. Codons that differ but encode the same amino acid are called synonymous. Likewise, mutations of the genetic code that do not change the amino acid are referred to as synonymous mutations.

As shown in the Figure 1.4, the degeneracy is not evenly distributed, some amino acids are only encoded by a single codon, while other are up to six-fold degenerate. When an amino acid is degenerate, the redundancy is often due to allowed variation in the third nucleotide. For example, the amino acid alanine requires the first two bases to be GC, while the third nucleotide can be anything. For many others, the degeneracy allows for either pyridine (C/U)

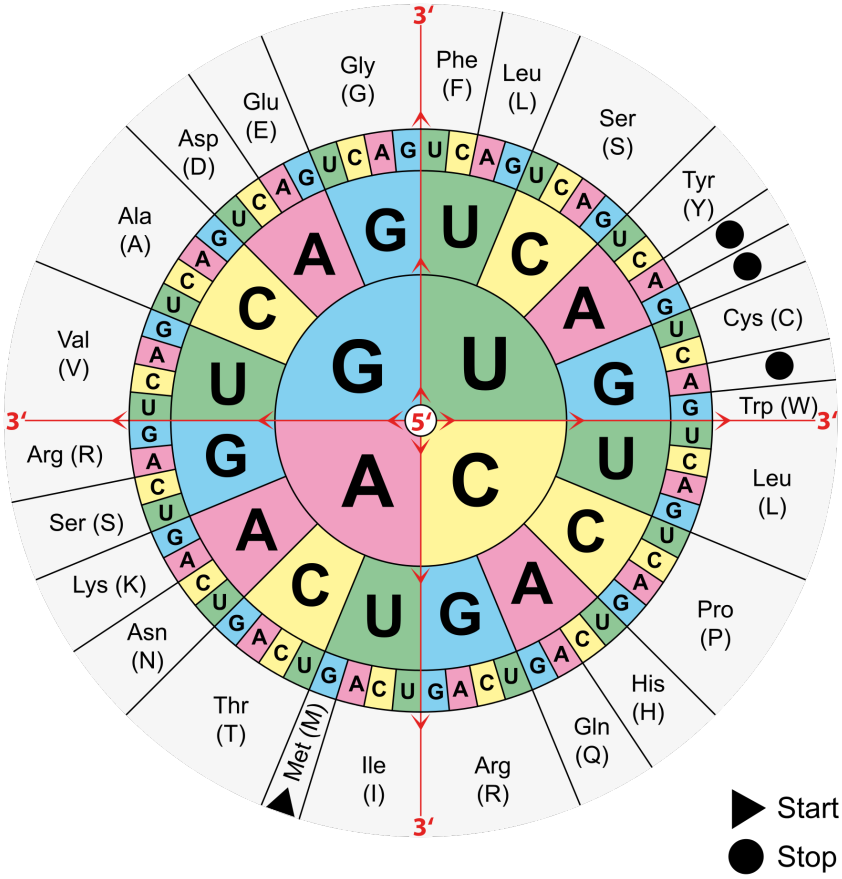


Fig. 1.4 The genetic code. The information within the mRNA sequence specifies a protein to be built via the genetic code. The nucleotides are grouped together into sets of three called codons, with each codon specifying an amino acid within the protein. In the diagram above, the code is deciphered by constructing a codon using a nucleotide from each of the rings, starting with the innermost ring (5') and moving outwards. The amino acid adjacent to the outermost selected nucleotide is that which would be incorporated into the protein given that codon. Dots are used to signify the stop codons that do not specify an amino acid and instead cause translation to cease. Likewise, an additional marker is used to highlight the start codon, which specifies methionine, as this signifies the start of the region to be translated within the mRNA. Note that methionine can be incorporated at other positions within the protein as well.

or either purine (A/G) to be the third nucleotide. As a result of this degeneracy, the central dogma, the flow of information between DNA, RNA and protein, is considered unidirectional between RNA and protein as many different mRNA sequences can give rise to the same protein sequence [12]. Conversely, given a protein sequence, one can not specify the exact series of codons in the mRNA that when translated led to the protein's creation.

1.4.2 Wobble base pairing

On top of the degeneracy built into the genetic codon, whereby multiple codons can encode for the same amino acid, the association of tRNA to the correct codon at the A-site in the ribosome is also degenerate. A given codon may pair successfully with multiple different anticodons (i.e., many different tRNAs), though akin to the degeneracy within the genetic code, this is typically restricted to certain variations of the third nucleotide [41, 42]. For example, in *S. cerevisiae*, the codon UUU is supplied by a tRNA with the anticodon GAA as the there exist no tRNA with the cognate anticodon AAA [43]. This variable matching is allowed due to space within the ribosome around the third nucleotide position allowing for small conformational adjustments. These adjustments, or wobbles, lead to the phenomena being referred to as wobble base pairing. While most wobble base pairing will result in the same amino acid due to the degeneracy of the genetic code, it is possible for wobble base pairing to result in a point mutation to the protein sequence. However, this is uncommon as the propensity for the association of cognate tRNA is significantly higher [42, 44]. Regardless, the degeneracy in codon-anticodon matching is utilised fully within nature as only in a few species are all 61 possible tRNAs present, one for each codon-anticodon pairing. For example, *H. sapiens* has only 56 different tRNA types, while in *E. coli* only 40 different tRNA types are present. As such, this variableness is essential to the successful translation of proteins in many forms of life [43].

1.5 Protein structure

Once a protein has been successfully translated, but before it can facilitate its role, it must adopt its optimal configuration or structure. As mentioned earlier, the structure of a protein is of crucial importance as it often defines its function within a cell; as such, the determination of an unknown protein's three-dimensional shape can give insight into how, where and what it interacts with [7]. To experimentally determine a protein's structure is currently both costly and time-consuming, which contrasts starkly with our ability to determine both nucleotide and protein sequences, which is quick, easy, and cheap. For example, in 2016 alone, approximately two billion sequences were sequenced and deposited in GenBank, the largest online collection of nucleotide and protein sequences [45]. While the Protein Data Bank, which stores nearly all known protein structures, expanded by only 10,000 entries over the same period [46]. This stark difference has led to a growing gap of many orders of magnitude between the number of sequences and structures that are publicly available; our knowledge of sequence space is increasingly dwarfing that of structure space. We have created a scenario where we increasingly know more about protein sequences, but not their shape and, hence, their function.

1.5.1 Primary structure

The structure of a protein is traditionally described in a series of levels, namely primary structure, secondary structure, tertiary structure and quaternary structure [10]. The first of these, primary structure, is merely the amino acid sequence of the various chains that construct a given protein. The chains are typically given from the 5' end, that which is translated first, to the 3' end. Note that akin to DNA and RNA chains, proteins have directionality given by the manner in which the peptide bond forms between the carboxyl and amino groups of neighbouring residues. The bonding results in an unused amino acid group remaining at the 5' end, which can also be referred to as the N-terminus, and a unused carboxyl group remaining at the 3' end, which can also be referred to as the C-terminus. Lastly, the primary structure also contains any

post-translational modifications, such as phosphorylations and glycosylations, as these cannot be inferred directly from the mRNA sequence.

1.5.2 Secondary structure

In the next layer, the secondary structure, local areas within the protein which have regular structure are defined. Most notable examples of this are α-helices and β-strands, both of which have a defined geometry given by the dihedral angles between bonded amino acids as well as interactions with other residues. While the primary structure focuses only on residues that are covalently bonded, these sub-structures are formed primarily due to strong hydrogen bonds. Hydrogen bonds are formed due to the attraction between the partial positive charge on a hydrogen atom and the partial negative charges on either oxygen or nitrogen atom. Note that the hydrogen bonds form between non-neighbouring residues which may not be close regarding the sequence overall, as is often the case for β-strands. These interactions result in the chain pinching and aligning to form these sub-structures. The remaining protein, that which does not form part of a regular substructure, are referred to as coils and connect together the regions with defined secondary structure. Coil regions account for 45% of a globular protein on average [47].

1.5.3 Tertiary structure

The tertiary structure defines the overall three-dimensional shape, or fold, that a protein chain takes. For globular proteins, this describes how the secondary structure elements, such as α-helices and β-strands, are compressed together to form a compact spherical protein. This compactness is driven both by interactions between residues as well as the hydrophobicity of certain residue types. The former of these, interactions between residues, includes hydrogen bonds alongside occasional salt bridges and disulphide bonds. The later, the hydrophobicity, refers to non-polar amino acids having unfavourable interactions with the surrounding water

such that the protein adopts an overall shape to shield them from the medium. Not all proteins are driven by these forces, most notably, disordered proteins exist and are defined by their lack of tertiary structure.

The tertiary structure of one region of a protein chain can form independently of other regions. As such, a single chain can form many distinct globular units within its overall tertiary structure. These units are called domains and are joined by regions named linkers. The specific arrangement of secondary structure elements found within a single domain is rarely unique to a particular protein. In fact, if two proteins have very similar sequences over a region of 35 residues or more (upwards of 30% sequence identity), then those regions are likely to fold into homologous structures [48]. Note that similarity in the sequence leading to a similarity in structure is not absolute, as even the change of a single amino acid can drastically alter the structure of the protein produced [49, 50].

1.6 Quaternary structure

The final level, the quaternary structure, refers to how multiple folded protein chains fit together to create larger complexes. For example, many proteins are dimers, formed from two protein chains with the highly similar sequences and structures.

1.7 Thermodynamic hypothesis

Having defined the various levels with which protein structure is described above, below we discuss how the protein shape forms. The thermodynamic hypothesis, which is also known as Anfinsen's dogma, was first stated in the 1970s and postulates that a protein will always fold to the most energetically favourable state given a suitable local environment (e.g., pH, temperature) [8]. It gained credence due to work by Anfinsen on the protein ribonuclease A, whereby he showed that the protein would reversibly denature (unfold) with the addition of

Urea and 2-mercaptoethanol, which combine to change the acidity of the medium and break any disulphide bonds [51]. The subsequent removal of these agents results in the protein refolding to a state of full biological activity and, as such, he showed that by simply varying the environment that the protein was placed in, the folded state, natured or denatured, could be controlled. Such behaviour would indicate that the amino acid sequence alone determines the final three dimension shape that a protein takes as only the interactions between these residues define the energetics at play within Anfinsen's experiment. Notably, the results of these experiments and the hypothesis drawn from them are closely related to the central dogma. Both suggest that all the information needed for the protein to fold is fully encoded within the amino acid sequence and, as such, there is no dependence on the exact sequence of the mRNA that produced it.

1.7.1 Levinthal's paradox

One issue that arises from the thermodynamic hypothesis is Levinthal's paradox, which is a thought experiment in which one calculates an approximate time taken for a protein to find its most energetically favourable state [52]. The calculation involves combining the number of possible configurations a protein can take with an estimate of the time spent to explore a single configuration. The number of configurations is approximated by noting that the position of adjacent residues can be described with respect to one another using only two angles [53]. If one approximates that these angles have only three stable configurations, the total number of configurations of an entire protein is given as 3^{N-1}, where N is the number of residues. The subtraction is due to their being $N-1$ bonds whose configuration need specifying between N residues. The time taken to explore one of these configurations is typically taken as being on the order of a nanosecond, which is the fastest time scale a distinct folding event has been observed [54]. Given this, for a 100 residue protein, the time that would be taken to explore all configurations is longer than the existence of the known universe

($3^{100-1} \times 10^{-9}$seconds $\approx 5 \times 10^{38}$seconds $> 13.8 \times 10^{9}$seconds) [55]. This paradox holds true even if one estimates the sampling time to be on the order of pico or femtoseconds, which is comparable to the frequency of visible light [56].

1.7.2 Resolving the paradox

The conclusion one draws from the Levinthal's paradox is that the majority of states must not be sampled when a protein folds. This inference is collaborated further by the existence of ultrafast folding proteins which can go from a denatured state to their globular configurations in mere microseconds [57]. Protein folding, therefore, must be guided such that the folded state is reached efficiently and without major detours. A simple conjecture is that local interactions, such as those between adjacent residues, both determine a significant portion of the final structure and rapidly form due to their proximity. This conjecture is likely true, at least in part, given that we can predict secondary structure to a high level of accuracy given just the amino-acid sequence of a protein [58]. Levinthal himself suggested that these local structure formations can then act as nucleation points around which the rest of the protein will then quickly fold [59]. This solution to the paradox in which each structure formation iteratively leads to further structure forming until the native state is reached is referred to as the folding funnel hypothesis [60]. The energetics of this are represented in Figure 1.5 in which each intermediate state has slightly lower energy, until the optimal fold, which is at has the lowest energy, is reached.

The issue with the above hypothesis is that it assumes that the protein starts unfolded and progressively folds through multiple transition states until reaching its final configuration, the native state. This viewpoint does not reflect the actuality of protein production within a cell, whereby a protein is translated sequentially, passed residue-by-residue through the ribosome tunnel, which stops most structure formation from occurring, and then emerges into the cellular pool [61]. In fact, the additive process of translation occurs at an approximate rate of five

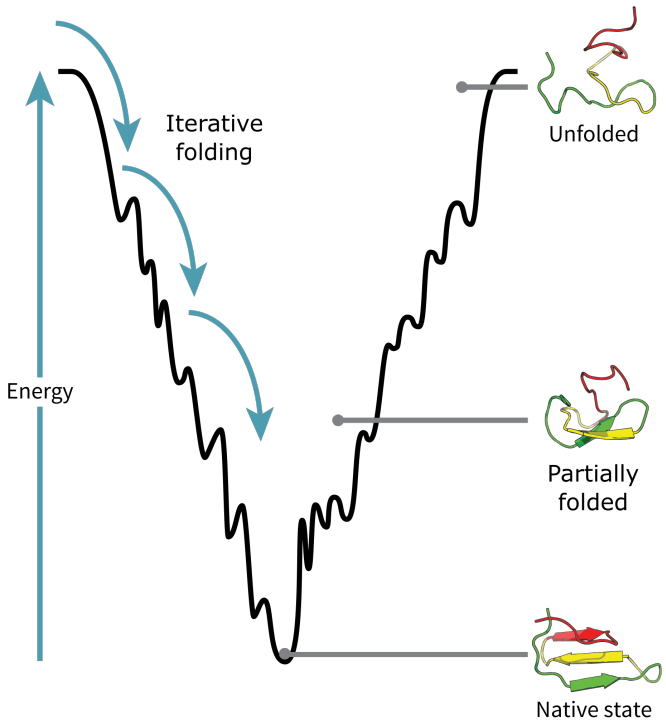


Fig. 1.5 The folding funnel hypothesis. The above diagram depicts the energetics of a protein folding in the manner described by the folding funnel hypothesis. The lower the energy of the protein's configuration, the more stable it is, hence a protein seeks to be in the lowest energy state, located at the bottom of the funnel. To reach the native state from the unfolded state at the top, the protein iteratively jumps between lower and lower energy states via small transitions in its configuration. With each transition, the structure becomes more and more energetically favourable, and the overall shape becomes more and more defined. Eventually, the base of the funnel is reached, at which point the protein has achieved its native conformation.

residues per second in eukaryotes and twenty residues per second in prokaryotes [62, 63]. These rates are comparable to that with which it has been observed that most secondary structure elements of proteins form (milliseconds), and is far faster, in general, than the folding of the entire protein chain (up to minutes) [64, 65]. Given this, it is unlikely that the protein would wait to fold until the whole chain is both translated and free from the ribosome. Furthermore, even if the chain was fully translated and separate before folding occurred, the cellular pool is overcrowded, filled with many other biomolecules that would interact with the emerging nascent chain while it is still under production [66]. These interactions could change the folding behaviour, likely in a deleterious manner that may even cause misfolding to occur. The folding

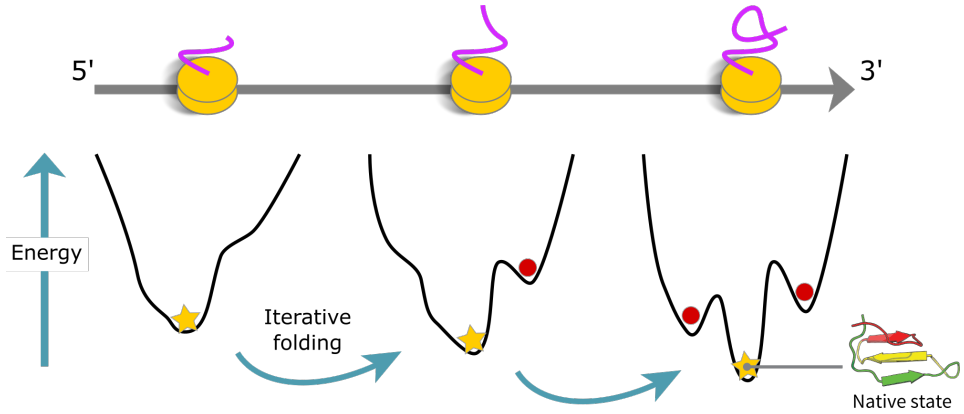


Fig. 1.6 Cotranslational protein folding. Shown are the energetics of a protein taking on its three-dimensional structure as it is still undergoing translation by the ribosome (yellow). Initially, when only a small fragment of the protein has been translated, the energy surface is simple and the optimal configuration, which is indicated by a star, can be located quickly. As more of the protein is produced, the surface becomes increasingly complex. However, the new optimal configuration is highly similar to the previous optimal configuration, that of the shorter chain, and, as such, can be reached by a small transition in the configuration that incorporates the new residues. Additionally, these small transitions between optimal states on more and more complex energy surfaces helps the growing protein avoid folding into wrong configurations that are highlighted by the large dots.

funnel hypothesis, as it is stated above, does not accommodate the practical nature of protein folding within a cell [67].

1.8 Cotranslational protein folding

From comparing the rate of translation to that of protein folding above, we note that proteins can fold, at least locally, faster than they are produced. Thus, it is reasonable to conjecture that a protein's secondary structure is formed cotranslationally, i.e., in tandem with the translation of the protein. If in fact, the whole protein chain folds cotranslationally, it avoids the flaws associated with the folding funnel hypothesis we noted above. Namely, it must no longer wait unfolded until the whole sequence is translated, which results in a lower likelihood that other biomolecules can interact in a deleterious manner with the nascent chain [68, 69]. We show a representation of cotranslational folding in Figure 1.6; the nascent chain emerges from the

ribosome, gradually folding via transitions into iteratively lower energy states that incorporate more and more of the partial chain. When the complete chain is full translated and free from the ribosome, the optimal fold can be reached via a short transition, thereby forgoing the need for sampling the entire computational space and resolving Leventhal's paradox. Also, the guided folding pathway provided by folding cotranslationally may help the protein avoid folding into non-native configurations, which can be considered as local minima on the complex energy surface available to the complete protein [70].

1.8.1 Experimental evidence

Various experimental evidence exists that shows support for the cotranslational folding hypothesis. Firstly, biological activity has been reported in various proteins that are still undergoing translation. For example, Kiho and Rich found that β-galactosidase is enzymatic while still bound to the ribosome [71]. Likewise, a protease sourced from the Semliki Forest capsid cleaves a precursor section from itself while still undergoing translation [72]. Lastly, various mutants of the protein rhodanese that were extended at the C-terminus showed that with a minimum extension of 23 residues, the protein could still be enzymatically active while undergoing translation [73]. Further evidence is based on the comparison of the folding pathway taken by proteins *in vivo* to the pathway taken *in vitro*. For example, the folding of firefly luciferase after natural synthesis differs in both the pathway that was taken and the rate of folding, which was far faster, than refolding of the protein from a denatured state [74]. Similarly, the bacterial luciferase β subunit was shown to reach its final state far faster undergoing natural synthesis than folding *in vivo* [75].

More notable is that cotranslational folding has even been actively observed in translating ribosomes using nuclear magnetic resonance (NMR) spectroscopy, cryoelectron microscopy (cryo-EM), and Forster resonance energy transfer (FRET). Using NMR, ribosomes stalled at various points along the translation of an Ig-like domain showed that partially folded

intermediates formed out of the exposed nascent chain [76]. Similarly, in studies of stalled *E. coli* ribosomes using cryo-EM, rudimentary globular structures were found within the widest section of the ribosome tunnel, which is located just prior to the exit, formed from the nascent chain that was still undergoing translation [33]. Even a whole zinc-finger domain, albeit a relatively small one, was shown to fold cotranslationally within the ribosome tunnel using cryo-EM [69]. In more recent years, FRET has allowed cotranslational to be monitored real-time, forgoing the use of stalled ribosomes. For example, HemK was shown to adopt a series of compact non-native transition states as it was being translated until the full chain emerged from the ribosome, at which point it quickly adopted the native state [77]. Similarly, the cystic fibrosis transmembrane conductance regulator was observed to fold sequentially using FRET, the subdomains discretely forming one after another upon emerging from the ribosome tunnel [78].

1.8.2 Computational evidence

In addition to the experimental observations, computational models also provide support for the cotranslational folding hypothesis. Simple lattice based models incorporating stepwise protein production found that the inclusion of cotranslational folding led to the optimal confirmation being reached far faster [79]. Furthermore, the model showed that cotranslational folding could bias the final configuration to favour one native state over another when they are of similar energy. In another study, a Go-model of a protein, in which each residue was represented by a single point and only the interactions between neighbouring residues in the experimental protein structure were considered, was folded cotranslationally from the N-terminus to the C-terminus [80]. The authors found that folding pathway could differ significantly to that taken when the full chain was folded simultaneously. Lastly, Ellis *et al.* found that in 94% of cases, building a protein structure model from the N-terminus to the C-terminus produced a more accurate prediction than the when modelling progressed in the reverse sense direction [81].

1.9 Modulation of the translation speed

Given all of the above evidence, it is reasonable to conclude that folding does, in fact, occur cotranslationally as the nascent chain emerges from the cell. Given this, one can then ask whether the manner in which the protein is produced influences the eventual product. If folding no longer occurs separately to the construction of the polypeptide, then any variation in the environment, whether that be from the coding mRNA sequence, the ribosomal RNA sequence, or the surrounding cellular environment, which may or may not be blocked from interacting with the nascent chain by a chaperone protein, may lead to changes in the protein produced [82]. In this research, we focused on the influence that the translation speed along the mRNA may have on the generated protein.

For the translation speed to affect the protein produced, it must vary along the length of a transcript as the ribosome progresses. Hence, different codons on a given transcript would have to differ in their respective translation rates. In fact, this has been measured and found true in a variety of species using experiments focussing on the translation speed along a specific gene [83–89]. For example, the relative translation rates of codons placed at the location prone to a frameshift, which is when the ribosome jumps forward a single nucleotide, were determined by quantifying the occurrence of the frameshift to non-frameshifted products; the frameshift found to occur proportionally to the time spent at the frameshift site [84]. This experiment, as well as the others, found that different codons on the same transcript could naturally vary in translation speed by up to an order of magnitude.

1.9.1 Codon usage bias

Further support for variation in the translation speed is given by analysing the usage of synonymous codons. Even though they produce the same amino acid, synonymous codons are not used in equal measure across a genome [90]. Instead, analysis of the regions in genomes that encode for proteins has shown that there exists a preference for the use of certain codons

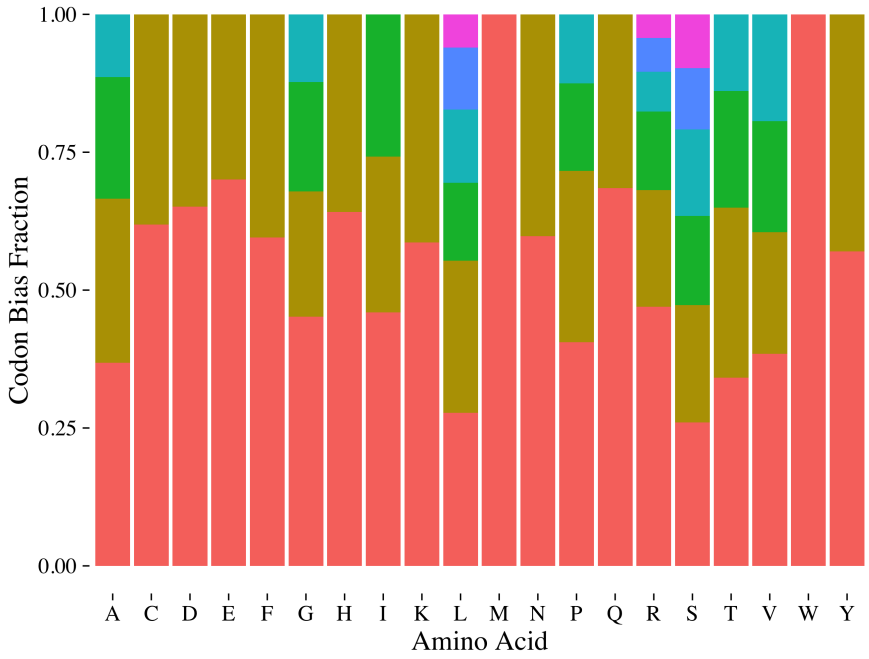


Fig. 1.7 Codon bias within the protein coding regions of *S. cerevisiae*. The codon bias of *S. cerevisiae* was calculated via analysing all protein coding regions of the genome as given by the Ensembl annotation R64-1-1. Along the x-axis, the twenty different amino acids are shown. The corresponding bar for each amino acid gives the relative fraction of the constituent codons that resulted in that amino acid. Note that the bars are set to total 100 percent and that the amino acids do not appear in equal numbers.

over other synonymous codons and that these preferences differ between species [91]. The preference for the use of one codon over another is referred to as the codon usage bias. As an example of how pronounced this bias can be, we show the preference of usage within each amino acid for *S. cerevisiae* in Figure 1.7. We see that for arginine, one codon is roughly eleven times more prevalent than that least used, while the smallest codon bias is found for tyrosine, which is encoded only by two synonymous codons, where one codon is only 30% more prevalent. The codon usage bias becomes even more pronounced if one only considers the most highly expressed genes [92, 93]. This suggests that codon bias may be caused by a selection pressure as these genes are translated more, which is reasonable if one hypothesises that the cell would wish to optimise translation to produce proteins as efficiently as possible [94]. Ergo, synonymous codons are likely translated at the different rates on average with

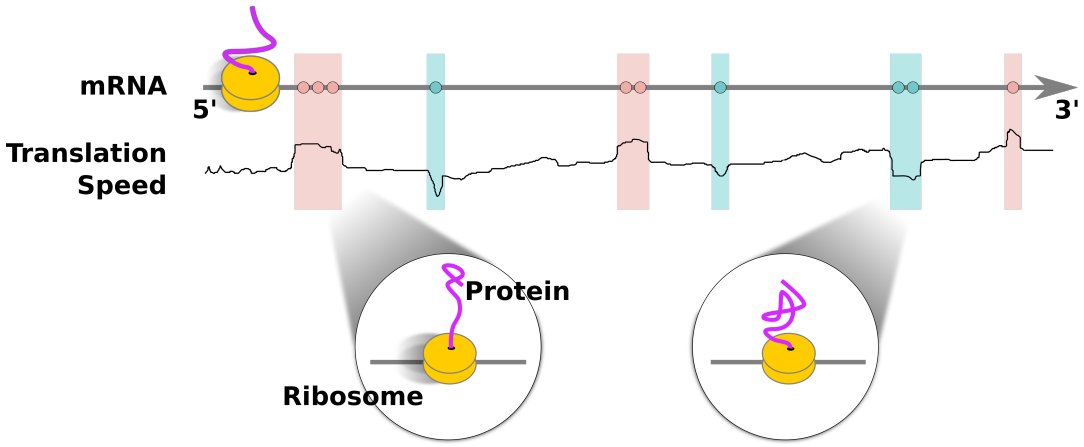


Fig. 1.8 How translation speed can affect the protein structure produced. The ribosome (yellow) passes along the mRNA, iteratively converting the sequence of codons into a protein. The translation speed for each codon differs, there existing both rapidly and slowly translated codons, these highlighted in red and blue respectively. Changes in translational speed via the use of these extreme codons leads to variation in the time available to the nascent peptide chain (purple) to explore the fold space. Rapidly translated codons cause a quick elongation, while slowly translated codons allow more time to search the fold space. Under such a scheme, the choice of codon, fast or slow, can act as an additional source of structural information.

those used more likely to be translated faster. This conclusion is further collaborated by noting the tRNA pool contains higher concentrations of those cognate to the most common codons [95, 96].

1.9.2 Relating translation speed to protein structure

Combining the variance in translational speed with the phenomenon of cotranslational folding gives the scenario outlined in Figure 1.8. Quickly translated codons cause fast elongation of the nascent protein chain reducing the time with which it can search the fold space. Conversely, slowly translated codons delay translation giving additional time to fold. In such a scheme, synonymous codons could produce the reported differences in proteins by changing the folding pathway and potentially the resultant protein structure. Notably, this can occur without altering the encoded sequence of amino acids.

Further, support for our hypothesis is garnered from experimental evidence which shows that the modulation of the translation kinetics can affect the protein produced. For example, synonymous mutations of a consecutive group of rarely used codons into those often used, resulted in a reduction in the ribosome pausing that occurred during translation of the CAT gene and caused a 20% decrease in activity [97]. Likewise, a study found that a single synonymous point mutation of a common codon into a rare codon within the MDR1 gene caused the protein to bind to a different substrate [98]. Lastly, Sander *et al.* constructed a *gedanken* protein consisting of three sequentially linked protein half-domains, A, B, and C, in which the middle half-domain (B) could bind to either the first (A) or last (C) half-domain, but not both, to create mutually exclusively proteins, AB or BC, which could be identified by their differing fluorescence signatures [88]. They found that modifying the linker region between the half-domains B and C domain, via synonymous switches, would alter the ratio of the two protein products formed as it varied the time available for AB to form, before having to compete with the formation of BC.

1.10 Summary

Above we have introduced the manner in which a cell produces a protein, regarding both the degenerate genetic code, which specifies the exact sequence of amino acids that bond together, and the elaborate machinery that reads the mRNA and facilitates the construction. Next, we defined the various levels at which the emergent protein's structure is described and discussed how the one-dimensional polypeptide chain folds into the three-dimensional biologically active shape. We showed that the manner in which a protein is produced could have an effect on the protein product, namely via the process of cotranslational folding, which may have a dependence on the translation speed of the encoding codon. Determining how to infer the translation speed and then ascertaining its relationship to the protein structure is the topic of the work enclosed in this thesis.

Chapter 2

Predicting the translation speed and its relationship to the protein structure

2.1 Introduction

From the evidence outlined in the introductory chapter, it is apparent that synonymous codons are not equivalent to one another. We discussed experimental evidence that showed how the synonymous switch of even a single codon could change the physical properties of the protein produced. We also suggested that this phenomenon might be due to variation in the translation speed between synonymous codons, which combined with cotranslational folding, could produce substantial changes in the protein structure without altering the encoding sequence. The research presented in this chapter focusses on our initial investigation, undertaken in the first year of doctoral study, into this hypothesis. It details how we mapped various theoretical estimates of the translation speed to experimental protein structures to unearth and quantify any speed-structure relationships. At the time of this work, no experimental transcriptome-wide measurement of the translation speed was readily available. As such, the research presented here, as well as that of previous investigations, could only be undertaken using these predictions

of the translation speed. In the following chapters, we will discuss ribo-seq, an innovative technique that we believe to be an experimental measure of the translation speed.

Translation is known to be one of the most energetically costly processes that a cell can undertake. An estimated 30% of the total available energy within the cellular pool is consumed performing protein synthesis alone [99]. As such, it is reasonable to assume that translation is optimised to be as efficient an undertaking as possible. Furthermore, the faster translation occurs, the sooner a given ribosome will complete translation, and as such, optimising the process would help avoid potential ribosome shortages [100]. Based on this assumption, a myriad of theoretical estimators of the translation speed have been constructed. Most estimators can be broadly classified into two types, namely, those based on the biases found within the genomes and those based on the cognate tRNA concentrations [101, 102]. The former normally make use of the codon bias, whereby the relative usage of codons within a synonymous set differ. It is assumed that the more common a codon is within the genome, the faster it is translated; the genome presumed to be optimised in a manner that produces proteins, on average, the quickest. Hence, the codons most used are those translated rapidly, while those used least may delay translation. Both the Frequency of optimal codons (Fop) and the Codon Adaption Index (CAI), two of the oldest estimators, are based on this concept [103, 104]. Measures based on the cognate tRNA concentrations assume that the rate-limiting step in translation is the association of the cognate tRNA with the ribosome. The higher the concentration of a specific tRNA, the quicker on average that the ribosome and the tRNA will associate, and hence the faster the corresponding codons will be translated. Notably, the relative concentrations are compared between all codons, rather than just those in a synonymous set. The tRNA adaptation index (tAI) is the most popular algorithm to use this approach. These two viewpoints are not orthogonal to each other; the cognate tRNA concentration and the codon bias having been found to correlate strongly to one another in various species thereby suggesting they are intrinsically linked [84, 93, 96]. In fact, the normalised Translational Efficiency (nTE) scale, a relatively

new estimator of the translation speed, combines the cognate tRNA concentrations and the codon bias, arguing that both the supply and demand must be considered [105].

Using these various theoretical estimators of translation speed, bioinformaticians have found correlations between the speed with which a given codon is translated and the protein structure it encodes. For example, it has been found with various estimators that residues contained within the protein core have a propensity for being encoded by the codons translated fastest [105–108]. The authors hypothesise that the faster a codon is translated, the higher its fidelity, and hence the protein core, which is more sensitive to mutations than the periphery, has a bias towards faster translation. Similarly, multiple works have reported that the initial codons of protein coding transcripts are translated slower than the remaining sequence [105, 109, 110]. A given explanation for this observation is that by retarding translation near the start codon, the mRNA helps prevent sequential ribosomes coming into contact during translation [109]. While these studies agree on the presence of the ramp, they disagree on the length over which the speed increase occurs. Furthermore, a more recent investigation suggests that the ramp is only present in secretory proteins to provide additional time for a polysome to be transported to the endoplasmic reticulum [111].

Disagreement between the studies is in fact relatively common in the literature and often coincides when their particular choice of theoretical estimators differs. For instance, domain boundaries have been stated as being both enriched and depleted by both quickly and slowly codons across a pair of studies which differed in their estimator choice [112, 113]. Similarly, there exists much disagreement as to the speed preferences of the various secondary structure types [105, 113–115]. For example, using nTE it was found that α-helices are depleted of quickly translated codons, while another study, which used CAI, reported that they are enriched with them [105, 115]. In general, nearly all inference based on these translation speed estimators seems to be highly dependent on one's selection of translation speed predictor.

Given that these contradictions in the literature appear to be a by-product of estimator choice, comprehending the underlying differences and compiling their respective inferred speed-structure biases would be of use. In this chapter, we compared four of the most well-known predictive measures of translation speed, CAI, tAI, the MinMax algorithm and nTE [104, 105, 116, 117]. We performed this comparison across ten diverse species, analysing the estimated speeds in terms of both the overall distributions and explicit comparison of coding transcripts. Next, we computed various speed to structure relationships inferred using each of the estimators across our set of species. We did this by overlaying each of the estimated speeds onto experimentally determined protein structures. Focusing on the speed to structure relationships reported in the literature, we collated the reported relationships to establish points of agreement and disagreement. In general, no clear consensus emerges between the predictive measures as to the genuine translation speed to protein structure biases, often the various measure giving contradictory results. Given this, we conclude that there is no evidence that any individual measure gives better predictions that the others. Moreover, it suggests that the speed-structure biases reported in the literature should be considered nuances of the respective algorithms unless collaborated by experimental evidence.

2.2 Materials and methods

2.2.1 JOY

JOY is a programme which calculates a one-dimensional sequence-like annotation of a protein's three-dimensional structure [118]. Each residue in the structure is annotated with its secondary structure, percentage accessibility, and any hydrogen bonds to other residues. The secondary structure annotation is calculated using SStruc, a variant of the DSSP algorithm [119, 120]. Four secondary structure annotations can be given, C, E, H, and P, which correspond to coils, β-strands, helices (α, 3_{10} and π), and residues with positive-ϕ angles respectively. The hydrogen

bonds are predicted using HBOND [118]. The percentage accessibility is calculated using the PSA programme [118]. Additionally, JOY also labels residues as either accessible or buried depending on whether their percentage accessibility is above or below 7% respectively. Throughout this thesis, we use JOY to extract one-dimensional structural annotations of proteins so that they may be mapped onto their respective coding mRNA sequences.

2.2.2 ClustalW

ClustalW calculates multiple sequence alignments based on the similarity between a set of given query sequences [121]. To create the alignment, ClustalW performs a pairwise alignment between pairs of sequences, which it then uses to infer a hierarchical clustering. Then, noting this clustering, the sequences are progressively aligned as a growing group to produce the multiple sequence alignment. ClustalW can also be used to perform simple pairwise alignment between two sequences. ClustalW has been superseded by Clustal Omega in recent years [122]. In this chapter, we used ClustalW to perform pairwise alignment of translated mRNA sequences to protein sequences. We used ClustalW over Clustal Omega primarily due to access to a python wrapper as well as the high similarity of our sequences (>90% sequence identity) meaning that a more advanced aligner was not required.

2.2.3 Protein Data Bank

The Protein Data Bank (PDB) contains three-dimensional structures of large biomolecules, most notable proteins, but also DNA, RNA, and whole viruses [46]. The database acts as a hub for structural biology, with nearly all publicly available protein structures contained within it. Throughout this thesis, we used the PDB to source experimental protein structures.

2.2.4 Structural Classification of Proteins - Extended

The Structural Classification of Proteins - Extended (SCOPe) database, and its predecessor the Structural Classification of Proteins (SCOP), are hierarchical classifications of the domains of protein structures within the PDB based on both sequence and structural homology [123]. There are four hierarchical levels of the classification, namely, and in order of increasing similarity, Class, Fold, Superfamily, and Family. Class is a broad classification of the overall secondary structure. Fold focuses on the number, arrangement and connection of secondary structure elements. At the superfamily level, domains are assumed to have a common, albeit distant, ancestor. At the lowest level, family, the sequence similarity is high and often domains in the same group will perform similar functions. In this thesis, we used SCOPe classifications to annotate protein structures with their domains from which we noted which proteins have regions with similar structure.

2.2.5 Genomic tRNA database

Throughout this thesis, we used the Genomic tRNA database (GtRNAdb) to source the copy numbers of tRNA genes for a given species [43]. The GtRNAdb contains counts of tRNA genes found across species' genome separated by anti-codon type. Counts are predicted using tRNAscan-SE, a programme that scans whole genomes for the characteristic sequence of tRNA genes, there being only a few codons difference between the various types. For some species, the tRNA concentration within the cell has been measured and found to have a strong linear relationship with the tRNA gene count [93, 96]. As such, the relative tRNA gene count is often used as a proxy for the relative tRNA concentration [116].

2.2.6 Coding Sequence and Structure

The Coding Sequence and Structure (CSandS) database, created by Saunders and Deane, contains 4406 non-redundant entries of one-to-one mappings between protein structures and

their coding mRNA sequences [113]. The CSandS database is based upon the Universal Protein Resource (UniProt) pdbtosp database, this a mapping of the sequences contained within UniProt, which itself is a repository of heavily annotated protein sequences, to protein structures within the PDB based on cross-referencing identifiers in the respective metadata [124, 125]. The same metadata identifiers are then also used to find the corresponding protein coding mRNA sequences in the European Nucleotide Archive (ENA) database, which is a vast database of nucleotide sequences [126]. These pairings of protein structure (and sequences) to mRNA sequence are then verified using BLAST, a tool which identifies regions with similar sequences [127]. Comparing all protein coding mRNA sequences from EMBL-ENA to the protein sequences of structures in PDB, only matches that appeared in both the metadata pairings and the BLAST sequence alignments are considered true. Additionally, only mRNA sequences that aligned with zero gaps were considered, as well as requiring 90% of the protein sequence to be contained within the mRNA sequence. Finally, redundant protein sequences, such as proteins with multiple identical chains, are removed to create a fully non-redundant set. All pairings are then annotated along the length of the sequence with their secondary structure as calculated using JOY. The domain regions, as based on the SCOPe database, and the source organism are also given as metadata for each entry. We use the CSandS database as a basis for the work undertaking in this chapter as it provides a high-quality pairing of mRNA sequences to protein sequences and their structures. In the Results, we detail a series of modifications that we made to the CSandS to address issues with the original construction.

2.2.7 Theoretical estimators of translation speed

Multiple predicted translation speed profiles were calculated for each entry within our dataset based on their respective mRNA sequences. Translation profiles were created using the CAI, tAI, MinMax and nTE algorithms, full details of which are given below [104, 105, 116, 117]. Alongside the sequence of a given entry, the codon bias found across the transcriptome, and the

Table 2.1 Reference genome annotations and expression datasets selected for each species

Species	Genome	Expression Data
Human	GCA_000001405.14	E-MTAB-1733
E. coli	GCA_000010245.1	E-MEXP-3811
Yeast	GCA_000146045.2	E-GEOD-20351
Mouse	GCA_000001635.4	E-MTAB-599
Cow	GCA_000003055.3	E-GEOD-19696
B. subtilis	GCA_000009045.1	E-GEOD-55202
Rat	GCA_000001895.3	E-GEOD-20344
S. flexneri	GCA_000183785.2	E-GEOD-12535
S. typhimurium	GCA_000006945.1	E-GEOD-35750
H. influenzae	GCA_000027305.1	GSE5061

relative expressions of each transcript are required. For the former, we sourced the cDNA of all transcripts from the NCBI genomes database [128]. For the latter, we sourced various RNA-seq experiments freely available online in either the Gene Expression Omnibus or the Expression Atlas dataset [129]. The exact study used is given in Table 2.1. For multicellular organisms, uncertainty arises as the expression is tissue specific. For our work, we have chosen to take the geometric mean of the expression for a given gene across all tissue types measured in the stated dataset. Lastly, for MinMax, a smoothing window is included within the algorithm; while the others, CAI, tAI and nTE, it is applied posthoc to the translation speed profile as an averaging kernel. The choice of window width should reflect the phenomena being investigated; larger windows allow for insight into changes that take place over the length of a transcript, while small windows allow the comparison of single residues. For our analyses, in which we are primarily focused on single residue biases, we made use of a three-codon window size unless otherwise stated.

Codon Adaption Index

One of the first theoretical estimators of translation speed was CAI, which was originally designed to give an overall statistic of the codon optimisation across a given gene [104].

The applied resolution was increased over time, and CAI is now more often viewed at the codon-level instead. The CAI for codon i is given by

$$\text{CAI}_i = \frac{f_i}{\max(f_j)} \qquad i, j \in [\text{synonymous codons for an amino acid}], \tag{2.1}$$

where f_i is the respective frequency of codon i, and $\max(f_j)$ is the frequency of the most prevalent codon in the corresponding synonymous codon set for the amino acid encoded for by codon i. These frequencies, known as codon usages, are calculated using only the top five percent most expressed protein-coding genes, as these are thought to optimise their translation to a greater extent. This assumption is supported by analysis showing that the preference for some synonymous codons to be used over others becomes more pronounced in genes which are expressed more [93].

The MinMax algorithm

Created by Clarke and Clark, MinMax is also based upon relative codon frequencies [117]. However, in contrast to CAI, MinMax calculates these frequencies using the entire genome, rather than just the most highly expressed genes. Using these frequencies, MinMax calculates all synonymous alternatives to a given segment of mRNA to produce the possible maximum, minimum, and average translation speed one could observe while maintaining the protein sequence produced. Comparing these values to those calculated using the codons present gives an estimate of optimisation and, hence, the relative translation speed. In Figure 2.1, this calculation is explained pictorially. The separate Max and Min values shown in the figure are then combined to give the MinMax value via

$$\text{MinMax} = \begin{cases} \text{Max}/100 & \text{if Max} \geq 0 \\ -\text{Min}/100 & \text{if Min} \geq 0 \end{cases}. \tag{2.2}$$

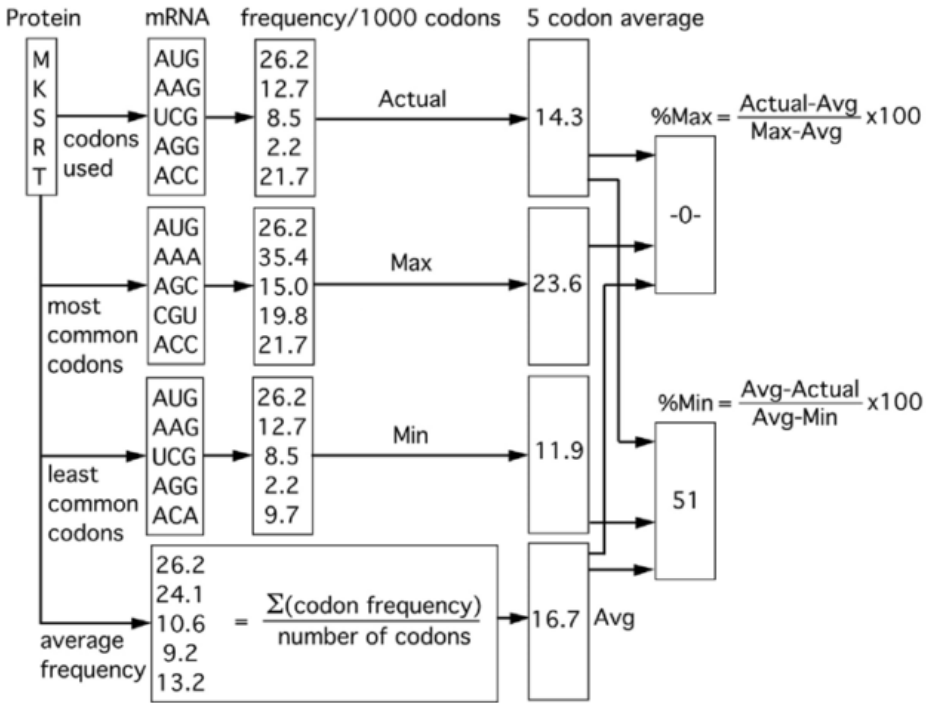


Fig. 2.1 The MinMax algorithm. The MinMax algorithm applied to the middle codon in an example five codon fragment. A Min value of 51 is given meaning that the sequence is approximately halfway between the most non-optimal sequence and the average. Window sizes can range from 4 to 30 codons. Image recreated and text reworked from Clarke & Clark (2008), *PLoS One*, **3**(10).

Note that the above formulation includes a division by 100 not present in the original construction. This division limits the range of possible values so as to be more comparable to those produced by the other algorithms.

tRNA Adaption Index

The tAI by dos Reis was one of the first estimators to be based on a physical rather than statistical metric [116]. It uses cognate tRNA concentrations in combination with the strength of codon-anticodon interactions to estimate a given codon's translation speed. To calculate the tAI values, the absolute adaptiveness value W_i of all codons must first be calculated, which is defined as

$$W_i = \sum^{n_i} (1 - s_{ij}) \text{tGCN}_{ij}. \tag{2.3}$$

for codon i. The sum is over all tRNAs that can pair with a given codon, s_{ij} is the strength of the pairing between the tRNA and codon, and tGCN_{ij} is the corresponding tRNA gene copy number. The concentration of more than one tRNA must be considered due to Crick's wobble pairing, whereby various tRNAs can match with a single codon due to the variable pairing of the third nucleotide of the codon and the isoreceptor [41]. However, alternate pairings do not associate with the same propensity as cognate tRNA, hence the variability of the pairing strength through parameter s_{ij}. Values of s_{ij} for all possible non-cognate pairing within eukaryotes and prokaryotes are given in Table 2.2. For all cognate pairings $s_{ij} = 0$. Lastly, the tRNA gene copy number, the number of times the gene appears across the genome, is used as an approximation for the tRNA concentration. This approximation is based on the assumption that each instance of a tRNA gene is expressed equally; hence, relative gene numbers is equivalent to relative concentrations. This assumption has been shown to hold for multiple species [96, 109]. We obtain tRNA gene copy numbers for our species of interest from the GtRNAdb.

Having determined the absolute adaptiveness values, the tAI for codon i is given by

$$\text{tAI}_i = \begin{cases} W_i / W_{\max} & \text{if } W_i \neq 0 \\ w_{\text{mean}} & \text{else} \end{cases}, \tag{2.4}$$

where $W_{\max}$ is the highest absolute adaptiveness value out of all codons and w_{mean} is the geometric mean of all non-zero absolute adaptiveness values.

normalised Translation Efficiency scale

The nTE scale is the newest estimator of translation speed we tested [105]. It combines both the statistical codon usage measurements and physical tRNA concentration to produce its estimates. It argues that a cell must balance supply and demand with regards to codon choice; if a codon

Table 2.2 Interaction parameter for pairing between codons and non-isoreceptor tRNAs

Anticodon:Codon	Eukaryotes: s_{ij}	Prokaryotes: s_{ij}
GNN:NNU	0.410	0.6294
ANN:NNC	0.280	0.4211
ANN:NNA	0.999	0.8773
UNN:NNG	0.680	0.6980
LNN:NNA	0.890	0.9500

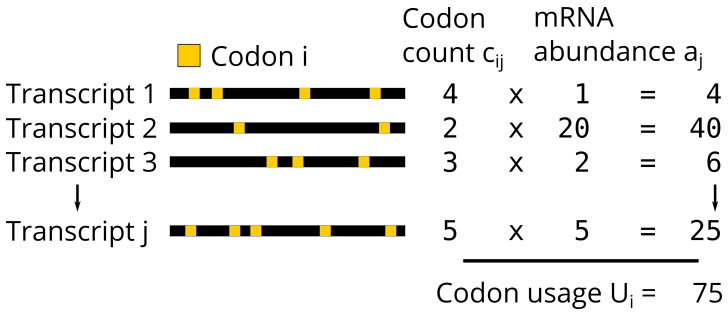


Fig. 2.2 Weighted codon usage. A schematic outlining the calculation of the weighted codon usage used within the nTE algorithm. Image reworked from Pechmann and Frydmann (2013), *Nat. Struct. Mol. Biol.*, **20**(2).

is overused, then its corresponding tRNA pool will be depleted and hence it will not associate with the ribosome as quickly. The nTE for codon i is given by

$$\text{nTE'}_i = \frac{\text{tAI}_i}{cu_i} \tag{2.5}$$

$$\text{nTE}_i = \text{nTE'}_i / \text{nTE'}_{\max}, \tag{2.6}$$

where tAI_i is the optimality value as defined by tAI (see above) and cu_i is the weighted codon usage of that codon.

The weighted codon usage cu_i is based upon the relative codon counts across transcripts when the expression is considered. There are more copies in the transcriptome of codons from highly expressed genes than those from genes expressed less. To account for the expression, the counts c_{ij} of each codon i on a given transcript j are then weighted by that transcripts

expression a_j. Summation over all transcripts in the transcriptome g followed by rescaling gives the weighted codon usages. This calculation is outlined in Figure 2.2 and is given by

$$U_i = \sum_{j=1}^{g} a_j c_{ij} \tag{2.7}$$

$$cu_i = U_i / U_{\max}. \tag{2.8}$$

2.2.8 Statistics

Optimal and non-optimal codons

Comparison of the translation speeds predicted by each of the estimators directly is often flawed, as depending on the width of the smoothing window used, the distribution of translation speeds predicted is highly non-normal (see Results). As such, a rank based metric provides a fairer comparison, reducing the speeds along a given transcript to ordinal values to remove the influence of the underlying sampling distribution. As such, we ranked the speeds predicted for each codon across each individual transcript, and then assigned the fastest and slowest 10% of codons within each transcript independently, as given by their estimated speeds, as optimal (fast) and non-optimal (slow) codons respectively. Those not assigned either of these labels were not assigned a category. For example, a 100 residue entry would have the ten residues with the highest predicted speed values labelled as optimal, and the ten residues with the lowest predicted speed values labelled as non-optimal. Note that the codons assigned to these classifications will likely differ for each of the predictive measures.

The above classification scheme mimics that employed by Pechmann *et al* [130]. However, where they made use of a 50% threshold, thereby labelling half their codons as optimal and half as non-optimal, we used only 10%, labelling the remaining codons as neither class. Our choice of only 10% for the codons to be classified stemmed from a believe that we would see a

clearer signal in our further analysis should only the codons which exhibit the most extreme translation speeds be considered. This is collaborated by analysis of our speed distributions (see Results) in which we noted that many codons exhibited values close to the median, and as such, very small changes would have led to reclassification if 50% thresholds, such that all codons were classified, were instead used.

Cochran-Mantel-Hanzel test

To measure the association of optimal and non-optimal codons to various facets of protein structure, we used the Cochran-Mantel-Hanzel (CMH) test unless otherwise stated [131]. The CMH test allows us to accommodate differences between transcripts via stratification. The validity of this stratification was not verified via the Breslow-Day test, as our methodology was chosen to mimic that implemented by Pechmann *et al* [130, 132]. The CMH test produces an odds ratio (OR) which indicates whether a given category (e.g. helices) is either over or under represented, namely when the $OR > 1$ or $OR < 1$ respectively.

Bonferroni correction

In our work, we test various associations of the translation speed to facets of the protein structure. Furthermore, we do this using four different estimates of the translation speed. Given this, it is important to account for multiple testing across our investigation; otherwise, the likelihood of Type 1 (false positive) errors is high. In this study, we use a Bonferroni correction, which can be summarised as dividing the accepted error rate (the p-value) by the number of comparisons made to establish an adjusted threshold for acceptance [133].

2.3 Results

2.3.1 Modifying CSandS

The CSandS database provided an excellent basis for our investigation. Namely, it provided a high-quality mapping of mRNA transcripts to experimental protein structures. However, in CSandS, the domain information is given in the format supplied by SCOPe, namely the domain type (e.g., 1.a.a.a.), the protein chain identifier, and the start and end residue identifiers. Hence to identify a domain within a CSandS entry, the residue-by-residue identifiers would need to be supplied along the length of the aligned protein sequence. These can not be inferred as residue identifiers are often not simple incremental values starting at one. Often numbering starts at a different value for various reasons and larger increments between adjacent residues may be observed, most often due to a gap in the protein backbone. Additionally, inserted residues are often labelled using an additional character (e.g., 13A) so as to preserve numbering along the remaining protein chain. As CSandS does not supply these residue identifiers, it is impossible to identify which residue is part of which domain correctly. Tangentially to the inability to identify domains, we also found a series of inconsistencies in the codon-residue pairings within CSandS, such that the sequences of some mRNA transcripts were not correctly aligned to their protein sequences. Most errors were associated with the misassignments of gaps in the protein sequence, which propagated along the remaining sequence such as to corrupt the entire entry.

To improve the quality of the CSandS database and address the above, we repeated the alignment of the protein sequence to the mRNA sequence using ClustalW. We extracted the protein sequence directly from the PDB structure linked to each CSandS entry rather than use that supplied as this let us retain the residue identifiers. As such, domains could now be correctly located post alignment. All alignments achieved a coverage of at least 80%, with 82% achieving 100% of the mRNA coverage. Our coverage is comparable to that reported in the original construction of CSandS [113, 134]. In addition to the realignment, we also recalculated

Table 2.3 Modified CSandS database entries. The number of entries within the CSandS separated by species. Each entry consists of the sequence (both protein and mRNA aligned to each other), secondary structure, percentage accessibility and predicted translation speeds for a single protein domain.

Species	Number of Entries
H. sapiens	1059
E. coli	885
S. cerevisiae	275
M. musculus	198
B. taurus	150
B. subtilis	129
R. norvegicus	108
S. flexneri	64
S. typhimurium	60
H. influenzae	55
Total	2983

the structural annotation using JOY. While the original secondary structure annotation could be transferred to the alignment with only minimal calculation, repeating the calculation provided an additional annotation of interest, namely the percentage accessibility of each residue in the structure.

After performing the alignment and additional annotation, proteins with multiple domains were split into separate entries based on the boundaries defined by SCOPe. In cases where a gap in the protein sequence existed between the boundary residues, due to either the alignment or unresolved residues in the structure, the excess mRNA sequence was not assigned to either domain. Similarly, excess residues at either the start or end of the protein sequence were discarded. Note that splitting after structural annotation allows for neighbouring domains to affect each other's solvent accessibility.

After splitting multidomain proteins, our modified CSandS database contained roughly 4300 entries. From this, we created a non-redundant set of entries by enforcing a 90% sequence identity threshold based on the protein sequence between the entries. This threshold removed approximately 20% of the entries, after which the database contained 3600 entries. We then

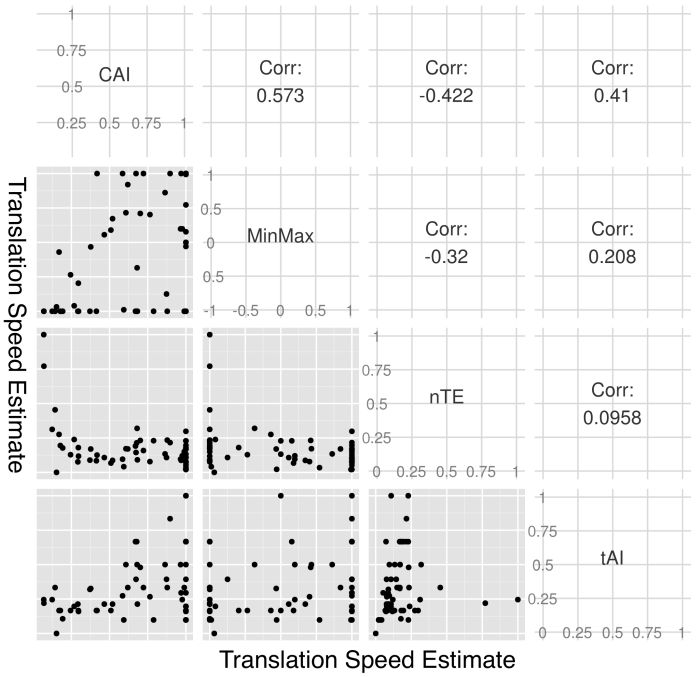
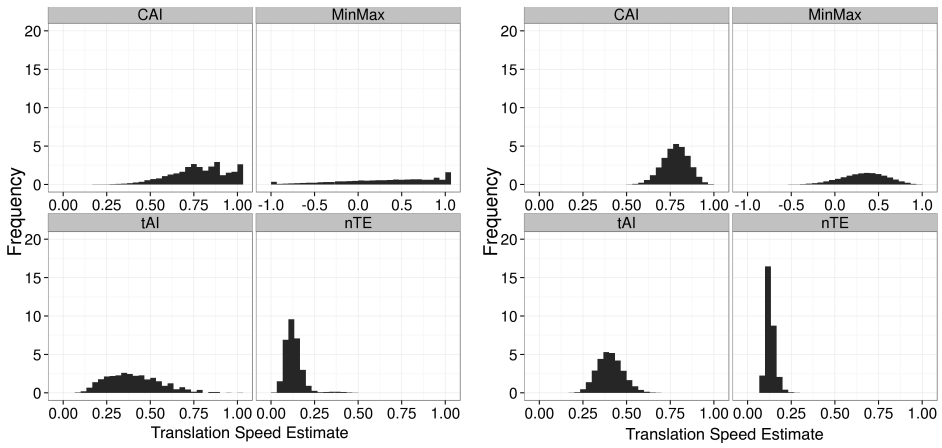


Fig. 2.3 Comparison of predicted translation speed for *B. subtilis*. The single codon values assigned by each of the predictive measures of translation speed for *B. Subtilis* are compared pairwise. The lower panel shows scatter plots of the predicted speeds for each codon, and the upper panel gives the corresponding Spearman correlations.

reduced the remaining dataset to only the ten most populous species in CSanS in preparation for predicting the translation speed using the theoretical estimators. The estimators require additional species-specific information to be provided on top of the mRNA sequence found within an entry. As such, it would be untenable to work with all species in the database. The top ten species collectively accounted for 84% of the remaining data and totalled 2983 entries. The organisms and their respective contributions are given in Table 2.3. Lastly, we added various predictions of the translation speed along a given sequence using the various estimators, details of which were provided above.

2.3.2 Comparison of predicted translation speeds

To understand the underlying differences between the predictive measures, we compared the values predicted by each algorithm for each codon type. Smoothing windows inherently reduce differences, so treating them as single codons help highlight how the estimators differ from one another. In Figure 2.3, the predictive translation speed assigned to each of the 61 non-stop codons is given for *B. subtilis*. Only tAI is found to gives an even coverage of the full range of possible translation speeds. However, some rare cases exist for which this is not true; for example, in *B. taurus*, the most common tRNA gene occurs roughly 1300 times across the genome, while the next most prevalent gene occurs roughly 400 times [43]. The large differences in copy numbers between tRNA genes lead to a significant bias in the single codon tAI values. Using CAI, values are biased towards high speeds, assigning multiple codons to the maximum translation speed ($\text{CAI}_i = 1$) due to every synonymous group having at least one codon assigned to this extreme. This assignment occurs even when only a single codon encodes for a given amino acid. For MinMax, a similar bias towards both the extremes is found for synonymous groups of at least two codons. For these cases, MinMax will always have one value assigned as the speed maximum ($\text{MinMax}_i = 1$) and one as speed minimum ($\text{MinMax}_i = -1$). When only a single codon encodes for a given amino acid, MinMax assigns it as neither quickly nor slowly translated ($\text{MinMax}_i = 0$). Lastly, nTE assigns the majority of codons across the various species as slowly translated. This bias is due to using the weighted codon usage which prejudices towards the codon selection observed in the most expressed genes. If a single codon rarely appears in the most expressed transcripts, then it will have an extremely low weighted codon usage and, as such, a very high nTE value. Then, due to scaling, if one unscaled nTE value dwarfs the others, the remaining codons will appear to be translated slowly. For example, we found that ignoring the top twenty most expressed transcripts resulted in the weighted codon usages to vary significantly in *B. subtilis* ($R^2 = 0.91$ when compared to original). For comparison, removing these transcripts when calculating other descriptions of



(a) Three codon window (b) Fifteen codon window

Fig. 2.4 Effect of window size on predicted translation speeds. The distribution of translation speeds generated using each predictive measure for all *B. subtilis* transcripts. On the left these are calculated using a three codon-wide window, on the right a fifteen codon-wide window is used. The x-axes are set to show the full range of translation speeds that could be observed using each algorithm.

codon usage caused minimal changes ($R^2 > 0.99$ when compared to original). These results suggest that there are inherent biases within the base construction of each estimator that should be considered when performing inference based on the speeds they predict.

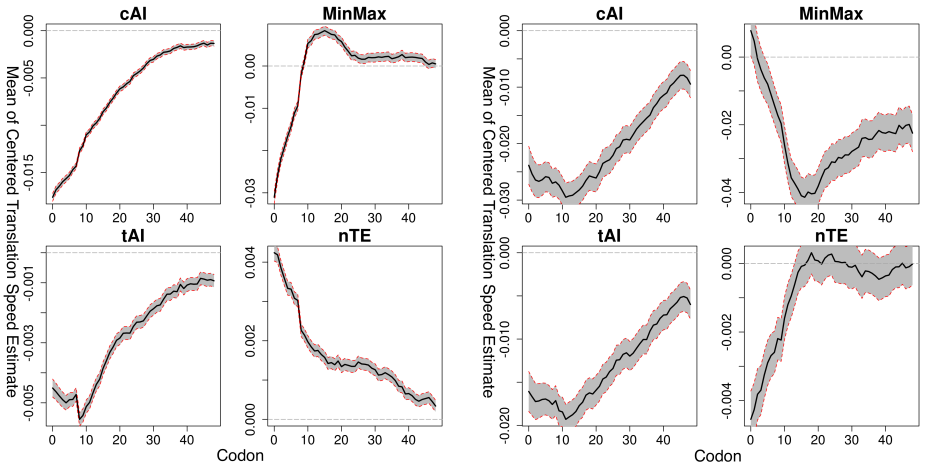
These underlying biases are masked in the translation speed profiles of transcripts due to the use of smoothing windows. An example of this is shown in Figure 2.4 where we have given the distribution of translation speeds generated using all coding transcripts for *B. Subtilis* using each predictive measure. We have calculated these distributions for both a three and fifteen codon-wide window. The three codon-wide window produces distributions that are highly non-normal with the underlying biases of the single-codon values still apparent. For example, the CAI distribution is heavily skewed towards quickly translated codons and MinMax produces a bimodal distribution. The fifteen codon-wide window removes these features; the resultant distributions are approximately normal. The *B. subtilis* transcriptome exhibits a strong bias towards fast codons using both CAI and MinMax. This is a consequence of their underlying assumption, that evolutionary pressure causes translation to be optimised and, therefore, using

Table 2.4 Correlation between predictive measures of translation speed. The coefficient of determination (R^2) between the single codon values assigned by each pairing of the algorithms is given for each species in our analysis. Bold and italics are used to highlight the largest and smallest coefficient respectively for each species.

	CAI MinMax	CAI nTE	CAI tAI	MinMax nTE	MinMax tAI	nTE tAI
B. subtilis	**0.323**	0.178	0.168	0.103	0.043	*0.009*
B. taurus	0.640	0.008	0.013	*1.2e-5*	9.4e-04	**0.931**
E. coli	**0.337**	0.123	0.233	0.092	0.090	*2.3e-08*
H. influenzae	**0.540**	0.175	0.169	0.188	0.071	*0.001*
H. sapiens	**0.569**	*0.002*	0.205	0.012	0.137	0.280
M. musculus	0.561	*0.002*	0.120	*0.002*	0.064	**0.716**
R. norvegicus	**0.565**	0.014	0.243	*0.004*	0.186	0.564
S. flexneri	**0.270**	0.132	0.056	0.094	*0.014*	0.129
S. typhimurium	**0.560**	0.103	0.158	*0.075*	0.077	0.077
S. cerevisiae	0.223	*0.004*	**0.585**	0.072	0.100	0.056
Average	**0.459**	0.074	0.195	*0.064*	0.078	0.276

the codon bias in their construction. In contrast, tAI and nTE suggest that the transcriptome exhibits a bias towards codons translated slower, which is at odds with the assumption that evolutionary pressure should cause translation to be optimised.

In Figure 2.3 we showed that CAI and MinMax were the most similar predictive measures for *B. subtilis*, while nTE and tAI were the most divergent. Expanding upon this, Table 2.4 gives the coefficient of determination (R^2) between each predictive measure for all species analysed. We find that CAI and MinMax are consistently the most similar, R^2 being largest for this pair of measures in seven out of ten of the species, as well as having the highest average R^2 of 0.459. This observation is unsurprising as both are based on descriptions of codon bias. In nine out of ten species, the lowest R^2 is between nTE and one of the other algorithms; this result highlighting the unique formulation of nTE compared to the other algorithms. An exception to these trends is found for *B. taurus* and *M. musculus*, where the highest R^2 is observed between nTE and tAI. This similarity is caused by these species having far more copies of a single tRNA gene than all others resulting in the majority of tAI values being small with one large common outlier. These large differences are incorporated into tAI, which in turn are used to calculate



(a) All species (b) *S. cerevisiae* only

Fig. 2.5 Predicted translation speed variation over initial region of transcripts. The mean within-gene-zeroed translation speed over the first 50 codons is plotted, where the initial codon has been ignored due to its propensity for AUG. The mean, shown by the black line, is calculated on values that have been produced using a 15 codon width window and the 95% bootstrapped confidence interval of the mean is shown in grey and highlighted by the red line. Any position whose value is below zero (dotted line) are translated slower than the average. In (a), the results are given for the entries sourced from all species combined. In (b), only those identified as sourced from *S. cerevisiae* have been used.

nTE. In general, however, the correlation between any two measures is poor, with only one comparison out of the 60 tested resulting in an $R^2 > 0.8$.

2.3.3 Association of translation speed to protein structure

Having shown that the estimators of translation speed differ significantly in their predictions, we next investigated how this divergence affects speed-structure biases inferred when the estimated translation speeds are paired with experimental protein structures. As discussed above, there exist many contradictions in the literature with regards to speed-structure biases, likely due to the differences noted above. We repeat their respective analyses, using all four of our predicted translation speeds instead of just one, so as to compile points of both agreement and disagreement in the inferred speed-structure biases.

One result that has been reported, using both tAI and nTE, is that the initial codons of protein coding genes are translated slower than the remaining sequence [105, 109]. We give the equivalent results found for each theoretical estimator in Figure 2.5. For this analysis, predicted speeds were calculated using a 15 codon-wide averaging window and shifted such that each entry's mean was zero. This zero-shift provides the normalisation required to allow comparison between entries sourced from different species. The initial codon was also removed due to its propensity to be AUG. We find that each of the algorithms, bar nTE, produces the reported ramp in translation speed. However, as noted in the literature, the length over which this increase occurs differs between the measures. Both tAI and CAI exhibit a slow increase in the translated speed, both plateauing around the 50^{th} codon. MinMax exhibits a sharp rise in translation speed over the initial 15 codons, followed a short but slower decline, and then plateauing around the 20^{th} codon. In contrast to the other predictive measures, and contradictory to the prior works, nTE exhibits a steady slope of decreasing translation speed over the first 50 codons [105].

The difference in behaviour over the initial transcript region using nTE to that reported could be due to the taxonomic diversity of our dataset. The prior work found the ramp using a dataset consisting of transcripts from *S. cerevisiae* and nine other closely related species [105]. To test, we repeated the analysis using only the transcripts from *S. cerevisiae* in our dataset. The results, which are also given in Figure 2.5, show that nTE clearly exhibits the reported ramp in translation speed over the initial ten codons. Furthermore, while CAI and tAI maintain a similar ramp to that seen before, the behaviour observed using MinMax changes drastically. Instead of the ramp in translation speed followed by a small dip, MinMax now declines in translation speed over the first 15 codons, before increasing. These results suggest that the ramp may only be present in certain organisms. However, given the inconsistencies between measures and the variance between species, it is inconclusive that a ramp in translation speed exists.

Another speed-structure bias that has been previously observed is that quickly translated codons are found at structurally sensitive buried sites. This observation has been found using both nTE and another measure based on codon bias [105, 106]. We tested for this association using our categorisations of buried and exposed as well as optimal that had been assigned to each codon. As stated previously, we corrected our significance levels due to performing multiple comparisons using the Bonferroni correction, noting both these comparisons and those discussed below for our adjustment. Using nTE, we found that optimal codons, those that are translated quickly, exhibit a preference for buried regions ($OR = 1.18;\ p < 10^{-16}$), as per the observations reported previously. In contrast, tAI indicated that optimal codons prefer exposed positions ($OR = 0.85;\ p < 10^{-16}$). Both CAI and MinMax reported no association. These results indicate that even though the literature may agree, other untested measures may still produce a contradictory result. We repeated the analysis for non-optimal codons finding that tAI, CAI and nTE all report that they are biased towards buried sites ($OR = 1.20$, 1.10, and 1.10 respectively; $p < 10^{-16}$). MinMax reported the reverse, finding that non-optimal codons prefer exposed sites ($OR = 0.92;\ p < 10^{-16}$). Note that for nTE, both optimal and non-optimal codons exhibit a preference for buried sites, a possibility due to the 80% of codons not labelled as either class. Such a result would suggest that those codons with translation speeds that differ significantly from the average prefer buried sites. Again, these results highlight the divergence in the reported observations found when using different predictive measures.

A source of contention within the literature is whether fast or slow codons are enriched or depleted at domain boundaries and linker regions. For example, Thanaraj *et al.* found, using a measure based on codon usage, that linker regions are enriched with the slowest codons [112]. In contrast, Saunders *et al.* found, using tAI, that the boundaries were deficient in slowest codons and enriched with fastest codons [113]. For comparison, we tested the enrichment within the first and last ten percent of each domain by length for the codons we categorised as optimal and non-optimal. This subset of the domain contains both the linker regions and the

termini of the sequences, so is not entirely analogous to the referenced studies. We found that tAI, MinMax and nTE all report enrichment by optimal codons in these regions ($OR = 1.03$, 1.02, and 1.07 respectively; $p < 0.05$) and CAI reported the null hypothesis. For non-optimal codons, tAI, CAI, and MinMax all report enrichment of the regions ($OR = 1.07$, 1.19 and 1.09 respectively; $p < 10^{-10}$) and nTE gave the null hypothesis. In general, we find that both optimal and non-optimal codons are enriched in these regions, with the relative odds ratios suggesting that non-optimal codons are more prevalent. Alternatively, our observation could merely be due to the ramp in translation speed discussed above. However, it has also been suggested that non-optimal codons are required in these regions to enable the correct folding of multidomain proteins. Slowing translation within the linker regions is thought to allow the prior domain to fold before the next domain is translated to prevent deleterious interactions.

The fastest and slowest translated codons have been reported to favour certain secondary structure types over others [105, 113–115]. Our results, shown in Figure 2.6, find that even though multiple significant associations are found, there is little agreement between the various predictive speed measures. For example, helices were found to be enriched with optimal codons using CAI, tAI, and MinMax with odds ratios of 1.08, 1.21 and 1.11 respectively ($p < 0.05$), while nTE reports the opposite with an odds ratio of 0.91 ($p < 0.05$). In general, the largest biases are observed using tAI, which could be an indication that it contains the most structurally relevant information out of the algorithms tested. Regardless, given the inconsistencies between measures, we are not able to state categorically that there is a link between translation speed and the secondary structure formed.

Lastly, it was also reported by Saunders and Deane using the tAI algorithm that transitions between secondary structures were accompanied by sudden changes in the translation speed [113]. For example, a significant dip in the translation speed was found when transitioning from a coil to a helix. As per their analysis, the translation speed of 16 codon-wide fragments centred on transitions was analysed with the additional requirement that each fragment must

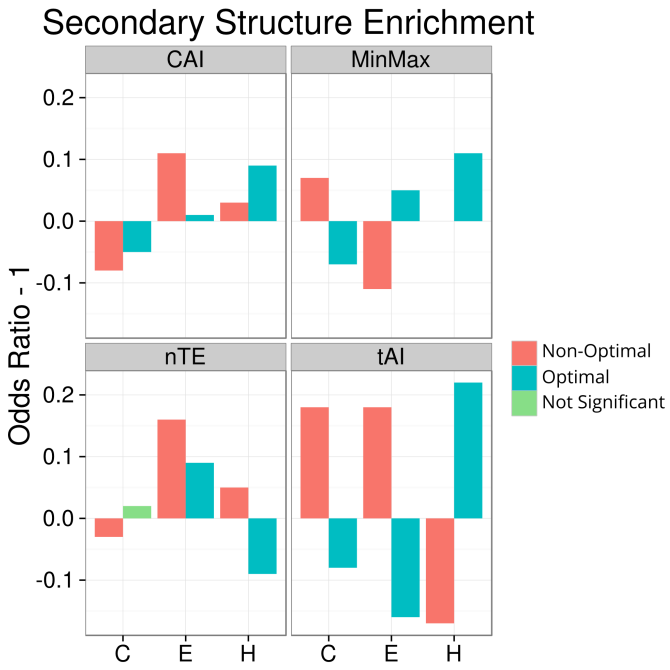


Fig. 2.6 Enrichment and depletion of optimal and non-optimal codons within each secondary structure type. For each predictive speed measures, the preferred secondary structure element for both optimal and non-optimal codons is given by their pooled odds ratios. The labels C, E, and H refer to coils, β-strands, and helices respectively. Under the null hypothesis, that the translation speed does not affect protein structure, we would expect no bias to be found and the odds-ratio to be near 1. Values above 1 indicate enrichment, while those below 1 signify depletion. MinMax appears to have no value for non-optimal codons within helices, but this is due to the value lying near 1.

have the same secondary structure for at least four codons on either side of the transition, e.g., XXXXXCCCCHHHHXXXXX for a coil to helix transition. The results for the coil to helix transition are given in Figure 2.7. We found that using tAI reproduces exactly the reported observations, though this is expected as their analysis was also based upon the CSandS database. Furthermore, the result may be reflective of the inherent translation speed biases of the respective secondary structure types; our prior analysis indicating that coils are enriched with non-optimal codons, and that helices are enriched with optimal codons. For all other algorithms, the presence of a change in translation speed at the transitions residue is unclear, some variation being observed in the region, but not specifically at the transition point. Considering other to

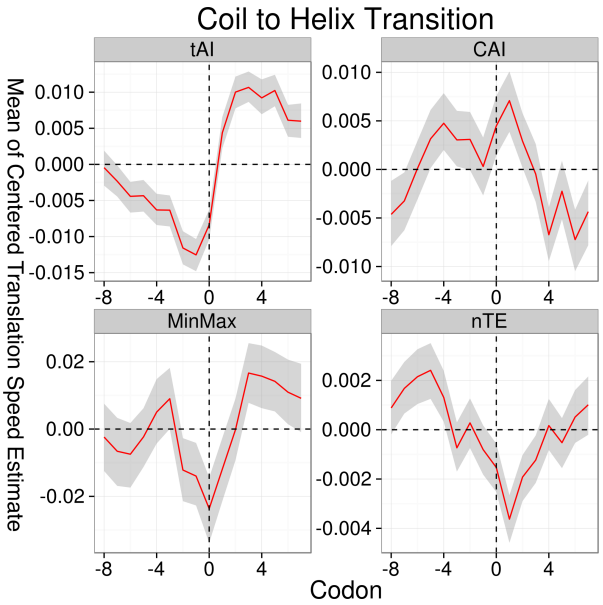


Fig. 2.7 Predicted translation speeds across coil to helix transitions. The mean predicted translation speed is shown in red with the 95% confidence interval of this value given by the surrounding grey area. The vertical and horizontal dashed lines indicate the transition point and the mean translation speed respectively.

and from coil transitions, we find that using tAI gives similar behaviours to those found by Saunders and Deane as expected; that there is a decrease in the translation speed just before any transition. However, using CAI, MinMax and nTE this was not found. Again, we find inconsistencies between the predictive measures of translation speed. Given this, it is uncertain whether the translation speed varies at transitions between secondary structure types.

2.4 Discussion

Many predictive measures exist to estimate the translation speed along a given transcript. In this chapter, we compared four well-known predictive measures of the translation speed to each other, namely CAI, tAI, MinMax and nTE. Looking at the values assigned to individual codons for *B. subtilis*, we found that measures based on comparing the codon usage within a given synonymous set, namely CAI and MinMax, contain inherent biases towards extreme values. Similarly, nTE was found to exhibit a bias for low values due to the infrequency of particular

codons within the most expressed genes. The effects of both these biases were lessened by smoothing the translation profiles. However, their underlying presence can still be seen to skew the distributions of predicted translation speeds computed across an entire transcriptome. Next, we compared the predictive measures to each other across all ten species. We found that CAI and MinMax were consistently the most similar and that nTE was the most unique. However, in general, the correlation between any two predictive estimators is poor, suggesting each of them are distinct from one another.

Each of the selected theoretical estimators of the translation speed has previously been used to comment on possible links between the speed with which a codon is translated and the protein structure it encodes for. Unsurprisingly, given the disparity between the estimators we showed, the studies using different predictive measures reported different observations, occasionally even stating contradictory biases. We repeated these past analyses, collating the speed-structure biases produced using each of the selected predictive measures. Doing so, showed that no previously reported speed-structure bias was unanimously agreed upon by all estimators. Even with observations which had no contradictory reports in the literature, such as the presence of a ramp of translation speed over the initial region of coding transcripts, we found that the predictive measures still did not agree. From these results, we summarised that no one measure outperforms the others at predicting the translation speed. Furthermore, given the contradictions found in the observation, it suggests that any speed-structure bias reported by these theoretical estimators of translation speed may be unsound.

2.5 Conclusion

In summary, we compared four well-known theoretical estimators of translation speed to each other, namely CAI, tAI, MinMax and nTE, as well testing and collating the speed-structure biases predicted by each. We found that the correlation between any two predictive measures is poor, the speeds predicted being highly dependent on the estimator chosen. Collating the

speed-structure biases, we found that the biases reported were also highly dependent on the predictive measure used, often contradictory biases reported between the estimators. Given this, it is inconclusive that any one measure gives better predictions of the translation speed than the others. Moreover, it suggests that any speed-structure biases reported by a predictive measure should be considered erroneous unless backed by experimental evidence proving its validity.

Chapter 3

Creating a database of ribo-seq translation profiles

3.1 Introduction

In the previous chapter, we discussed theoretical estimators of translation speed. We compared four well-known estimators, both in terms of the speeds they predicted for given codons and their predictions of translation speed to positions in the protein structure. We found that there was little similarity between the measures and concluded that this divergence was the probable source of contention in literature, whereby different studies have reported different speed to structure biases. At the time I started my doctoral studies, the majority of experimental measures which could provide an objective test of the various estimators were based on measuring the relative translation rates of codons within only a single gene [83–89]. For example, Curran and Yarus obtained the relative translation rates of codons placed at the location of a competing frameshift, which is assumed to occur at a constant rate, by quantifying the occurrence of the frameshift to non-frameshifted products [84]. Likewise, Sander *et al.* created a protein that consisted of three half-domains connected by linkers, in which the middle half-domain could form a whole protein with the first or last half-domain independently. Measuring the ratio of

the two products formed while varying the codons used in the linker regions gave an estimate of the relative translation rate of the codons [88]. While these experiments did provide an objective measurement of the translation rate of codons across a given gene, extrapolating any comparison to the whole genome is tenuous. As such, when limited in experimental scope to gene-level, no one estimator could be deemed objectively better or worse than any of the others.

Ribo-seq, also known as ribosome profiling, is a technique that measures the location of translating ribosomes across the whole transcriptome [135, 136]. From the experimental output of ribo-seq, one can infer the translation speed along each gene, and hence it provides an experimental metric with which the theoretical estimators of the translation speed can be compared. [137, 138]. First introduced in 2009, ribo-seq is an expansion of RNA-seq, in which the RNA from a sample is broken into small chunks called reads, measured using next-generation sequencing (NGS), and then mapped to the source's transcriptome via alignment [139, 140]. The relative numbers of reads aligned to each gene give a measure of their relative expression. Ribo-seq is similar, though prior to sequencing, first the translation within a sample is halted using either inhibition by drugs, or, ideally, flash freezing. Then, only those fragments of RNA protected by a ribosome at the point of cell lysis are sequenced. This subsetting is enabled via the addition of nuclease to the cell lysate to both break the mRNA transcripts into reads and degrade all mRNA not protected by a ribosome. Following multiple purification steps, the protocol yields a solution containing only mRNA fragments sourced from locations undergoing translation at the point of cell lysis alongside contaminating rRNA which can be filtered out post-sequencing. Mapping these fragments back to their respective transcripts gives the position of the translating ribosomes to near codon-level resolution. Then, from the relative numbers of ribosomes at positions across a given transcript, the relative speed of translation along that transcript can be inferred as the slower the translation, the more ribosomes should be found at a given site. A broad overview of the method is presented in Figure 3.1.

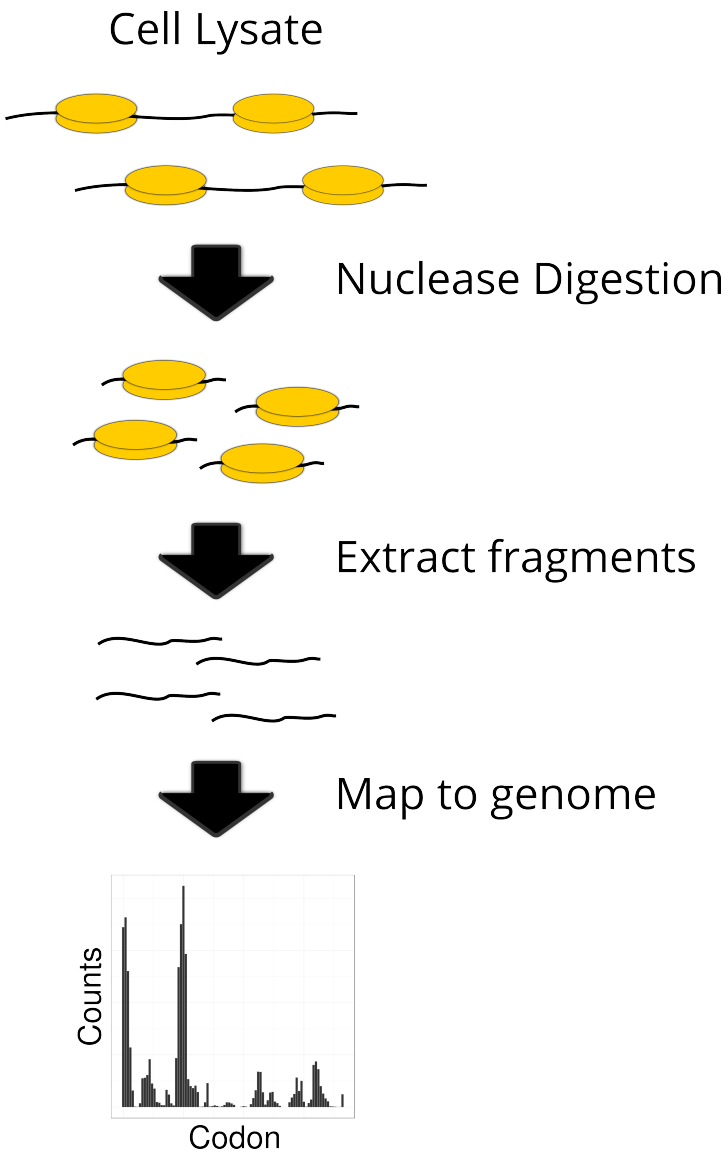
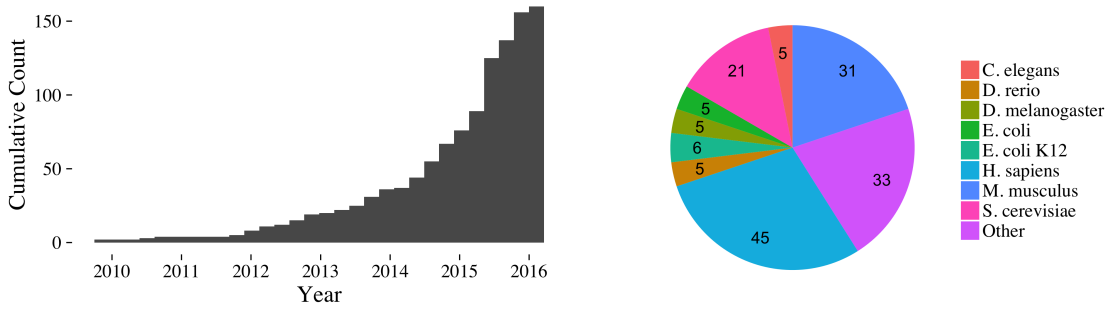


Fig. 3.1 Ribo-seq experimental protocol overview. First, a cell is lysed, and the polysomes (mRNA with ribosomes attached) are separated from the other lysate. Nuclease is then added to the sample to digest all mRNA not shielded by a ribosome, while also separating ribosomes attached to the same mRNA strand. Subsequent removal of the ribosomes leaves behind only the mRNA fragments which were undergoing translation at the point of cell lysis. These fragments can then be mapped back to the genome or transcriptome to give the location of the translating ribosome at nucleotide level resolution. The relative counts of aligned reads at each position gives a transcriptome-wide measurement of the translation occurring within the cell. From this we can infer the translation speed associated with any given codon in any given gene.



(a) Availible studies by year deposited

(b) Availible studies by species

Fig. 3.2 Ribo-seq datasets within the SRA database as of 19^{th} January 2016. The studies were located by performing a combined search for the terms Ribo-seq, Ribosome reads, and Ribosome profiling across all possible fields. This search returned 4124 different experiments sourced from 156 different studies. On the left, we show the growth in the total number of datasets over time. On the right, we separate the available studies by species, grouping 23 different species for which less than five studies existed under the label Other.

Since its introduction, ribo-seq has been used to investigate an array of phenomena associated with translation not previously accessible [136]. For example, many non-standard translation initiation sites have been located upstream of AUG codons on transcripts thereby identifying new open reading frames and potential proteins [135, 141]. Similarly, the Shine-Dalgarno sequence has been shown to interact with the ribosome in such a manner to delay and pause translation [142]. Likewise, the translation of non-cognate pairings of tRNA anti-codons to mRNA codons slows translation, potentially revealing a method in which translation can be finely tuned [143]. This latter result is drawn from the widely held assumption that the relative ribo-seq counts along a given transcript are inversely related to the relative translation speed [137, 138, 143–152]. Given this assumption, ribo-seq may provide the first objective experimental transcriptome-wide measure of translation speed against which the theoretical estimators of translation speed can be compared. Furthermore, it could replace the theoretical estimators directly as a measure of translation speed for establishing the relationship between the translation speed and the protein structure.

At the time this work was undertaken, another study had already attempted to establish how well the theoretical estimators predicted the translation speed as given by ribo-seq [137]. While some correlation was reported, the study was limited in scope, investigating only one or two species, basing results on only a few ribo-seq studies, and compared the ribo-seq data to a single estimator [137]. The former two flaws were not faults of the study referenced, but overall limitations due to the lack of ribo-seq data available at the time. As shown in Figure 3.2, in the years following this study, this constraint was lifted as the number of publicly available studies grew substantially. Alongside this overall growth, the range of species that ribo-seq data is available for also expanded. Given this new found availability, we wanted to perform a larger comparison of ribo-seq to the theoretical estimators, both in terms of the number of species and the number of estimators. However, to accomplish this comparison, an unprecedented amount of ribo-seq data across a large number of species would need to be collated. In this chapter, we discuss our collation of ribo-seq data. We describe how ribo-seq experiments were identified and sourced from the NCBI-SRA database, then discuss in depth how we reprocessed each study to remove their inherent biases. In total, we processed 57 experiments spanning 24 studies across five species. Other comparable databases also emerged during the span of this work [153, 154]. We discuss the construction of these competing databases and highlight the differences when compared to our own. Finally, we compare a single ribo-seq dataset that is present in each of three databases to give an estimate of their similarity to each other.

3.2 Materials and methods

3.2.1 Study selection

The datasets we selected to form our database are listed in Table 3.1 found below. Here, we discuss our considerations in making these selections. Our first decision was to constrain ourselves to only those species with sufficient data available. We selected the five most populous

Table 3.1 Ribo-seq experiments selected to form our database. The Study and Run IDs refer to the SRA ascension codes. The label Lab refers to the principal investigator for the experiment, which was noted from the corresponding paper. The labels Size and Reads refer to the file size of the FASTA corresponding file and the number of reads it contains. Cleaning specifies which protocols were used to process each dataset. Lastly, Offset species the number of nucleotides between the 5' end of a read and the ribosome P-site. The offset is not present for some entries as they failed to reach this stage in processing.

Study ID	Run ID	Lab	Size	Reads	Species	Cleaning	Offset
SRA049309	SRR407630	A. Fire	1.4G	15296402	C. elegans	Tail	12
SRA049309	SRR407632	A. Fire	1.2G	13061646	C. elegans	Tail	12
SRA049309	SRR407645	A. Fire	1.2G	12639508	C. elegans	Tail	12
SRA055804	SRR522883	A. Fire	20G	174693758	C. elegans	Linker	12
SRA055804	SRR522897	A. Fire	20G	172256552	C. elegans	Linker	12
SRA249088	SRR1945004	S. Leidel	2.3G	23677484	C. elegans	Tail	12
SRA249088	SRR1945005	S. Leidel	2.7G	27850343	C. elegans	Tail	12
SRA082464	SRR869826	J. Roberts	2.5G	21836708	E. coli	Barcodes & Tail	-13
SRA082464	SRR869827	J. Roberts	1.3G	10944563	E. coli	Barcodes & Tail	-14
SRA190684	SRR1613263	K. Fredrick	2.8G	23558119	E. coli	Linker	-7
SRA190684	SRR1613265	K. Fredrick	2.4G	20020207	E. coli	Linker	-7
SRA190684	SRR1613266	K. Fredrick	2.2G	18140727	E. coli	Linker	-7
SRA209025	SRR1693437	B. Palsson	882M	9173033	E. coli	Linker	-16
SRA209025	SRR1693438	B. Palsson	1.7G	14832542	E. coli	Linker	-19
SRA297169	SRR2340141	A. Buskirk	4.8G	39351578	E. coli	Linker	-16
SRA297169	SRR2340143	A. Buskirk	4.8G	39177034	E. coli	Linker	-16
SRA297169	SRR2340144	A. Buskirk	2.3G	18366041	E. coli	Linker	-16
SRP010825	SRR407274	J. Weissman	21G	193193015	E. coli	Linker	-16
SRP010825	SRR407275	J. Weissman	19G	171654193	E. coli	Linker	-16
SRP010825	SRR407276	J. Weissman	23G	204841933	E. coli	Linker	-16
SRP010825	SRR407277	J. Weissman	9.3G	86856374	E. coli	Linker	-16
ERA390450	ERR690838	M. Bjarklund	3.2G	30886959	H. sapiens	Tail	
ERA390450	ERR690839	M. Bjarklund	3.4G	32022179	H. sapiens	Tail	
ERA390450	ERR690840	M. Bjarklund	2.6G	276685896	H. sapiens	Tail	
SRA020150	SRR057511	J. Weissman	430M	4807481	H. sapiens	Linker	12
SRA020150	SRR057512	J. Weissman	1.7G	18618732	H. sapiens	Linker	12
SRA020150	SRR057526	J. Weissman	2.1G	20366374	H. sapiens	Linker	12
SRA049309	SRR407637	A. Fire	3.0G	34308259	H. sapiens	Tail	13
SRA049309	SRR407638	A. Fire	3.0G	33862770	H. sapiens	Tail	13
SRA049309	SRR407643	A. Fire	2.7G	30991437	H. sapiens	Tail	13
SRA049309	SRR407644	A. Fire	1.7G	18711151	H. sapiens	Tail	13
SRA056377	SRR618770	S.-B. Qjian	1.1G	9742062	H. sapiens	Tail	12
SRA056377	SRR618771	S.-B. Qjian	5.1G	44827310	H. sapiens	Tail	12
SRA072609	SRR810100	R. Agami	20G	166137136	H. sapiens	Linker	13
ERA358090	ERR601610	I. Brierley	4.6G	38543800	M. musculus	Linker	12
SRA043500	SRR315616	J. Weissman	1.8G	19449857	M. musculus	Linker	10
SRA043500	SRR315617	J. Weissman	2.3G	24822364	M. musculus	Linker	10
SRA043500	SRR315618	J. Weissman	2.4G	25726565	M. musculus	Linker	10
SRA043500	SRR315619	J. Weissman	2.4G	25600179	M. musculus	Linker	10
SRA051495	SRR458756	V. Kim	2.3G	21824511	M. musculus	Tail	13
SRA056377	SRR618774	S.-B. Qjian	3.0G	25947180	M. musculus	Barcodes	5
SRA060331	SRR606203	C. Burge	20G	195019680	M. musculus	Tail	12
SRA008252	SRR014374	J. Weissman	429M	4601558	S. cerevisiae	Tail	13
SRA008252	SRR014375	J. Weissman	442M	4737798	S. cerevisiae	Tail	13
SRA008252	SRR014376	J. Weissman	353M	4221683	S. cerevisiae	Tail	13
SRA008252	SRR014377	J. Weissman	354M	4237384	S. cerevisiae	Tail	13
SRA008252	SRR014378	J. Weissman	352M	4211315	S. cerevisiae	Tail	13
SRA008252	SRR014379	J. Weissman	394M	4758893	S. cerevisiae	Tail	13
SRA008252	SRR014380	J. Weissman	440M	5322596	S. cerevisiae	Tail	13
SRA008252	SRR014381	J. Weissman	452M	5468170	S. cerevisiae	Tail	13
SRA096677	SRR948553	A. Shteyman	4.4G	41214129	S. cerevisiae	Linker	13
SRA096677	SRR948555	A. Shteyman	4.0G	38244257	S. cerevisiae	Linker	12
SRA097097	SRR951828	H. Frase	14G	136829250	S. cerevisiae	Tail	18
SRA104387	SRR1002819	B. Futcher	7.8G	66929528	S. cerevisiae	Linker	13
SRA107920	SRR1015436	W, Gilbert	771M	8914589	S. cerevisiae	Tail	13
SRA107920	SRR1015437	W, Gilbert	1.7G	16704810	S. cerevisiae	Tail	13
SRA107920	SRR1015438	W, Gilbert	1.8G	18129494	S. cerevisiae	Tail	12

species as per their respective number of ribo-seq studies within the Sequence Read Archive (SRA) as found by our keyword search discussed above (see Figure 3.2). Note that this differs from the experiment count, as one study may feature multiple ribo-seq runs, e.g., replicates. Furthermore, many of the studies returned by this search were false positives; the most common reason being that the investigation was, in fact, performed using RNA-seq. The species selected were *H. sapiens*, *M. musculus*, *S. cerevisiae*, *E. coli* and *C. elegans*. For each of these species, we selected up to five studies that we would attempt to process for inclusion in our database. For *C. elegans* we were not able to gather a full five studies as only three were identified. All ribo-seq experiments associated with a study that fit the below criteria were downloaded. Further, we treated each run within an experiment separately instead of merging as is normal with RNA-seq. For RNA-seq, it has been shown that the variation between two runs on the same sample is less than the technical error, so merging runs is valid and provides larger outputs. No work exists, to our knowledge, that shows this holds true for ribo-seq, and, as such, we treat runs independently. As the number of experiments varies between the studies, the number of experiments associated to each of the five species is uneven. In total, we selected 23 studies which contained a combined total of 57 ribo-seq experiments.

Further to the broad selection criteria above, we tried to enforce the following rules to help select appropriate studies and experiments within them for inclusion in our database. Depending on the number of available studies for a given species, some criteria were loosened such that five studies could be selected.

- **Sourced from different laboratories**

  To reduce the effect of experimental bias across our ribo-seq datasets, we strived, where possible, to select sets of studies performed by different laboratories for any single species. Laboratory independence was achieved for all bar *C. elegans*, for which we sourced three studies from two laboratories.

- **Only studies from published work**

  So that only data from a traceable source was used in our analysis, we did not consider any study listed in the SRA for which a corresponding paper could not be identified. Note that the relevant paper is rarely given in the metadata associated with a study in SRA and, as such, this normally involved manual identification.

- **Experiment performed on wild type cells**

  Many studies focus was the identification of differences in ribo-seq measurements between a wild type and a variant. For example, a sample with a genetic mutation or a sample exposed to a hazardous environment. These variants have been shown to express genes and translate differently to the wild type and should not be considered for inclusion in our dataset. Given this, for single-celled species, we only selected experiments performed on wild types. For multi-cellular organisms, various cell types were present across the available studies. While this heterogeneity will cause differences in translation, we do not expect these to effect the genome-wide translation speed relationships we seek to infer using this database. As such, we did not constrain ourselves to a single cell type for a given species.

- **Inability to identify specific runs**

  Given the lack of metadata for some studies within the SRA, we were not always able to determine how the experiments within the SRA match to the experiments discussed in the corresponding paper. It is essential that the experiments associated with different strains or extreme conditions can be identified such that we can fulfil the prior selection criterion. As a precaution, studies affected by this lack of metadata were not selected.

- **Translation in cell lysate ceased via flash freezing or cycloheximide**

  For ribo-seq to capture a true representation of the translation occurring in a cell, ribosomes should halt translating at the moment the cells are lysed [155]. However, this

is not always feasible, and more commonly a sample is pretreated with a drug based translation inhibitor immediately prior to lysis. The inhibitor interacts with the ribosome to disable its ability to translate and, as such, causes the ribosome to remain stationary [155]. However, in recent years, drug-based inhibitors have been shown to cause various bias in ribo-seq, different drugs causing the enrichment and depletion of ribosomes at numerous locations [141, 156, 157]. For example, pretreatment with either harringtonine or lactimidomycin has been found to cause an accumulation of ribosomes at initiating sites, neither drug found to immobilise ribosomes in the initiation phase fully [141, 156]. Ideally, we would like for our datasets not to use any drug based inhibitor, instead, identifying ribo-seq experiments performed using flash freezing followed by cryogenic pulverisation; a methodology that simultaneously lyses the cells and causes translation to cease [155]. Unfortunately, at the time this database was created too few samples existed in which flash freezing was used due to the increased experimental difficulty. Instead, cycloheximide is most common translation inhibitor used in ribo-seq studies [155]. As such, we relaxed this constraint and allowed datasets in which cycloheximide was used as the translation inhibitor. However, it should be noted that cycloheximide has also been reported to alter the ribosome density across the length of the transcript [157].

3.2.2 Processing overview

At the time this database was created, ribo-seq had no standardised pipeline to process the experimental output [158]. As such, there are key differences in the processing of the experimental output between each of the studies we selected to form our database. These differences would result in biasing the produced read counts and, as such, we could not use the processed counts for a given study if available. We deemed it necessary that each experiment within our database should be reprocessed from the raw data to remove these processing biases. Processing the raw output from ribo-seq experiments involves the steps outlined both below and in Figure 3.3:

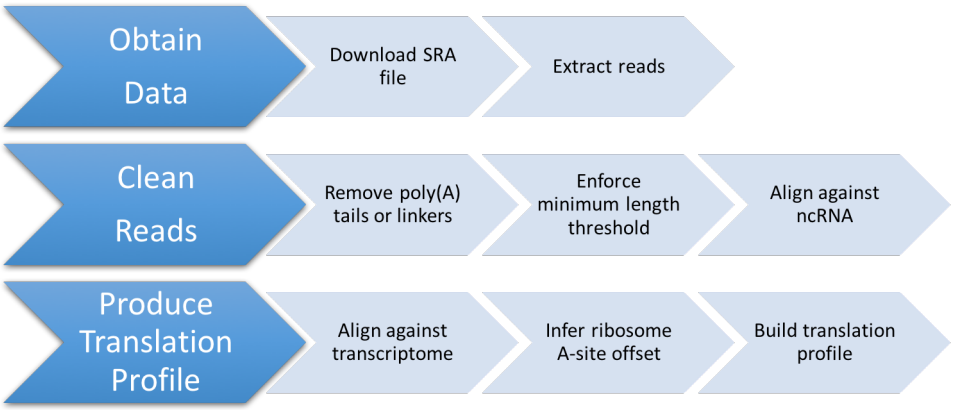


Fig. 3.3 Ribo-seq processing overview. Depicted broadly is the required processing steps to obtain translation profiles from the raw experimental output of ribo-seq. The steps are separated into three groups, which are executed in the order of top to bottom in the diagram above. Within a group, the steps are performed from left to right. From the translation profiles generated, the relative translation speeds of codons can be inferred.

- **Obtaining raw reads**

  Raw (unprocessed) reads were downloaded directly from the NCBI SRA database in SRA file format [126]. They were then converted into FASTA format using fastq-dump from the SRA toolkit [126]. While extracting the reads in FASTQ format was possible for some selected experiments, and would result in better alignments later in the protocol, using only sequences in the FASTA format allowed us to maintain consistency across our collated datasets. Additionally, the FASTA format allows for easier manipulation of the reads throughout our processing pipeline.

- **Cleaning the reads**

  Raw reads often contain additional nucleotides due to, for example, barcodes, PCR linkers, and polyadenylation [155]. These impurities are study specific and depend on the experimental procedure used. Failure to remove these nucleotides would result in

reads that cannot align to the transcriptome later in the protocol. Details of our cleaning methods are discussed in detail below.

- **Enforcing a minimum read length**

  Reads should not be too short as the shorter a given read, the more likely it is to align to multiple locations [159]. By enforcing a minimum read length, the protocol biases towards reads that uniquely align which are considered to be of higher value. In addition, short reads that can align to multiple locations are more computationally intensive to align, so removing them via the enforcing of minimum length threshold decreases the overall processing time of the protocol significantly. For our pipeline, we required reads to have a minimum length of 20 nucleotides after cleaning. This length is shorter than most other ribo-seq studies ($\approx$ 25 nucleotides), but crucially still long enough that the ribosome's A-site can still be inferred from the distribution of alignments (see below). This slight drop in minimum read length reflects our decision to maximise reads successfully passing through each stage of our protocol.

- **Removing non-coding RNA**

  Subtractive hybridisation is a common step in the experimental protocol used to filter ribosomal RNA (rRNA) from a sample [155]. Nevertheless, rRNA still makes up a significant proportion of the measured reads in most ribo-seq experiments [155]. While rRNA reads should not, in general, align with the transcriptome, if not removed, their inclusion would introduce significant computational overhead in the alignment. In contrast, aligning all remaining reads against the rRNA sequences is a far smaller task, which leads to an overall reduction in the total processing time. We removed rRNA reads by performing an alignment using Bowtie against the reference ncRNA for the associated species as given by Ensembl. Any read that aligned was deemed a contaminant and removed. Further details of the alignment can be found below.

- **Aligning of reads**

  The remaining reads are then aligned, again using Bowtie, against the transcriptome to give the location of translating ribosomes at the time of cell lysis. Further details of the alignment are provided below. This step is consistently the bottleneck with regards time in the overall processing of a ribo-seq dataset. While dependent on numerous factors, including the size of the dataset and the fraction of reads reaching this stage, the final alignment step can take upwards of multiple days.

- **Inferring the ribosome P-site**

  The above alignment gives the location within the transcriptome that a segment of mRNA sequence matches that of a given read. However, we are interested in the specific codon undergoing translation by the ribosome that protected a given read, rather than the entire protected segment. It has been found that for a given read, one can infer the P-site of the ribosome using an offset from either the 5' or 3' end of the alignment. This offset is experiment specific and must be inferred from the full alignment. Details on how this is achieved are given below.

- **Building a translation profile**

  Once an offset for the P-site is determined, a translation profile can be constructed from the alignment. In general, this involves assigning a value of one for each read whose A-site lies at a given nucleotide. However, this assignment is not simplistic, as one must consider both splicing and the treatment of reads that aligned to multiple locations. We discuss these details in more depth below.

Using the above pipeline, we reprocessed all the datasets we had selected. As to be expected, a substantial proportion of the reads did not make it through this process. In Figure 3.4, we show the fractions of successful reads for each experiment. Also, we give the breakdown of the other stages and the percentage of reads that failed.

3.2.3 Cleaning reads

As mentioned above, raw reads often contain additional unwanted nucleotides that must be removed prior to alignment. The nature of these impurities can usually be ascertained from the accompanying papers, which will either state them explicitly or give the experimental method, from which they can then be determined. Once the specific impurities contained in the reads are ascertained, the appropriate cleaning methods can be applied. For some studies, the impurities could not be determined from the text, and in these situations, we attempted to infer them via analysis of the reads. If inferable, we assumed that only a poly(A) tail should be trimmed (see below). The specific cleaning actions taken for each experiment are listed in Table 3.1. Below we detail the various impurities in the reads that were present across our selected studies and the action that was taken to remove them. These measures were applied via custom python scripts.

- Barcodes are sometimes attached to each read such that various subsamples can be identified after sequencing. The simplest example of this would be when multiple experiments have been sequenced simultaneously. Barcodes can be found at either end and are typically three or four nucleotides long. Within our compiled datasets, only study (SRA056377) had barcodes situated at the at the 5' end of each read, for which we removed the barcode and any other proceeding nucleotides. Study (SRA082464) had barcodes placed at the 3' end, so this was removed as well as any following nucleotides.

- Sequencing of ribo-seq reads involves polyadenylation, which is the addition of multiple adenine residues to the 3' end of a read to create a poly(A) tail [155]. Such additions obfuscate a given reads terminus if it already ends with any number of adenines, and, as such, has the potential to remove information from a dataset. This information loss can be avoided via the use of preadenylated linkers, which are discussed below. We removed poly(A) tails from the reads in experiments that did not make use of linkers, allowing for one mismatch within the tail. For example, XXXXAAATAA would trim up to the last

X where X is not adenine. This cleaning was applied to 11 out of the 23 studies in our database.

- As mentioned above, to avoid the obfuscation due to the polyadenylation involved in sequence the reads, a preadenylated linker can be attached to each read [155]. The linkers used in a given experiment are typically stated in the accompanying paper, though when not, we tested the relevant datasets for the inclusion of some of the most common linkers. If a linker was present in the reads, it was trimmed as well as any following nucleotides. If no linker was found, the read was removed from the dataset. Identifying linkers was not straightforward as often only a segment from the 5' end of the linker was present in the raw data, the linker often extending past the maximum length of the reported reads. As such, we only required a partial match to the given linker at the 3' end of a read. We enforced a dynamic minimum overlap length necessary for a positive match; the length set such that our cleaning had an error rate of 0.1%, i.e., one in 1000 linker matches will be due to a random nucleotide segment matching the 5' end of the linker. Preadenylated linkers were removed from the reads in 11 studies in our database.

Figure 3.4 gives the breakdown of where reads failed in the processing protocol. Combined into the fraction removed by cleaning are reads that were too short following any of the actions listed above. The average fractions of reads remaining after the two most used actions are applied, namely removing either poly(A) tails or preadenylated linkers, were 97.6% and 76.5% respectively. As is observed, poly(A) tail removal should remove very few reads, as removal can only occur when the procedure results in a read becoming too short. In contrast, the removal of linkers requires matching a segment of the read sequence to the 5' end of the linker. The average fraction of reads remaining for linker removal is also affected by study SRA020150, an outlier with only 15.1% of reads remaining post-cleaning. If we instead report the median, poly(A) tail and linker removal remaining fractions would be 98.4% and 83.1% respectively.

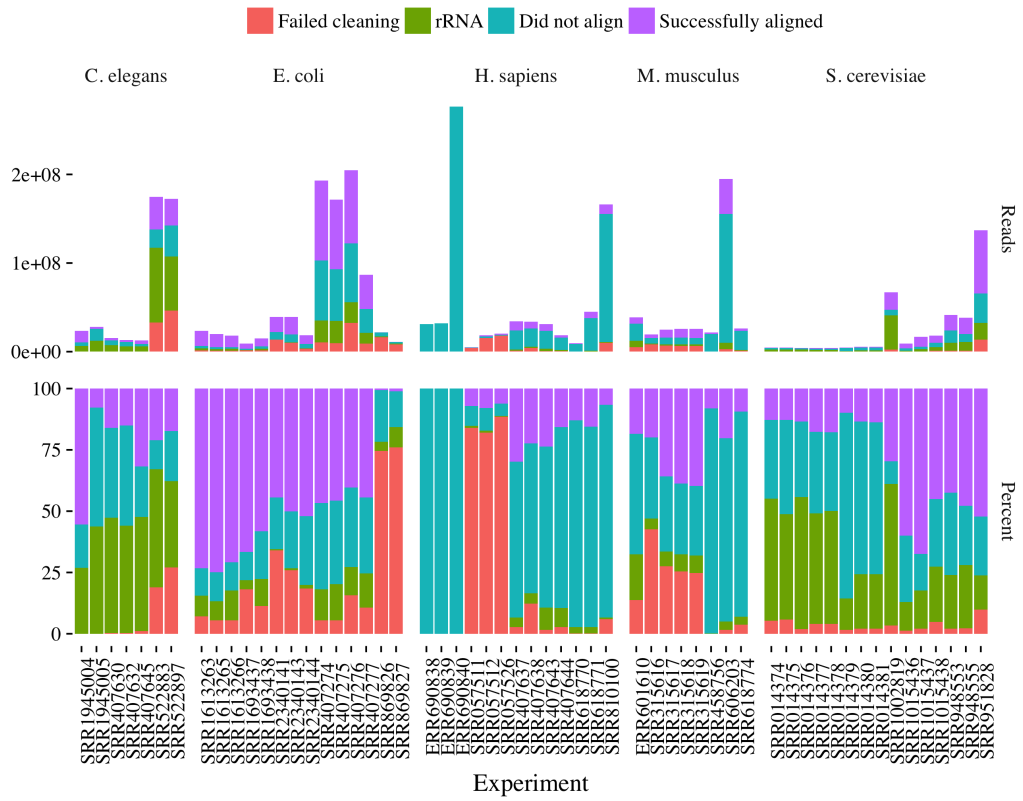


Fig. 3.4 Processing statistics of our ribo-seq database. The number of reads (top), and the fraction they represent (bottom), that were either removed at a processing step or successfully made it through the entire processing protocol. It should be noted samples that contain a large fraction of reads labelled as failing to align to the transcriptome are indicative of the cleaning stages being unsuccessful. The experiments, which are grouped by species, are given along the x-axis and labelled by their respective runIDs. Further details on the different experiments can be found in Table 3.1.

Lastly, two studies made use of barcodes, SRA056377 and SRA082464. The remaining fraction of reads for the former was 24.7%, while the latter was 96.4%.

3.2.4 Aligning reads

The protocol stated above includes two alignments, that to rRNA sequences to remove contaminating reads, and that to the transcriptome to determine the location of translating ribosomes. For both of these alignments, we used Bowtie (v1.1.1) set to default parameters [160]. Bowtie is a sequence alignment tool made specifically for the alignment of short segments of nucleotides

to larger sequences, such as transcripts or genomes. Bowtie was created for use with RNA-seq, and optimised to be both ultrafast and memory-efficient for such alignment problems. The alignment of short nucleotide segments to large ones is equivalent to the alignment problem posed by ribo-seq, and, as such, Bowtie has been adopted for processing the data outputted by ribo-seq experiments as well [e.g., 141–143, 149, 152, 154, 155, 161–165]. Bowtie is not splice aware in order to maximise speed, which means alignments using these tools must be performed against the transcriptome for Eukaryotes, rather than the genome, due to splicing. Unfortunately, splicing results in a transcriptome that features highly redundant sequence segments due to the existence of multiple isoforms. As such, the short nucleotide segments can align to multiple transcripts due to the same exon being used multiple times. Bowtie can be made splice-aware by using the TopHat, a wrapper to Bowtie that as a whole is splice-aware [166]. However, both computational speed and memory usage are sacrificed to achieve this.

For the first of our alignments, we aligned against the ncRNA of the source species as given in Ensembl. The ncRNA set contains the sequence for all rRNA, the main contaminant, as well as tRNA, microRNA, and many others. We did not allow for any mismatches between the reads and the ncRNA to maximise the number of reads taken forward in the protocol. Any read that aligned to this set of sequences was removed. Across the experiments in our database, an average of 31.3% of the reads remaining after cleaning aligned to ncRNA. However, broken down into the experiments associated with each species, we see significant variations. For example, on average 48.5% of cleaned reads in *C. elegans* are identified as being from ncRNA, while only 23.6% for *M. musculus*. These values are lower, in general, than those reported elsewhere, which normally are around 60%, but can approach even 80% [167]. However, this is a reflection of our disallowing of any mismatches in this alignment step, which is non-standard [153, 155].

The second alignment involves the set of cleaned and decontaminated reads being aligned against the transcriptome to give the translating ribosome's locations. As the transcriptome,

we used a modified set of cDNA transcripts based on the respective Ensembl annotations for each species. First, any transcripts associated with mitochondrial genes were removed. Next, the remaining transcripts were extended by pairing the cDNA with the given location in the genome and adding on the proceeding 60 nucleotides to the transcript. For a small fraction of transcripts, we could not ascertain the start codon, and as such, these were discarded from the cDNA set ($< 1\%$). This extension is required to allow reads to align near the 5' end of a transcript, which is a crucial region for determining the ribosome's A and P-site (see below). For the alignment using Bowtie, we allowed up to two mismatches between a read and any transcript. On average 71.4% of the remaining reads successfully aligned, which within the various species ranged from 52.2% in *E. coli* to 88.6% in *H. sapiens*.

3.2.5 Inferring the ribosome P-site

The second alignment gives the location in the transcriptome that the sequence matches that of a given read. From this 20 to 30 nucleotide segment, we would ideally infer the codon undergoing translation by the protecting ribosome at the point of cell lysis. Inferring the translating codon, and, hence, achieving nucleotide level accuracy, has been shown to be achievable even for reads which are not the width of the ribosome (approximately 28 nucleotides). This level of accuracy is partially due to the experimental method trimming reads precisely at the nucleotide immediately upstream of the ribosome in eukaryotes and the nucleotide immediately downstream in prokaryotes. The difference between eukaryotes and prokaryotes is due to the differing digesting enzyme used in their respective experimental protocols. Specifically, MNase, used in ribo-seq experiments involving eukaryotes, has been shown to cut the mRNA tightly at the 5' side of the ribosome and with variability on the 3' side. Similarly, for experiments involving prokaryotes, ribonuclease (RNase) is used, which instead cuts tightly on the 3' side, and with variability on the 5' side.

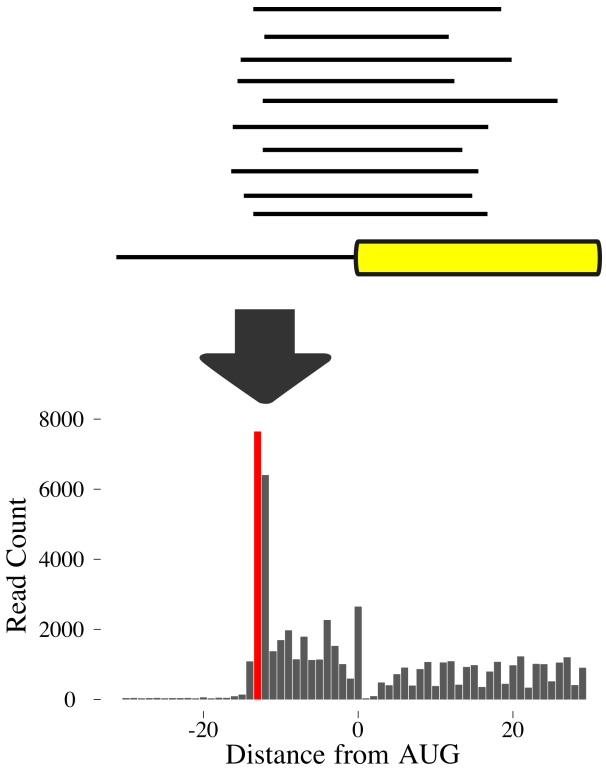


Fig. 3.5 Inferring the ribosome P-site. The likelihood to find a ribosome at the start codon is higher than elsewhere on any given transcript due to the small ribosomal subunit awaiting the association of the large ribosomal unit at this location before translation can begin. This bias results in a significant number of reads within any given ribo-seq experiment being associated with the start codon. Above, we show an example of the start codon enrichment for a eukaryote, in which the 5' end resides on a consistent nucleotide, on the start codon or nearby due to error, while the 3' end is variable. Summating the reads by their 5' nucleotide location produces the plot seen above, which was created using experiment SRR014374 on *S. cerevisiae*. We find a significant peak upstream of the start codon, highlighted in red, which represents the 5' end of the fragments shielded by initiating ribosomes on the start codon. The difference between this location and the start codon is the offset required to infer the first nucleotide of the P-site within the ribosome that protected a given read at the point of cell lysis.

The start codon is typically used as the small ribosomal subunit must await the association of the large ribosomal unit at this location before translation can begin. As such, near the start codon, a peak in the number of aligned reads will be found from which the P-site of the initiating ribosomes can be inferred as the distance between these points.

The tight trimming alone provides nucleotide resolution to the experiment. However, it does not provide enough information to infer the codon undergoing translation from the aligned

location of a given read. The noted accuracy must be combined with a position that ribosomes spend proportionally more time on, such that an increased number of reads compared to the transcript, in general, will be aligned to a nearby point. The start codon is typically used as the small ribosomal subunit must await the association of the large ribosomal unit at this location before translation can begin. As such, near the start codon, a peak in the number of aligned reads will be found from which the P-site of the initiating ribosomes can be inferred as the distance between these points. This distance or offset can then be applied to any aligned read to give P-site of the protecting ribosome. The A-site, the location of the codon undergoing translation, is then three additional nucleotides downstream. We show a model example of this inference for eukaryotes in Figure 3.5.

In our protocol to infer the correct offset, we used the reads that aligned in the extended 5' UTR as well as the first 60 nucleotides of the coding sequence across all transcripts. The transcripts were aligned by their respective start codons and then combined to give a summary translation profile. A single transcript does not provide enough data to infer the offset, hence the need to combine the profiles of many transcripts. Additionally, we base this inference only on reads that aligned to a unique location. We identify a peak in our aligned reads for each experiment from which we can derive an offset to be applied to each read to give the expected P-site. The offset is calculated with respect to the 5' end of an aligned read in eukaryotes, and the 3' end of the aligned read in prokaryotes. The offsets determined for each experiment in our database was given in Table 3.1.

Building a translation profile

Having calculated the appropriate shift to apply to each read to give the associated ribosome's A site, we built translation speed profiles for each mRNA transcript based on the process data from each experiment. First, we took only the reads that aligned to a single location and built an initial profile by assigning a value of one to a nucleotide for each read associated (alignment

+ offset) with it. From the initial profile, a local read density profile was built by applying an average smoother of ten nucleotide width. Reads with multiple alignment locations were then added to the initial profile, the probability of which location they contributed towards proportional to the local read densities at their respective associated nucleotides. After every read had been assigned, the UTR regions were then removed from each profile so that only the protein coding sequence remains. Finally, to ensure that correct assignment of the A-site had occurred, we checked the respective number of counts assigned to the first, second or third nucleotide of codons. If the first nucleotide was not assigned more than the others, we took this as an indication that the wrong offset had been used [168]. In such cases, we shifted the profile so that this criterion was fulfilled.

Database statistics

Above we have outlined how translation profiles were produced for 52 experiments out of the 57 experiments we had selected. For each of the five remaining experiments, the majority of reads failed to be successfully processed and as such translation profiles could not be built. Specifically, these experiments are SRR869826 and SRR869827 from study SRA082464 on *E. coli*, and experiments ERR690838, ERR690839, and ERR690840 from study ERA390450 on *H. sapiens*. The 52 successfully processed experiments represent a total of 2.4 billion reads which have a combined FASTA file size of 260GB. Out of these, we successfully aligned 777 million reads (32%). The mean depth and coverage for each of the successfully processed datasets were calculated, and the respective results are shown in Figure 3.6. The depth is the average number of reads per nucleotide on a transcript, while the coverage is the fraction of nucleotides on a transcript that one or more reads associated to. Both these measures give insight into the quality of the experiments within our database, though are also dependent on the conservativeness of processing method with regards to cleaning and aligning. As seen from the accompanying figure, the depth and the coverage are extremely related to one another.

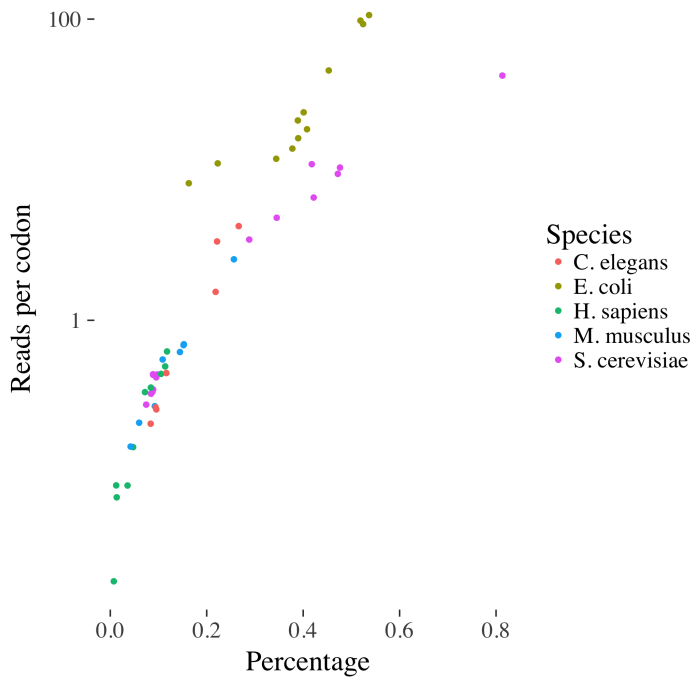


Fig. 3.6 Quality statistics of our ribo-seq database. The average depth (y-axis) and coverage (x-axis) calculated across all transcripts for each experiment in our database is shown where we have coloured the points by species. The depth is the average number of reads per nucleotide on a transcript, while the coverage is the fraction of nucleotides with which one or more reads associated to it. Both these measures give insight into the quality of the experiments within our database.

3.3 Other ribo-seq databases

3.3.1 Overview

GWIPS-viz

The Genome-Wide Information on Protein Synthesis visualization [sic] (GWIPS-viz) database contains ribo-seq experiments for various species which are available for viewing alongside data from RNA-seq [153]. The GWIP-viz database was released in 2014, and since then has steadily grown, both in terms of the number of experiments contained within it and the range of species that are represented. As of October 2016, there are hundreds of ribo-seq experiments available that cover 26 diverse species including eukaryotes, prokaryotes, and viruses. For each

experiment contained in the database, read counts along the whole genome are available for direct download in BigWig format. The user can specify at the point of download whether the reads count should be based on the initial nucleotide of their alignment, or whether the counts should be at the inferred A or P-site of the ribosome that protected the read. For each study, a global aggregate file consisting of the summation of reads from all experiments undertaken is also available. Likewise, for each species, a global aggregate of the reads from all studies is given.

As with our database, the GWIPS-viz database reprocesses all experiments contained within it such as to remove any biases that may exist for a given study. First, the reads are cleaned by trimming adapter sequences and poly(A) tails from the 3' end of each read where required. Trimming was performed using Cutadapt version 1.1. Reads shorter than 25 nucleotides after cleaning were discarded. Cleaned reads were then aligned against the rRNA reference sequences using Bowtie to identify contaminating ncRNA. Three mismatches were allowed in the alignment between the read sequence and the rRNA sequence. Reads that successfully aligned to the rRNA were removed. Finally, the remaining reads were aligned against the genome and the transcriptome using RUM, the parameters of which are not stated (see below for details regarding RUM). Only reads that aligned to a unique location in either alignment contribute to the counts available to download in the GWIPS-viz database. As mentioned above, these counts are based on the 5' end of the alignments. For eukaryotes, the ribosome P-site associated with a fragment is 12 nucleotides from the 5' end of a read, and, hence, the ribosome A-site is 15 nucleotides. For prokaryotes, these locations are inferred from the 3' end of the alignment, namely the ribosome P-site is 15 nucleotides from the 3' end and the A-site is 12 nucleotides.

RPFdb

The Ribosome Protected mRNA Fragments database (RPFdb) was released in 2015 and contains 777 ribo-seq samples sourced from 82 studies across eight species as of October 2016 [154]. For each sample, the Reads Per Kilobase of transcript per Million (RPKM) for each gene is available for direct download. The RPKM is essentially a measure of the density of reads on a given transcript and is closely related to the depth. In addition to the RPKM values, the RPFdb also allows a user to view the counts of aligned reads at each nucleotide along the corresponding genome via their genome browser. No information is given as to whether the counts are based on the 5' or 3' end of the aligned reads, or instead based on the inferred ribosome A or P-sites. A genome browser allows one to compare the counts from the various studies alongside other genome annotations such as the location of genes. These counts are reported in the accompanying paper as being normalised, though it is not stated either in the paper or online the manner in which this is done. From the genome browser, one can download the counts in a variety of formats, though only for small gene length snippets of the genome at once, thereby making it infeasible for genome-wide analysis. However, a publicly accessible folder can be found at sysbio.sysu.edu.cn/rpf_data_bam/ from which genome-wide normalised counts can be downloaded.

Akin to both ours and the GWIPS-viz database, the samples within the RPFdb have been reprocessed so as to remove any inherent study based processing bias. First, the raw reads were trimmed to only the first 26 nucleotides, additional nucleotides at the 3' end presumed to be either from a linker or a poly-A tail. No mention is given as to the treatment of reads shorter than 26 nucleotides. Once trimmed, reads were then aligned to the genome using STAR, an alternative alignment software to Bowtie. One mismatch was allowed in the alignment of a read to the genome. No further description is given as to how a read aligned to multiple locations is handled. As such, we assume that each aligned location is treated in a manner indistinguishable to those from reads that uniquely aligned, i.e., each location attributed as

having a read associated to it. Lastly, any sample in which the number of uniquely mapped reads was less than one million was discarded. This criterion means that not all experiments of a given study within the RPFdb may be present.

3.3.2 Processing differences

GWIPS-viz

There are some minor differences in the processing between the GWIPS-viz database and our own. First, the GWIPS-viz database enforces a minimum length threshold of 25 nucleotides for a read, as opposed to the 20 nucleotide limit we enforced. Subsetting to longer reads means that a higher fraction will align uniquely to the genome or transcript later in the protocol. However, this preference for a unique alignment is partially redundant as they discard any reads that align to multiple locations. The more stringent threshold also reduces the computational complexity later on, as fewer reads will be considered in the alignment against the genome, likely quicking the overall pipeline. Another small difference is allowing three mismatches between the reads and the ncRNA contaminants, whereas we allowed none. We choose our threshold to maximise the number of reads taken forward to the alignment stage, whereas they have favoured minimising contaminating reads. As just mentioned, the removal of more reads reduces the computational complexity in the final alignment, again quickening the overall pipeline.

The most significant difference between our protocol and that used to the GWIPS-viz database is the use of RUM, a splice-aware alignment tool, over Bowtie, in order give the read counts along the length of the genome rather than the transcriptome [169]. RUM, created in 2011, uses Bowtie first to align against both the transcriptome and the genome independently, before merging the two alignments. Any reads that failed to align are then attempted again using the Blast-like Alignment Tool (BLAT) [170]. In comparison to our alignment, performing splice-aware alignment allows different transcripts that share an exon to be handled more rigorously.

Namely, they will not be reported as aligning to multiple locations. We overcame this issue using our proportional random assignment method discussed previously. The advantage of our approach is that it separates, albeit heuristically, the reads associated with a shared exon to each of the transcripts that contain it. In contrast, aligning against the genome may lead to scenarios in which genes that undergo splicing may contain significant jumps or drops in their read counts due to exons being used in different ratios. For example, consider a gene with exons A, B, and C, that encodes for two splice variants, ABC and AC. In such a scenario, exon B will have fewer ribosomes associated with it than either exon A or C.

The remaining difference is the simplification of the ribosome A and P-site offset to standard values for eukaryotes and prokaryotes, rather than being calculated independently for each experiment as per our methodology. The chosen offsets are roughly inline (± 1 nucleotide) with the values that emerged when we performed our offset analysis (see Table 3.1). As mentioned above, the validity of an offset choice can be measured by analysing the periodicity of the counts as the majority of inferred A-sites should be located on the first nucleotide of a codon if correct [168]. There is no indication that this check was performed during the construction of the GWIPS-viz database.

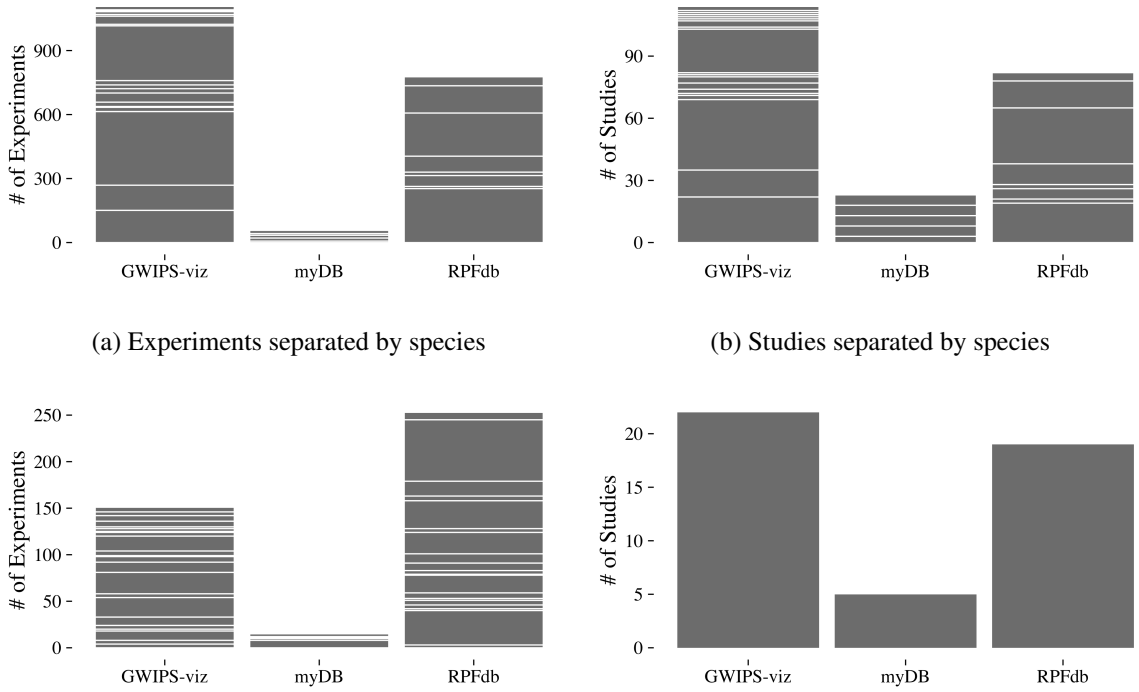
RPFdb

The protocol used to create the RPFdb includes many differing steps to those chosen to create our own, some of which we express apprehension at their usage. Our most prominent concern is the lack of any alignment of reads against rRNA such as to remove contaminants from a given sample. This step is standard procedure in nearly all ribo-seq protocols as up to 90% of the reads in any given sample may be from these contaminating sources [167]. We highlighted these contaminations previously in Figure 3.4, where we showed that between 25% to 50% of the cleaned reads in each experiment processed was classified as a contaminant. Not considering the significant increase in the computational complexity, one explanation for the RPFdb not

explicitly removing these contaminants is that they align their reads against the genome, making use of the splice-aware alignment provided by the aligner STAR. We discussed the pros and cons of using splice-aware aligners above with regards to the RUM aligner used to construct the GWIPS-viz database. By aligning against the genome, rather than the transcriptome, reads can align to the source rRNA genes of contaminants and, as such, an independent removal step in the protocol may not be needed. However, any protein coding gene which contains a segment of a similar sequence to the contaminants will be attributed many false reads. Conversely, if the contaminants were removed, the same gene would have the genuine reads sourced from this segment labelled as contaminants and removed. However, the advantage of the latter case over the former is that there is orders of magnitude difference in the number of mislabelled reads. Furthermore, these mislabelled reads are not randomly distributed, they will drastically alter the read counts for various genes and, as such, minimising mislabelled reads should be considered paramount.

Another issue is the trimming of each read to 26 nucleotides, an arbitrary threshold that could result in a significant amount of data loss. Specifically, it hinges on whether the creators discard reads shorter than 26 nucleotides, or if shorter reads are included. If the former, we can draw from our processing that up to 40% of reads may have been removed unnecessarily. If the latter, it begets the question of whether they enforced a minimum length threshold, and if so, for what length? Further to this, it is unclear why identification of linkers or poly(A) tails was not attempted. Neither is complex and would result in the substantially more accurate trimming of reads than the approach taken constructing the RPFdb. Only if these methods failed should a coarse trimming, such as the 26 nucleotide threshold used here, be considered.

Lastly, the RPFdb treats the multiple alignments of a single read as the equivalent to those reads that uniquely aligned. Such an assumption is inherently flawed as it attributes more value to reads that lack specificity than those with which the exact location can be determined. This outcome is the opposite to what should be sought. Additionally, and in conjunction



(a) Experiments separated by species

(b) Studies separated by species

(c) S. cerevisiae experiments separated by study

(d) S. cerevisiae studies

Fig. 3.7 Comparison of the size and diversity of each database. The number of experiments and the number of studies are given, both for all species combined, and only for those performed on *S. cerevisiae*, our species of interest for further work. The white lines splitting the various bars shows the contribution from the different species or studies in each database as noted by the subcaptions.

with the previous critical point, if reads shorter than 26 nucleotides are included, due to a lack of minimum length threshold, then these smaller reads are far more likely to align to multiple locations across a genome. As such, these short reads will artificially inflate the counts significantly.

3.3.3 Database comparison

To objectively compare the database sizes, we compiled the number of studies and experiments in each as well as noting the source species. The results are given in Figure 3.7, where we show the breakdown both in terms of the number of studies and experiments overall. We also

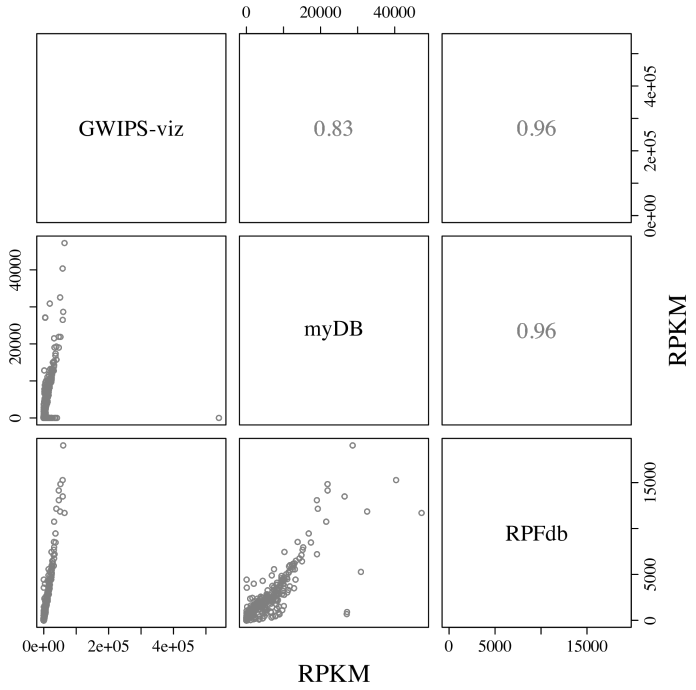


Fig. 3.8 Comparison of the Ingolia 2009 *S. cerevisiae* in rich conditions ribo-seq dataset. The databases are compared pairwise, the RPKM values for each gene within the dataset plotted in the lower panels with the corresponding Spearman correlations given in the upper panels. Genes not present in both datasets are removed from the corresponding comparison. As both our database (myDB) and the RPFdb contained multiple experiments labelled as *S. cerevisiae* in rich conditions, while the GWIPS-viz database contained only one, the RPKM values for both of the former were calculated via taking the mean across the relevant datasets.

show the breakdown when only the experiments performed on *S. cerevisiae* are considered as this becomes our species of interest for the work undertaken in future chapters. We find that across all of these breakdowns, the GWIPS-viz database performs the best, containing far more studies and experiments than either our database or the RPFdb.

Next, to check the quality of the databases with regards to the respective processing, we compared a dataset that was present in each. Shown in Figure 3.8, we compared a subset of experiments from the Ingolia 2009 study, namely, only the experiments associated with the rich, non-starved, condition. We compared the RPKM values of the genes rather than the explicit profiles as we could not determine the how the count for each codon had been assigned for the RPFdb (see above). The relevant experiments from both our database and the RPFdb

were averaged by gene, whereas the GWIPS-viz database provides only a single file that we presumed was the average of the relevant experiments. Genes not present in both datasets are removed from the corresponding comparison. From the Spearman correlations, we note that our database and the GWIPS-viz database appear the most dissimlar ($\rho = 0.83$), with RPFdb correlating well to both ($\rho = 0.96$ and $\rho = 0.96$ respectively). Regardless, the level of correlation between each of the databases is such that we cannot identify any whose processing method may be flawed at the gene-level.

3.4 Conclusion

In summary, we created a database containing a considerable amount of ribo-seq data which was reprocessed to remove as much inherent study based bias as possible. We have given an overview of our protocol and discussed various considerations that should be taken into account when working with ribo-seq. Over the period with which we undertook this task, two other similar databases were produced, the GWIPS-viz database and the RPFdb. We discuss the differences between these databases and our own at length before comparing the data contained in each. We compared them both in term of the number of experiments contained and the similarity of the data once processed. From this, we concluded that the GWIPS-viz database was constructed in a comparable, if not better, manner to our own database. Furthermore, the GWIPS-viz database dwarfed our own construction both in terms of the number of studies it contained and the diversity of species present. In contrast, the RPFdb was found to be the least rigorous in the manner with which it processed data. Given this, for our future work, we decided to source processed ribo-seq datasets from the GWIPS-viz database rather than continue research with our own.

Chapter 4

Identifying reproducible ribo-seq data at the codon-level

4.1 Introduction

In the previous chapter, we introduced ribo-seq, an experimental protocol that involves the targeted transcriptome-wide sequencing of only the mRNA protected by a ribosome at the point of cell lysis [135]. From the alignment of these protected fragments, we can infer the positions of actively translating ribosomes, from which we can then estimate the relative translation speed of the codons on a given mRNA transcript. We created an extensive database of these experiments performed on *S. cerevesiae* and compared it two other recently developed ribo-seq databases, the GWIPS-viz database and the RPFdb. The GWIPS-viz database was deemed to have a comparable, if not better, pipeline to process ribo-seq data from its raw state than that used to construct our database. In addition, it also contained a large number of datasets, both on *S. cerevesiae* alone, and more so when all other species are considered. Given this, we concluded that it was advantageous to continue our research using the GWIPS-viz database. With the GWIPS-viz database supplying a suitable extensive collection of ribo-seq experiments, and hence translation profiles, we next wanted to compare these profiles to the theoretical

estimators of translation speed. The work in this chapter follows on from Chapter 2, where we found that there was little similarity in the predictions given by the different estimators. Ribo-seq provides the first experimental transcriptome-wide objective measure by which these estimators can be compared.

Comparison of the translation profiles produced by ribo-seq to those predicted by theoretical estimators, as well as other properties of the mRNA sequence, has been performed previously [137, 138, 142, 147, 150, 151]. Such comparisons provide insight into the determinants of ribosomal velocity. For example, the concentration of cognate tRNA was found not to correlate to the average ribo-seq count for each codon type within *E. coli*, suggesting that the association of tRNA with the ribosome is not a bottleneck to translation [142]. However, the authors noted that selectively starving the *E. coli* of specific amino acids caused the observed ribo-seq counts found on the corresponding codon throughout the transcriptome to increase significantly. Instead, they reported that a major determinant of the translation speed within prokaryotes was the presence of anti-Shine-Dalgarno sequences. This result was endorsed by another study performed on *S. cerevisaie* that found that ribo-seq counts correlate neither to the cognate tRNA concentrations nor the stability of the mRNA secondary structure [151]. In addition, they reported that codons that encode for positively charged amino acids cause ribosomal pausing, suggesting that the amino acid interacts with the ribosome tunnel to retard translation. Dana and Tuller performed a comprehensive set of comparisons of ribo-seq translation profiles to various theoretical estimators across a range of species in a series of papers [137, 138, 147]. While they concurred that the average ribosome count does not correlate with the tRNA concentration or the codon usage bias, as given by the tAI and CAI respectively, they found that other features, most notably the skewness of the respective distributions for each codon type, do correlate with these estimators [147]. Given this, they showed that if the distribution of ribo-seq counts for a given codon type are modelled as an exponentially modified Gaussian, a direct correlation could be obtained between the Gaussian component and the cognate tRNA concentration [137, 138].

In each of the above studies, ribo-seq was described and then used as if it achieves codon-level resolution. Analysis of RNA-seq has shown that each fragment can be successfully traced back, via sequencing, to the high-quality source transcript [140]. Furthermore, analysis of ribo-seq has shown that from a fragment's aligned position on a transcript, the location of the A-site of the ribosome that protected the fragment at the point of cell lysis can be inferred [135, 171]. Specifically, the A-site is approximately 15 nucleotides from the 5' end of the fragment in Eukaryotes, and 12 nucleotides from the 3' end of the fragment in Prokaryotes [135, 171]. As shown in the construction of our database in the previous chapter, these values often vary slightly between different species and studies. Recently, however, the reproducibility of ribo-seq has been called into questioned. First, a comparison of various ribo-seq datasets in 2014 showed that technical sequence biases exist in the generation of fragments such that the 5' end exhibits a bias for particular nucleotides [161]. When this bias was accounted for, the incorporation of proline into the emergent protein was reported to be the major determinant of ribosomal velocity over the factors discussed above. Further differences, above technical biases, were highlighted by a study that compared the translation profiles of replicates from the same study at the codon-level [163]. They found that while the total number of fragments associated with each transcript is consistent, the aligned locations of the fragments along the transcript differed, especially for transcripts that were under-sampled. In other words, the translation profile changed shape between replicates.

This finding calls into question the previous results based upon ribo-seq data at the codon-level. It also may be the source of some contention currently in the literature discussed above, whereby different ribo-seq experiments have reported differing results [137, 138, 142, 145, 147, 151]. In this chapter, we analyse the similarity between a total of 107 experiments sourced from 15 different studies on *S. cerevisiae* at both the transcript-level and the codon-level. At the transcript-level, we find that most ribo-seq experiments appear to be extremely similar to one another, thereby indicating that the data is reproducible. However, when compared at

the codon-level, we find that different ribo-seq experiments are rarely in agreement. Upon further investigation, we found that transcripts that were highly sampled were, in general, able to achieve a higher degree of codon-level similarity. Given this finding, we created a simple statistical model to predict whether a given transcript was adequately sampled and used this to establish a threshold above which a ribo-seq translation profile is likely to be reproducible. We then used our codon-level comparisons between the experiments to prune the collated data of all transcripts that were not reproduced in multiple studies. Merging the remaining data, we calculate high-quality reproducible translation profiles for 2601 out of the 6692 transcripts expressed by *S. cerevisiae* (38.87%). Finally, we compared these high-quality translation profiles at the codon-level to those predicted by the most well-known theoretical translation speed estimators.

4.2 Material and methods

4.2.1 GWIPS-viz ribo-seq datasets

The GWIPS-viz database, which was discussed in depth in the previous chapter, contains various ribo-seq experiments processed in a systematic manner to remove some study specific biases. For this work, all ribo-seq datasets available as of 14^{th} July 2016 for *S. cerevisiae* were downloaded from the GWIPS-viz website as BigWig files [153]. These files contained the counts of reads aligned to each nucleotide across the whole genome. The counts are based on the inferred A-site of the ribosomes associated with each read. Aggregate datasets, both global and study specific ones, were discarded, leaving only those of the individual experiments within each study. Counts were extracted using bwtools for the cDNA region of each mRNA transcript as given by the sacCer3 Ensembl annotation of *S. cerevisiae* [172, 173]. Reduction to only the coding sequence (CDS) was then performed using Python. Profiles were changed from nucleotide-level resolution to codon-level resolution; each codon assigned the average of

the counts from its three constituent nucleotides. In total, we collated 107 ribo-seq experiments sourced from 15 independent studies. We list the individual experiments and their source studies in Table 4.1 where we have also noted the total number of reads, which experiments can be considered control or wild-type in their respective studies, and whether the experimental protocol involved pretreatment by cycloheximide. Pretreatment by cycloheximide has been shown to alter the observed translation profiles [157]. Further detail on the studies and the ribo-seq processing can be found both in the previous chapter and at the GWIPS-viz website (http://gwips.ucc.ie/).

4.2.2 Ribo-seq depth

We define the depth as the average number of counts per codon on a given transcript. The depth is also used as a proxy for the relative expression of transcripts and, as such, is used in the normalisation of the translation profile (see below).

4.2.3 Ribo-seq coverage

We define the coverage as the percentage of codons along the length of a given transcript with non-zero counts, i.e. had one or more reads assigned to them.

4.2.4 Transcript-level correlation

To test the similarity between datasets at the transcript-level, we compare ribo-seq depths of all transcripts using the Spearman's rank correlation.

4.2.5 Codon-level correlation

To test the similarity at the codon-level, we compare the fragment counts along the same transcript using the Spearman's rank correlation. As this gives a correlation for each transcript,

Table 4.1 Studies and their experiments collated to produce our dataset. The ID label is used to refer to specific studies throughout this text. The Study and Experiment labels are those used on the GWIPS-vis website. Control and Pre.CHX labels refer to the methodology employed to generate each of the datasets. The Read label is the total read count for each dataset and acts as an estimate of their relative size.

ID	PMID	Study	Experiment	Control	Pre.CHX	Reads
0	25340754	Albert 2014	BY_FP_RiboProElong	Y	Y	73230195
1	25340754	Albert 2014	hybrid_FP1_RiboProElong	N	Y	32536787
2	25340754	Albert 2014	hybrid_FP2_RiboProElong	N	Y	43528721
3	25340754	Albert 2014	RM_FP_RiboProElong	Y	Y	134723353
4	25043188	BaudinBail 2014	Cter_RP	N	N	78810156
5	25043188	BaudinBail 2014	PSi+_YPD_RP	N	N	10771609
6	25043188	BaudinBail 2014	psi-_YPD_RP	Y	N	8355506
7	25043188	BaudinBail 2014	PSI_RP	N	N	13929222
8	22194413	Brar 2012	RiboPro_A14201	N	Y	12781314
9	22194413	Brar 2012	RiboPro_gb15	Y	Y	29328970
10	22194413	Brar 2012	RiboPro_Mata_a	N	Y	4568610
11	22194413	Brar 2012	RiboPro_ndt80	N	Y	25783776
12	22194413	Brar 2012	RiboPro_Trad	N	Y	41310878
13	23045643	Ger 2012	RiboPro_30minperoxide	N	Y	50255901
14	23045643	Ger 2012	RiboPro_5minperoxide	N	Y	54384623
15	23045643	Ger 2012	RiboPro_Untreated	Y	Y	78822141
16	25056308	Gerashchenko 2014	aurin_noCHX_RiboProElong	N	N	5279082
17	25056308	Gerashchenko 2014	complete_SD_media_RiboProElong	Y	Y	7557305
18	25056308	Gerashchenko 2014	edeine_noCHX_RiboProElong	N	N	1934584
19	25056308	Gerashchenko 2014	heatshock_1x_CHX_RiboProElong	N	Y	2275312
20	25056308	Gerashchenko 2014	heatshock_noCHX_RiboProElong	N	N	2805953
21	25056308	Gerashchenko 2014	oxidative_100x_CHX_RiboProElong	N	Y	23026398
22	25056308	Gerashchenko 2014	oxidative_1_16x_CHX_RiboProElong	N	Y	4478391
23	25056308	Gerashchenko 2014	oxidative_1_4x_CHX_RiboProElong	N	Y	5711861
24	25056308	Gerashchenko 2014	oxidative_1_64x_CHX_RiboProElong	N	Y	2076799
25	25056308	Gerashchenko 2014	oxidative_1x_CHX_RiboProElong	N	Y	4959083
26	25056308	Gerashchenko 2014	oxidative_8x_CHX_RiboProElong	N	Y	7199976
27	25056308	Gerashchenko 2014	oxidative_noCHX_RiboProElong	N	N	9225501
28	25056308	Gerashchenko 2014	SD_no_aminoacids_RiboProElong	N	Y	1705285
29	25056308	Gerashchenko 2014	SD_no_aminoacids_plus_His_Met_Leu_RiboProElong	N	Y	4904299
30	25056308	Gerashchenko 2014	small_subunit_RiboProElong	N	Y	18609
31	25056308	Gerashchenko 2014	unstressed_100x_CHX_RiboProElong	Y	Y	22241145
32	25056308	Gerashchenko 2014	unstressed_1_16x_CHX_RiboProElong	Y	Y	5825908

Continued overleaf

Table 4.1 Studies and their experiments collated to produce our dataset. The ID label is used to refer to specific studies throughout this text. The Study and Experiment labels are those used on the GWIPS-vis website. Control and Pre.CHX labels refer to the methodology employed to generate each of the datasets. The Read label is the total read count for each dataset and acts as an estimate of their relative size.

ID	PMID	Study	Experiment	Control	Pre.CHX	Reads
33	25056308	Gerashchenko 2014	unstressed_1_4x_CHX_RiboProElong	Y	Y	5517306
34	25056308	Gerashchenko 2014	unstressed_1_64x_CHX_RiboProElong	Y	Y	1522434
35	25056308	Gerashchenko 2014	unstressed_1x_CHX_RiboProElong	Y	Y	5896294
36	25056308	Gerashchenko 2014	unstressed_8x_CHX_RiboProElong	Y	Y	6051593
37	25056308	Gerashchenko 2014	unstressed_noCHX_RiboProElong	Y	N	21484928
38	24581494	Guydosh 2014	Guydosh14_dom34K0_short_footprints_RiboPro	N	N	177945
39	24581494	Guydosh 2014	Guydosh14_dom34KO_3_AT_RiboPro	N	N	50691338
40	24581494	Guydosh 2014	Guydosh14_dom34KO_CHX_RiboPro	N	N	10339544
41	24581494	Guydosh 2014	Guydosh14_dom34KO_CHX_GMP_PNP_RiboPro	N	N	3514336
42	24581494	Guydosh 2014	Guydosh14_dom34KO_diamide_RiboPro	N	N	8598612
43	24581494	Guydosh 2014	Guydosh14_dom34KO_disome_footprints_RiboPro	N	N	289565
44	24581494	Guydosh 2014	Guydosh14_dom34KO_glucose_starvation_RiboPro	N	N	3214786
45	24581494	Guydosh 2014	Guydosh14_dom34KO_GMP_PNP_RiboPro	N	N	4937063
46	24581494	Guydosh 2014	Guydosh14_dom34KO_high_Mg_RiboPro	N	N	5618059
47	24581494	Guydosh 2014	Guydosh14_dom34KO_no_additive_RiboPro	N	N	12903396
48	24581494	Guydosh 2014	Guydosh14_dom34KO_suppressor_tRNA_RiboPro	N	N	8657504
49	24581494	Guydosh 2014	Guydosh14_hbs1KO_RiboPro	N	N	9956273
50	24581494	Guydosh 2014	Guydosh14_ski2Ko_RiboPro	N	N	5703615
51	24581494	Guydosh 2014	Guydosh14_wild_type_3_AT_RiboPro	N	N	11819215
52	24581494	Guydosh 2014	Guydosh14_wild_type_CHX_RiboPro	Y	N	11114275
53	24581494	Guydosh 2014	Guydosh14_wild_type_CHX_GMP_PNP_RiboPro	N	N	5908352
54	24581494	Guydosh 2014	Guydosh14_wild_type_diamide_RiboPro	N	N	8271669
55	24581494	Guydosh 2014	Guydosh14_wild_type_disome_footprints_RiboPro	N	N	158002
56	24581494	Guydosh 2014	Guydosh14_wild_type_GMP_PNP_RiboPro	N	N	3135442
57	24581494	Guydosh 2014	Guydosh14_wild_type_high_Mg_RiboPro	N	N	9032145
58	24581494	Guydosh 2014	Guydosh14_wild_type_no_additive_RiboPro	Y	N	13488935
59	24581494	Guydosh 2014	Guydosh14_wild_type_short_footprints_RiboPro	N	N	209123
60	24581494	Guydosh 2014	Guydosh14_wild_type_suppressor_tRNA_RiboPro	N	N	53386401
61	19213877	Ing 2009	RiboPro_Rich	Y	Y	3672275
62	19213877	Ing 2009	RiboPro_Starved	N	Y	2354422
63	24842990	Lareau 2014	3AT_treatment_RiboProElong	Y	N	504888
64	24842990	Lareau 2014	Anisomycin_RiboProElong	Y	N	235164
65	24842990	Lareau 2014	Anisomycin_1A_RiboProElong	Y	N	516870

Continued overleaf

Table 4.1 Studies and their experiments collated to produce our dataset. The ID label is used to refer to specific studies throughout this text. The Study and Experiment labels are those used on the GWIPS-vis website. Control and Pre.CHX labels refer to the methodology employed to generate each of the datasets. The Read label is the total read count for each dataset and acts as an estimate of their relative size.

ID	PMID	Study	Experiment	Control	Pre.CHX	Reads
66	24842990	Lareau 2014	Anisomycin_1B_RiboProElong	Y	N	843607
67	24842990	Lareau 2014	Cycloheximide_RiboProElong	Y	Y	4804865
68	24842990	Lareau 2014	Untreated_RiboProElong	Y	N	4414775
69	26887592	Nissley 2016	yeast_S288C_Riboseq_RiboProElong	Y	N	34626016
70	25538139	Pop 2014	RP_Pop_ACA_K_profile	N	N	486539
71	25538139	Pop 2014	RP_Pop_AGG_OE_profile	N	N	10568242
72	25538139	Pop 2014	RP_Pop_AGG_QC_profile	N	N	6853948
73	25538139	Pop 2014	RP_Pop_WT_profile	Y	N	38127378
74	25538139	Pop 2014	RP_Pop_WT_URA_profile	N	N	9976768
75	26122911	Sen 2015	ribo_ded1cs_1_15_deg_RiboProElong	N	Y	2150626
76	26122911	Sen 2015	ribo_ded1cs_2_15_deg_RiboProElong	N	Y	3886102
77	26122911	Sen 2015	ribo_ded1ts_1_37_deg_RiboProElong	N	Y	5401976
78	26122911	Sen 2015	ribo_ded1ts_2_37_deg_RiboProElong	N	Y	4472599
79	26122911	Sen 2015	ribo_tif1ts_1_30_deg_RiboProElong	N	Y	2699124
80	26122911	Sen 2015	ribo_tif1ts_1_37_deg_RiboProElong	N	Y	1860604
81	26122911	Sen 2015	ribo_tif1ts_2_30_deg_RiboProElong	N	Y	1574701
82	26122911	Sen 2015	ribo_tif1ts_2_37_deg_RiboProElong	N	Y	1979158
83	26122911	Sen 2015	ribo_wildtype_DED1_1_15_deg_RiboProElong	N	Y	5260920
84	26122911	Sen 2015	ribo_wildtype_DED1_1_37_deg_RiboProElong	Y	Y	4517928
85	26122911	Sen 2015	ribo_wildtype_DED1_2_15_deg_RiboProElong	N	Y	6445534
86	26122911	Sen 2015	ribo_wildtype_DED1_2_37_deg_RiboProElong	Y	Y	4641195
87	26122911	Sen 2015	ribo_wildtype_TIF1_1_30_deg_RiboProElong	N	Y	2019288
88	26122911	Sen 2015	ribo_wildtype_TIF1_1_37_deg_RiboProElong	Y	Y	6458020
89	26122911	Sen 2015	ribo_wildtype_TIF1_2_30_deg_RiboProElong	N	Y	3254595
90	26122911	Sen 2015	ribo_wildtype_TIF1_2_37_deg_RiboProElong	Y	Y	4441940
91	24476825	Subtelny 2014	Cerevisiae_RPF_RiboProElong	Y	N	296844
92	26276635	Young 2015	Young15_rli1_depletion_RiboPro	N	N	40991670
93	26276635	Young 2015	Young15_rli1_depletion_3AT_RiboPro	N	N	40663607
94	26276635	Young 2015	Young15_rli1_depletion_dom34_RiboPro	N	N	7056569
95	26276635	Young 2015	Young15_rli1_high_dom34_RiboPro	N	N	14416026
96	26276635	Young 2015	Young15_wild_type_RiboPro	Y	N	26343644
97	26276635	Young 2015	Young15_wild_type_3AT_RiboPro	N	N	37434430
98	25119046	Zid 2014	BYRibo_Glu15_RiboProElong	N	Y	2510683

Continued overleaf

Table 4.1 Studies and their experiments collated to produce our dataset. The ID label is used to refer to specific studies throughout this text. The Study and Experiment labels are those used on the GWIPS-vis website. Control and Pre.CHX labels refer to the methodology employed to generate each of the datasets. The Read label is the total read count for each dataset and acts as an estimate of their relative size.

ID	PMID	Study	Experiment	Control	Pre.CHX	Reads
99	25119046	Zid 2014	BYRibo_logphase_RiboProElong	Y	Y	7204733
100	25119046	Zid 2014	EYRibo_Glu15_RiboProElong	N	Y	2728017
101	25119046	Zid 2014	EYRibo_logphase_RiboProElong	Y	Y	1083510
102	26492405	Zinshteyn 2013	delta_elp3_Ribosome_Footprint_RiboProElong	N	Y	7288288
103	26492405	Zinshteyn 2013	delta_ncs2_Ribosome_Footprint_RiboProElong	N	Y	6135281
104	26492405	Zinshteyn 2013	delta_ncs6_Ribosome_Footprint_RiboProElong	N	Y	14777404
105	26492405	Zinshteyn 2013	delta_uba4_Ribosome_Footprint_RiboProElong	N	Y	15930646
106	26492405	Zinshteyn 2013	WT_Ribosome_Footprint_RiboProElong	Y	Y	16654618

we then take the median value out of all gene-gene comparisons to summarise the correlation between two experiments. To test whether the Spearman's rank correlation was applicable, we calculated the autocorrelation for various lags to establish the independence of codons on the same transcript. We found no significant long-range interactions; hence, treating each codon as an independent measurement is valid.

4.2.6 Smoothed ribo-seq translation profiles

Smoothed ribosome profiles were generated by applying an averaging kernel of five codon-width across each transcript. The first and last two codons of each transcript were trimmed from the smoothed profiles.

4.2.7 Thresholded ribo-seq translation profiles

Thresholded datasets are generated by enforcing a coverage threshold of 50% to remove sparse transcripts, i.e., half of the codons on any given transcript must have a non-zero count.

4.2.8 Model training and testing

Models were built to predict whether a given codon-level comparison of a single transcript would achieve a high correlation. As predictors, we used the lower value for both the depth and coverage of the compared transcripts. For example, if a transcript had a depth and coverage of 20 reads per codon and 60% in study 1, and 18 reads per codon and 70% in study 2, then the values taken would be 18 reads per codon and 60% for the depth and coverage respectively. Logit models were built in R using the "speedglm" package and trained on a biased sampling of the codon-level comparisons in which all the positive classifications were taken alongside an equal number of negative classifications that were selected randomly. The unbiased data has a prevalence of 5.052% and 1.548% in comparison of studies within and outwith of the same study respectively, while this biased sample has a prevalence of 50%. To determine whether a

given model was influenced by the selection of the biased training set, we retrained models using nine other random samplings. As shown in Figure 4.4, if the confidence intervals of the coefficients were found to overlap, the model was deemed independent of the training set and stable. Optimal models were then calculated by averaging the coefficients of the ten models. Testing was performed against all codon-level comparisons, though those within and outwith of the same study were treated independently, and the F1 score used to select the optimal cutoff. The F1 score is a measure of a model's accuracy that gives equal weighting to both the precision, the fraction of true positives out of those labelled positive, and the recall, the fraction of all positives correctly identified. Mathematically, it is given as

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{4.1}$$

4.2.9 Existing translation speed estimators

We predicted the translation profile of transcripts using various traditional theoretical estimators, namely the Codon Adaption Index (CAI), the tRNA Adaption Index (tAI), the MinMax algorithm and the normalised Translational Efficiency (nTE) scale [104, 105, 116, 117]. Each algorithm is given in full within Chapter 2. The codon bias for CAI, MinMax and nTE was calculated using the CDS regions of *S. cerevesiae* transcripts specified in the sacCer3 Ensembl annotation [173]. The relative expression of each transcript needed for the calculation of CAI and nTE was taken as the relative ribo-seq depth, this having been shown to correlate strongly with their relative expression levels [135]. The relative tRNA concentrations were estimated using the relative copy number of tRNA genes, the counts of which were obtained from the Genomic tRNA Database [43].

4.2.10 Normalisation of ribo-seq translation profiles

To extract summary statistics from ribo-seq data, such as the average count for a given codon type, the translation profile of each transcript must be normalised such that they are comparable to one another. Direct comparison is often not valid, due to differences in the expression and initiation rates which may bias the relative prevalence of fragments from various transcripts over others [149, 174]. Several normalisation techniques have been suggested and used to account for these differences, though the most common method involves dividing the translation profile of each transcript by their respective ribo-seq depths, this a proxy for the expression. This technique, which we will refer to as conventional normalisation, has been shown to perform better than more complicated procedures on simulated data, both with and without various forms of added noise [149].

The only technique shown to outperform conventional normalisation was the Ribo-seq Unit Step Transformation (RUST) method [149]. The RUST method involves transforming a given translation profile into a series of ones and zeros in an attempt to minimise the influence of extreme values. A value of one is assigned to codons with a count greater than the depth of the transcript they reside on, while zeros represent codons with counts equal to or less than the depth. Each codon is also assigned the expected RUST score for the transcript they reside on, which is merely the summation of the observed RUST values divided by the transcript length. Summary statistics are then calculated as the ratio between the sum of the observed RUST values against the sum of the expected RUST values for a given property, such as codon type. For the work presented in this Chapter, we use both conventional and RUST normalisation.

4.3 Results

4.3.1 Reproducibility of ribo-seq data

We investigated the reproducibility of ribo-seq by comparing the translation profiles of analogous mRNA transcripts across multiple studies. To perform this comparison, we collated 107 ribo-seq experiments sourced from 15 independent studies on *S. cerevesiae* from the GWIPS-viz database (see Table 4.1) [153]. We compare the ribo-seq profiles of the various experiments to each other, both at the transcript and codon-level and separating comparison to within and outwith of the same study, to quantify the reproducibility. Comparisons within a given study have been investigated previously in a smaller study, with the translation profiles of replicates being reported to differ [163]. Comparisons outwith of the same study have not been investigated previously. In doing so, lets us comment on the reproducibility one would obtain when comparing any two ribo-seq datasets.

In Figure 4.1, we show the difference between comparing ribo-seq datasets both at the transcript and codon-level. The transcript-level comparison uses the depths of all transcripts and is the method most commonly used for comparing ribo-seq experiments [175]. The codon-level comparison is based on the similarity of the translation profile of each transcript explicitly. The correlation between studies is strikingly different between these comparison methods. Using transcript-level comparison, a high level of similarity is found, which would indicate that ribo-seq data is reproducible. In contrast, using the codon-level comparison, most datasets are found to be dissimilar overall. Subsetting the comparisons to those of experiments within the same study or those from different studies establishes to what degree experimental differences and technical biases may have influenced these results. While experiments from the same study are slightly more correlated at both the transcript-level and codon-level, the overall trend that ribo-seq experiments are similar when compared at the transcript-level while differing at the codon-level remains.

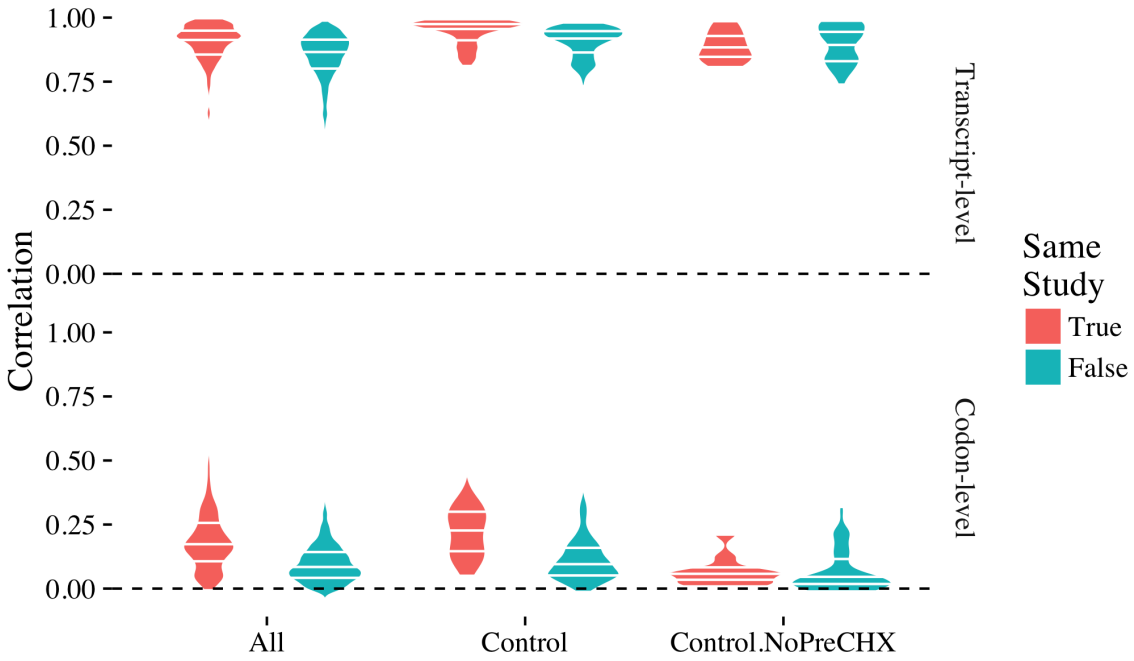


Fig. 4.1 The distribution of transcript-level and codon-level correlations between ribo-seq experiments for all datasets, only control/wild-type datasets, and only control/wild-type datasets whose protocol did not include pretreatment by cycloheximide. The correlations have been grouped by whether the experiments compared are from the same study. The horizontal lines on each distribution mark the upper and lower quartiles, as well as the median. The codon-level correlation indicates the similarity between experiments when the translation profiles of each transcript are compared explicitly, while the transcript-level correlation indicates the similarity when the transcripts are averaged.

These results may be in part due to the diverse set of biological conditions that many of the experiments have been carried out under. For example, the translation profile can change significantly when the host organism is starved of both oxygen and glucose [176]. Likewise, pretreatment with cycloheximide prior to cell lysis, which is a very common inclusion in the ribo-seq experiment protocol, has been shown to change the resultant translation profiles [157]. Given this, we tested whether removing these known biased datasets would change our observations. The results, which are also given in Figure 4.1, show that the divergence between transcript and codon-level comparisons is maintained when only control or wild-type experiments are considered, and when datasets pretreated with cycloheximide are removed.

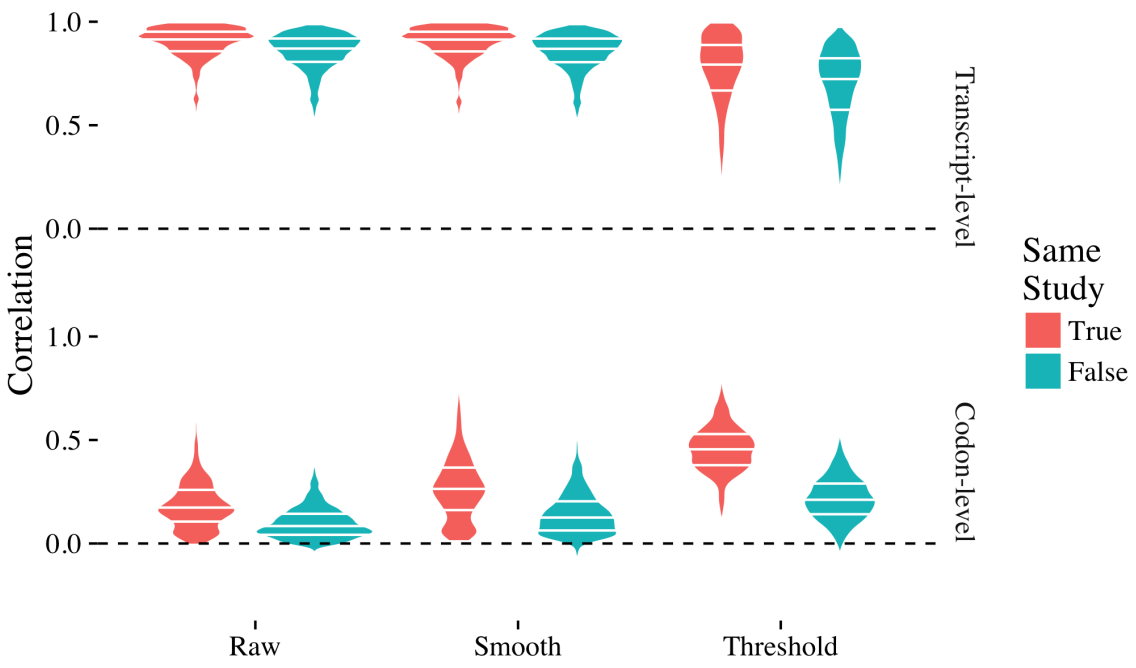


Fig. 4.2 The distribution of transcript-level and codon-level correlations between ribo-seq experiments for the unaltered raw datasets, the smoothed datasets, and the thresholded dataset. The correlations have been grouped by whether the experiments compared are from the same study. The horizontal lines on each distribution mark the upper and lower quartiles, as well as the median. The codon-level correlation indicates the similarity between experiments when the translation profiles of each transcript are compared explicitly, while the transcript-level correlation indicates the similarity when the transcripts are averaged.

Possible causes of this lack of reproducibility at the codon-level include misalignment of fragments and the presence of sparsely sampled transcripts within datasets. Both of these were tested for, the former by smoothing each profile to reduce the effect of slight misalignments, the latter by applying a threshold to the data such that sparsely sampled transcripts are removed. The effect of these modulations on both the transcript and codon-level comparisons are given in Figure 4.2 alongside the comparisons for the unaltered datasets. Smoothing the profiles is found to reduce the disparity slightly between the transcript-level and codon-level comparison, though this could be merely due to the general effect of smoothing to reduce differences. In contrast, removing the sparse transcripts significantly altered the observed distributions of transcript-level and codon-level correlations. At the transcript-level, the values are reduced, indicating

that transcripts with low expression heavily influenced the observed reproducibility. At the codon-level, a jump in the correlation is observed, the same transcripts with low expression found to exhibit more variation in their translation profiles than most.

The underlying cause of the changes caused by smoothing and thresholding can be seen in Figure 4.3 in which we show the codon-level correlations for each transcript (rather than summarising via the median) between a pair of experiments not sourced from the same study. It highlights that the majority of transcripts that are found to be dissimilar between two datasets are those in which at least one of the transcripts compared are poorly sampled. The figure also shows that some transcripts do have similar translation profiles at the codon-level, as given by the existence of high correlations. Such correlations indicate that it is possible for a translation profile observed in a given ribo-seq experiment to be reproduced elsewhere. This observation would not be expected if the codon-level comparison was not a valid measurement of the reproducibility.

4.3.2 Identifying adequate sampling of transcripts

Given that inadequate sampling may be contributing to the lack of reproducibility at the codon-level, we attempted to identify transcripts within our compiled dataset that could be considered as sampled sufficiently. The profiles of such transcripts should be reproducible across multiple studies and thus can be regarded as high-quality. The codon-level comparisons from above can identify such transcripts by assuming that reproduction of a translation profile, as given by a high correlation, only occurs when both of the compared transcripts are of high-quality. We set a threshold of $\rho_{codon} > 0.7$ above which comparisons were deemed to be between high-quality transcripts. Combined, 5.052% and 1.55% of our codon-level comparisons between transcripts achieved this threshold for those within and outwith of the same study respectively.

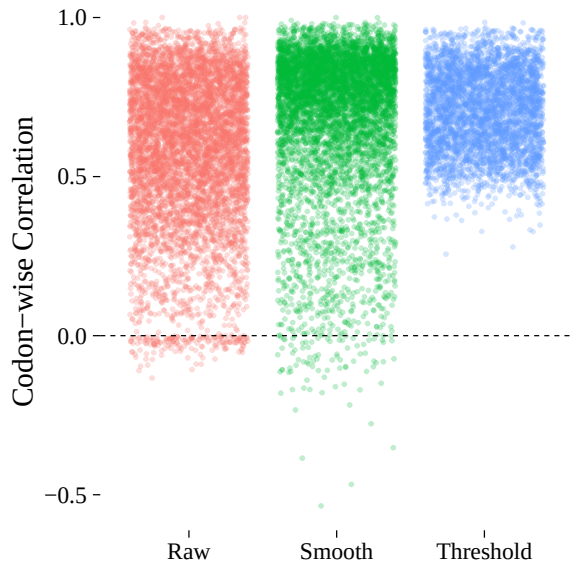


Fig. 4.3 The median of the codon-level correlations between experiment 39 and 93 (see Table 4.1) was the highest out of any pair of compared experiments that were outwith of the same study. The codon-level correlations for all transcripts compared between these experiments are shown below for the raw, smoothed, and thresholded datasets. They clearly show how enforcing a threshold that removes sparse transcripts predominately removes comparisons that result in low correlations.

Predicting high-quality transcripts

Using the above definition of a high-quality transcript, simple statistical logit models were built using the depth and coverage as inputs to predict whether a given comparison would achieve a correlation above or below the classification threshold. If such models proved sufficiently accurate, they would be able to generate depth and coverage based thresholds such that high-quality transcripts could be predicted within each study independently and, hence, reduce the need for comparison to other studies. Various predictive models were trained using the protocol outlined in the materials and methods section above. The models alongside their respective Akaike Information Criterion (AIC) values are given in Table 4.2 [177]. All models significantly outperformed the null model ($p < 10^{-10}$). Furthermore, from comparison of the AIC, we found that $\text{logit}(P(\rho_{codon} > 0.7)) = \beta_0 + \beta_1 \log(depth_{min})$ significantly outperformed the others tested ($p < 10^{-10}$), where $P(\rho_{codon} > 0.7)$ is the probability of a codon-level correlation being above

Table 4.2 Various logit models tested for their ability to identify high-quality transcripts from comparisons of two ribo-seq datasets. A correlation above 0.7 indicates a comparison between high-quality transcripts as the translation profiles are similar. Same study shows whether the model was trained on comparisons of transcripts from experiments sourced from the same study or those outwith.

Same Study	Model	AIC
TRUE	$(cor > 0.7) \sim log(depth.min)$	280732.8
TRUE	$(cor > 0.7) \sim depth.min$	323430.3
TRUE	$(cor > 0.7) \sim cov.min$	376637.1
FALSE	$(cor > 0.7) \sim log(depth.min)$	818866.4
FALSE	$(cor > 0.7) \sim depth.min$	998482.8
FALSE	$(cor > 0.7) \sim cov.min$	1114078.4

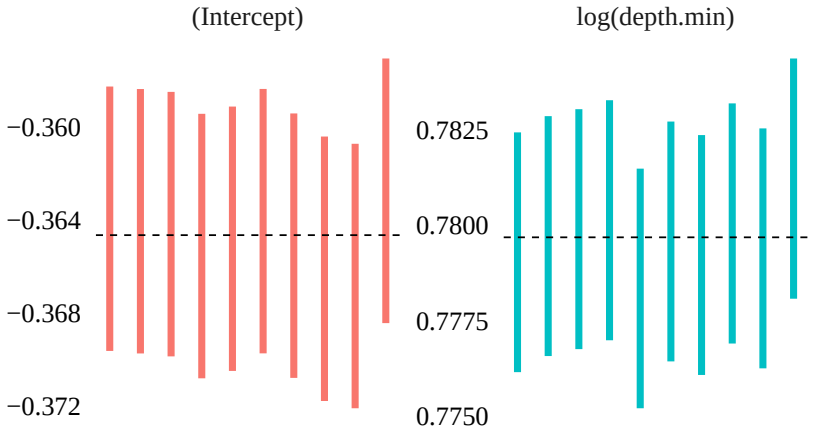
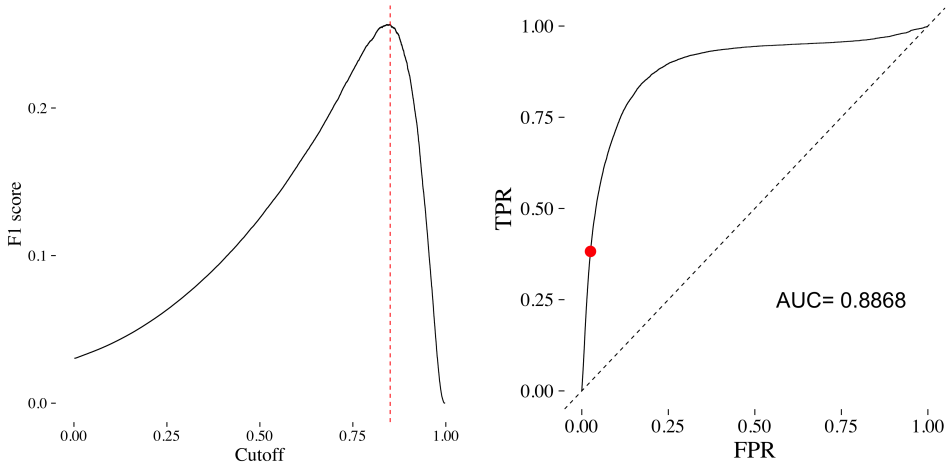


Fig. 4.4 Confidence intervals of the coefficients for the ten different random samplings used to train our model. The average of the coefficients is given by the dashed line and are those taken as the consensus model. As the confidence intervals overlap, we infer that the coefficients are not biased by our random sampling and the model is stable.

0.7, β_0 and β_1 are the fitted coefficients, and $depth_{min}$ is the lowest depth between the transcripts compared. This model performed best for both predictions within and outwith of the same study. Models combining the depth and the coverage were attempted but suffered from multicollinearity. Likewise, models based on the first component of a principal component analysis of these predictors did not give an improved model.

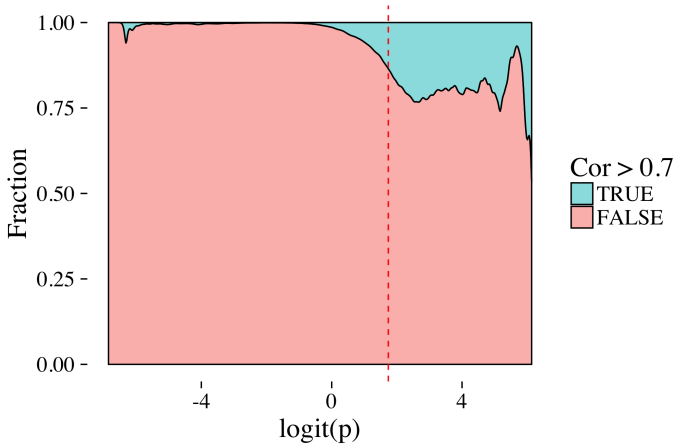
We focused on the comparisons outwith of the same study going forward as these exhibit lower levels of reproducibility and, hence, will generate more stringent depth thresholds to define high-quality. The values for the coefficients of the corresponding model, β_0 and β_1, for each of the ten randomly sampled training sets are shown in Figure 4.4. As the confidence intervals on the coefficients fitted to each training set overlap, we concluded that the coefficients are stable and not biased by the random samplings. We then averaged the β_0 and β_1 coefficients fitted for each sampling to establish a consensus model.

The predictive power of the chosen model is shown in Figure 4.5 in which we have given the F1 score against the various thresholds, the corresponding ROC curve and the fraction of true positives to true negative at various thresholds. We used the F1 score to identify the optimal threshold for our classification as it gives little weight to the true negatives which are not of interest. The F1 score is maximised at a threshold of 0.850, which corresponds to a depth threshold of 2.696 reads per codon. The accuracy, precision, and recall at this threshold are 0.966, 0.193, and 0.382 respectively. The high accuracy, the fraction of true positives and negatives, is misleading, as the low prevalence rate of comparisons between high-quality transcripts (1.547%), means that merely the labelling of easily identified negatives will result in an inflated value. The precision and recall indicate that approximately 20% of our predicted positives are true positives and that this fraction accounts for roughly 40% of all true positives.



(a) F1 score vs cutoff

(b) ROC curve



(c) Predictions

Fig. 4.5 Various measures as to the predictive power of the selected model in its ability to identify comparisons between high-quality transcripts. The maximum of the F1 score is taken as the cutoff threshold and is marked by the dashed red line or as a red dot. The F1 score is a combined measure of the precision, the fraction of predicted positives that are true positives, and recall, the fraction of true positives that are predicted positives. The ROC curve, shown in (b), shows the true positive rate (TPR) and the false positive rate (FPR) as the threshold is varied. The outcome if one were to randomly assign classifications proportionally is marked using a dashed black line. Lastly, in (c), we show the fraction of true positives against true negatives as the cutoff, which is based on the depth, is varied. The dashed red line, which signifies our threshold, indicates where our model would predict a comparison as being between high-quality transcripts.

While the chosen model was fitted on the comparisons of translation profiles for the same transcript sourced from different studies, the depth threshold generated can be applied to the transcripts within each study independently. In other words, any transcript in a study with a depth greater than 2.696 can be deemed high-quality, and, hence, would be reproduced if sampled sufficiently in another study. For example, applying this threshold to our 107 collated ribo-seq datasets predicts 166847 transcripts as high-quality. In this predicted high-quality subset, 5881 out of the possible 6692 transcripts expressed in *S. cerevisiae* were represented (87.88%). Notably, this abstraction of the model allows identification of high-quality transcripts that did not produce a high codon-level correlation due to lack of a high-quality partner from another study. Of those predicted as high-quality, 207 transcripts were only present in a single experiment.

Directly using codon-level comparisons

The above predictive model and corresponding depth threshold are useful to apply to individual datasets in order to identify which transcripts are of high-quality. However, our collated ribo-seq datasets provide enough data that we can use the codon-level comparisons directly to determine exactly which translation profiles were unquestionably reproduced. If reproduced, both the compared transcripts must be of high-quality. Akin to above, we used a threshold of $\rho_{codon} > 0.7$ to establish which translation profiles were reproduced and only considered comparisons between experiments outwith of the same study. This labelled 1.548% of our codon-level comparisons outwith of the same study as being between high-quality transcripts.

4.3.3 Creating a high-quality ribo-seq dataset

Using those transcripts labelled as high-quality from the codon-level comparisons, we wished to build a database of reproducible transcription profiles. However, further analysis of the comparisons that achieved the high-quality threshold, found that many involved two transcripts

that had extremely low coverage; the few non-zero positions found to align. As our aim in producing this dataset was to provide a suitable comparison to the theoretical estimators of translation speed, as well as other properties of the mRNA sequence, sparse profiles provide little worth. Given this, correlations between extremely sparse transcription profiles were discarded by removing any comparison which involved a transcript with coverage less than 10%. This coverage threshold removed 8.364% of those codon-level comparisons that were considered as being between high-quality transcripts.

Any further analysis on this high-quality subset may be biased due to the differing number of translation profiles for each transcript present. Given this, all the translation profiles for each transcript were merged to give a consensus profile by taking the mean for each codon over all copies. Hence, a transcript's consensus translation profile in the high-quality set is based on the merger of anywhere between two and 107 translation profiles. If no profiles of a given transcript met our criteria for high-quality, the transcript is not represented. Our high-quality merged ribo-seq dataset contained profiles for 2601 out of the 6692 transcripts (38.87%) specified in the sacCer3 Ensembl annotation. All translation profiles in this high-quality dataset are available for download and viewing at http://opig.stats.ox.ac.uk/resources.

4.3.4 Comparison to traditional estimators of translation speed

Translation profiles can also be predicted using theoretical estimators. These theoretical estimators typically use either the relative frequency of occurrence in the genome or physical properties, such as the cognate tRNA concentration, to give a prediction of the relative translation speed for a given codon. The translation profiles generated by some of these estimators have been compared to those from ribo-seq previously, and a correlation between them was observed [137, 147]. However, the variation at the codon-level between ribo-seq studies shown above suggests that the actual relationship may differ to that reported.

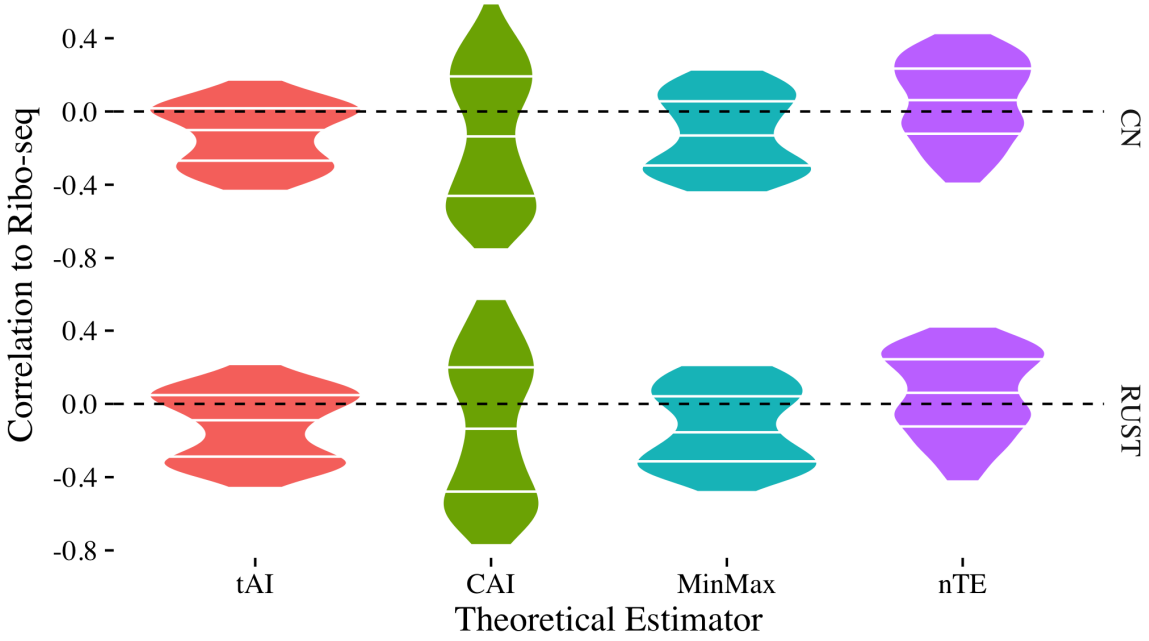


Fig. 4.6 Traditional theoretical estimators of translational speed compared to 107 ribo-seq experiments. The distribution is formed from the correlations calculated between the estimator's predicted value and the normalised averaged ribo-seq count for each codon type within each of the experiments. The correlations are computed for both conventional normalisation (CN) and RUST normalisation. The horizontal lines on each distribution mark the median and the quartiles. A negative correlation is sought, as ribo-seq finds higher counts on slowly translated codons, while the estimators give higher values to codons deemed rapidly translated.

In Figure 4.6, we show the distribution of correlations between the values assigned to each codon type by several estimators and their normalised ribo-seq count for each of the 107 experiments. We compare to four different well-known estimators, CAI, tAI, MinMax, and nTE, to values produced when the datasets have been normalised using both conventional and RUST normalisation. Details on the estimators and the normalisation methods can be found in the Methods. In reverse to the ribo-seq data, higher values given by any of these estimators predicts a faster translation rate. Given this, a negative correlation between the profiles generated by ribo-seq and the estimators is sought. For all estimators, the Spearman's correlation to the ribo-seq values is found to vary widely depending on the experiment used, regardless of normalisation. Both large positive and negative correlations are reported, and no

clear trends are apparent. This behaviour remains even if start codons are removed from the dataset due to their inherent bias. For example, comparing the conventional normalised ribo-seq values of Dataset 27 to CAI produces a correlation of -0.754, while comparing Dataset 31 gives a correlation of 0.591. We conclude that the performance of the theoretical estimators is highly dependent on the ribo-seq experiment chosen.

Comparison of these estimators to only high-quality transcripts should provide better inference as low-quality transcripts, which are highly variable, would have contributed significantly to the variation seen above. As above, we compared the normalised values of each codon type in our high-quality dataset to those given by the estimator with a negative correlation being sought. We give the results in Figure 4.7 for both normalisations. Focussing on the comparison to conventional normalised values, we find that of the estimators, CAI performs the best, with a Spearman correlation of -0.194, while nTE performs the worst, with a correlation of 0.187. When compared to RUST normalised values, the relative ordering of the estimators does not change, while the correlations only change marginally. Additionally, and as found above, these observations do not change if the start codon is not considered.

Akin to the ribo-seq data, we also compared the high-quality ribo-seq profiles and theoretical estimators at the codon-level explicitly. While we found that some estimators correlate reasonably to normalised ribo-seq values when summarised, as reported elsewhere, they could perform quite differently when estimating the translation profiles of transcripts independently instead. For each transcript in the dataset, translation profiles using the four different estimators named above were generated and compared to the merged ribo-seq profile (see Figure 4.8). The median correlations are -0.026, -5.457×10^{-4}, -0.010, and -0.011 for tAI, CAI, MinMax and nTE respectively. These values indicate that while a small amount of the variability is still captured by the estimators, it is significantly less than that reported in the mean value comparison, Crucially, it is not of a significant enough level to be considered predictive. Hence,

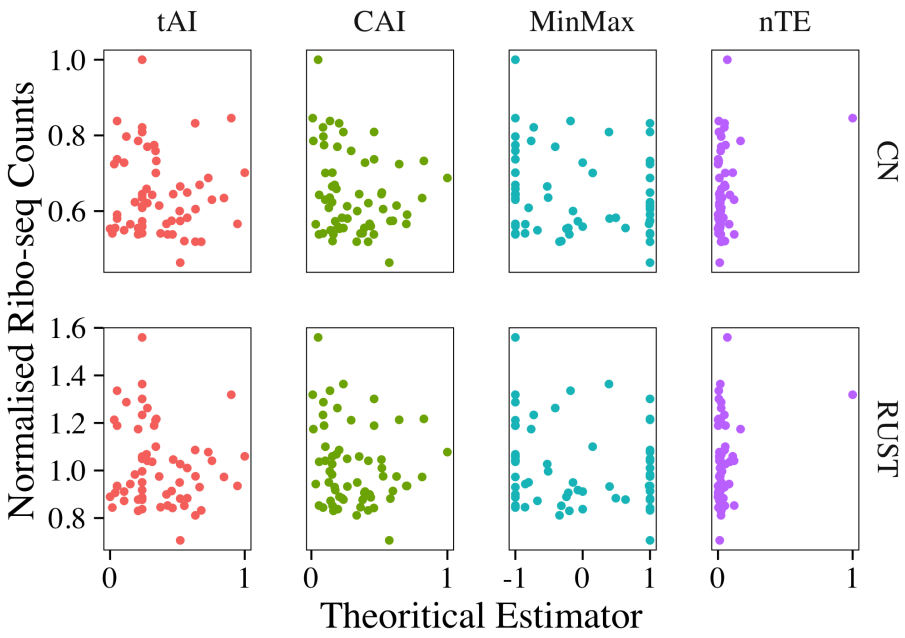


Fig. 4.7 Comparison of our high-quality merged ribo-seq dataset to traditional theoretical estimators of translation speed averaged over codon type. Shown are the normalised ribo-seq counts compared to the theoretical estimator's values for each codon with the normalised counts generated both with conventional normalisation (CN) and RUST. A negative correlation is sought, as ribo-seq observes higher counts on slowly translated codons, while the estimators assign higher values to codons deemed quickly translated.

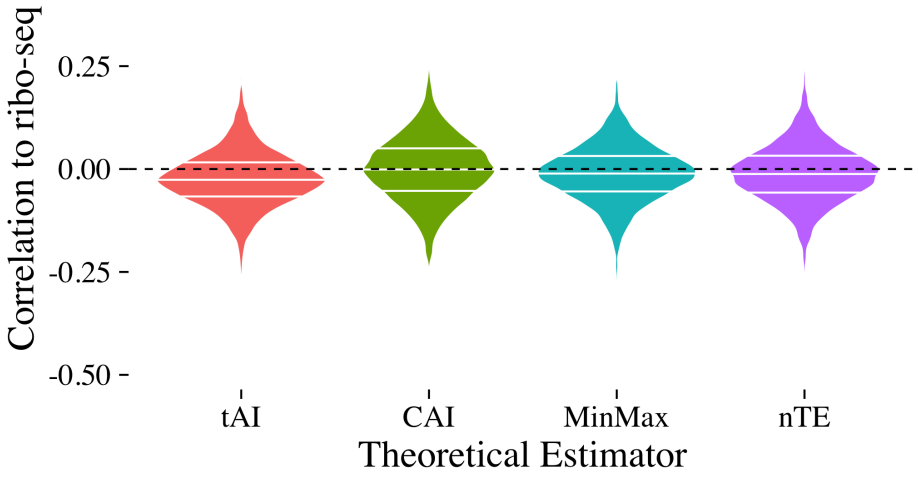


Fig. 4.8 Comparison of our high-quality merged ribo-seq dataset to traditional theoretical estimators of translation speed at the codon-level. Shown are the distributions of correlations when comparing the translation profile, as given by high-quality ribo-seq and that predicted by the estimator, of each transcript explicitly at the codon-level. The horizontal lines on each distribution mark the upper and lower quartiles, as well as the median. A negative correlation is sought, as ribo-seq observes higher counts on slowly translated codons, while the estimators assign higher values to codons deemed quickly translated.

our results suggest that current theoretical translation speed estimators do not capture the translation landscape.

4.4 Discussion

Ribo-seq is the first transcriptome-wide experimental measure of translation and has been widely reported as having codon-level accuracy. However, ribo-seq's reproducibility has traditionally been tested by comparing the depths of transcripts, which is a summary statistic, and therefore does not inform as to the similarity at the codon-level. Here we carried out a comparison of the translation profiles of transcripts explicitly to truly measure the codon-level similarity of datasets. We compared over 100 different experiments to one another; the data from each treated in a systematic manner to remove post-processing differences. To date, this is the largest comparison of ribo-seq experiments. We showed that for all but a few transcripts that have been heavily sampled, the translation profile of a transcript at the codon-level is not reproducible. This result is in agreement with a similar smaller study that investigated only replicates across a handful of studies [163].

Given that there is such disparity between the translation profiles, any attempts to test a hypothesis based on codon-level translation profiles would result in observations that would be greatly experiment dependent. In fact, this disparity between experiments may be the source of some of the contention that already exists within ribo-seq literature. In order to perform a codon-level analysis, it is first necessary to prune ribo-seq datasets to only those transcripts that are high-quality, i.e., those for which the translation profile is reproducible across multiple experiments. We built a model to identify transcripts that are of high-quality and used our codon-level comparisons of the various datasets to train said model. This model established a depth threshold of 2.696 reads per codon above which one can consider a transcript to be reasonable high-quality. Across our collated datasets only 166847 transcripts achieved this threshold. These predicted high-quality transcripts were not distributed evenly between the

collated experiments; experiments with a higher total read count, and hence those which were more sampled overall, contained a larger fraction than those with lower total read counts.

Using our codon-level comparisons, we next pruned our collated dataset to only high-quality transcripts with profiles reproduced across multiple studies. We then compared the profiles of this subset to those predicted by traditional theoretical estimators of translation speed. Similar comparisons have been performed previously and reported that some of the variability of ribo-seq is captured by the estimators [137, 138, 147]. However, these results were based on summarising by codon type, rather than an explicit comparison of the experimental and predicted profiles at the codon-level. We first compared the estimators to each experiment separately using values averaged by codon type. For every estimator, a vast range of correlations was found, spanning both negative and positive values. This observation is a clear example of how inference based on codon-level ribo-seq data is highly experiment dependent. When the same comparison was performed using our dataset of high-quality transcripts, some estimators, most notably CAI, correlated slightly ($\rho = -0.194$). However, when we performed explicit comparison at the codon-level, the correlation was severely decreased.

4.5 Conclusion

In summary, we collated and compared over 100 ribo-seq experiments and found that the translation profiles of transcripts are not reproducible for all but a few highly sampled transcripts. Given this, we created a simple model to generate a threshold to the sampling above which transcripts can be considered high-quality, i.e., those with reproducible translation profiles. We then compared both the individual ribo-seq experiments and our high-quality dataset to traditional theoretical estimators of translation speed. For the former, we found the correlation to be highly dependent on the ribo-seq experiment. For the latter, only weak correlations were observed. Given this, we concluded that the translation landscape is not adequately described by any current theoretical estimator.

Chapter 5

Relating ribo-seq data to the protein structure produced

In the previous chapters, we discussed ribo-seq, an experimental technique that provides a transcriptome-wide measurement of the translation speed [135, 155]. However, by comparing an unprecedented number of ribo-seq datasets sourced from various studies of *S. cerevisiae*, we showed that the method struggles with reproducibility at the codon-level for all but the most sampled transcripts. Identifying such transcripts, namely those whose translation profiles were similar across multiple datasets, we combined them to form a high-quality dataset that should yield more reproducible outcomes when undergoing further analysis. Using this high-quality dataset, we compared the translation speeds of codons as given by the ribo-seq data to that predicted by theoretical estimators of the translation speed. We found that while correlation is observed when the ribo-seq data is summarised by codon type if instead compared explicitly within each transcript instead, the estimators do not capture the behaviour of translating ribosomes.

From the above observations, we realised that many of the relationships probed for in Chapter 2, namely, those relating the speed with which a codon is translated and the protein structure produced, may be erroneous. Given this, we repeated these analyses, replacing the

translation profiles generated by the theoretical estimators with those from our high-quality experimental ribo-seq dataset. Other works have already shown using ribo-seq data that translating ribosomes do exhibit biases. For example, ribosomes have been found to aggregate, and hence are presumed to translate slower, at the N-terminus, at areas of highly stable mRNA secondary structure, and when proceeding codons encode positively charged residues [144, 146, 151, 164].

Bias with regards to the emergent protein structure have previously been commented on by Lopez and Pazos, who tested whether various structural and functional features of proteins are correlated with the variance found across aligned ribo-seq profiles of *E. coli* [178]. They found that most features, both structural and functional, were marked by a considerable translation pause near their initial codons. However, to perform their analysis, they made use of only a single ribo-seq dataset in which reads were attributed to the central residues, rather than the 3' alignment method discussed in Chapter 3. Also, no normalisation of profiles was performed, extreme ribo-seq values (outliers) were purged, and the protein features were based on UniProt entries, of which the majority are based upon predictions [124, 171]. Given this, their observations may be affected by technical biases associated with their protocol.

To our knowledge, no prior work has combined high-quality, reproducible experimental ribo-seq data with experimental protein structures. We show that variation in translation speed, as given by the ribo-seq data, is related to many aspects of a translated protein's structure. For example, a given codon's translation speed was found to bias its placement within the resultant structure with regards to both the solvent accessibility and the secondary structure it formed. These results suggest that the translation speed, possibly modulated through codon choice, may have an effect on the protein structure produced.

5.1 Materials and methods

5.1.1 Ribo-seq translation profiles

In the previous chapter, we generated a dataset that contains high-quality ribo-seq translation profiles for some of the coding transcripts in *S. cerevisiae*. We use this dataset as the basis for the analyses undertaken within this Chapter. The dataset contains counts along the transcripts for the number of ribosomes whose A-site aligned on a given codon at the point of cell lysis [135, 155]. From these counts, the relative translation speed of codons on a given transcript can be inferred. The constituent datasets were all sourced from the GWIPS-viz database, which uses RUM, a splice aware aligner, in its protocol to align the reads against the relevant genome [153]. As such, exons used disproportionately may lead to inherent jumps and decreases in the profiles across splice junctions within transcripts [179]. Given this, we discarded any transcripts which shared an exon with another. After this pruning, we were left with 3656 transcripts with translation profiles which contained a combined total of 1323590 codons.

5.1.2 Ribo-seq normalisation

As discussed in the previous chapter, to compare the translation profiles of different transcripts they must first be normalised such that differences in both the expression and initiation rates are accounted for [147, 149, 174]. For the work presented in this chapter, we make use of the Ribo-seq Unit Step Transformation (RUST) method whereby a translation profile is translated into a series of zeros and ones [149]. This procedure is designed to minimise the effects of extreme values found at sites prone to ribosome pausing on inference. A value of one is assigned to codons with a count greater than the arithmetic mean count of the transcript they reside on, while zeros represent codons with counts equal to or less. However, while the proposed method to compare RUST value is via the comparison of the summated observed RUST score to the summated expected RUST scores, here, instead, we perform categorical

inference on these values [149]. Coarsely, zeros can be considered as codons rapidly translated and ones as those slowly translated.

5.1.3 Protein structure annotation

For each transcript within our dataset, we attempted to identify its corresponding experimental protein structure from the Protein Data Bank (PDB) [46]. Each transcript's translated sequence was compared against the compiled sequences of all experimental protein structures within the PDB annotated as being sourced from *S. cerevisiae*. The PDB sequences were based on the SEQRES sequence given within each PDB structure, and the comparison was performed using the Basic Local Alignment Search Tool (BLAST) with an *E*-value threshold of 10^{-5} [127]. The best hit, as given by the BLAST bit score, between the transcript sequence and the sequence from any given structure was taken as the correct pairing. If a pairing existed, the secondary structure annotation and the solvent accessibility were calculated using JOY [118]. Four secondary structure annotations can be given, C, E, H, and P, which correspond to coils, β-strands, helices (α, 3_{10} and π), and residues with positive-ϕ angles respectively. From the solvent accessibility, residues with under seven percent solvent accessibility were classified as buried, while all other residues were classified as accessible. To map these annotations, which were assigned to the PDB structure sequence, to the transcript sequence, we performed pairwise alignment of the structure sequence to the SEQRES sequence. This alignment combined with that of the transcript to SEQRES sequences from BLAST allowed for the structural annotation to be mapped onto the transcript sequence. The pairwise alignment and the subsequent mapping was performed using Python (v2.7) and the Biopython package [180]. A corresponding structure was found for 498 transcripts, of which 124118 codons had a structural annotation. Note that not all codons within a transcript which had a successful pairing with a structure may be attributed a structural annotation as not all residues may be present in the experimental structure. The absence of residues can be due to the experiment only

focussing on a segment of the protein or the residue not being observed within the experiment [181].

5.1.4 Domain annotation

For each of the transcript-structure pairings described above, we sourced the domain annotation for the structure, and hence also the transcript, from the SCOPe (v2.5) database [123]. SCOPe annotations classify every residue in a given PDB entry as part of a domain. As such, a domain annotation was attributed to each codon within our pairings which had been assigned a structural annotation. Note that not all structures have domain annotations in SCOPe. A domain annotation was found for 333 transcript-structure pairs, which allowed us to annotate 79534 codons as part of a domain.

5.1.5 Linker annotation

As our domain annotation labels every residue within the protein structure, no linker regions between neighbouring domains are explicitly defined. We chose to label the linker region on multidomain proteins as the ten residues either side of the junction between neighbouring domains. Assigning a total of twenty residues as the linker is a conservative threshold with respect to ensuring that the whole linker region is contained as analysis of linkers found within experimental protein structures show that they average 10 amino acids in length [182]. If the junction occurs at a gap within our domain annotation, the unannotated codons do not contribute to these counts. Our dataset contains 52 multidomain proteins, from which 58 linker regions were defined. The number of linkers exceeds the number of multidomain proteins due to the presence of six proteins with three domains in their respective structures.

5.1.6 Secretory proteins

Identification of secretory proteins in our dataset was performed using the UniProtKB reviewed set [125]. The UniProtKB reviewed set contains a manual annotation for each transcript stating whether or not they are secretory proteins, as well as which residues form the signal peptide at the N-terminus. Using the latter information, we took note of where the junction between the signal peptide and the remaining protein chain was located. Out of the transcripts which we generated translation profiles for, 200 were identified as being secretory proteins.

5.1.7 Statistics

All statistical analysis was performed in *R* (v3.2.4) [183]. To perform the Cochran-Mantel-Haenszel (CMH) chi-squared test, we used the default packages. To calculate odds ratios (OR), we used the "vcd" or the "epitools" packages [184, 185]. Confidence intervals given for ORs are those representing 95% confidence ($p = 0.05$). To create multinomial log-linear models, we used the "nnet" package which fits them via a neural network [186]. Models were fitted on a training set consisting of 80% of the available data selected randomly, with the remaining 20% of data then used to assess the model's accuracy.

5.2 Results

5.2.1 Association of translation speed to the termini

It has been reported by multiple studies that the initial segment of an mRNA transcript is translated slower on average than the rest of the protein chain [109, 110, 187–189]. One explanation given for this phenomena is that the ribosome requires weak mRNA structure, which is easy to unfold, near the N-terminus to begin translating efficiently. Weak mRNA structure can be caused by the inclusion of rare codons, which are thought to be translated

slower than those commonly used [110, 187, 189]. Another suggested explanation is that the delay helps prevent ribosome traffic jams forming [109, 188]. Ribosome traffic jams are when sequential ribosomes collide such that the trailing ribosome becomes limited in its translation rate by the progression of the leading ribosome [190]. An initial delay gives additional time to the proceeding translating ribosome to progress further along the transcript, which decreases the likelihood that they collide during translation [109, 188]. A recent study stated that this ramp is in fact only present in secretory proteins, the authors suggesting that it instead acts as a signal to promote other key processes, such as membrane translocation and protein processing [111]. As to the C-terminus, no prior works have commented on whether it exhibits unique behaviour with regards to the translation speed.

In Figure 5.1, we show the enrichment by slowly translated codons, as given by the OR, in the regions adjacent to both the N and C-termini. Transcripts are aligned by their respective termini and the OR for each codon calculated with respect to all other codons. We observe both initial and terminal spikes in the enrichment due to the propensity for ribosomes to be placed on either the start and stop codon respectively which combines with the small degree of error in the alignments. Bar these spikes, an approximately forty codon region exists at the N-terminus which is heavily enriched by slowly translated codons (OR$\approx$2). This region is then followed by a steady decrease in the enrichment until roughly the 200^{th} codon, at which point it plateaus (OR$\approx$0.8). The switch from favouring slowly translated codons to rapidly translated codons occurs roughly at the 150^{th} codon. At the C-terminus, no plateaus are found in the proceeding codons. Instead, only a steady decline in the enrichment occurs across the entire region, with the last codons exhibiting a strong preference for quickly translated codons (OR$\approx$0.6).

These observations may be by-products of either aligning the transcripts by their termini combined with the variance in transcript lengths within our dataset. An overview of this length variance is given in Figure 5.2, in which we give the distribution of transcript lengths found within our dataset. To test whether the various lengths affected the termini enrichment, we also

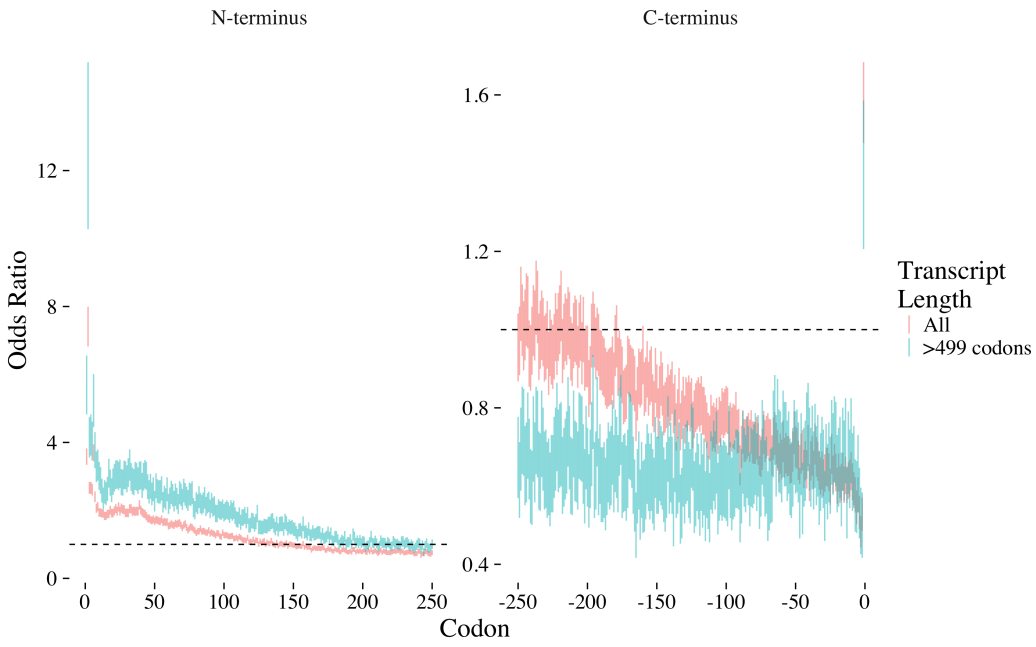


Fig. 5.1 The enrichment by slowly translated codons at the transcript termini. The enrichment by slowly translated codons within the initial and final 250 codons of mRNA transcripts has been calculated by aligning all transcripts by their respective start or final codons and then calculating the OR with respect to all other codons. The enrichment is also shown overlaid for when one only considers transcripts of length 500 codons or more. The N-terminus plot is shown with the start codon labelled as codon zero and progressing from left to right. The C-terminus plot shows the final codon as codon zero and progresses from right to left. The bars at each position represents the confidence interval on the OR. Values above one, which is highlighted via the dotted line, show a preference for slowly translated codons at a given position, while values below one show a preference for rapidly translated codon.

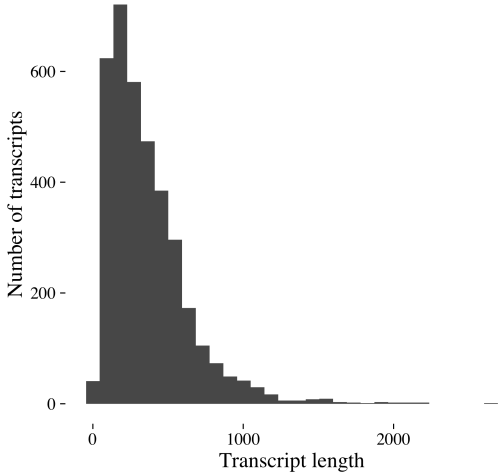


Fig. 5.2 Distribution of transcript lengths within the ribo-seq dataset. Out of the 3656 transcripts present, 848 transcripts had lengths of 500 codons or more.

calculated the enrichment when we only consider transcripts of length 500 or more (N=848). The result, also plotted in Figure 5.1, shows that at the N-terminus, the overall behaviour is unchanged, though there is a greater degree of enrichment overall when only the longer genes are considered. In contrast, at the C-terminus, the steady decline witnessed when all genes are considered has been replaced by a plateau over the entire region. This would suggest that the steady decline was merely due to the presence of shorter genes whose initial and final 250 codons overlapped. Given these results, we would conclude there is an enrichment of slowly translated codons in the initial region, notably over the first forty codons, before slowly decreasing. This result is in agreement with many of the prior works discussed above [109, 110, 187–189].

Next, we tested whether this observation is only present for the secretory proteins within our dataset as was reported elsewhere [111]. In Figure 5.3, we show the enrichment by slowly translated codons at the N-terminus for both secretory proteins and non-secretory proteins overlaid. As above, we also perform the analysis for when transcripts of all lengths are considered and when only transcripts of at least 500 codons are considered. Firstly, the enrichment of non-secretory proteins appears unchanged from that seen in Figure 5.1. Secondly, the averaged translation profile of secretory proteins differs significantly to that of the non-secretory proteins. The initial fifty codons are found to be enriched by slowly translated codons, which is then followed by a plateau along the remaining transcript in which neither slowly nor quickly translated codons are preferred. This is in contrast to the enrichment seen for non-secretory proteins, whereby the enrichment by slowly translated codons decreases steadily along the initial 200 codons approximately. Such a profile would indicate that the translation of the initial codons of secretory proteins differs significantly to the remaining length. These observations are found regardless of if one only considers longer transcripts. We postulate that this sudden drop in the prevalence of slowly translated codons may be due to differences

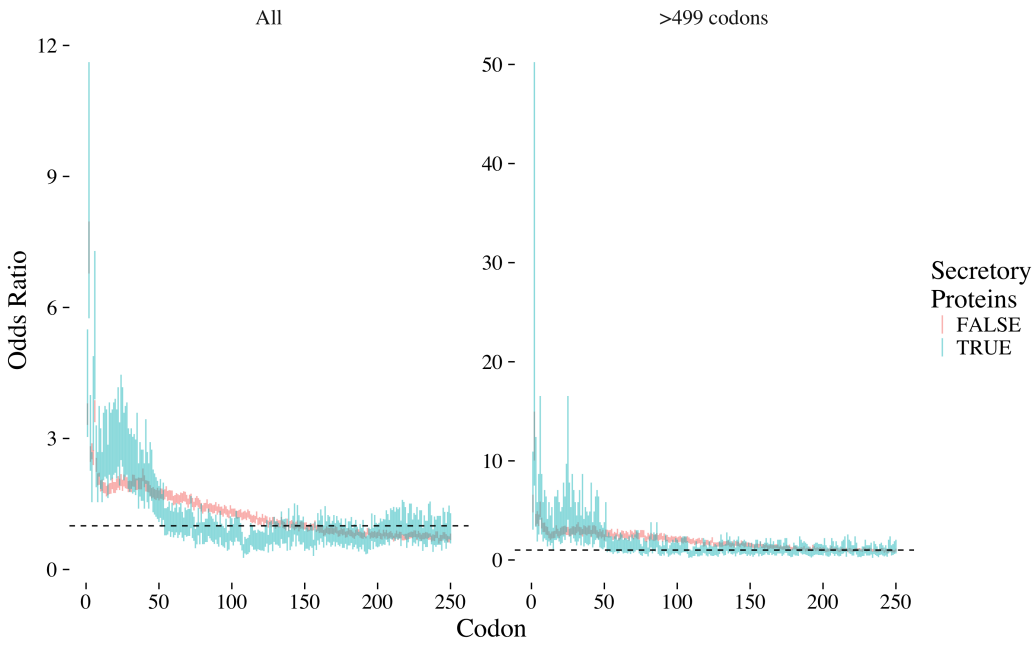


Fig. 5.3 The enrichment by slowly translated codons at the N-terminus of transcripts that encode for secretory proteins. The enrichment is calculated across the initial 250 codons of transcripts that have been aligned by their start codons. On the left, the enrichment for both secretory proteins (blue) and non-secretory proteins (red) is shown for when all transcripts are considered. On the right, the same analysis is presented, but only transcripts with lengths of at least 500 codons have been considered. The odds ratios are calculated with respect to all other codons, with the vertical bars indicating the confidence intervals on the OR. Values above one, which is highlighted via the dotted line, show a preference for slowly translated codons at a given position, while values below one show a preference for quickly translated codons.

between the signal peptide and the remaining peptide chain. Given this, we also tested the enrichment across this junction and found no indication that this was the cause.

5.2.2 Association of translation speed to the protein structure

It has been suggested that the translation speed of a codon influences the protein structure that it forms through the phenomenon of cotranslational folding [e.g., 68, 70, 105, 108, 145, 191–193]. This hypothesis has been tested in the past using various theoretical estimators of translation speed, as well as experimental studies that focused on a single protein and computational models of protein folding. These studies have collectively suggested many relationships between the

translation speed and the protein structure. However, as the theoretical estimators have been shown not to correlate well at the codon-level to the translation speed, as given by ribo-seq data, these relationships may be erroneous. As such, we tested many of the reported speed to structure relationships using ribo-seq data in place of the estimators to give the translation speed of a given codon. The results of our inferences are given in Table 5.1 in which we give the CMH-OR and its significance when the data is stratified by amino acid type. The stratification is essential as different amino acid exhibit different preferences for structural features. For example, proline is extremely constrained in the confirmations it can take due to its cyclic structure. As such, it is rarely found within either α-helices or β-strands, as it struggles to form the required geometry [194]. Given this, and also noting the encoding codons have different distributions within the ribo-seq data, accounting for the structural biases associated with each amino acid is essential [137, 138, 147]. To give further insight, Table 5.1 also contains the OR for each amino acid type separately.

Solvent accessibility

Previous studies have found that the fastest translated codons are found to exhibit a preference for sites within a protein's core [105–108]. A common hypothesis to explain this observation is that the faster the codon is translated, the higher its fidelity and hence a preference for usage at sites where errors would be structurally disruptive, most notably the protein core. The results in Table 5.1 support these hypotheses, finding that the exposed areas exhibit a significant preference for slowly translated codons (CMH-OR=0.96, $p < 0.05$), or conversely, rapidly translated codons prefer buried sites. From the separate ORs for each amino acid type, we note that this effect is most prominent in methionine, tryptophan, and phenylalanine (OR=0.82, 0.84, and 0.86 respectively), all of which have hydrophobic side chains which already bias them towards being place within the protein core [195]. This suggests that the translation speed may be used to further bias the placement of these hydrophobic residues.

Table 5.1 The enrichment of various structural elements by slowly translated codons. The relationship tested for in a given column is indicated by the header A/B, which stands for testing the enrichment of element A by slowly translated codons in comparison to element B. For each association tested, we give both the OR for each individual amino acid type as well as the CMH-OR for the overall hypothesis test where we have stratified the data by amino acid type. We also give the confidence interval and p-value for the CMH-OR value. Values above one for comparison A/B indicate that slowly translated codons preferentially encode for element A over element B.

AA	Stat	1st.Dom/2nd.Dom	Linker/Domain	Buried/Access	E/C	H/C	P/C	H/E	P/E	P/H
A	OR	2.12	0.72	0.93	0.95	1.15	1.02	1.22	1.08	0.89
C	OR	1.87	2.57	1.04	1.22	1.25	1.45	1.03	1.19	1.16
D	OR	1.66	1.74	0.87	0.85	0.85	1.16	1.00	1.36	1.35
E	OR	1.11	1.44	0.99	0.80	0.88	1.01	1.10	1.27	1.15
F	OR	1.92	0.77	0.86	0.98	0.90	0.77	0.92	0.79	0.86
G	OR	1.71	1.60	1.03	1.18	1.04	1.04	0.88	0.88	1.01
H	OR	1.52	1.11	0.98	0.88	0.83	0.72	0.94	0.82	0.87
I	OR	1.57	1.07	0.95	1.01	0.90	0.63	0.89	0.62	0.70
K	OR	1.82	0.95	1.01	0.95	0.93	0.98	0.97	1.03	1.06
L	OR	2.07	1.11	1.04	1.04	1.03	1.48	0.99	1.42	1.44
M	OR	1.77	1.67	0.82	0.92	0.70	0.61	0.76	0.66	0.87
N	OR	1.83	0.97	0.89	0.84	0.91	0.96	1.08	1.14	1.06
P	OR	2.08	0.56	0.92	0.91	1.14	0.92	1.25	1.01	0.81
Q	OR	1.66	0.87	0.94	1.10	0.95	0.94	0.86	0.85	0.98
R	OR	2.25	0.79	1.05	0.79	0.97	0.93	1.23	1.18	0.96
S	OR	1.67	1.03	0.89	0.79	0.94	1.03	1.19	1.31	1.10
T	OR	1.95	1.22	0.94	0.97	0.83	0.66	0.86	0.68	0.80
V	OR	2.17	0.96	1.01	0.81	0.81	0.84	1.00	1.04	1.04
W	OR	1.10	2.00	0.84	0.96	1.10	0.61	1.15	0.64	0.56
Y	OR	1.40	0.98	0.96	1.00	1.00	1.16	1.00	1.16	1.16
	CMH	1.77	1.06	0.96	0.93	0.94	1.00	1.00	1.04	1.05
	CMH.lower	1.64	0.94	0.93	0.90	0.92	0.94	0.97	0.97	0.98
	CMH.upper	1.90	1.20	0.98	0.97	0.97	1.06	1.04	1.12	1.12
	CMH.p	0.00	0.31	0.00	0.00	0.00	0.96	0.90	0.27	0.15

Linkers

It has been reported that linker regions in multidomain proteins may be enriched by slowly translated codons as a mechanism to allow the N-terminal domain to complete folding before the C-terminal domain emerges from the ribosome [112, 196]. It is hypothesised that without such a delay, the two sections of the nascent chain could interact which may lead to misfolding. However, conflicting results exist in the literature, stating either that a translation speed bias does not exist in this region, or, in opposition to above, that the region exhibits a preference for fast translating codons [113, 114]. The result in Table 5.1 does not prove or disprove this hypothesis, instead reporting the null hypothesis, that there is no inherent preference for fast or slowly translated codons within linkers regions in comparison to domain regions. The null hypothesis may be due to the variability we see across the ORs of the amino acids independently leading to low confidence in the CMH-OR. For example, cysteines in linker regions are far more likely to be slowly translated (OR=2.57), while in contrast prolines in linker regions exhibit a bias for fast translation (OR=0.56). This variability may also be, in part, due to the small number of linker regions within our dataset (N=58). As such, some residues are only observed a handful of times, most notably Tryptophan, which occurs only eight times within our defined linkers.

Domains

Past studies have found that in general the domains nearer the N-terminus both fold faster and are smaller than their C-terminus counterparts [197]. These observations would suggest that N-terminal domains may also be translated faster. Experiments comparing γ B-crystallin to its circularly permutated variant (the N and C-terminal domains exchanged) support this; the authors reporting that the ribosome spends more time on the natural C-terminal domain than the natural N-terminal domain regardless of the domain ordering [198]. We tested for the enrichment by slowly translated codons in N-terminal domains in comparison to the C-terminal

domains in the 46 proteins with exactly two domains in our dataset. From the result given in Table 5.1, we find that the first domain is significantly enriched by slowly translated codons when compared to the second domain. In fact, this bias is the largest we observe out of all our speed-structure inferences. This result, however, can not be disentangled from our previous observation that the translation speed increases across the entire length of a transcript. This gradual increase along the length would cause the bias between sequential domains we observe. Nevertheless, whether the domain biases cause the translation ramp, or vice-versa, or even both are present independently, cannot be stated within the confines of our analysis.

Secondary Structure

Many previous works have commented that translation speed differs between different secondary structure elements; however, disagreement exists about what the preferences are exactly [e.g., 105, 113–115, 178]. For example, Pechmann *et al.* found that within *S. cerevisiae*, β-strands are enriched with quickly translated codons, while α-helices are depleted of them [105]. In contrast, Zhou *et al.* found that both α-helices and β-strands prefer quickly translated codons, while coil regions prefer slowly translated codons [115]. We tested for such associations, comparing the enrichment of the various secondary structure elements to one another. The results are given both in Table 5.1 and Figure 5.4. They show that slowly translated codons significantly enrich coil regions in comparison to both helices and β-strands. Comparison of helices and strands gives the null hypothesis, suggesting that they are equally prevalent in these structures. Comparison of either of these three structural elements to positive-ϕ residues, also reports the null hypothesis, which suggests that there is not enough statistical power on this secondary structure type to perform inference. From these results, we can infer a translation speed ordering as helices $\approx$ β-strands $>$ coils. Finding that coils are slowest structural element translated is in agreement with prior work, where it is suggested that it allows the nascent chain more time to fold cotranslationally the regions prior to the coil. Note that this result is

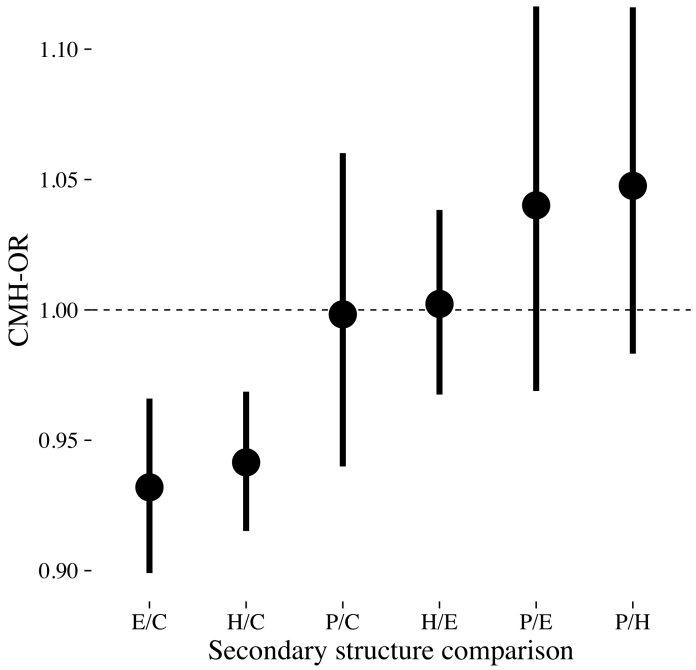


Fig. 5.4 Enrichment of protein secondary structure elements by slowly translated codons. The CMH-OR is shown for comparisons between different secondary structure elements, namely helices, β-strands, coils and residues with positive-ϕ angles, labelled as H, E, C and P respectively. The label A/B indicate testing for the enrichment by slowly translated codons in the element A when compared to element B. Stratification for the CMH-OR is by amino acid type. The line range represents the confidence interval. Values above one for comparison A/B indicate that slowly translated codons preferentially encode for element A over element B.

entangled with the solvent accessibility bias we found above, as coils are more likely to be found outside of the protein core [199].

Along a helix

The above results show that fast and slowly translated codons exhibit different preferences with regards to what structural elements they encode. However, the results do not indicate whether codons which encode the same secondary structure type are further distinct from one another. Existing work suggests this may be the case, Pechmann *et al.* reporting that at the N-terminus of α-helices there exists a distinct pattern of fast and slowly translated codons [105]. Note that helices are the only secondary structure type that can commonly form within the ribosome

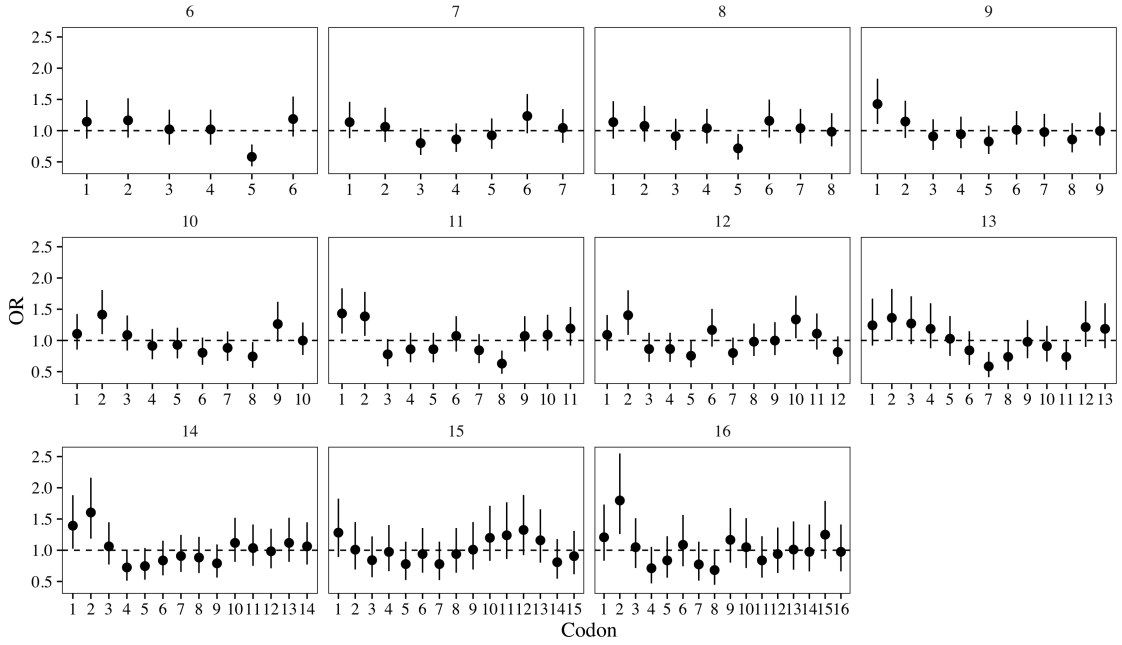


Fig. 5.5 The enrichment of helices by slowly translated codons along their length facetted by helix length. The OR for a given codon position has been calculated with respect to codons at other positions within helices of the same length. The plot shows the OR for each codon position within each helix length, the codons of each helix length grouped together. Each facet represents a different helix length, while the line range represents the confidence interval on these values. Values above one, which is highlighted via the dotted line, show a preference for slowly translated codons at a given position within helices of the given length, while values below one indicate a preference for quickly translated codons

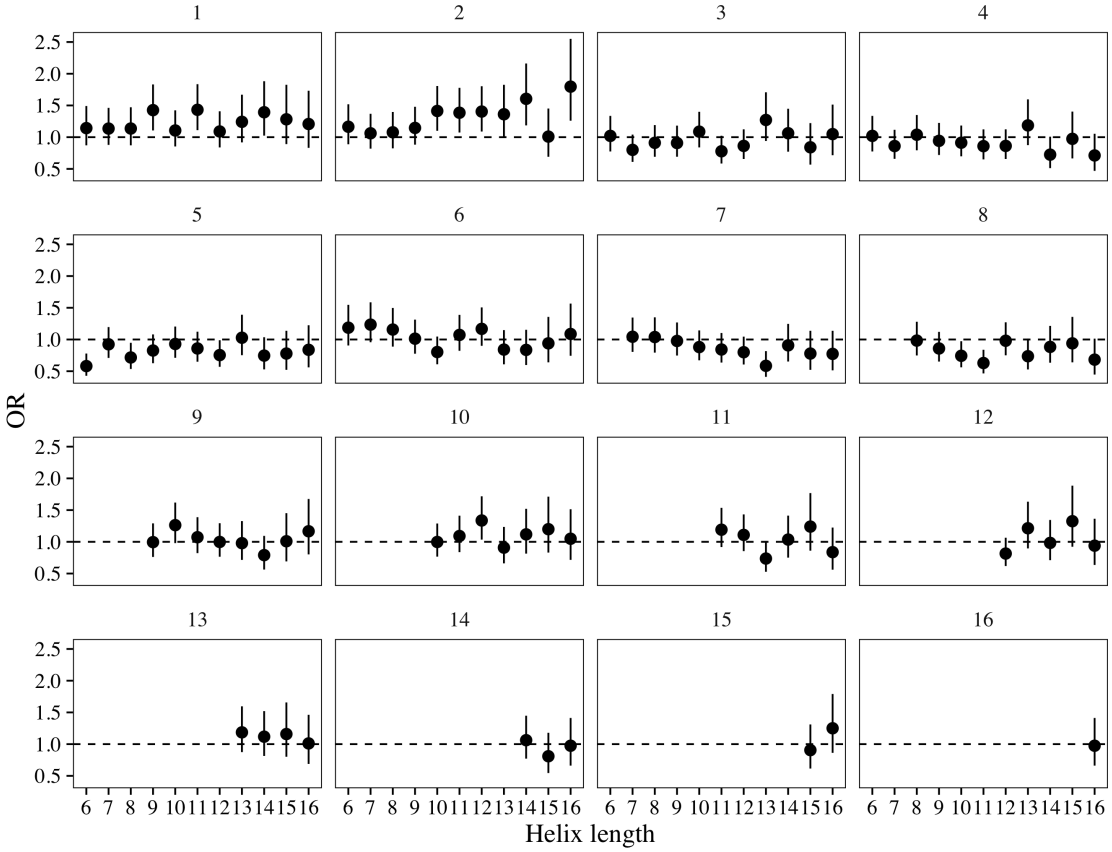


Fig. 5.6 The enrichment of helices by slowly translated codons along their length facetted by position. The OR for a given codon position has been calculated with respect to codons at other positions within helices of the same length. The plot shows the OR for each codon position within each helix length, grouped by position relative to the helix start, e.g. all the initial codons together, all the second codon together, e.t.c. Each facet represents a different position along the helix, while the line range represents the confidence interval on these values. Values above one, which is highlighted via the dotted line, show a preference for slowly translated codons at a given position within helices of the given length, while values below one indicate a preference for quickly translated codons

tunnel and, as such, are more likely to exhibit finer translation biases [200]. We investigated whether differences existed within helices by testing for the enrichment of slowly translated codons at each position along their length. To perform this analysis, we first extracted all helices within our dataset before subsetting to only those between six to sixteen amino acids long. Collectively, this amounted to 2764 helices, with those eight residues long being the most common (N=310). We then calculated the OR at each position in comparison to only the other codons from helices of the same length. Note that this differs from the previous tests, as we do not attempt to account for differences in amino acid composition along the helix length.

In Figure 5.5 and 5.6, we give the results of this analysis, showing the OR at each position for each helix length. In the former, the helices of each length are plotted independently, while in the latter, the helices were aligned by their initial codon, and then the codons at each position grouped, e.g. all the initial codons together, all the second codon together, e.t.c. We find that for all lengths, both the first and the second codon exhibit a preference for slowly translated codons, though only for some helix lengths is this significant. This enrichment indicates that translation may pause before helices are translated, or that once a helix begins translating the remaining section is rapidly translated. Alternatively, these observations could be related to helix capping which biases towards the use of certain motifs in the sequence at the start and end of helices [201]. Note that this enrichment is in addition to the biases reported above when comparing helices to other secondary structure types. As with the translation ramps, we also tested whether aligning the helices by their final codon uncovered any bias at the C-terminus. The results, which are not shown, did not find any significant associations.

5.2.3 Using translation speed to improve secondary structure predictions

Above we have shown that there are many associations between the speed with which a codon is translated and the resultant structural element it encodes for. Given this, we tested the relevance of this work in improving the structure predictions by seeing if including the speed

Table 5.2 Coefficients and their significance for the multinomial secondary structure prediction model using amino acid sequence as the sole predictor. Values of zero for the significance indicate the p-value is less than the machine precision.

Predictor	Strand (coef)	Strand (p)	Helix (coef)	Helix (p)	Pos.Phi (coef)	Pos.Phi (p)
(Intercept)	-0.76	0	0.47	0	-2.5	0
Amino Acid (C)	0.34	2.4×10^{-5}	-0.59	0	-0.035	0.84
Amino Acid (D)	-0.78	0	-0.9	0	0.51	7.1×10^{-8}
Amino Acid (E)	-0.33	8.3×10^{-9}	-0.17	1×10^{-5}	0.079	0.46
Amino Acid (F)	0.54	0	-0.26	1.3×10^{-8}	-0.24	0.064
Amino Acid (G)	-0.14	0.021	-0.93	0	2.9	0
Amino Acid (H)	-0.15	0.034	-0.66	0	0.55	5.6×10^{-6}
Amino Acid (I)	0.99	0	-0.092	0.029	-1.2	1.1×10^{-12}
Amino Acid (K)	-0.19	0.00043	-0.32	2.2×10^{-16}	0.46	3.1×10^{-6}
Amino Acid (L)	0.48	0	0.076	0.043	-0.78	1.5×10^{-9}
Amino Acid (M)	0.32	5×10^{-5}	-0.061	0.32	0.21	0.19
Amino Acid (N)	-0.73	0	-0.96	0	0.99	0
Amino Acid (P)	-1.4	0	-1.7	0	-2.2	0
Amino Acid (Q)	-0.22	0.0011	-0.17	0.00038	0.2	0.099
Amino Acid (R)	-0.041	0.51	-0.18	0.00012	0.54	8.5×10^{-7}
Amino Acid (S)	-0.41	2.1×10^{-14}	-0.86	0	-0.1	0.32
Amino Acid (T)	0.06	0.26	-0.8	0	-0.59	3.4×10^{-6}
Amino Acid (V)	1	0	-0.29	4.7×10^{-12}	-0.93	6.3×10^{-10}
Amino Acid (W)	0.41	3.6×10^{-6}	-0.16	0.026	-0.032	0.88
Amino Acid (Y)	0.44	6.9×10^{-14}	-0.32	1.4×10^{-11}	-0.32	0.024

Table 5.3 Coefficients and their significance for the multinomial secondary structure prediction model using amino acid sequence and translation speed as predictors. Values of zero for the significance indicate the p-value is less than the machine precision.

Predictor	Strand (coef)	Strand (p)	Helix (coef)	Helix (p)	Pos.Phi (coef)	Pos.Phi (p)
(Intercept)	-0.74	0	0.49	0	-2.5	0
Amino Acid (C)	0.34	2.2×10^{-5}	-0.59	0	-0.036	0.84
Amino Acid (D)	-0.77	0	-0.89	0	0.51	8.8×10^{-8}
Amino Acid (E)	-0.32	1.7×10^{-8}	-0.17	2×10^{-5}	0.077	0.47
Amino Acid (F)	0.54	0	-0.26	1.2×10^{-8}	-0.24	0.065
Amino Acid (G)	-0.13	0.03	-0.92	0	2.9	0
Amino Acid (H)	-0.15	0.041	-0.66	0	0.55	6.1×10^{-6}
Amino Acid (I)	0.98	0	-0.092	0.028	-1.2	1.1×10^{-12}
Amino Acid (K)	-0.19	0.00051	-0.32	2.2×10^{-16}	0.46	3.3×10^{-6}
Amino Acid (L)	0.48	0	0.076	0.043	-0.78	1.5×10^{-9}
Amino Acid (M)	0.32	5.8×10^{-5}	-0.063	0.31	0.21	0.19
Amino Acid (N)	-0.73	0	-0.96	0	0.99	0
Amino Acid (P)	-1.4	0	-1.6	0	-2.2	0
Amino Acid (Q)	-0.22	0.00087	-0.17	0.00029	0.2	0.097
Amino Acid (R)	-0.039	0.53	-0.17	0.00014	0.54	8.8×10^{-7}
Amino Acid (S)	-0.41	2.4×10^{-14}	-0.86	0	-0.1	0.32
Amino Acid (T)	0.059	0.26	-0.81	0	-0.59	3.5×10^{-6}
Amino Acid (V)	1	0	-0.29	4.9×10^{-12}	-0.93	6.3×10^{-10}
Amino Acid (W)	0.41	4.2×10^{-6}	-0.17	0.024	-0.031	0.88
Amino Acid (Y)	0.44	3.9×10^{-14}	-0.32	2.5×10^{-11}	-0.32	0.024
RUST (1)	-0.058	0.0038	-0.051	0.0015	0.021	0.51

classification (fast/slow) improved predictions of the secondary structure above using the amino acid sequence alone. Using the amino acid sequence alone has shown to give an accuracy of roughly 50%, which is considerably lower than the most advanced methods which have an accuracy of >80% [58, 202]. Further details can be found in the Methods section, but broadly, we created multinomial logistic regression models, fitting to a training set which consisted of 80% of the codons with a secondary structure annotation. These models were then tested against the remaining 20% to assess their accuracy. The coefficients and their significance are shown in Tables 5.2 and 5.3 for the sequence only and the sequence+speed models respectively. We found that including the speed classification in addition to the amino acid sequence decreased the residual variance significantly ($p < 10^{-4}$). Furthermore, the coefficients associated with speed classification were significant in separating all bar residues with positive-ϕ angles from the coils; coils used as the baseline. However, when the accuracy was assessed on the training set via a confusion matrix, we found that both models were equally predictive (46.1% of predictions were correct). In fact, the predictions for every codon were unchanged between models. This suggests that the additional information provided by the speed classification either improves the predictions of codons whose secondary structure was already correctly predicted by the sequence or does not change the predictions of those wrongly classified enough for them to be correctly predicted.

5.3 Discussion

An active area of investigation is whether there exists a relationship between the translation speed of a given codon and the protein structure it encodes for. Previous attempts to establish this relationship have been undertaken using various theoretical estimates of the translation speed which make predictions of the speed using either the codon bias or a physical property, such as the tRNA concentration. However, we showed in the previous chapter that these theoretical estimators do not adequately predict the translation speed as inferred from experimental

ribo-seq data. Therefore, the existing results that showed links between the translation speed and the protein structure may be erroneous. In this chapter, we have examined the association of translation speed, as given by ribo-seq, to various aspects of the protein structure. This is the first time, to our knowledge, that translation speed to protein structure relationship has been tested using only experimentally derived speeds and experimentally derived structures.

To perform our analysis, we classified every codon in our dataset as either quickly or slowly translated and tested for their association to various structural features while accounting for differences in the amino acid composition. We found that slowly translated codons exhibit significant preferences for the N-terminus of a transcript, the first domain over the second domain in multi-domain proteins, and for exposed positions over buried sites. We also noted that linker regions exhibit no biases with regards to translation speed in comparison to the surrounding domains. Next, we compared the translation speed preference of the various structural elements finding that coils are more enriched by slowly translated codons than either helices or β-strands, which exhibit similar preferences. Naturally, many of theses associations are interconnected, most notably that preference for slowly translated codons within the first domain and nearer the N-terminus. As such, further work is required to see which biases remain when accounting for others. Nevertheless, that any such biases exist at all suggests that the manner in which a protein is produced affects the resultant protein structure.

Given our results, we next investigated whether the translation speed could be used to improve protein structure prediction. We built simple models that attempted to predict the secondary structure of a given residue using either the amino acid sequence alone or in combination with our translation speed classification. The results showed that the latter model which included the translation speed classes performed significantly better than the simpler model when the residual variance was compared. However, if instead the overall accuracy of the models was compared, no difference between them was reported. Given this, there may be information contained within the translation speed pertinent to the protein structure, albeit

marginally. As such, further studies should attempt to incorporate translation speeds into more advanced secondary structure prediction or even three-dimensional protein structure prediction.

5.4 Conclusion

In summary, we found various associations between the translation speed, as given by ribo-seq data, and the protein structure produced. This is the first time that these relationships have been investigated on a transcriptome-wide scale using both experimentally derived translation speeds and protein structures. Having found these associations, we assessed whether the protein's secondary structure could be predicted better when the translation speed is considered in addition to the protein sequence. Analysis showed that predictions were significantly improved, albeit only marginally. That such a relationship exists betweens the protein structure and the translation speed suggests that the manner in which a protein is produced affects resultant protein structure. This conclusion provides strong support for the cotranslational folding hypothesis.

Chapter 6

Conclusion and future work

6.1 Conclusion

In recent years, many experimental studies have produced results that show synonymous mRNA sequences may be translated into proteins with different physical properties [e.g., 88, 97, 98, 203]. Such outcomes suggest that there is information contained in the mRNA sequence pertaining to the structure of the encoded protein above and beyond mere specification of the amino acid sequence. Within this thesis, we have investigated whether the relative translation speed of a codon in an mRNA transcript may act as this additional information. We tackled two central and currently outstanding questions related to this hypothesis, namely:

1. What is an appropriate transcriptome-wide measure of the translation speed?

2. Does the speed with which a codon is translated influence the protein structure it encodes for?

Both of these questions have been researched before by others, primarily via the construction of estimators which attempt to predict the translation speed from the mRNA sequence using either the codon bias or physical properties, such as the tRNA concentration [e.g., 101, 102, 105, 106, 112, 113]. However, we compared four of the most well-known estimators in Chapter

2 and found that they these differ significantly to one another in their predictions of the translation speed. As such, the relationships we inferred when we related the different estimates to the protein structure encoded varied wildly. This variance combined with there being no experimental measure, historically, with which they could be compared objectively meant that determining which estimator performed best proved infeasible.

In Chapter 3, we introduced Ribo-seq, an experimental method that is promoted as the first experimental transcriptome-wide measurement of the translation speed [135, 136]. By sequencing only the fragments of RNA that are protected by a ribosome at the point of cell lysis, the relative translation speed of codons on a transcript may be inferred from the relative number of reads that align to them [137, 138]. Given the relevance to our work, we constructed an extensive database containing an unprecedented amount of ribo-seq data which was reprocessed to remove any inherent study based biases. Simultaneously to our efforts, two other databases were also produced, the GWIPS-viz database and the RPFdb, both of which we compared to our own [153, 154]. We found that the GWIPS-viz database was generated using an analogous pipeline to that used in our own database's construction. However, we deemed it superior overall due to the sheer quantity of studies and diverse range of species it contained.

Having decided to continue our work using the ribo-seq datasets available in the GWIPS-viz database, we analysed the reproducibility of the ribo-seq method in Chapter 4. Another study had indicated that the translation profiles of replicates within a ribo-seq experiment could differ drastically and, as such, we wished to know the extent of this issue across different studies [163]. We compiled over 100 ribo-seq experiments from 15 different studies and found that the translation profiles of transcripts are not reproducible for all but the most highly sampled. We then subsetted our collated datasets to only the translation profiles that were reproduced in more than one study, before averaging this subset to form a high-quality dataset. Our high-quality ribo-seq dataset should, in theory, give reproducible results.

Next, we compared the translation profiles within our high-quality dataset to those produced by the theoretical estimators investigated in Chapter 2. We found that some of the estimators performed adequately when the translation speed values were averaged by codon type. However, when compared explicitly at the codon-level, none of the estimators could capture the variability observed within the ribo-seq translation profiles. We concluded that no estimator could be considered a suitable proxy for the translation speed from which we inferred that the speed-structure biases found in Chapter 2 were likely erroneous.

In Chapter 5, we repeated the analysis in Chapter 2 in which we attempted to relate the translation speed to the protein structure produced. However, we replaced the estimated translation speeds with the translation profiles of our high-quality ribo-seq dataset. This is the first time, to our knowledge, that the speed-structure relationships have been investigated on a transcriptome-wide scale using both experimentally derived translation speeds and experimentally derived protein structures. We found that the translation speed, as given by ribo-seq data, does in fact bias the protein structure produced. For example, codons that encode coil regions are translated slower on average than codons that encode other secondary structure types. Relating these results back to our central questions stated above, we would conclude that ribo-seq is an appropriate large-scale measurement of translation speed and that the speed with which a codon is translated does have a tangible effect on the protein structure produced.

6.1.1 Limitations and improvements

We concluded in Chapter 3 that the GWIPS-viz database was better than our database primarily due to its sheer size. One issue, however, that may have affected the subsequent work undertaking using data drawn from it, is that the GWIPS-viz database generates translation profiles from the aligned reads using a common offset of 15 nucleotides [153]. This offset is applied universally to all datasets and without regards to each read's specific length. As discussed in Chapter 3, this may introduce small errors when inferring the position of the ribosome A-site

on the order of one or two nucleotides [147]. Consequentially, the overall codon-level accuracy will be negatively affected, which, in turn, will have diminished the reproducibility observed in Chapter 4 and stymied the speed-structure biases unearthed in Chapter 5. The drop in accuracy is somewhat mitigated by us summating the counts across the constituent nucleotides of a given codon to give that codon's count in a translation profile. This effect, however, could have been entirely avoided if the optimal offset for each read length within each dataset was determined. Determining these offsets would require that all datasets were reprocessed from the raw state as the final alignments from which the translation profile is constructed are not available for the GWIPS-viz database. Given this, our database would likely prove more appropriate starting point if this improvement was to be enacted. If done, we would expect improved signal in the subsequent analyses undertaken.

An alternative database improvement would be expanding the range of species contained, such that high-quality ribo-seq datasets were available for multiple organisms. While this study does represent the first time that the relationship between translation speed and protein structure has been investigated on a transcriptome-wide scale using purely experimental data, the conclusions drawn in the final chapters are limited scope due to the analyses focusing on only a single species, *S. cerevisiae*. Given this, caution should be taken if extrapolating these results to other organisms or assuming them to be general properties of protein production. Most notably, prokaryotic ribosomes translate significantly faster than eukaryotic ribosomes, and, as such, assuming the relationships unearthed here would be unchanged when considering the former is precarious [204, 205]. On a technical note, the analysis presented in Chapter 5 is not easily transferred to more complex Eukaryotes that feature higher amounts of splicing. Exons used in differing quantities will produce jumps or drops in the associated number of ribosomes along the length of transcripts built from them. Handling the effect of splicing on the translation profiles generated by ribo-seq is currently an open question within the field.

6.2 Future work

6.2.1 Conservation of translation speed

Ribo-seq, and the translation speeds inferred from its counts, have been analysed with respect to many features of the coding sequence [e.g., 144, 146, 151, 164]. However, no work, to our knowledge, has explored whether the translation profiles of transcripts with similar sequences are also alike. Measuring the conservation of the translation speed with respect to the conservation of the sequence would provide insight into the importance of the translation speed relative to the coding sequence. For example, it is still unknown the extent to which a mutation, synonymous or otherwise, to the coding sequence will change the translation speed on average. If single mutations do not appear to alter the translation speed profile greatly, the profile could potentially act as a fingerprint to a given group of transcripts with a similar sequence.

The conservation of translation speed observed with respect to the sequence conservation is measured by comparing the translation profiles of homologous sequences. Similarly, the conservation of translation speed occurring with respect to structural conservation is measured via comparing the translation profiles of structurally homologous proteins. Our work already indicates that the translation speed provides structural information above and beyond specification of the amino acid and, as such, the comparison of translation profiles of structurally similar proteins should highlight which of the biases are of most importance. Crucially, structural similarity can occur without high levels of sequence similarity. As such, comparisons of translation profiles between structurally homologous proteins with low sequence identity, facilitated by structural alignment, could help clarify the degree to which the translation speed provides structural information beyond specification of the amino acid. For example, we would expect a higher level of similarity in the translation profiles within the members of a structurally homologous group than between those with differing structures.

6.2.2 Incorporating translation speed into protein structure prediction

Given that we unearthed multiple relationships between the translation speed of a given codon and the protein structure it encodes for, it is not unreasonable to predict that *de novo* protein structure prediction could be improved by the inclusion of the translation speed. We performed a preliminary investigation in Chapter 5 in which we found that including the translation speed marginally improved secondary structure prediction above using the amino acid alone. A more in-depth investigation as to the benefit of the translation speed within three-dimensional *de novo* protein structure prediction would be of significant interest to the community.

A possible avenue for this investigation is to modify the protein folding prediction software SAINT to include the information provided by the translation speed [206]. SAINT is a protein structure prediction tool that mimics cotranslational folding by sequentially attaching protein fragments to a growing nascent chain to construct a protein. At each step, SAINT allows for either the chain to be extended by another fragment, or the fold space of the current nascent chain to be explored to find a better global confirmation. The ratio between attachment operations and time spent searching the fold space at each step is currently proportional to the current nascent chain length. If one can use translation speed to direct variation of this ratio during translation, it may result in improved predictions. Specifically, codons that are translated slowly should increase the number of operations spent searching the fold space, while those rapidly translated should cause more immediate attachment of the next fragment. If the inclusion of translation speed leads to an improvement in the predictions, it confirms the influence that the translation speed had the on the protein structure and provides further evidence for the cotranslational folding hypothesis.

References

[1] Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K., and Linn, S. (2004) Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem.*, **73**, 39–85.

[2] Sontag, E. M., Vonk, W. I. M., and Frydman, J. (2014) Sorting out the trash: The spatial nature of eukaryotic protein quality control. *Curr. Opin. Cell Biol.*, **26**, 139–146.

[3] Verghese, J., Abrams, J., Wang, Y., and Morano, K. A. (2012) Biology of the heat shock response and protein chaperones: Budding Yeast (Saccharomyces cerevisiae) as a model system. *Microbiol. Mol. Biol. Rev.*, **76**, 115–158.

[4] Land, M., et al. (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–61.

[5] Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.*, **30**, 1083–1090.

[6] Jeffery, C. J. (2003) Moonlighting proteins: Old proteins learning new tricks. *Trends Genet.*, **19**, 415–417.

[7] Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.

[8] Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

[9] Dunker, A. K., et al. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.

[10] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) *Molecular biology of the cell*. Garland Science, 4th edn.

[11] Crick, F. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.*, **12**, 138–166.

[12] Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.

[13] Shapiro, J. A. (2009) Revisiting the central dogma in the 21st century. *Ann. N. Y. Acad. Sci.*, **1178**, 6–28.

[14] Watson, J. D. and Crick, F. H. (1953) A structure for deoxyribose nucleic acid. *Nature*, **171**, 737–8.

[15] Watson, J. D. and Crick, F. H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.

[16] Crick, F. H. (1954) The complementary structure of DNA. *Proc. Natl. Acad. Sci. U. S. A.*, **40**, 756–8.

[17] Larralde, R., Robertson, M. P., and Miller, S. L. (1995) Rates of decomposition of ribose and other sugars: implications for chemical evolution. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 8158–8160.

[18] Joyce, G. F. (2002) The antiquity of RNA-based evolution. *Nature*, **418**, 214–221.

[19] Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., Vendeix, F. A., Fabris, D., and Agris, P. F. (2011) The rna modification database, rnamdb: 2011 update. *Nucleic Acids Res.*, **39**, D195–D201.

[20] Moore, P. B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.

[21] Hurowitz, E. H. and Brown, P. O. (2003) Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. *Genome Biol*, **5**, R2.

[22] DeHaseth, P. L., Zupancic, M. L., and Record, M. T. (1998) RNA polymerase-promoter interactions: The comings and goings of RNA polymerase. *J. Bacteriol.*, **180**, 3019–3025.

[23] Ponting, C. P. and Hardison, R. C. (2011) What fraction of the human genome is functional? *Genome Res.*, **21**, 1769–1776.

[24] Doolittle, W. F. (2013) Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 5294–300.

[25] Palazzo, A. F. and Gregory, T. R. (2014) The case for junk DNA. *PLoS Genet.*, **10**, e1004351.

[26] Peters, J. M., Vangeloff, A. D., and Landick, R. (2011) Bacterial transcription terminators: the RNA 3'-end chronicles. *J. Mol. Biol.*, **412**, 793–813.

[27] Warner, J. R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.

[28] Bartel, D. P. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

[29] Guhaniyogi, J. and Brewer, G. (2001) Regulation of mRNA stability in mammalian cells. *Gene*, **265**, 11–23.

[30] Matlin, A. J., Clark, F., and Smith, C. W. J. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–98.

[31] Lynch, K. W. and Maniatis, T. (1996) Assembly of specific sr protein complexes on distinct regulatory elements of the drosophila doublesex splicing enhancer. *Genes Dev.*, **10**, 2089–2101.

[32] Mignone, F., Gissi, C., Liuni, S., and Pesole, G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.

[33] Gilbert, R. J. C., Fucini, P., Connell, S., Fuller, S. D., Nierhaus, K. H., Robinson, C. V., Dobson, C. M., and Stuart, D. I. (2004) Three-dimensional structures of translating ribosomes by Cryo-EM. *Mol. Cell*, **14**, 57–66.

[34] Wool, I. G. (1979) The structure and function of eukaryotic ribosomes. *Annu. Rev. Biochem.*, **48**, 719–754.

[35] Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.

[36] Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S., and Stormo, G. (1981) Translational initiation in prokaryotes. *Annu. Rev. Microbiol.*, **35**, 365–403.

[37] Voorhees, R. M. and Ramakrishnan, V. (2013) Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.*, **82**, 203–236.

[38] Jenni, S. and Ban, N. (2003) The chemistry of protein synthesis and voyage through the ribosomal tunnel. *Curr. Opin. Struct. Biol.*, **13**, 212–219.

[39] Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2012) Termination and post-termination events in eukaryotic translation. *Adv. Protein Chem. Struct. Biol.*, **86**, 45–93.

[40] Yoshida, T., Wakiyama, M., Yazaki, K., and Miura, K.-i. (1997) Transmission electron and atomic force microscopic observation of polysomes on carbon-coated grids prepared by surface spreading. *Japanese Soc. Electron Microsc.*, **46**, 503–506.

[41] Crick, F. H. (1966) Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.

[42] Varani, G. and McClain, W. H. (2000) The G·U wobble base pair. *EMBO Rep.*, **1**, 18–23.

[43] Chan, P. P. and Lowe, T. M. (2009) GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.

[44] Ogle, J. M. and Ramakrishnan, V. (2005) Structural insights into translational fidelity. *Annu. Rev. Biochem.*, **74**, 129–177.

[45] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.

[46] Berman, H. M., Westbrook, J. D., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

[47] Zhang, C. T. and Zhang, R. (1999) Skewed distribution of protein secondary structure contents over the conformational triangle. *Protein Eng.*, **12**, 807–10.

[48] Dukka, B. K. (2016) Recent advances in sequence-based protein structure prediction. *Brief. Bioinform.*, p. bbw070.

[49] Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 11963–8.

[50] Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2009) A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 21149–54.

[51] Anfinsen, C. B. and Haber, E. (1961) Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, **236**, 1361–1363.

[52] Levinthal, C. (1969) How to fold graciously. *Mossbauer Spectrosc. Biol. Syst. Proc. a Meet. held Allert. House, Monticello, Illinois*.

[53] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.

[54] Ballew, R. M., Sabelko, J., and Gruebele, M. (1996) Observation of distinct nanosecond and microsecond protein folding events. *Nat. Struct. Biol.*, **3**, 923–926.

[55] Planck Collaboration (2015) Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.*.

[56] Halliday, D., Resnick, R., and Walker, J. (2011) *Fundamentals of physics*. Wiley.

[57] Gelman, H. and Gruebele, M. (2014) Fast protein folding kinetics. *Q. Rev. Biophys.*, **47**, 95–142.

[58] Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015) JPred4: A protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.

[59] Rooman, M., Dehouck, Y., Kwasigroch, J. M., Biot, C., and Gilis, D. (2002) What is paradoxical about Levinthal paradox? *J. Biomol. Struct. Dyn.*, **20**, 327–329.

[60] Dill, K. a. and Chan, H. S. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.*, **4**, 10–19.

[61] Marino, J., Von Heijne, G., and Beckmann, R. (2016) Small protein domains fold inside the ribosome exit tunnel. *FEBS Lett.*, **590**, 655–660.

[62] Ross, J. F. and Orlowski, M. (1982) Growth-rate-dependent adjustment of ribosome function in chemostat-grown cells of the fungus Mucor racemosus. *J. Bacteriol.*, **149**, 650–653.

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

[64] Roux, P., Ruoppolo, M., Chaffotte, A. F., and Goldberg, M. E. (1999) Comparison of the kinetics of S-S bond, secondary structure, and active site formation during refolding of reduced denatured hen egg white lysozyme. *Protein Sci.*, **8**, 2751–60.

[65] Guzman, I. and Gruebele, M. (2014) Protein folding dynamics in the cell. *J. Phys. Chem. B*, **118**, 8459–8470.

[66] Zhou, H. X. (2013) Influence of crowded cellular environments on protein folding, binding, and oligomerization: Biological consequences and potentials of atomistic modeling. *FEBS Lett.*, **587**, 1053–1061.

[67] Clark, P. L. (2004) Protein folding in the cell: Reshaping the folding funnel. *Trends Biochem. Sci.*, **29**, 527–534.

[68] Deane, C. M., Dong, M., Huard, F. P. E., Lance, B. K., and Wood, G. R. (2007) Cotranslational protein folding - Fact or fiction? *Bioinformatics*, **23**, i142–i148.

[69] Nilsson, O. B., et al. (2015) Cotranslational protein dolding inside the ribosome exit tunnel. *Cell Rep.*, **12**, 1533–1540.

[70] Wang, E., Wang, J., Chen, C., and Xiao, Y. (2015) Computational evidence that fast translation speed can increase the probability of cotranslational protein folding. *Sci. Rep.*, **5**, 15316.

[71] Kiho, Y. and Rich, A. (1965) A polycistronic messenger RNA associated with beta-galactosidase induction. *Proc. Natl. Acad. Sci. U. S. A.*, **54**, 1751.

[72] Melancont, P. and Garoff, H. (1987) Processing of the Semliki Forest virus structural polyprotein: role of the capsid protease. *J. Virol.*, **61**, 1301–1309.

[73] Kudlicki, W., Chirgwin, J., Kramer, G., and Hardesty, B. (1995) Folding of an enzyme into an active conformation while bound as peptidyl-tRNA to the ribosome. *Biochemistry*, **34**, 14284–14287.

[74] Frydman, J., Erdjument-Bromage, H., Tempst, P., and Hartl, F. U. (1999) Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. *Nat. Struct. Mol. Biol.*, **6**, 697–705.

[75] Fedorov, A. N. and Baldwin, T. O. (1999) Process of biosynthetic protein folding determines the rapid formation of native structure. *J. Mol. Biol.*, **294**, 579–586.

[76] Cabrita, L. D., Hsu, S.-T. D., Launay, H., Dobson, C. M., and Christodoulou, J. (2009) Probing ribosome-nascent chain complexes produced in vivo by NMR spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 22239–22244.

[77] Holtkamp, W., Kokic, G., Jager, M., Mittelstaet, J., Komar, A. A., and Rodnina, M. V. (2015) Cotranslational protein folding on the ribosome monitored in real time. *Science*, **350**, 1104–1107.

[78] Kim, S. J., Yoon, J. S., Shishido, H., Yang, Z., Rooney, L. A., Barral, J. M., and Skach, W. R. (2015) Translational tuning optimizes nascent protein folding in cells. *Science*, **348**, 444–448.

[79] Morrissey, M. P., Ahmed, Z., and Shakhnovich, E. I. (2004) The role of cotranslation in protein folding: a lattice model study. *Polymer (Guildf).*, **45**, 557–571.

[80] Senturk, S., Baday, S., Arkun, Y., and Erman, B. (2007) Optimum folding pathways for growing protein chains. *Phys. Biol.*, **4**, 305–16.

[81] Ellis, J. J., Huard, F. P. E., Deane, C. M., Srivastava, S., and Wood, G. R. (2010) Directionality in protein fold prediction. *BMC Bioinformatics*, **11**, 172.

[82] Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., and Ulrich Hartl, F. (2013) Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.*, **82**, 323–355.

[83] Han, Y., David, A., Liu, B., Magadán, J. G., Bennink, J. R., Yewdell, J. W., and Qian, S.-B. B. (2012) Monitoring cotranslational protein folding in mammalian cells at codon resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 12467–12472.

[84] Curran, J. F. and Yarus, M. (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J. Mol. Biol.*, **209**, 65–77.

[85] Goldman, D. H., Kaiser, C. M., Milin, A., Righini, M., Tinoco, I., and Bustamante, C. (2015) Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo. *Science*, **348**, 457–460.

[86] Borg, A. and Ehrenberg, M. (2015) Determinants of the rate of mRNA translocation in bacterial protein synthesis. *J. Mol. Biol.*, **427**, 1835–1847.

[87] Pedersen, S. (1984) Escherichia coli ribosomes translate in vivo with variable rate. *EMBO J.*, **3**, 2895–2898.

[88] Sander, I. M., Chaney, J. L., and Clark, P. L. (2014) Expanding Anfinsen's principle: Contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.*, **136**, 858–861.

[89] Sørensen, M. A. and Pedersen, S. (1991) Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.*, **222**, 265–280.

[90] Sueoka, N. (1962) On genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. U. S. A.*, **48**, 582–&.

[91] Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., and Sockett, R. E. (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, **33**, 1141–1153.

[92] Supek, F. (2016) The code of silence: Widespread associations between synonymous codon biases and gene function. *J. Mol. Evol.*, **82**, 65–73.

[93] Quax, T. E. F., Claassens, N. J., Söll, D., and van der Oost, J. (2015) Codon bias as a means to fine-tune gene expression. *Mol. Cell*, **59**, 149–161.

[94] Plotkin, J. B. and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, **12**, 32–42.

[95] Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728–730.

[96] Percudani, R., Pavesi, A., and Ottonello, S. (1997) Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J. Mol. Biol.*, **268**, 322–330.

[97] Komar, A. A., Lesnik, T., and Reiss, C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, **462**, 387–391.

[98] Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., and Gottesman, M. M. (2007) A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, **315**, 525–528.

[99] Buttgereit, F; Brand, M. D. (1995) A hierarchy of ATP-consuming processes in mammalian cells. *Biochem. J.*, **312 (Pt 1**, 163–167.

[100] Bosdriesz, E., Molenaar, D., Teusink, B., and Bruggeman, F. J. (2015) How fast-growing bacteria robustly tune their ribosome concentration to approximate growth-rate maximization. *FEBS J.*, **282**, 2029–2044.

[101] Gingold, H. and Pilpel, Y. (2011) Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.*, **7**, 481.

[102] Deane, C. M. and Saunders, R. (2011) The imprint of codons on protein structure. *Biotechnol. J.*, **6**, 641–649.

[103] Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurence of the respective codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.*, **151**, 389–409.

[104] Sharp, P. M. and Li, W.-h. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.

[105] Pechmann, S. and Frydman, J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.*, **20**, 237–43.

[106] Zhou, T., Weems, M., and Wilke, C. O. (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.*, **26**, 1571–1580.

[107] Lee, Y., Zhou, T., Tartaglia, G. G., Vendruscolo, M., and Wilke, C. O. (2010) Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, **10**, 4163–4171.

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

[109] Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.

[110] Goodman, D. B., Church, G. M., and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science*, **342**, 475–479.

[111] Mahlab, S. and Linial, M. (2014) Speed controls in translating secretory proteins in eukaryotes - an evolutionary perspective. *PLoS Comput. Biol.*, **10**, e1003294.

[112] Thanaraj, T. A. and Argos, P. (1996) Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.*, **5**, 1973–1983.

[113] Saunders, R. and Deane, C. M. (2010) Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.*, **38**, 6719–6728.

[114] Brunak, S. and Engelbrecht, J. (1996) Protein structure and the sequential structure of mRNA: alpha-Helix and beta-sheet signals at the nucleotide level. *Proteins Struct. Funct. Genet.*, **25**, 237–252.

[115] Zhou, M., Wang, T., Fu, J., Xiao, G., and Liu, Y. (2015) Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.*, **97**, 974–987.

[116] dos Reis, M., Savva, R., and Wernisch, L. (2004) Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.*, **32**, 5036–5044.

[117] Clarke, T. F. and Clark, P. L. (2008) Rare codons cluster. *PLoS One*, **3**, e3412.

[118] Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

[119] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

[120] Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.*, **241**, 132–145.

[121] Larkin, M. A., et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

[122] Sievers, F., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

[123] Fox, N. K., Brenner, S. E., and Chandonia, J. M. (2014) SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–9.

[124] Bateman, A., et al. (2015) UniProt: A hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

[125] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., Poux, S., Bougueleret, L., and Xenarios, I. (2016) Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.

[126] Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–21.

[127] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

[128] O'Leary, N. A., et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

[129] Kapushesky, M., et al. (2011) Gene Expression Atlas update - a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, p. gkr913.

[130] Pechmann, S., Chartron, J. W., and Frydman, J. (2014) Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat. Struct. Mol. Biol.*, **21**, 1100–1105.

[131] Mantel, N. (1963) Chi-Square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.*, **58**, 690–700.

[132] Breslow, N. E., Day, N. E., et al. (1980) *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.*, vol. 1.

[133] Dunn, O. J. (1961) Multiple comparisons among means. *J. Am. Stat. Assoc.*, **56**, 52–64.

[134] Saunders, R. and Deane, C. M. (2010) Protein structure prediction begins well but ends badly. *Proteins Struct. Funct. Bioinforma.*, **78**, 1282–1290.

[135] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.

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

[137] Dana, A. and Tuller, T. (2014) The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.*, **42**, 9171–9181.

[138] Dana, A. and Tuller, T. (2015) Mean of the typical decoding rates: A new translation efficiency index based on the analysis of ribosome profiling data. *G3*, **5**, 73–80.

[139] Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

[140] Consortium, S. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–14.

[141] Lee, S., Liu, B., Lee, S., Huang, S.-X. X., Shen, B., and Qian, S.-B. B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E2424–32.

[142] Li, G.-W., Oh, E., and Weissman, J. S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–541.

[143] Stadler, M. and Fire, A. (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, **17**, 2063–2073.

[144] Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., and Ziv-Ukelson, M. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.*, **12**, R110.

[145] Yang, J.-R., Chen, X., and Zhang, J. (2014) Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.*, **12**, e1001910.

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

[147] Dana, A. and Tuller, T. (2014) Properties and determinants of codon decoding time distributions. *BMC Genomics*, **15**, S13.

[148] Qian, W., Yang, J. R., Pearson, N. M., Maclean, C., and Zhang, J. (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.*, **8**, e1002603.

[149] O'Connor, P. B. F., et al. (2016) Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.*, **7**, 12915.

[150] Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., and Koller, D. (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.*, **10**, 770.

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

[152] Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *Elife*, **3**, e03735.

[153] Michel, A. M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P. B. F., Heaphy, S. M., Mullan, J. P. A., Donohue, C. A., Higgins, D. G., and Baranov, P. V. (2014) GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.*, **42**, D859–64.

[154] Xie, S. Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., Liu, Y., Ren, J., and Xie, Z. (2016) RPFdb: A database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **44**, D254–D258.

[155] Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.*, **7**, 1534–50.

[156] Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

[157] Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., and Press, W. H. (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in Yeast. *PLoS Genet.*, **11**, e1005732.

[158] Michel, A. M., Mullan, J. P., Velayudhan, V., O'Connor, P. B., Donohue, C. A., and Baranov, P. V. (2016) RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol.*, **13**, 316–9.

[159] Li, W., Freudenberg, J., and Miramontes, P. (2014) Diminishing return for increased Mappability with longer sequencing reads: implications of the k-mer distributions in the human genome. *BMC Bioinformatics*, **15**, 2.

[160] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

[161] Artieri, C. G. and Fraser, H. B. (2014) Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.*, **24**, 2011–21.

[162] Stadler, M., Artiles, K., Pak, J., and Fire, A. (2012) Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. *Genome Res.*, **22**, 2418–2426.

[163] Diament, A. and Tuller, T. (2016) Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol. Direct*, **11**, 24.

[164] Del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z. (2015) Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet.*, **11**, e1005613.

[165] Stadler, M. and Fire, A. (2013) Conserved translatome remodeling in Nematode species executing a shared developmental transition. *PLoS Genet.*, **9**, e1003739.

[166] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

[167] Gerashchenko, M. V., Lobanov, A. V., and Gladyshev, V. N. (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 17394–9.

[168] Michel, A. M. and Baranov, P. V. (2013) Ribosome profiling: A Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip. Rev. RNA*, **4**, 473–490.

[169] Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., and Pierce, E. A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). **27**, 2518–28.

[170] Kent, W. J. (2002) BLAT - The BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

[171] Woolstenhulme, C. J., Guydosh, N. R., Green, R., and Buskirk, A. R. (2015) High-Precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep.*, **11**, 13–21.

[172] Pohl, A. and Beato, M. (2014) bwtool: A tool for bigWig files. *Bioinformatics*, **30**, 1618–1619.

[173] Yates, A., et al. (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.

[174] Li, G. W., Burkhardt, D., Gross, C., and Weissman, J. S. (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, **157**, 624–635.

[175] Gerashchenko, M. V. and Gladyshev, V. N. (2016) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.*, p. gkw822.

[176] Andreev, D. E., O'Connor, P. B. F., Zhdanov, A. V., Dmitriev, R. I., Shatsky, I. N., Papkovsky, D. B., and Baranov, P. V. (2015) Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biol.*, **16**, 90.

[177] Akaike, H. (1998) Information theory and an extension of the maximum likelihood principle. *Sel. Pap. Hirotugu Akaike*, pp. 199–213, Springer.

[178] López, D. and Pazos, F. (2015) Protein functional features are reflected in the patterns of mRNA translation speed. *BMC Genomics*, **16**, 513.

[179] Juntawong, P., Girke, T., Bazin, J., and Bailey-Serres, J. (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E203–12.

[180] Cock, P. J. A., et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

[181] Tompa, P. (2009) *Structure and function of intrinsically disordered proteins*. CRC Press.

[182] George, R. A. and Heringa, J. (2002) An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng. Des. Sel.*, **15**, 871.

[183] R Core Team (2016) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[184] Developer, T. J. A. (2012) *epitools: Epidemiology tools*.

[185] David Meyer Achim Zeileis and Hornik, K. (2016) *vcd: Visualizing Categorical Data*.

[186] Venables, W. N. and Ripley, B. D. (2002) *Modern applied statistics with S*. Springer, fourth edn.

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

[188] Tuller, T. and Zur, H. (2015) Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.*, **43**, 13–28.

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

[190] Zhang, S., Goldman, E., and Zubay, G. (1994) Clustering of low usage codons and ribosome movement. *J. Theor. Biol.*, **170**, 339–54.

[191] Ciryam, P., Morimoto, R. I., Vendruscolo, M., Dobson, C. M., and O'Brien, E. P. (2013) In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E132–40.

[192] Saunders, R., Mann, M., and Deane, C. M. (2011) Signatures of co-translational folding. *Biotechnol. J.*, **6**, 742–751.

[193] Chaney, J. L. and Clark, P. L. (2015) Roles for synonymous codon usage in protein biogenesis. *Annu. Rev. Biophys.*, **44**, 143–166.

[194] Otaki, J. M., Tsutsumi, M., Gotoh, T., and Yamamoto, H. (2010) Secondary structure characterization based on amino acid composition and availability in proteins. *J. Chem. Inf. Model.*, **50**, 690–700.

[195] Chothia, C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**, 338–339.

[196] Makhoul, C. H. and Trifonov, E. N. (2002) Distribution of rare triplets along mRNA and their relation to protein folding. *J. Biomol. Struct. Dyn.*, **20**, 413–20.

[197] Jacob, E., Unger, R., and Horovitz, A. (2013) N-terminal domains in two-domain proteins are biased to Be shorter and predicted to fold faster than their C-terminal counterparts. *Cell Rep.*, **3**, 1051–1056.

[198] Komar, A. A. and Jaenicke, R. (1995) Kinetics of translation of gammaB crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. *FEBS Lett.*, **376**, 195–198.

[199] Lodish, H. (2008) *Molecular cell biology*. W. H. Freeman.

[200] Tu, L. W. and Deutsch, C. (2010) A folding zone in the ribosomal exit tunnel for Kv1.3 Helix Formation. *J. Mol. Biol.*, **396**, 1346–1360.

[201] Aurora, R. and Rose, G. D. (1998) Helix capping. *Protein Sci.*, **7**, 21–38.

[202] Garnier, J., Osguthorpe, D., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

[203] Spencer, P. S., Siller, E., Anderson, J. F., and Barral, J. M. (2012) Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J. Mol. Biol.*, **422**, 328–335.

[204] Boehlke, K. W. and Friesen, J. D. (1975) Cellular content of ribonucleic acid and protein in Saccharomyces cerevisiae as a function of exponential growth rate: calculation of the apparent peptide chain elongation rate. *J. Bacteriol.*, **121**, 429–433.

[205] Dennis, P. P. and Nomura, M. (1974) Stringent control of ribosomal protein gene expression in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, **71**, 3819–23.

[206] de Oliveira, S. H. P. (2015) *Biologically inspired de novo protein structure prediction*. Ph.D. thesis.