The Role of Rare Codons in Protein Expression

Charlie Harrison

July 13, 2015

A thesis submitted in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy

of

University College London

I, Charlie Harrison, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

That the flow of information from gene sequence to protein sequence depends on the translation of a code that could literally be described as digital is a truly incredible feat of nature. However, the process of translation is a noisy, stochastic, kinetic process that depends on many factors. The redundancy in the genetic code allows the transmission of additional, analogue information by varying some of these factors. How organisms use the redundancy is termed codon usage, and rare codons are those that are typically shunned in favour of other synonymous options. Synonymous variations to the codon usage pattern of a gene have been linked to disease, and can have huge effects on the functionality and quantity of protein produced from a gene, but the nature of these variations is complex and poorly understood. In some cases, rare codons appear to have a beneficial influence on expression. This thesis investigates the phenomenon of rare codons and attempts to elucidate their evolutionary role in optimal gene expression. It begins with the design of a novel statistical algorithm, which is used to generate a dataset of interesting genetic locations. The dataset is the subject of a hypothesis-driven investigation to discover meaningful biological correlates, and this is complemented by experimental work, to attempt to provide conclusive validation of the approach.

Acknowledgements

Here is my thesis. It was hard, and it may not change the world after all, but I stuck it out, it's finished, and I'm proud of that. It wouldn't be complete without acknolwedging the vital help of a few people that got me through: Prof. John Ward, for acting as my primary supervisor. Prof. David Jones, for going far above and beyond the obligations of a secondary supervisor. Dr. Domenico Cozetto, for his help with the development of the algorithm and much invaluable advice about the project. Dr. Markus Gerschater and Sean Ward, for proposing the project and guiding the early stages. My small army of proof-readers: Zena Hadjivasiliou, Sophie Atkinson, Stuart Tetchner, Tomasz Kosciolek, Federico Minnecci, Benji Lichman, Jack Jeffries, and Greg Staw. All my friends and family, for their encouraging and sustaining words at dark times, and for taking an interest at lighter ones. Mum and Dad, for everything.

To Mel. You are so wonderful.

Contents

List of Figures

List of Tables

1 Introduction

To obtain a fully functional, correctly structured protein, a cell must pass information through three physical states (DNA, mRNA, protein) via two transitions (transcription and translation) – this is the “central dogma” of molecular biology (Crick, 1970). The nucleotide sequence of a gene uniquely determines the sequence of amino acids that make up the protein it produces. However, the process is more complex and less deterministic than this description allows.

The folding of a protein from a one dimensional peptide sequence into one or more functional domains arguably represents a third transition into a fourth state. Proteins do not infallibly attain their native structure upon translation, sometimes misfolding or forming aggregates. The mapping from protein sequence to structure is not injective either; while single amino acid changes can destroy the structure or function of a protein, most locations of most proteins are quite tolerant to substitutions, especially when the two residues have similar physicochemical properties (Ohta, 1973).

Because of this flexibility in amino acid sequence and the high level of redundancy in the genetic code, the constraints placed on the DNA sequence of a gene by the target protein structure are loose (Itzkovitz and Alon, 2007). This flexibility allows neutral variations that do not influence the propagation of information to take hold. There are also "non-canonical" characteristics of each stage that fall outside the realm of the genetic code as defined in the central dogma, but do affect the efficiency of information transfer and are therefore acted on by natural selection.

1.1 An Introduction to Codon Usage

An early finding in the field of genetic analysis established that although there are a great many synonymous coding options available to organisms, they are not used equally (Ikemura, 1981a). Even within synonymous sets of codons that are translated to the same amino acid residue, codons are selected with bias, and the bias varies between species (Gustafsson et al., 2004; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). Codon usage describes this bias and its variations, and there are numerous interesting findings that imply a selective role.

Within synonymous sets, organisms display strong preferences for some codons over others. These preferences can be observed in the frequency of usage, either across the whole genome or in a limited sample of genes. Preferred codons are used more frequently and are therefore described as common, as opposed to rare.

In endogenous genes, the degree of codon usage bias is positively correlated with expression level. Genes containing a higher proportion of common codons are transcribed more (Goetz and Fuglsang, 2005), and their protein products are more abundant (Tuller et al., 2007; Le Roch et al., 2004). These relationships have been shown to hold for heterologous genes that have been artificially added to a cell's genomic portfolio. Designing a gene so that its codon usage matches the preferences of the host organism, within the extensive space of synonymous coding options, often increases its expression level (Gustafsson et al., 2004). The translation rate of individual codons is variable and, broadly speaking, this appears to correlate with the usage frequencies of the codons, so that more commonly used codons are translated faster (Pedersen, 1984). Finally, codon usage frequencies are correlated with tRNA abundances, so that tRNAs that decode more common codons are present in greater abundances (Ikemura, 1981a, 1982).

This amounts to the most common explanation for codon usage bias: common codons are translated faster because they are decoded by more abundant tRNAs. Faster translation is generally preferable, because it allows the cell to respond more rapidly to environmental cues and ultimately to grow and multiply more quickly. However, this description does not capture the complexity of the selective landscape. The gene design strategy described above, where each amino acid is encoded by the most commonly used codon in the host organism, is far from infallible and recombinant expression is very often problematic (Purvis et al., 1987; Lavner and Kotlar, 2005; Angov et al., 2008; Welch et al., 2009; Agashe et al., 2013).

Section 1.2 describes the flow of genetic information from gene sequence to functional protein in more detail, introducing some of the factors that influence the specific sequence of a gene along the way. We then return to codon usage and the main subject of this thesis in more detail.

1.2 The Flow of Genetic Information

This section describes how information is transferred from DNA to functional proteins, and introduces indirect and biased influences on the process. Some of these influences are non-selective, and others are the result of the ability of genes to transfer additional, analogue information alongside the digital content that describes the amino acid sequence.

1.2.1 DNA

Although organism fitness depends on many phenotypic characteristics, the signal upon which evolution acts is genomic DNA. Random forces also constantly perturb DNA composition. Physiochemical modification of nucleotides, and occasional erroneous or biased synthesis or replication of DNA, alter genomic nucleotide content and thus codon usage. Meanwhile, the cell has manifold mechanisms in place to correct errors arising from these effects, but these are themselves occasionally biased (Marais, 2003; Nakken et al., 2010). This section discusses biases in occurrences of mutations and the cellular mechanisms that repair them, and the codependencies between these processes and the DNA sequence.

Mutations and Stability

Mutations are an unavoidable consequence of noise in the metabolic processes of a cell and the aggressive environmental influences to which it is exposed. They can take the form of changes to single nucleotides, called point substitutions, or insertions or deletions of one or more bases in one or both strands – generically termed replication slippage (Hartl and Jones, 1998).

Point substitutions occur with different frequencies between particular pairs of bases, depending on the molecular structure of the nucleotides. Some combinations of chemical modifications to bases can transform them into other bases entirely. For example, cytosine can transition to thymine by two chemical modifications, methylation and deamination. Deamination can be induced by UV radiation, enzymatic action or simply contact with water (Duncan and Miller, 1980; Lindahl, 1993). In isolation this modification produces uracil, and RNA base which is readily identified and excised by DNA repair systems. However, if this happens concurrently with methylation, the nucleotide produced is a thymine, creating a mismatch between the two strands in which the correct base is less readily distinguished. Methylation is widely used in gene regulation (Wolffe and Matzke, 1999) and methylated cytosine is particularly susceptible to deamination (Nakken et al., 2010), making this kind of mutation relatively likely. Additionally, each of the bases has alternative isomers that are capable of binding to bases other than their proper complement while still allowing the DNA to form a stable structure. For example, thymine in its regular "keto" form binds to adenine, but in its (rarer)"enol" form, with one hydrogen atom bound to the aromatic ring in a different position, it is capable of forming three hydrogen bonds and complementing guanine. Chemical modifications, for example, bromination, can create base analogues which shift between alternative isomers more easily (Anthony JF Griffiths et al., 2000). These modified bases can become fixed as mutations in some of the offspring after replication (Francino and Ochman, 2001).

G-C content, the proportion of complementary guanine-cytosine base pairings, is a useful summary statistic for comparing the genomes of different organisms. Mutation rates across all species of bacteria are biased towards increasing the proportion of complementary A-T pairings (Hershberg and Petrov, 2010). Despite this universal bias genomic G-C content varies extremely widely across bacteria, in the range 20-80% (Chen and Texada, 2006), and deviates from the equilibrium levels predicted from lab-measured mutation rates (Hildebrand et al., 2010). Cells must invest considerable metabolic resources in maintaining this discrepancy, but the reasons for this are unclear.

The code appears to be structured in such a way that the impact of point mutations is minimised. Synonymous codons are grouped so that most substitutions at the third codon position do not affect the polypeptide sequence. When point mutations do result in an amino acid change it is usually a like-for-like swap in terms of the properties of the two amino acids (Gilis et al., 2001). The identity of the stop codons also appears to reduce the probability of mutations that give rise to non-sense errors – the premature termination of translation, resulting in an incomplete and usually useless polypeptide at high cost to the organism (Goodarzi et al., 2004; Gilchrist and Wagner, 2006). Since the three stop codons include just two G-C nucleotides between them, this has been suggested as one reason for the preference for high G-C codons in many bacteria (Schmid and Flegel, 2011).

Mutations are non-uniform with respect to context as well as base composition, and can target specific motifs with prejudice. A well-known example is the under-representation of CpG tandems (Bird, 1980; McVean and Hurst, 2000; Nakken et al., 2010). DNA methyltransferases target the cytosine in the CpG dinucleotide, which is readily deaminated to form thymine, leading to a four-fold scarcity of CpG motifs compared with expectation (Greenbaum et al., 2014). The mutability of DNA also depends on the local stability of its molecular structure, because this determines its accessibility to mutagens (Nakken et al., 2010). This is influenced partly by G-C content, which determines the number of hydrogren bonds between complementary strands, but more strongly by the stacking interactions between adjacent bases in the helical structure (Yakovchuk et al., 2006).

Repair Mechanisms

Cells employ numerous systems to detect and correct mutations. These repair systems can be more complex than the replication systems themselves (Anthony JF Griffiths et al., 2000). Cohorts of enzymes detect mismatches between complementary strands, accidental breaks in one or both strands, and unpaired loops. Offending regions are excised by ligases, and breaks are repaired by polymerases. Methylation of the two strands is used to determine which strand has the correct sequence; since a lot of errors occur during synthesis, and methylation is a post-synthetic modification, the strand which is less methylated is modified to match the other (Modrich and Lahue, 1996; Kunkel and Erie, 2005). The systems are extremely active, catching upwards of 99.9% of mutations before they become fixed through replication (Harr et al., 2002; Alberts et al., 2008). Even a small bias in such an active process would influence genomic nucleotide content over evolutionary time scales (Harr et al., 2002).

Some systems that have been studied in detail do indeed reveal biases. Lu and Chang (1988) studied the mismatch repair system of *Escherichia coli* by transfecting cells with viral DNA containing restriction sites with deliberate mismatch errors. By exposing the DNA to restriction enzymes and assaying the products, they were able to deduce whether or not the mismatch had been correctly repaired. They found that some mismatch combinations were repaired more efficiently than others. They also varied the methylation states of the two strands, and measured the dominant nucleotide in an A-G mismatch by combining two restriction sites so that the direction of repair could be deduced from the fragments obtained. They discovered that the A-G mismatch was repaired with bias towards replacing the adenine with a cytosine, regardless of the methylation states of the two strands, although the degree of bias did vary with methylation. Jones et al. (1987) found that the efficiency of the mismatch repair mechanisms in *E. coli* is positively correlated with the G-C content of the surrounding 4-10 base pairs. This is another possible source of genomic G-C increase, because fixed mutations can lead to gene silencing which relaxes other pressures on nucleotide content, allowing that portion of DNA to diverge. Although mismatch repair systems vary slightly in mechanism, the same bias towards increasing G-C content applies in all studied organisms (Marais, 2003)

Another example of bias in DNA repair is in mismatches in series of short tandem repeats (STRs) in *Drosophila melanogaster*. Harr et al. (2002) compared the efficiencies of repair of slippage errors of STRs in wild-type *D. melanogaster* with that of a spel1$^{-/-}$ mutant strain lacking a functional repair system. They deduced that the repair system that corrects errors in repeat regions has a bias that tends to reduce the number of repeats. Moreover, the repair system is significantly more efficient at correcting slippages in (AT)^n regions than in (GC)^n regions. The authors postulated that both of these factors have had an influence on the *Drosophila* genome.

DNA replication can also alter nucleotide composition, through biased gene conversion. This is a genetic recombination event associated with sexual reproduction, where two homologous sections of DNA are combined, but it has also been reported in *E. coli* (Yamamoto et al., 1992). The underlying cause is again the mismatch repair system. When homologous genes are crossed over and recombined, exchanging a single strand, mismatches are often generated, and these are repaired asymmetrically. The effect on genome evolution has not been properly quantified, but biased gene conversion is thought to have played a significant role particularly in the development of certain eukaryotic genomes (Galtier and Duret, 2007; Duret, 2002; Harrison and Charlesworth, 2011).

1.2.2 Transcription

The process of transcription impacts neutrally on genomic coding regions. Mutation and repair have been shown to act differently on the coding and non-coding strands through several mechanisms. During transcription the non-coding strand is exposed, whereas the coding strand is shielded by RNA polymerases and the nascent RNA strand. This means the non-coding strand is more vulnerable to mutagens, such as UV radiation (Hendriks et al., 2010), that can cause lesions, deamination and other spontaneous mutations as described in Section 1.2.1. These mutational signals are strand-specific and proportional to expression levels (Beletskii et al., 2000), but are probably not dominant forces in shaping codon usage for two reasons. First, both these mechanisms produce a mutational bias that acts to increase the level of C→T transitions in the coding strand (Francino and Ochman, 2001), whereas codon usage patterns in highly expressed genes vary greatly between organisms and do not necessarily display this bias (Ikemura, 1985). Second, a study of the comparative codon bias in eukaryotic intronic sequences and the associated exons reveals distinct differences, despite the transcription of these regions being completely coupled (Duret and Mouchiroud, 1999). However, the contribution of this effect to overall patterns in all kingdoms of life has not been properly quantified and remains a confounding influence on analysis of selective forces (Duret, 2002).

Gene coding also imposes DNA sequence constraints not directly pertaining to the encoded polypeptide, both in and around the coding region. Specific promoter and repressor sequences upstream of initiation sites play a critical role in gene expression regulation, either by recruiting or preventing the binding of RNA polymerase. Slight variations in these sequences affect the affinity with which they bind the appropriate regulatory elements (Gustafsson et al., 2004; Jana and Deb, 2005).

The coding sequence of the gene itself is linked to transcriptional efficiency, acting as a further regulatory control via transcript level (Le Roch et al., 2004; Trotta, 2011). At the same time, location-specific transcriptional pauses can be programmed into the mRNA sequence. These pauses are thought to be very common, occurring approximately once every hundred nucleotides, and are implicated in numerous regulatory processes (Larson et al., 2014). These include recruitment of transcriptional and translational co-factors (Artsimovitch and Landick, 2002), correct folding of functional RNA (Pan et al., 1999), and proper termination (Weixlbaumer et al., 2013). Some repair mechanisms are triggered by the stalling of RNA polymerases during transcription, so spontaneous mutations that occur on the transcribed strand are corrected with greater efficiency (Francino and Ochman, 2001).

1.2.3 mRNA

An mRNA strand must possess certain features to facilitate the initiation of translation. In eukaryotes, the 5' end of the mRNA is capped with a sequence of modified nucleotides, and the ribosome detects this and looks for a nearby AUG start codon. In prokaryotes, initiation relies upon a sequence motif called the Shine-Dalgarno (SD) sequence. This is a series of five to eight nucleotides that must be present about eight bases upstream of the start codon, which binds part of the 16S RNA in the small ribosomal subunit. The lack of a cap in prokaryotic translation allows for polycistronic genes – multiple proteins encoded in a single mRNA strand (Alberts et al., 2008). In eukaryotes, the same term is used to describe sets of genes under a single promoter, but the transcribed mRNA is spliced and the new strands are capped before initiation (Blumenthal, 1998), so no constraint is placed on the coding segment of the mRNA.

The SD sequence is not entirely discriminate, and variations to it affect the strength of complementary binding to the ribosome and thus the efficiency of initiation (Gygi et al., 1999; Tuller et al., 2007). Ribosome resources are limited and initiation is the rate-limiting step in protein expression, so these variations can have a significant impact on protein production (Chu et al., 2011). At internal locations in the coding sequence SD-like motifs are generally under-represented, and have been implicated in frame-shifting (Berg and Silva, 1997) and ribosome stalling (Li et al., 2012).

Strands of mRNA are capable of forming complex structures that are stabilised through hybridisation between different parts of the strand. The ribosome has at least two mechanisms of helicase activity that help to unwind RNA secondary structure and ensure a basal rate of elongation, but translation can be hampered by excessive secondary structure occurring mid-way through a gene (Qu et al., 2011). If the secondary structure obscures the Shine-Dalgarno sequence or the start codon it can prevent initiation altogether (Plotkin and Kudla, 2011). Secondary structure in mRNA strands, especially near the 5' end, has been shown to be a powerful determinant in gene expression levels and there is strong selection against it (Kudla et al., 2009; Tuller et al., 2010b; Allert et al., 2010; Goodman et al., 2013). Genes also appear to be more conserved at sites where a synonymous mutation would disrupt mRNA secondary structure (Gu et al., 2014).

Finally, prokaryotic genomes are extremely dense, containing about 87% protein coding DNA on average (Rogozin et al., 2002). Genes are located in close proximity or even overlapping on the genome, and this can lead to conflicting evolutionary pressures on segments of neighbouring genes (Eyre-Walker, 1996). For example, the need to avoid mRNA secondary structure around the initiation region of one gene may shape the coding sequence of the 3' end of the upstream gene.

1.2.4 Translation

Translation accounts for over 60% of cellular ATP consumption in bacteria (Russell and Cook, 1995) and is the most complex stage of the gene expression process, involving a host of molecular components. The ribosomal complex is among the largest macromolecules in cells, comprising 52 proteins and 4.5 kilobases of RNA in prokaryotes, and 79 proteins and 6.8 kb of RNA in eukaryotes. Translation also requires initiation, elongation, and termination factors, tRNAs and their associated aminoacyl-tRNA synthetase (aaRS) enzymes, as well as the mRNA and amino acids themselves.

tRNA Abundance

In general, it is advantageous to produce proteins as quickly as possible - so much so that in single-celled organisms, protein production rate is almost synonymous with fitness (Sharp et al., 2010). Since initiation of translation is the rate-limiting step in gene expression, ribosome time is at a premium. Completing translation more rapidly helps to ease this bottleneck, making regulation more dynamic, which allows the cell to develop faster and respond to environmental cues with greater agility. It also reduces the amount of mRNA needed, and thus the metabolic cost of upregulating genes (Liljenström and von Heijne, 1987). Modulating the elongation rate of specific genes could also provide the cell with another mechanism to manipulate the relative quantities of proteins produced, giving it finer control (Parmley and Huynen, 2009).

As mentioned previously, cytosolic abundances of tRNA vary between tRNA species, and their abundances correlate broadly with the frequency of use of the corresponding codons, especially in highly expressed genes (Ikemura, 1981b; Percudani et al., 1997; Lavner and Kotlar, 2005). Elongation is a kinetic chemical process so its rate depends on the concentrations of the components. Location of a cognate aa-tRNA ternary complex is the rate-limiting step in elongation, much slower than GTP hydrolysis and translocation, peptidation, or aminoacylation of the discharged tRNA (Johansson et al., 2008). There is evolutionary pressure for the cell to upregulate the most commonly used tRNA species, and to improve the efficiency of elongation by altering the gene sequence to match the tRNA pool (Akashi, 2003; Shah and Gilchrist, 2011). The selective forces driving this co-adaptation are stronger for more highly expressed genes (Bulmer, 1991).

The abundance of the cognate tRNA is not the only determinant of translation rate. Near-cognate tRNAs carrying the wrong amino acid can also associate with the A-site. The ribosome undergoes a conformational change upon tRNA binding that closes the decoding centre tightly around the codon-anticodon pair, attempting to force it into a hybridised conformation (Khade et al., 2013), and if the tRNA does not match the codon it is rejected (although there is some flexibility in this; see below). Prior to the conformational shift a non-cognate tRNA with low-affinity binding may spontaneously dissociate; this is known as kinetic proofreading (Blanchard et al., 2004). The proofreading mechanisms help to ensure fidelity but incur a time penalty, as a cognate tRNA cannot locate in the A-site while it is occupied. The elongation rate is therefore a function not only of the abundance of the cognate tRNA, but of the whole tRNA pool, because near-cognate and non-cognate tRNAs compete and interfere with location of the correct tRNA (Fluitt et al., 2007; Zouridis and Hatzimanikatis, 2008). The abundances of near-cognate tRNAs are correlated, which somewhat mitigates the advantage of translation by an abundant tRNA (Shah and Gilchrist, 2011).

Wobble Pairing

The aforementioned redundancy in the genetic code adds to the complexity and energetic cost of translation. The ability to translate the full complement of codons necessitates a large number of tRNA genes and species which need to be transcribed and charged by specific aaRSs. The presence of extra tRNAs in the cellular milieu also complicates selection of the cognate tRNA during elongation. Cells mitigate these difficulties by utilising non-canonical nucleotide pairings in codon-anticodon interactions. This is known as wobble pairing. It allows tRNAs to decode multiple codons, and single codons to be decoded by multiple tRNAs (Yokoyama and Nishimura, 1995; Grosjean et al., 2010). The conformational shift at the ribosome that accompanies tRNA binding closes the decoding centre tightly around the first two complementary nucleotide pairs but allows more flexibility around the third base pair, thus limiting the scope of wobble and maintaining accuracy (Demeshkina et al., 2012).

Wobble pairing enables the cell to fully utilise the genetic code with a reduced complement of tRNA species. Organisms across the kingdoms use various strategies, many of them involving post-transcriptional modification of the anticodons of selected tRNA species, to translate all 61 sense codons with as few as 28 different tRNA genes (Grosjean et al., 2010). The energy saving appears come at a price as bacteria with fewer unique tRNA genes have lower growth rates (Rocha, 2004). The affinity of the codon-anticodon pairing affects the time it takes to locate a cognate tRNA, and also the rate of dissociation, so wobble paired codons are translated slower on average (Spencer et al., 2012; Sabi and Tuller, 2014). What proportion of a tRNA species is modified, how the tRNA pool is shared between its cognate codons, and how the altered binding affinities affect the kinetics of elongation remain open questions.

Steric tRNA Compatibility

Another physical effect that has been postulated to interfere physically with the action of translation is the steric compatibility of tRNA species at the A- and P-sites of the ribosome. Certain consecutive pairs of codons are dramatically under- or over-represented in the *E. coli* genome, even after correcting for amino acid associations and nucleotide level patterns (Irwin et al., 1995; Boycheva et al., 2003). This has been attributed to selection against physically incompatible permutations of tRNAs that hamper translational efficiency (Gutman and Hatfield, 1989). The way in which selection acts on these pairs is unclear and there is conflicting evidence about their effect. One study found that the discrepancy between actual and expected pair use was inversely correlated with local elongation rate, so that over-represented pairs are translated slower than under-represented ones (Irwin et al., 1995), but an attempt to reproduce this finding failed (Cheng and Goldman, 2001). Yet another study found that eukaryotic genes show autocorrelation in the tRNAs required to decode a gene, so that once a codon corresponding to a specific tRNA species has been used in a gene it is likely to be used again for subsequent occurrences of the same amino acid. Increasing the degree of tRNA autocorrelation in recombinant genes also increased the translation rate (Cannarrozzi et al., 2010). It was proposed that the mechanism behind this was tRNA recycling, or slow diffusion of tRNAs relative to recharging, but this is unconfirmed and the findings have not been corroborated in prokaryotes.

Accuracy

Occasionally the proofreading mechanisms of the translational machinery fail or a tRNA is charged with an incorrect amino acid resulting in a missense error, which can render the protein non-functional by altering the functional site or eventual structure. Misfolded proteins place a significant burden on the cell; they are a waste of resources, consume further metabolic energy as they have to be degraded (Arslan et al., 2011), and often cause toxicity directly by forming aggregates (Bucciantini et al., 2002). Missense error rates have been estimated as low as 3×10^{-7} in a cell free system analogous to *E. coli* (Johansson et al., 2008), and as high as 10^{-2} in *Bacillus subtilis* (Meyerovich et al., 2010). Generally error rates are stated as between 10^{-3} and 10^{-4} (Drummond and Wilke, 2008), amounting to at least one error in approximately one in seven proteins of average length (300 residues) in *E. coli*.

Error rates very likely vary between residues and the way they are encoded, so that translational accuracy shapes genetics. It is postulated that if a tRNA is present in higher concentration it is more likely to be correctly located by the corresponding codon before a near-cognate codon matches, although again this is mitigated by the correlation between the abundances of near-cognate tRNAs. Genes that code for more divergent proteins tend to use fewer common codons, and positions that are highly conserved on the amino acid level are often encoded by common codons (Akashi, 1994; Stoletzki and Eyre-Walker, 2007). Amino acid divergence across homologues suggests a greater tolerance to variation. Stoletzki and Eyre-Walker (2007) also discovered that codon bias is positively correlated with gene length and generally increases along the length of a gene, suggesting that as more resources are invested in the production of a protein, the pressure to finish it correctly becomes greater.

Local Variations in Elongation Rate

High translation rate as a bulk property is a selective advantage, but elongation rate is subject to local variations and there are cases where a transient reduction in elongation rate appears to improve the functional yield of protein. Many of these cases involve co-translational folding, which is discussed in detail in the next section, but there are other explanations. The prevalence of rare codons in the 5' region of many genes is fairly well-documented. As rare codons are generally thought to be translated slowly, one explanation for this is that the elongation rate is gradually ramped up as the gene is processed, in order to reduce ribosome density downstream and thus reduce the possibility of ribosomes disrupting each other mid-translation and potentially jamming (Tuller et al., 2010a; Navon and Pilpel, 2011). Where the phenomenon has been noted in proteins that are exported via the *sec*-pathway, the opposite interpretation has been advanced; that, in combination with high initiation rates, slow translation of the N-terminal region results in dense polysome structure, which increases the efficiency of recycling of the chaperone involved in the export process (Power et al., 2004; Zalucki et al., 2009). The idea of a translational ramp has been refuted (Charneski and Hurst, 2014), and it is perhaps more likely that N-terminal rare codons are the result of selection against mRNA secondary structure around the initiation region (see Section 1.2.3).

1.2.5 Protein

A polypeptide sequence is not guaranteed to attain the functional, native structure of a properly folded protein. Properties of both the sequence and target structure create a dependence on the translational machinery and the particulars of the translational process. The protein sequence can affect the rate of translation independently of the tRNA pool by altering the way the nascent chain emerges from the ribosome. A series of positively charged amino acids in the nascent chain is capable of slowing translation through electrostatic interactions with the negatively charged interior of the ribosomal exit tunnel (Lu and Deutsch, 2008). It has even been suggested that the normally untranslated poly(A) tail found in eukaryotic mRNA may act as protection against accidental frameshifts. If the stop codon is missed then the poly(A) region would translate into a string of positively charged lysines, which could stall translation and signal the ribosome complex for degradation (Charneski and Hurst, 2013). Ribosomal proteins that line the exit tunnel also relay signals to the surface upon interaction with nascent chains, aiding the recruitment of co-factors (Kramer et al., 2009).

Some proteins are capable of adopting the correct functional conformation upon re-folding from a denatured state (Anfinsen, 1973). This suggests that the native state is the globally stable thermodynamic minimum; mathematical models have indicated that protein sequences are expected to evolve towards more globally stable functional states (Govindarajan and Goldstein, 1998). However, refolding frequently does not yield a high proportion of functional protein (Fedorov and Baldwin, 1997; Huang et al., 2012). This is partly because the cell contains chaperones and other cofactors that aid protein folding, but can also be partly attributed to fact that protein structure develops concomitantly with and not independently of, translation.

Co-translational Folding

Proteins begin to fold long before they have been detached from the ribosome. Portions of the nascent chain can adopt helical secondary structure even before emerging from the exit tunnel (Wilson and Beckmann, 2011). Selection of each cognate tRNA takes in the order of tens to hundreds of milliseconds, whereas folding and unfolding can occur in less than ten milliseconds for simple proteins (O'Brien et al., 2012). This allows plenty of time for the nascent protein to explore the structural space before translation has completed. Once exposed to the cytosol the movement of the nascent chain is partially constrained by chaperone proteins and the surface of the ribosome itself, but it is still free to fold independently of the as-yet-untranslated portion – so much so that N-terminal domains can be fully functional before translation is completed (Nicola et al., 1999). This is key for protein maturation, because it allows proteins to fold in a more directed way, limiting the expanse of conformational space to be explored by the nascent chain. The full conformational space is prohibitively vast, so that proteins cannot feasibly visit even a large fraction of possible conformations in any reasonable timescale (Levinthal, 1968). The size of the conformational space increases with nascent chain length, so an immature protein has fewer possible structures to explore and can select the correct intermediate more easily.

Limited exploration of possible conformations carries the interesting corollary that the native structure of a protein is not necessarily the global free energy minimum for the sequence. Rather, it may simply be locally stable, with neighbouring conformations that impose a large energetic barrier to unfolding, so that the activation energy required to attain the global minimum is prohibitively high (Sohl et al., 1998; Baker, 1998; Baskakov et al., 2001). Such protein structures are said to fold under kinetic rather than thermodynamic control; the population of states is determined by the energy barriers between them, rather than the energies of the states themselves. Given sufficient time, the globally stable, non-native structure may become more heavily populated, but in shorter timescales the local minimum dominates (Baker, 1998; Huard et al., 2006; Fleishman and Baker, 2012).

If the nascent chain can adopt intermediate conformations that direct the folding pathway, and the relative populations of alternative conformations depends on their energetics and on the time available, then elongation rate can have an influence on the final conformation of the protein. A stable folding intermediate with a relatively high activation energy may become more populated by pausing translation at a key stage, driving the protein down the correct folding pathway (Purvis et al., 1987; O'Brien et al., 2012, 2014; Gloge et al., 2014). This is a critical point, because it means that myriad properties of a gene and protein can influence the final structure by modulating the elongation rate and thus the amount of time available for the emerging portion of the nascent chain to adopt an intermediate conformation.

1.3 Focus on Codon Usage

The process of producing fully functional proteins from genes is extremely complex with many contributing and confounding factors. Codon usage is a tractable, transparent, readily available and universal signal arising from this process, and it has garnered a lot of interest because of its apparent biological relevance. The most highly expressed endogenous genes in many organisms have the highest degree of codon bias (Ikemura, 1981a, 1982; dos Reis et al., 2004), and selecting the most frequently used codons in the genome of the host species when designing recombinant genes can produce dramatic increases in the amount of protein produced in a given time (Gustafsson et al., 2004; Kudla et al., 2009). Fast translation is the major determinant of selection in prokaryotes (Sørensen and Pedersen, 1991; Sharp et al., 2010), and so the cell evolves to translate common codons quickly and to use fast-translated codons more often (Plotkin and Kudla, 2011).

However, there is compelling evidence that the actual fitness landscape for protein translation is a good deal more subtle. Despite the weak selective advantage that can be conferred by a change to the rate of elongation at a single codon (Liljenström and von Heijne, 1987; Bulmer, 1991; Sharp et al., 2010), several experiments have identified synonymous changes to only a handful of codons that have a dramatic impact on the functionality of the protein product. Moreover, many of these cases show that swapping rare codons for more “optimal" ones actually hampers functional expression. This seemingly disproportionate effect is likely due to the need for the elongation rate of some proteins to be modulated at specific stages of their synthesis to aid co-translational folding (see Section 1.2.5), and this modulation comes from the coding sequence (Varenne et al., 1984).

Sander et al. (2014) demonstrated that translation rate variations arising from differences in synonymous codon usage can have a direct influence on co-translational folding. They designed a protein consisting of three “half-domains”. The N- and C-terminal half-domains are each capable of hybridising with the central half-domain to form different fluorescent proteins. This interaction is competitive, and the result can be determined from the fluorescent properties of the folded peptide. The investigators found that re-folding the protein from a denatured state produced a 50-50 mix of the two fluorophores, whereas folding co-translational production of the protein results in 2-fold enriched N-terminal protein. Substitution of rare codons in the C-terminal region enhances the imbalance. This shows that the protein forms co-translationally, and that the folding outcome can be influenced by altering the codon usage of the transcript. Kimchi-Sarfaty et al. (2007) found that removing a patch of rare codons actually changed the substrate specificity of a drug transporter; a structural mechanism was implicated by the altered sensitivity of the protein product to a specific protease. Komar et al. (1999) performed a similar modification on a gene sequence and achieved an increase in expression levels, but a 20% reduction in enzyme activity per mg of *E.coli* chloramphenicol acetyltransferase, a single domain enzyme. Zhang et al. (2009) went further and showed that the same effect is obtained when a multidomain *E. coli* model protein, SufI, is translated in the presence of an excess of the rare tRNA species corresponding to the patch of rare codons, strongly suggesting a link with elongation kinetics. Another experiment used mutant ribosomes whose elongation rates could be modulated, and found that slowing the elongation rate improved the functional yield of a complex eukaryotic protein expressed in *E. coli* (Siller et al., 2010). One recent study used homology to look for regions of conserved rare codons and concluded that they were present in most genes (Widmann et al., 2008).

Direct measurements of translation rates of individual codons are extremely challenging. *In vitro* methods have been developed, using advanced fluorescence microscopy techniques (Uemura et al., 2010; Chen et al., 2013) or direct measurement of force fluctuations during translation on an mRNA hairpin attached to an optically-trapped bead (Wen et al., 2008). These experiments are impressive and revealing, but they must be conducted in carefully constructed environments that bear little relation to the cytosol, rendering the rate measurements biologically irrelevant. Early *in vivo* methods were based on specially designed inserts attached to a reporter gene (Bonekamp et al., 1989; Sørensen and Pedersen, 1991). These findings suggest no correspondence between tRNA levels and translation rates, and only a broad correspondence between codon frequencies and translation rates. The constructs involve long strings containing a large number of repeats of the same codon and as such are rather artificial, and moreover the experiments have only been performed for a small number of codons. More recently a promising technique based on RNA sequencing has been used to give high-throughput readouts of the positions of ribosomes on translating mRNA strands (Oh et al., 2011). This approach is known as "ribosome profiling" or "ribosome footprinting". One major study using the technique found that codon usage and tRNA abundance bore no relation to ribosome occupancy, and that translational pausing was driven by Shine-Dalgarno-like sequences (Li et al., 2012). However, the technique is still in development and has a few outstanding questions. Particularly, the data analysis does not appear to account for initiation rates, which are the major determinant of overall translation rates (Bulmer, 1991; Gilchrist and Wagner, 2006; Chu et al., 2011) and depend non-linearly on mRNA levels and the precise sequence of the initiator region (Michel and Baranov, 2013). Data on cytosolic tRNA abundances would support the development and wider application of more sophisticated models of translation, but such data is difficult to obtain and only exists for a handful of species (Ikemura, 1981a, 1982; Dong et al., 1996; Kanaya et al., 1999; Zaborske et al., 2009). Further, current assay methods do not show all modifications or the proportion of a tRNA species that is charged, and cannot reflect dynamic regulation of the tRNA pool. The number of copies of tRNA genes is often used as a proxy for experimental data, but this is a blunt and imprecise measurement given the many sources of variation between this and active cytosolic abundances. Furthermore, when measurements of the translation rates of individual codons have been attempted, they do not appear to correspond to the relative abundances of their tRNA species (Bonekamp et al., 1989). One study found a more than three-fold variation in the translation rates of two codons predicted to be translated by the same tRNA (Sørensen and Pedersen, 1991).

The evolutionary pressures facing an organism are engraved in its codon usage preferences. Without accurate direct measurements of translation rates that can be compared across related proteins, codon usage is the best signal available for assessing non-canonical signatures of selection on translation. When a gene is encoded in a way that defies the typical preferences, it is difficult to determine whether this is a deliberate response to an unusual selective landscape or simply the result of the random fluctuations arising in genetic drift. This requires a specially designed approach to assessing codon usage at specific genetic locations. A great many computational methods have been developed in the past to try to assess the degree of codon usage bias in genes and organisms, with widely varying approaches and assumptions. The next section describes some of these methods.

1.4 Measures of Codon Usage Bias

Numerous attempts have been made to quantify the phenomenon of codon usage bias. Approaches vary widely in their scope, reliance on background information, and whether and how they correct for gene length and amino acid composition. This section describes some of the most prevalent metrics. [Note: notation has been adjusted from the original references for consistency where necessary]

1.4.1 Frequency of Optimal Codons

In one of the studies of codon usage bias, Ikemura (1981b) developed the notion of optimal and non-optimal codons based on their measurements of tRNA levels in *E. coli*. This binary classification could be considered crude in the light of more recent developments. The metric predicted an optimal codon for each amino acid according to rules based on their usage frequencies, the identity of the codon-anticodon pairing, and the measured abundances of tRNAs. The frequency of optimal codons F_{op} is simply the proportion of codons in a gene classified as optimal:

$$F_{op} = \frac{n_{optimal}}{n_{total}}$$

This ranges between 0 and 1 and is completely independent of gene length or amino acid content.

1.4.2 Effective Number of Codons

The effective number of codons (Wright, 1990) is a measure of how fully a gene uses the genetic code, based on an older method of assessing genetic variation in an organism with a number of copies of a certain gene. Its theoretical range is from 20, representing selective use of just one codon for every amino acid, to 61, for even use of every codon available for each amino acid, which is supposed to give an intuitive idea of the range of coding options taken. It is calculated from individual gene sequences based only on the universal genetic code, and corrects for amino acid content and gene length.

Each amino acid A has a synonymous set of k codons, and the number of instances of each codon in the gene is given by $n_{i=1:k}$. The total number of instances of A in the corresponding protein is $N = \sum\limits_{i=1}^{k} n_i$. The proportional occurrence of each codon is then $p_i = n_i/N$. The evenness of the selection between codons can be calculated as

$$\hat{F}_A = \frac{N \sum_{i=1}^{k}(p_i^2) - 1}{N - 1}$$

[see Nei and Tajima (1981) for an analogous derivation of this relating to population genetics]

This has the range $[1/k, 1]$. The $\hat{F}_A$ values are split into five groups according to the value of k, as dictated by the genetic code. The number of members in each of these groups is divided by the group average, and summed to give

$$\hat{N}_c = 2 + \frac{9}{\bar{\hat{F}}_{k=2}} + \frac{1}{\bar{\hat{F}}_{k=3}} + \frac{5}{\bar{\hat{F}}_{k=4}} + \frac{3}{\bar{\hat{F}}_{k=6}}$$

In the case that an amino acid is not present in the sequence, the averages are adjusted to compensate. This measure is an intuitive way of measuring the skew of a single gene, but the failure to utilise the context of a gene or any information about the organism it comes from is a significant limitation to its usefulness.

1.4.3 Relative Synonymous Codon Usage

Relative synonymous codon usage (RSCU) was first used as an independent measure by Sharp et al. (1986). It generates a value for each codon representing, in Sharp and Li's own words, “the observed number of occurrences divided by that expected if usage of synonymous codons was uniform". Mathematically this can be expressed as

$$\mathrm{RSCU}_{ji} = \frac{n_{ji}}{\frac{1}{k_j}\sum_{i=1}^{k_j} n_{ji}}$$

where $j \in [1, 20]$ represents the amino acid, $k_j \in \{1, 2, 3, 4, 6\}$ is the number of synonymous codons for that amino acid, $i \in [1, k_j]$ indicates the codon within the synonymous set. A gene is then represented by a set of 59 values (one for each codon that has a synonymous alternative), which have a hypothetical range of 0 to between 2 and 6, depending on the number of codons in the synonymous group. This is of limited use in isolation, but is of significant relevance because the equation forms the basis of the Codon Adaptation Index, an important codon bias measure – see below.

1.4.4 %MinMax

The %MinMax algorithm Clarke and Clark (2008) is location-specific method that uses a moving window to give local measures of codon usage that vary throughout a gene, rather than generating an aggregate figure for the entire gene. It is based on comparison of the usage frequency of codons in a window of a gene with the minimum, maximum and average usage frequency values given the amino acid composition.

For a window of size n, the Max, Min, and Actual terms are calculated by averaging the per-thousand usage frequencies of the most frequent and rarest codons for the amino acids in the window, and the codons used in the actual gene, respectively. The Avg term is calculated by averaging the average of the synonymous sets of the amino acids across the window. The Actual term is then compared to the average, and %MinMax is calculated as

$$\%\text{MinMax} = \begin{cases} \dfrac{\text{Actual} - \text{Avg}}{\text{Max} - \text{Avg}} \times 100 & \text{if Actual} > \text{Avg} \\ -\dfrac{\text{Avg} - \text{Actual}}{\text{Avg} - \text{Min}} \times 100 & \text{if Actual} < \text{Avg} \end{cases}$$

This yields a figure in the range $[-100, 100]$, where negative figures indicate rarer-than-average codon usage. The idea of comparison with some measure of expectation is given the amino acid sequence is good, but the average used is not weighted by expectation – it is a simple geometric average that does not account for the fact that codons with high usage frequencies by definition occur more often than those with lower usage frequencies, skewing the averages of observed windows. This results in there being many more high-usage regions than low, and the overall average is high.

1.4.5 Codon Adaptation Index

Also developed by Sharp and Li (1987), the Codon Adaptation Index (CAI) was for a long time the gold standard metric. It retains interest, although it has to some extent been taken over by the related tRNA Adaptation Index (tAI – see below). It uses statistics on codon usage from a highly expressed set of genes as the basis for its measure of bias. It corrects for gene length and, to an extent, amino acid composition, although the possible range of values is constrained by the number of synonymous coding options.

First, a set of genes that are highly expressed in the organism of interest must be selected. RSCU values (see above) are calculated for each codon within synonymous groups. Relative adaptiveness values w_i are then calculated for each codon by dividing by the maximum value in the synonymous set:

$$w_{ji} = \frac{\mathrm{RSCU}_{ji}}{\max_i[\mathrm{RSCU}_{ji}]}$$

The CAI of a gene is then calculated as the geometric mean of the w_{ji} values for each codon in the gene of interest,

$$\mathrm{CAI} = (\prod_{p=1}^{L} w_p)^{1/L}$$

where L is the length of the gene and w_p is the relative adaptiveness value for the codon at position p. The value given is effectively a measure of how well a gene matches the set of highly expressed genes in the cell. One major drawback is that it relies upon the availability of expression data for an organism, which precludes large-scale evolutionary analysis. Also, like the other measures discussed so far, it produces a single summary measure of codon bias across the gene and so does not account for local variations.

1.4.6 tRNA Adaptation Index

Inspired by the Codon Adaptation Index, the tRNA Adaptation Index (tAI) was described by dos Reis et al. (2004). Like CAI it uses relative adaptiveness values for each codon, but these are based not on codon usage but on a prediction of the tRNA pool. First, the absolute adaptiveness values for each codon (ignoring synonymous sets) is calculated as

$$W_i = \sum_{q=1}^{r_i} (1 - s_{iq}) t_{iq}$$

where $i \in [1, 61]$ represents the codon; r_i is the number of tRNA species that can decode codon i; t_{iq} is the gene copy number for the qth tRNA species that decodes codon i; and s_{iq} is a value representing any reduction in binding affinity between codon i and tRNA q resulting from a wobble pairing (see Section 1.2.4). These are normalised to give relative adaptiveness values for each codon,

$$w_i = \begin{cases} W_i/\max_i[W_i] & \text{if} \quad W_i \neq 0 \\ W_{mean} & \text{if} \quad W_i = 0 \end{cases}$$

where W_{mean} is the geometric mean of the non-zero W_i values. These values give an estimate of the adaptation of each codon to the cognate tRNA pool. The geometric mean of the individual codon values can be taken as a gene-wide summary statistic.

This is a sophisticated heuristic that attempts to measure the adaptiveness of a gene to the cellular environment, but there are problems with its formulation. Firstly, the tRNA binding affinities are based on a set of wobble rules that are not universal or necessarily complete (Yokoyama and Nishimura, 1995). The selection of tRNA isoacceptors from the cognate pool can also depend on local nucleotide context, not just the codon itself (Irwin et al., 1995). Second, the estimates for tRNA abundances rely on a presumed correlation with gene copy count, which may broadly hold but is crude at best, because of the highly discrete nature of copy count data and the fact that tRNA abundances are not fixed but are dynamically regulated in response to variations in conditions (Andersson and Kurland, 1990). Third, estimates of individual tRNA abundances do not account for tRNA competition or other selective forces influencing the adaptiveness of individual codons.

1.4.7 Normalised Translation Efficiency

Building on tAI, Pechmann and Frydman (2012) developed the normalised translation efficiency measure (nTE_i) that takes into account the relative abundance of transcripts in the cytosol as well as the abundances of the tRNAs themselves.

The usage U_i of codon i is calculated as the summed occurrences of codon i in the genome, weighted by transcript abundance a_j

$$U_i = \sum_{j=1}^{G} a_j c_{ij}$$

where G is the number of transcripts in the genome. This is normalised by the maximum value to give the relative codon codon usage

$$cu_i = \frac{U_i}{\max_k(U_k)}$$

which takes a value between 0 and 1. The normalised translation efficiency is then calculated as the ratio of the w_i values, defined in tAI, to these cu_i terms, again normalised by the maximum.

$$nTE'_i = \frac{w_i}{cu_i} nTE_i = \frac{nTE'_i}{\max_k(nTE'_k)}$$

This is intended to give a measure of how often each codon is translated in comparison to the supply of cognate tRNAs, approximating the load on the translating pool of tRNAs. This is a clever measure that considers another aspect of codon usage, but it suffers from the same drawbacks as tAI, compounded by the requirement for noisy measurements if transcript abundances, which are even more dynamic than tRNA species.

1.4.8 Rare Codon Rich Regions

Introduced by Widmann et al. (2008), this is one of a small number of methods that uses an approach based on alignments of homologous sequences to amplify the signal. Codon frequencies are counted as the number of occurrences per thousand codons in the whole genome of a species (these frequencies are obtained from an external database, presented in Nakamura et al., 2000). Starting with alignments based on homologous proteins, frequencies of codons in every column are multiplied together to calculate the column frequency

$$F_i = \prod_{j=1}^{n} f_j$$

This column frequency is compared to the products of all possible synonymous combinations of codons for the column and given a score according to the proportion of synonymous combinations with lower products. Columns are then grouped according to the score in brackets with a width of 0.2 (e.g. columns in group 1 have scores <0.2). Alignments are assessed with windows of nine columns according to a sum of the number of columns in the window with scores in group 1 or 2, weighted so that group 2 columns are worth 0.6 times as much as group 1 columns. Windows with a score above 1.8 (i.e. with three or more group 2 columns, two or more group 1 columns, or some equivalent combination) are classified as rare codon rich regions

(RCRRs).

Widmann et al. (2008) applied their method to alignments built for 16 homologous protein families containing 7-10 sequences with low overall sequence identities. The very high computational cost of their implementation, requiring explicit computation of all possible codon frequency combinations, places constraints on the alignment size and the number of families to which it can be applied, and the authors state an expected false positive rate of 4 %.

1.4.9 Sherlocc

Sherlocc (Chartier et al., 2012) is another in the class of methods that uses protein sequence alignments. The authors looked at pre-built alignments of protein families in the Pfam database (Punta et al., 2012), and obtained codon usage frequencies for the corresponding species from the Kazusa database (Nakamura et al., 2000). They calculated average codon usage frequencies over windows of seven positions in the alignment. At this point the implementation of their method becomes unclear; the text suggests that a *p*-value is obtained by fitting the average window scores to an extreme value distribution, but it appears to be more akin to a ranking statistic rather than a *p*-value.

1.5 Scope of the Thesis

Codon usage is an important aspect of recombinant gene design and modulation of endogenous gene expression, and an “optimal” codon usage pattern does not simply consist of the most common codons (Gustafsson et al., 2004; Plotkin and Kudla, 2011). There are cases where rare codons appear to aid gene expression and enhance the functional efficacy of the protein product, sometimes at a cost to production levels per mRNA. Identifying important rare codons would allow more effective optimisation of genes for heterologous expression, could shed light on disease-associated synonymous mutations, and would be of use in the design of fusion proteins and *de novo* protein design.

The body of experimental evidence supporting the existence of beneficial rare codons is limited and disparate, which poses a challenge to the development of accurate predictive methods. An expansive library of synonymous genes with well-quantified fitness profiles would greatly aid computational studies, but such data would be arduous and expensive to produce and is unlikely to become available in the near future. Despite the wealth of metrics for codon usage that have been proposed and implemented, there is no single preferred method. This is partly because different methods suit different applications, depending on how much data is available for the organism of interest and whether a local or gene-wide measure is appropriate. Another major reason is that no existing method has been demonstrated to work predictively. Making an *ab initio* assessment of codon usage and following it up with supporting experimental work that demonstrates that the regions identified are functionally relevant should be the gold standard for which codon usage metrics aim.

Through the many direct and indirect effects previously described (see Section 1.2), codon usage has an influence on the rate of translation. Though it is undoubtedly a simplification, the broad trend appears to be that rarely used codons are translated slower than common ones. Rare codons may instigate a translational pause, which could aid protein folding or some other aspect of expression. The noisy nature of evolution results in rare codons, and possibly translational pauses, that do not aid – and possibly even hamper – expression. A more specific sample of rare codons that offer a selective advantage could be obtained through the use of evolutionary conservation to filter out examples arising from noisy, non-selective mutations. What role, if any, these such rare codons might play in the expression of functional proteins would then need to be tested with a combination of experimental and computational work.

The goal of this thesis is to develop and verify a method for detecting selectively advantageous rare codons. The first target was to develop a sound statistical method capable of yielding a location-specific measure of codon usage in bacterial genes. This was subsequently verified with a combination of statistical analysis of other biologically relevant signals, and experimental work based on varying the identified locations. The investigation is described in the coming chapters. Chapter 2 describes the algorithm developed and the reasoning behind it. Chapter 3 details the analysis of the data generated. Chapter 4 describes the experimental work that was conducted in an attempt to verify the algorithmic findings. Finally, Chapter 5 presents conclusions and looks to future opportunities in the field.

2 Algorithm Development

There are numerous published methods designed to assess codon usage. These vary widely in their precise intentions, and no single one is universally accepted either for statistical merits or biological relevance. The record of publications developing new methods (see Section 1.4), and the lack of experimental evidence supporting any method, demonstrates this lack of consensus. The intention of this investigation is to measure the statistical signal of codon usage, and subsequently to measure the importance of the effect with experimental work conducted in *Escherichia coli* (see Chapter 4). The focus is on prokaryotic organisms, for a number of reasons. Translation in prokaryotes is faster overall, so there is more potential for beneficial translational pausing and more scope for elongation rate modulation (Angov et al., 2008; Siller et al., 2010); it involves fewer cofactors, such as chaperones, that may modulate folding in alternative ways; and there are fewer RNA regulatory elements, such as miRNAs binding and splice sites, that place unrelated constraints on gene sequence (Plotkin and Kudla, 2011).

This chapter presents the reasoning involved in, and the results of, the process of designing a codon usage assessment algorithm, and discusses the application of the algorithm.

2.1 Requirements

Four fundamental requirements for an algorithm were laid out:

1. **Use homology to enhance the evolutionary signal**

   The aim was to develop an algorithm to identify regions of genes that appear to have rare codons by virtue of some selective pressure, rather than due to the random fluctuations of codon usage effected by mutational drift. The most effective way to do this is to use homology to highlight regions of evolutionarily-related proteins that are consistently encoded with rare codons. Hence the method should be based on alignments of homologous proteins, rather than single sequences.

2. **Assess the codon usage in a local region**

   Rare codons have been shown to occur preferentially in clusters (Clarke and Clark, 2008; Parmley and Huynen, 2009; Zhang et al., 2009). Because of the stochastic nature of translation, a functionally relevant translational pause is likely to consist of multiple consecutive rare codons in order to reliably take effect. The algorithm should avoid overfitting to the noisy underlying signal, and be capable of detecting significant non-consecutive rare codons, by smoothing codon usage frequencies over a local region. This suggests a sliding window approach.

3. **Correct for amino acid composition**

   Amino acids appear in proteomes with greatly varying frequencies. Amino acid composition is a dominant constraint on protein evolution; it has a direct effect on the structure and function, and is correlated with the abundance and solubility of the protein product (Price et al., 2011). The metabolic load of synthesising a protein is influenced by the complexity of its constituent amino acids, placing an additional non-linear evolutionary pressure on the precise residue content (Akashi and Gojobori, 2002). The intention of the measure being developed here is to isolate codon usage from this complex set of pressures, which necessitates eliminating the influence of the underlying frequencies of the amino acids. Further, amino acids have structural preferences which may cause misleading correlations in later analysis if the effect of amino acid selection is not properly removed. Therefore some degree of normalisation based on the amino acid sequence is required to measure the frequencies of the codons independently from the frequencies of the amino acids.

4. **Yield a measure of statistical significance**

   To facilitate informed study of the results, the algorithm should yield a valid *p*-value, based on the comparison of observations with expectations under some null or background distribution.

These four requirements determine a basic approach: for a protein of interest, identify a set of homologous proteins. Build an alignment of the homologous set, and use the alignment as the framework for analysis of codon usage frequencies in genes. A measure of codon usage is applied to the alignment using a sliding window heuristic. The window smoothes the codon usage signal, the need for which is explained in requirement 2 above. The degree of smoothing can be varied by varying the size of the window. The window approach also allows the algorithm to be robust to gaps, as explained in Section 2.1.1.

It should be noted that alignments of homologous proteins have been used as the basis for codon usage analysis previously in methods named RCRR (Widmann et al., 2008) and Sherlocc (Chartier et al., 2012). The methods employed in these papers have their own shortcomings, which are discussed in Section 2.5.5.

The measure developed here is designed to represent the relative codon usage frequencies within synonymous groups, removing as far as possible the bias arising from the amino acid sequence. The central principle is a comparison of the codon usage frequencies observed in a window against an expected distribution, derived from the usage frequencies of codons in the synonymous set. This approach addresses requirements 3 and 4: correcting for amino acid composition, and yielding a measure of statistical significance.

2.1.1 Handling Gaps

To analyse codon usage across multiple sequence alignments, it is necessary to establish a heuristic for dealing with gaps in the alignments. Under a naive approach, regions of low alignment quality are more susceptible to noise and so more likely to be erroneously identified as containing a significant proportion of rare codons. To correct for this effect, the sliding window approach was modified to include a fixed number of contiguous residues from each sequence (see Figure 2.1). Under this heuristic the number of residues on which a column depends is precisely determined by the number of sequences aligned at that position.

It could be said that this heuristic results in comparisons between unaligned positions. This is acceptable, because the aim is to compare the process of translation between versions of the protein. A contiguous set of residues and the codons that encode them represents series of translational events. We are aiming to compare these translational events, rather than conserved regions. The alignment provides an anchor point based on the broad correspondence of structural and functional features between proteins, around which we can base our translational comparison. This anchoring is desirable for future analyses.

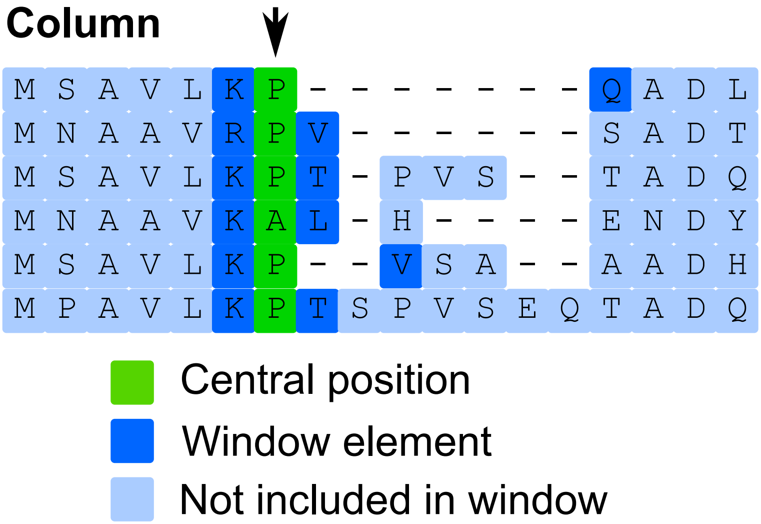

Column
M S A V L K P - - - - - - - Q A D L
M N A A V R P V - - - - - - S A D T
M S A V L K P T - P V S - - T A D Q
M N A A V K A L - H - - - - E N D Y
M S A V L K P - - V S A - - A A D H
M P A V L K P T S P V S E Q T A D Q
Central position
Window element
Not included in window

Figure 2.1: Windows are composed of contiguous sets of residues from each sequence, but not necessarily from the same set of alignment positions. The properties of the whole window are assigned to the central position. Contiguous regions of a protein can be thought of as a series of translational events – the addition of an amino acid onto the nascent chain, and everything that entails. Each window represents a set of series of translational events, anchored around the biochemical framework of the alignment.

2.2 Assumptions and the Null Hypothesis

The process of hypothesis testing can be described as estimating the probability that the observed strength of a phenomenon has arisen by chance from a system whose behaviour is in fact neutral on average, with random fluctuations over some distribution. To estimate the probability, it is necessary to construct a model for the distribution of the assumed random behaviour.

The phenomenon that we observe here is the codon usage frequency. Specifically, we chose to measure against a background of the usage frequencies across all coding sequences in a genome. In contrast, some previous measures have attempted to define optimal codons in a more prescriptive way. One approach that is widely used in the literature is the selection a set of highly-expressed set of genes – often ribosomal constituents – as being exemplary of optimal codon usage, and measure codon usage in other genes against this reference (Sharp and Li, 1987; Sharp et al., 2005; Hildebrand et al., 2010; Wang et al., 2011). Other methods have used predictions about tRNA levels and their cognate codons (dos Reis et al., 2004; Fluitt et al., 2007).

These methods are problematic for several reasons. In the case of expression-level based methods, the need for detailed data drastically reduces the availability of organisms, which limits the use of homology to add statistical power. Even when the data are available, defining high-expression genes is subjective, since specific mRNA levels can vary by at least two orders of magnitude depending on environmental conditions (Ishii et al., 2007). Some genes that are undetectable in ordinary log-phase growth are present in thousands of copies per cell under stress conditions. In the NCBI Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geoprofiles/`), about 98% of *E. coli* genes are in the top percentile of expression levels in at least one experiment. Further, the rationale is that a highly-expressed gene is under stronger selection and is therefore more likely to be effectively optimised. Pathogen-response and antibiotic resistance genes may be expressed or at low levels but when called upon are critical to survival. Intracellular growth inhibitors and other toxicity factors can offer a strong selective advantage in certain conditions. The tRNA-based methods make questionable assumptions about the stability of tRNA levels (Gingold et al., 2012) and their correspondence with translation rate (Bonekamp et al. 1989; Stadler and Fire 2011; see Section 1.2.4). They also make simplifications with regards to complementary codon-anticodon behaviour (Yokoyama and Nishimura, 1995; Ran and Higgs, 2010). Finally, both sets of methods assume that the differences in codon usage arise only from differences in the magnitude of the selective pressures. There is no accounting for qualitative differences in the nature of the pressures on the selected set versus the rest of the genome. Differences in the nucleotide composition of genes relative to intergenic regions provides evidence of selection on codon usage even in genes that show little apparent codon bias (Hershberg and Petrov, 2009, 2012). The goal of this work is to identify areas in genes that consistently defy genome-wide patterns, suggesting that the universal pressures that dictate these overall patterns are overridden by a more localised factor. The appropriate background is, therefore, species-specific, genome-wide codon usage frequencies.

As explained above, we are seeking an aggregate measure of codon usage over a window of an alignment. The null hypothesis is that codon usage frequency is independent of any property of the gene itself, or the protein it encodes. In testing this hypothesis, we make three main simplifying assumptions about how codons are selected. The first assumption is explained in 2.1, requirement 2: that amino acid selection is a dominant constraint, such that codon selection is dependent on the occurrence of a particular amino acid. The second assumption is that codon selections within a gene are independent. The selection of a codon is assumed to be independent of those selected in neighbouring positions. Although there is some evidence that sequential codons do influence one another (Skewes and Welch, 2013; Cannarrozzi et al., 2010), the strength of the effect on different species' genomes, and whether or not it plays a role in translation (Yakovchuk et al., 2006), are unknown.

The third assumption is that codon selections between organisms are independent. Like all assumptions, this is invalid in the strictest sense. Building alignments requires the identification of homologous proteins that are by definition not fully independent in their amino acid composition. In a sense, what we seek is biological independence between genes – that is, within the confines of the amino acid sequence of the proteins they encode, that genes are fully adapted to their cytosolic environments. It has been shown that horizontally-transferred genes adapt to conform to the codon usage pattern of the host species relatively quickly (Lawrence, 1999; Skewes and Welch, 2013). This implies that similarities in codon usage patterns between species can be attributed to similarities in the cellular environment rather than a lack of divergence from a mutually inherited gene sequence. If a single gene is transferred to two closely-related species and fully adapts to each environment, the two copies may be similar in nucleotide composition and codon usage. In this case the assumption of independence of codon selection is not violated. The question then becomes one of evolutionary independence between organisms; specifically, at what phylogenetic distance are two species sufficiently diverged to be considered independent? This is difficult to quantify, since measures of phylogenetic distance are generally based on the nucleotide-level similarity between sets of genes, so clearly codon usage patterns will be similar in species that are determined to be evolutionarily close. Again, strictly speaking, no two organisms are truly independent, because all species are likely derived from a common ancestor (Theobald, 2010) and so have traversed some portion of a shared evolutionary path. The cytosolic environments of all species, and thus the evolutionary forces imposed on genes, are on some level co-dependent.

To resolve this impasse, we construct a database that is non-redundant at the species level (see Section 2.5.1 for details of the construction of the database used). Taking all the arguments above, we consider that the assumption of inter-species independence is largely reasonable and hugely simplifying. Although we concede that it may undermine the statistical precision of the calculated p-values, constituting a relaxation of requirement 4 above, it should not affect the indicative power of the results. We are conservative in selecting thresholds and limits in the application of the method (see Section 2.5). Further, if homology at the gene level was the cause of a considerable portion of the detectable signal, this should be readily distinguishable through examination of the degree of codon conservation across the alignments at positions classified as rare and non-rare. This analysis is presented in Section 2.5.4.

2.3 Codon Frequencies as Random Variables

Having established the null hypothesis, we can begin to interrogate the model and estimate the probability that the genes we observe in nature were generated by a model that behaves according to the null hypothesis. The notation introduced in the following text is tabulated for reference in Table 2.1.

	Single residue	Sequence	Alignment window
Amino acids	a_i	$A_j = \{a_i\}_{i \in 1:z}$	$\mathbb{A} = \{A_j\}$
Coding frequency instances	f_i	$F_j = \{f_i\}_{i \in 1:z}$	$\mathbb{F} = \{F_j\}$
Coding frequency functions	$\mathrm{Fr}_j(a_i) = f_i$	$\mathrm{Fr}_j(A_j) = \sum f_i$	$\mathrm{Fr}(\mathbb{A}) = \sum\limits_{A_j \in \mathbb{A}} \sum\limits_{a_i \in A_j} f_i$
Satisfactory sets	$\mathrm{r}_c = \{f \mid f \leq c\}$	$\mathrm{R}_c = \left\{F_j \mid \sum\limits_{f_i \in F_j} f_i \leq c\right\}$	$\mathbb{R}_z = \left\{\mathbb{F} \mid \sum\limits_{F_j \in \mathbb{F}} \sum\limits_{f_i \in F_j} f_i \leq z\right\}$
Subscripts	Codon	$k \in 1 : n(a_i)$	
	Residue	$i \in 1 : z$	
	Species/sequence	j	
Constants	Synonymous set size	$n(a_i)$	
	Window size	z	
	Number of sequences	N	

Table 2.1: Notation reference

To attach a metric to codon selection we use the codon usage frequency f, which is simply the rate of occurrence of a given codon per thousand codons in the whole genome. For an amino acid a_i, there is a corresponding set of synonymous codons. Codons appear in different genomes with different frequencies. For the purposes of the metric, we can think of codons purely in terms of their frequencies. We can then define species-specific functions $\text{Fr}_j(a_i) = f$ (where the subscript j indicates the species) that map amino acids onto random variables whose values represent codon frequencies. Under the null model, the mapping is random, with a probability mass function that can be derived from the underlying frequencies by normalising by the sum of the frequencies in the synonymous set.

$$\Pr(\text{Fr}_j(a_i) = f) = \frac{f}{\sum\limits_{k=1:n(a_i)} f_k}$$

where $n(a_i)$ is the size of the synonymous set for a_i, so that k iterates over the possible frequency values.

For example, the amino acid alanine (Ala) can be encoded by codons GCA, GCC, GCG and GCU. Let us denote *E. coli* K-12 MG1655 as species $S1$. In the genome of $S1$, the codons for Ala occur with frequencies 20.22, 25.83, 34.21 and 15.25 per thousand. These frequencies are the support values of the function $\text{Fr}_{S1}(\text{Ala})$, giving us the probability mass function (PMF)

$$\Pr(\mathrm{Fr}_{S1}(\mathrm{Ala}) = f) = \begin{cases} 0.21, & f = 20.22 \quad (\mathrm{GCA}) \\ 0.27, & f = 25.83 \quad (\mathrm{GCC}) \\ 0.36, & f = 34.21 \quad (\mathrm{GCG}) \\ 0.16, & f = 15.25 \quad (\mathrm{GCU}) \end{cases}$$

In *Bacillus subtilis*, which we will denote species $S2$, the same codons occur with different frequencies, and we can evaluate the PMF as

$$\Pr(\mathrm{Fr}_{S2}(\mathrm{Ala}) = f) = \begin{cases} 0.28, & f = 21.1 \quad (\mathrm{GCA}) \\ 0.22, & f = 16.5 \quad (\mathrm{GCC}) \\ 0.26, & f = 19.8 \quad (\mathrm{GCG}) \\ 0.24, & f = 18.6 \quad (\mathrm{GCU}) \end{cases}$$

The probability of obtaining a frequency that is less than or equal to some defined limit c can be found by summing the probabilities of the events that meet that condition. We can express this as

$$\Pr\left(\mathrm{Fr}_j(a_i) \leq c\right) = \sum_{f \in \mathrm{r}_c} \Pr\left(\mathrm{Fr}_j(a_i) = f\right)$$

where $\mathrm{r}_c = \{f \in \mathrm{supp}(\mathrm{Fr}_j(a_i)) \mid f \leq c\}$ is the satisfactory set.

2.3.1 Convolution

The first part of this section introduces the principle of convolution in a general form using generic notation, before moving back to the specific problem of codon usage.

The sum of two or more random variables is known as their convolution, and its probability mass function (PMF) can be defined from the mass functions of the individual variables. For two independent discrete random variables X_1 and X_2 with distinct probability distributions $\Pr(X_1)$ and $\Pr(X_2)$, define $\mathrm{m}(X_i)$ as the smallest support value of X_i. The convolution $Z = X_1 + X_2$ is also a discrete random variable whose probability mass function can be calculated from

$$\begin{aligned}
\Pr(Z = z) &= \Pr(X_1 + X_2 = z) \\
&= \sum_{s=k}^{z-k} \Pr(X_1 = s, X_2 = z - s) \\
&= \sum_{s=k}^{z-k} \Pr(X_1 = s) \Pr(X_2 = z - s)
\end{aligned}$$

where k is the smallest possible increment to Z, i.e. $\min\Big(\mathrm{m}(X_1), \mathrm{m}(X_2)\Big)$ (Evans and Leemis, 2004). The cumulative distribution function (CDF) is given by

$$\begin{aligned}
\Pr(Z \leq z) &= \Pr(X_1 + X_2 \leq z) \\
&= \sum_{t=0}^{z} \Pr(X_1 + X_2 = t) \\
&= \sum_{t=0}^{z} \left(\sum_{s=k}^{t-k} \Pr(X_1 = s) \Pr(X_2 = t - s) \right)
\end{aligned}$$

This principle can be further extended to multiple independent variables, and although the notation can get a little unwieldy the principle is readily intuited from the fundamentals of probability theory governing the behaviour of independent events.

Moving away from the general form and back to our problem domain, we can define an aggregate measure of the codon usage frequency over a region of a single gene as the sum of the frequencies of the individual codons. If $A_j = \{a_i\}$ represents a sequence of amino acids in species j, we can define $F_j = \{f_i\}$ as a sequence of frequencies corresponding to a possible coding sequences for A_j (see Table 2.1 for a summary of the notation). We can extend the definition of our species-specific frequency function so that $\mathrm{Fr}_j(A_j) = \sum_{a_i \in A_j} \mathrm{Fr}_j(a_i)$. Rarer coding of A_j indicates a lower $\mathrm{Fr}_j(A_j)$. If we also set out an extended definition of a satisfactory set as the set of satisfactory coding sequences, $\mathrm{R}_c = \{F_j | \sum_{f_i \in F_j} f_i \leq c\}$, we can write the cumulative probability function succinctly as

$$\Pr\left(\mathrm{Fr}_j(A_j) \leq c\right) = \sum_{F_j \in \mathrm{R}_c} \prod_{f_i \in F} \Pr\left(\mathrm{Fr}_j(a_i) = f_i\right)$$

To examine an alignment requires further extensions to the notation. For a region of an alignment consisting of a set of sequences from different species $\mathbb{A} = \{A_j\}$, we can construct a set of sequences of coding frequencies $\mathbb{F} = \{F_j\}$. There are many possible different $\mathbb{F}$ sets, each corresponding to a different combination of codons across the whole region of the alignment. By extending the frequency function Fr so that $\mathrm{Fr}(\mathbb{A}) = \sum_{A_j \in \mathbb{A}} \sum_{a_i \in A_j} \mathrm{Fr}_j(a_i)$ (i.e. the sum of all species-specific frequencies across the alignment region); and the satisfactory set concept, so that

$\mathbb{R}_z = \{\mathbb{F} | \sum_{F_j \in \mathbb{F}} \sum_{f_i \in F_j} f_i \leq z\}$, we can finally write the cumulative probability function for the coding frequency of an alignment region as

$$\Pr\left(\mathrm{Fr}(\mathbb{A}) \leq z\right) = \sum_{\mathbb{F} \in \mathbb{R}_z} \prod_{F_j \in \mathbb{F}} \prod_{f_i \in F_j} \Pr\left(\mathrm{Fr}_j(a_i) = f_i\right)$$

We can use this result to achieve our stated aim: calculate the probability that the rareness of the coding observed for a region of an alignment was generated by a system behaving according to the null model. Positions of an alignment that display rarer than expected coding (this is discussed further in Section 2.5) will be referred to henceforth as CRPs – conserved rare positions.

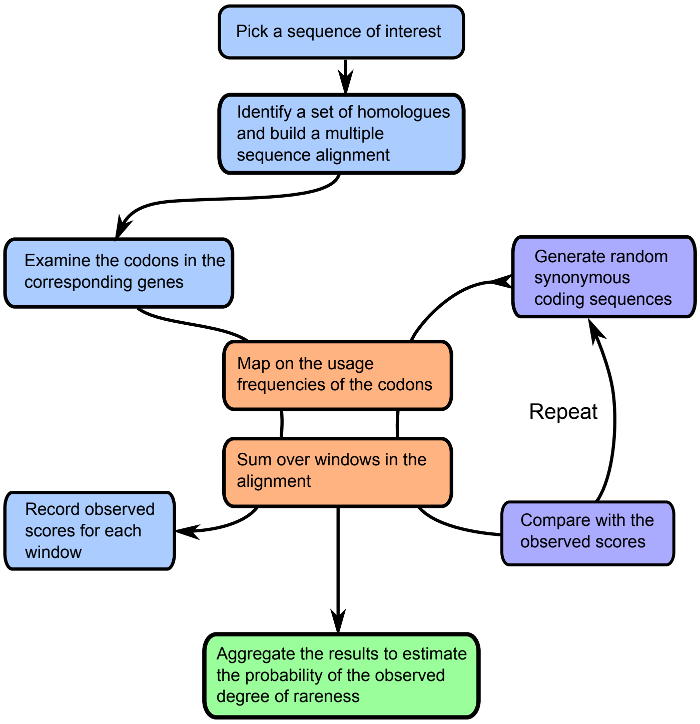

Pick a sequence of interest
Identify a set of homologues and build a multiple sequence alignment
Examine the codons in the corresponding genes
Generate random synonymous coding sequences
Map on the usage frequencies of the codons
Repeat
Sum over windows in the alignment
Record observed scores for each window
Compare with the observed scores
Aggregate the results to estimate the probability of the observed degree of rareness

Figure 2.2: An outline of the algorithm in flowchart form.

Figure 2.2 outlines the algorithm. The algorithm has several desirable properties.

Raw frequencies are taken into account, so codons that are very common or very rare in the context of the genome have more influence than average values, but the probabilities of occurrence are adjusted to correct for the amino acid composition. For example, the two codons encoding cysteine, the rarest multiply-encoded amino acid in the *E. coli* genome, occur with frequencies 5.11 and 6.46 per thousand. If a cysteine is present in a protein one of these two codons must be used to encode it, with probabilities of approximately 0.44 and 0.56 respectively, so the background reflects the certainty of obtaining a low score from that residue. Just 3.6% of leucine residues are encoded by the rarest cognate codon, CUA, which is nearly fourteen times rarer than the most common leucine codon, CUG. Organisms with low levels of bias contribute to the statistic by reducing the significance of a pattern in accordance with their own biases. This is desirable, because the presence of that gene is evolutionary evidence against a universal requirement for rare codons. Finally, we make very few assumptions about the mechanism of codon usage selection and how it might affect translation. We are looking only for a conserved pattern that defies the prevailing genomic preferences. We wish to remain agnostic to the nature of any selective force we might detect, because that allows reasonable testing of numerous hypotheses further down the line.

2.4 Implementation

This section discusses some technical aspects of implementing the algorithm outlined in the previous section.

2.4.1 Window Size Selection

Selecting the appropriate window size poses a challenge, as we seek to assume as little as possible about the mechanisms at work. Although some previous methods have used windows of 7 or more codons (Power et al., 2004; Widmann et al., 2008; Parmley and Huynen, 2009; Chartier et al., 2012), kinetic studies of folding and translation rates suggest that even single codon changes could influence folding pathways (O'Brien et al., 2012) and there is some experimental evidence to support this claim (Kimchi-Sarfaty et al., 2007; Tsai et al., 2008). Other studies have used a wide range of window sizes to try to gain insight into the mechanisms (Clarke and Clark, 2008; Saunders and Deane, 2010), but the computational expense of the method designed here prohibits this approach on large samples. Selecting too large a window could miss small, precise events as well as cases where rareness is not localised but is accumulated across the length of a gene. The use of alignments should reduce noise, making smaller window sizes more workable, and a smaller window makes the measure more sensitive. A smaller window also gives greater specificity to the identified regions, which would be an advantage when designing variants for experimental investigation. As a compromise between smoothing, sensitivity and computational expense, a window size of 3 was selected for this study. Section 2.5.3 discusses the effect of varying this parameter on a sample of alignments.

2.4.2 Explicit Computation and Expense

Explicitly computing the convolution using a brute-force approach requires evaluation of the probabilities and sums of all possible coding sequences. This depends exponentially on the number of amino acids in the window; formally, it has order $O(\bar{n}_j^{zN})$, where z is the window size as above, N is the number of sequences, and $\bar{n}_j$ is the expected number of synonymous codons per amino acid, which can be calculated for each organism as

$$\frac{\sum_{a\in A}(n(a)\times Gen_j(a))}{G_j}$$

where A is the set of all twenty amino acids, and $Gen_j(a)$ and G_j are respectively the number of occurrences of amino acid a and the total number of residues in the proteome of organism j. This figure is 3.8 in *E. coli*. Consider a perfect (gap-free) alignment of ten sequences, each of length $l = 100$ residues, from species with the same levels of codon usage bias; using a window size of $z = 3$, there are $l-(z+1)/2 = 98$ windows. Computing the probability convolution for every possible window composition would require the calculation of $98 \times 3.8^{(3\times 10)} \approx 2.4 \times 10^{19}$ distinct sums and probabilities, a prohibitively large number.

Evans and Leemis (2004) describe an algorithm for explicitly computing the PMFs and CDFs of the convolution of two discrete random variables with arbitrary supports, which exploits the order in the support sets and uses a technique similar to dynamic programming to traverse combinations in increasing order. It is conceptually possible to scale this to multiple variables, effectively adding extra dimensions to the dynamic programming-style array to be traversed. Additional computational savings could be made by updating the computed combinations with each iteration of the window position. However, this algorithm would be very memory intensive and its implementation complex, and for large numbers of variables with small support sets the savings would probably not be sufficient to render it viable. We therefore seek an approximation to the distribution rather than an explicit evaluation.

2.4.3 Monte-Carlo Simulation

The overall distribution of the coding frequency convolution for each window is a complex combination of simple components. This makes it well suited to Monte-Carlo analysis, where the probability mass function is repeatedly simulated instead of explicitly computed.

This was implemented as follows: for every protein sequence in an alignment, random synonymous coding sequences were generated according to the standard genetic code, with codons picked according to their probabilities in the relevant species under the null model (see Section 2.2). The random coding sequences were mapped back onto the alignment, and the summed codon frequencies $\mathrm{Fr}(\mathbb{A})$ (see Table 2.1) were computed for each window. Comparing these random scores to the observed frequencies $\mathbb{F}_o$ gives an estimate of the cumulative probability $\mathrm{Fr}(\mathbb{A}) \leq \mathbb{F}_o$.

The random sampling was done 10,000,000 times. To avoid wasting computation on windows of no interest, two intermediate steps were added – at 1000 and 100,000 samples – after which any window with one or more random scores less than or equal to the native observed score was not sampled any further. This allows estimation of the p-values of the rarest windows to a resolution of 1×10^{-7}.

2.4.4 Algorithm Runtime

For a given window size, the runtime of the algorithm is initially dependent on the product of the number of sequences in the alignment with the alignment length. However, because of the filtering strategy that ignores positions with poor alignment quality or common overall codon usage, the dependency on alignment length is lost after a few iterations. Runtime is then a function of the number of sequences in the alignment and the number of CRPs. Figure 2.3 shows that runtime is well explained by the product of these two parameters.

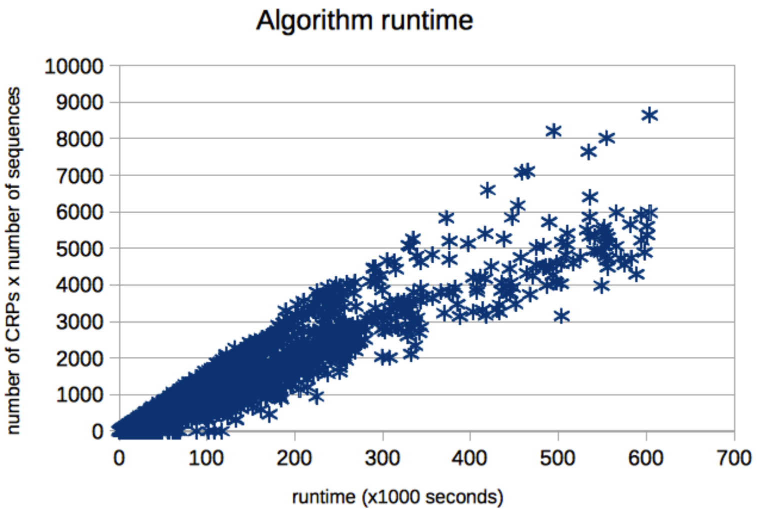


Figure 2.3: Scatter plot of the runtime of the algorithm for alignments in the dataset with a window size of 3. The runtime for a given alignment correlates closely with the product of the number of CRPS and the number of sequences in the alignment. Runtimes ranged from 2 seconds, for small alignments with no CRPs, to almost a week for the largest, most CRP-rich alignments.

2.5 Application

As mentioned previously, the intention is to complement the statistical data with experimental evidence. The molecular mechanisms of translation are extremely complex and although many particulars are universally conserved, there is significant variation between the phylogenetic kingdoms (Voigts-Hoffmann et al., 2012; Novoa et al., 2012) and it is simplest and probably best-understood in prokaryotes. *Escherichia coli* is by far the most studied prokaryotic model organism. Practical reasons, such as ease of growth and transformation, added to the availability of structural and metabolic data, make it a logical choice for investigation. This dictates that the algorithm be applied to a dataset centred around *E. coli*. For the purposes of investigating factors related to the mechanism of translation, it also makes sense to limit the search for homologous proteins to the prokaryotic realm.

This section discusses the application of the method, including the development of a database, the assignment of significance thresholds, and some examination of conservation as a potential null explanation of the calculated results.

2.5.1 Database Construction

The algorithm described in the previous section was designed to identify evolutionarily conserved patterns of rare codon usage using alignments of homologous protein sequences. The advantage of an alignment-based method is that the additional sequence information can be leveraged to reduce noise and enhance signal, but this comes at a price: such methods are inherently susceptible to latent biases in the database. Thus careful database selection is a key element of the process.

The database used was derived from a set of 1005 full prokaryotic genomes derived from the NCBI GenBank (Benson et al., 2008). Species containing genes not translated by the standard bacterial coding table (NCBI Table 11; `http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi`), and those with fewer than 500 genes across all chromosomes and plasmids, were discarded.

The set of species was then pruned to be completely non-redundant at the species level. Codon frequency tables were built for every organism in the database. Where multiple strains from a single species were present, the average codon usage frequency table was calculated, and the strain with the minimum squared Euclidean distance from the average was selected. All other strains from that species were removed from the database. A single exception was made for *Escherichia coli*, where the strain preserved was K-12 MG1655, because it was considered of the most biological interest and it, or a close derivative of it, would likely be used for subsequent laboratory work. Plasmids marked with taxon IDs other than those of the species they occur in were removed. This left a final database of 678 species, with about 2.3 million genes between them. Figure 2.4 illustrates the effect of removing redundancy at the species level. The two heatmaps represent the 64-dimensional codon usage space; an organism can be represented as a point in this space according to its genomic codon usage. The space can be projected onto two dimensions for visualisation using multidimensional scaling. The axes in the heatmaps are orthogonal linear combinations of the usage frequencies of each codon. The colour indicates the density of organisms in a particular region of the codon usage space. Removing redundant strains eliminates the dark red areas of high density and results in a much more uniform coverage of the space.

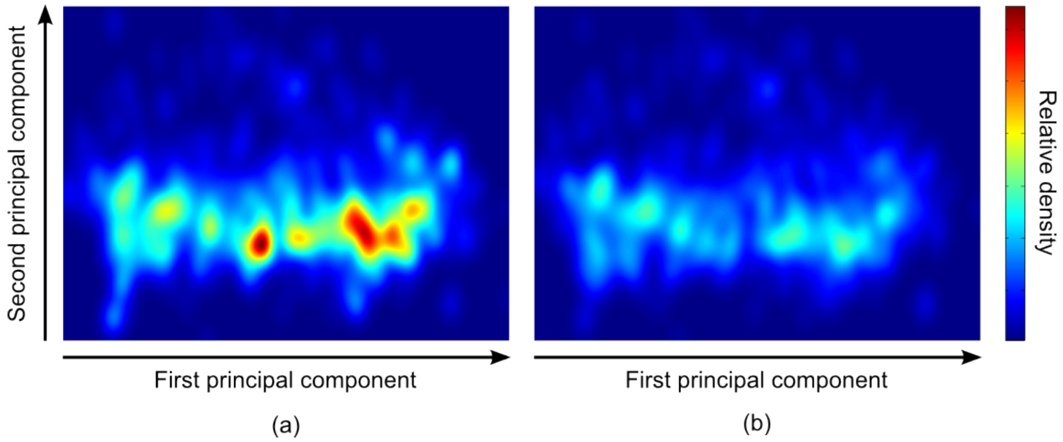


Figure 2.4: Two-dimensional representations of codon usage space. The figures show heat maps indicating the distributions of organisms in codon usage space (a) with and (b) without redundancy at the species level. Nonmetric multidimensional scaling was performed on matrices representing prokaryotic genomes in the 64-dimensional codon usage space; one matrix contained about 1000 strains including strains from the same species, the other contained only the strains selected for the species-level nonredundant database (Kruskal's normalised stress: (a) 0.0611, (b) 0.0628). The plots show the result of density estimation on the resulting two-dimensional projections. The axes are orthogonal linear combinations of of the usage frequencies of the 64 codons. The colour indicates the density of strains on a linear scale. It is clear that a good deal of redundancy is removed and the overall codon usage space is populated almost uniformly after removing redundant strains.

This allowed for the construction of homologous sets of proteins for sequences of interest. A set of homologues for every gene in the *E. coli* K-12 MG1655 genome was drawn from the database. Homologue identification was done with the `blastp` utility from the BLAST+ software package (Camacho et al., 2009), using an E-value threshold of $< 10^{-5}$. To preserve alignment quality, homologues whose lengths differed from the seed by more than ±10% were discarded from the search results. To prevent the inclusion of paralogues (multiple genes from the same species), the first hit from each unique species was included and any subsequent hits were discarded.

Families of less than five homologues were excluded altogether from subsequent analysis. Protein sequence alignments were then constructed using MUSCLE (Edgar, 2004). This yielded a set of 3890 alignments. The process was repeated using *E. coli* chloramphenicol acetyltransferase 3 (CAT-3) as the seed gene as it is of interest in the literature (Komar et al., 1999), adding a further alignment (see Section 2.6.2). Five of these alignments were subsequently removed upon mapping to Uniprot identifiers as they had become obsolete, leaving a final total of 3886 alignments.

Codon frequencies for each species were derived manually from the database of sequences, rather than taken from another codon frequency database (e.g. Kazusa – Nakamura et al. (2000)), by counting the occurrences of all codons in the set of coding sequences for that species and dividing by the total number of codons in the set.

2.5.2 Defining Rareness

The method described above generated p-values for the rareness of the codon usage observed in three-codon windows across 3886 alignments of homologous proteins.

Alignment positions with more than 20% gaps, and positions where the *E. coli* seed sequence was not aligned, were ignored. Positions containing the initiator methionine were also ignored because of the disproportionate reduction in coding options. This left 1,162,109 positions. Because the method used a Monte-Carlo sampling technique with ten million random samples, p-values were resolved to 1×10^7 (see Section 2.4.3). Each p-value is generated from a distinct cumulative distribution, which depends on the precise amino acid composition of the alignment window.

Interestingly, the overall distribution of p-values was heavily skewed to the extremes of 1 and 0 (see Figure 2.5). This suggests that at the majority of positions there is an evolutionary pressure in one of two directions – either towards “optimal”, high-frequency codons, or in the opposite direction, towards rare codons. It should be noted that we cannot be as confident of specific p-values at the high-end, because of the way the test was implemented to save computation, but the trend is clear.

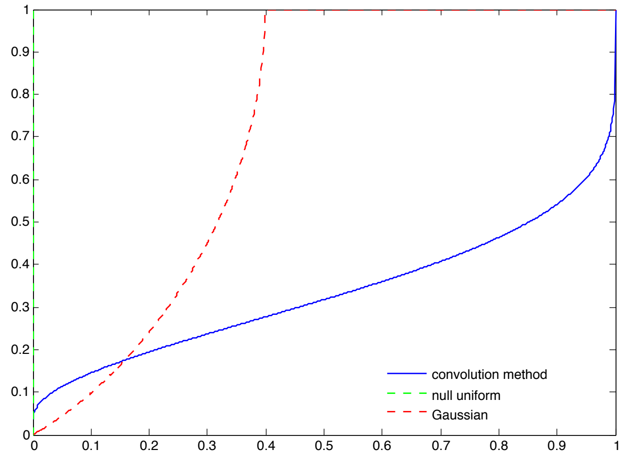


Figure 2.5: The distribution of p-values is heavily skewed towards the extremes, suggesting that at all positions codons are “optimised" in one of two directions. The blue line represents the proportion of p-values that are less than or equal to the corresponding x value. The dotted green vertical line, overlapping the y axis in this visualisation, indicates the null uniform distribution that was used to assign a significance threshold – only p-values to the left of this line were considered significant. This is plotted to give an indication of the stringency of the selected threshold. The dotted red line shows the behaviour of p-values in a Gaussian distribution, which are often produced by random processes.

To assign a significance threshold for rareness we adopted a false discovery rate approach, conservatively setting the number of expected false discoveries to 1 against a uniform null distribution. The p-value threshold for significance is therefore $1/n$, where n is the number of positions for which p-values were calculated, giving a significance threshold of approximately 8.6×10^{-7}. Under this scheme 13,615 positions (1.17%) across 2336 alignments (60.1%) were classified as conserved rare positions (CRPs). The alignment positions were mapped back onto the *E. coli* seed sequences for the purposes of examining the correspondence of CRPs with protein features. The *E. coli* entry almost always carries the most detailed and highest-quality annotations, and more auxiliary information about the species and the cytosolic environment is available to support the analysis.

2.5.3 Effect of Window Size

To measure the effect of the window size parameter on the output of the algorithm, an experiment was run on a sample of proteins. Ten proteins were selected, each containing ten CRPs at a window size of 3, and the algorithm was reapplied with window sizes of 5, 7, 9, 11, 15, and 19. The results are displayed in Figure 2.6. In the sample, the number of significant positions follows one of two patterns over the range of window sizes tested: it either stays fairly constant, or increases as the window size is increased. The proteins in which the number of significant regions increases all have regions interspersed with CRPs separated by 3-10 amino acids. In larger windows that overlap a number of CRPs the rareness is aggregated, making it quite likely that the whole window will be significantly rare. The larger the window size, the greater the number of positions that are affected by this. Furthermore, the subsequent analysis of the positions of CRPs would be made less precise by using a larger window size.

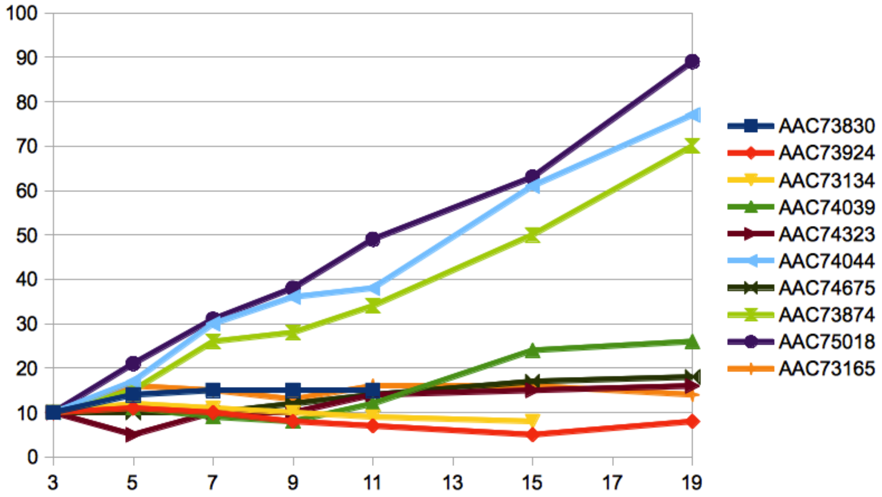


Figure 2.6: The number of CRPs as a function of window size. A sample of ten proteins was selected, each containing ten CRPs at a window size of three. In the majority of proteins, the number of CRPs stays reasonably constant. Where the number of CRPs increases with increasing window size, it appears to be a consequence of closely interspersed CRPs having an aggregate effect on a large number of overlapping windows.

2.5.4 Conservation

Amino acid and codon conservation at CRPs was examined largely as a potential null effect. Sets of closely related species with similar cytosolic environments will produce similar optimally-adapted genes, even when the codon usage in the genes is independent. Alternatively, genes that had been recently transferred to a number of members of a clade and not had time to adapt to the new cytosolic environments might give rise to CRPs when the effect was not due to a selective advantage. In both cases, the effect should be distinguishable by high levels of sequence conservation. Conservation of codons and amino acids was measured by normalised Shannon's entropy, so that the range is between 0 and 1 with lower values indicating higher conservation. Although positions with zero codon entropy were found to be slightly enriched in the set of CRPs versus other positions (odds ratio = 1.485; $p = 0.005$ under Fisher's exact two-tailed test), the average entropy is higher in CRPs than other positions (0.389 versus 0.369, $p << 10^{-10}$ under Mann-Whitney U test). We also checked the numbers of sequences and unique organisms represented in alignments containing CRPs, and found that alignments with CRPs are more diverse than those without. Given this surprisingly strong evidence against homology being the source of the CRP signal, it was considered safe to proceed with the interpretation of the results.

2.5.5 Comparison with Other Alignment-Based Methods

We know of two previously published metrics that take a similar approach to identifying conserved rare codons: the RCRR method (Widmann et al., 2008) and Sherlocc (Chartier et al., 2012). Both methods are described in Section 1.4.

RCRR

RCRR shares another conceptual similarity with our method in the comparison with the distribution of possible scores. Their implementation requires explicit computation of synonymous codon usage combinations, which scales very poorly (see Section 2.4.2). Presumably for this reason, the comparison score is calculated for specific columns only, placing a heavy burden of conservation on the exact position of rare codons. Taking the product of codon usage frequencies in the column as the test statistic, instead of their sum, counterbalances this by lowering the requirement for conservation in the column, because a mix of small and large values is scored lower than a set of intermediate values under this scheme. However, the measure considers alignment columns separately, so unconserved regions of rare codon usage in a small proportion of sequences can generate positive RCRRs erroneously. This compromise contributes to the high false positive rate of 4 %. The method presented here uses the summed codon frequencies as the test statistic, which is more conservative, and is implemented with a smaller window size (see Section 2.4.1).

A direct comparison of the results generated by the RCRR and CRP algorithms is difficult because the results are presented for just 16 homologous families built by the Lipase Engineering Database. Only two of these contain *E. coli* proteins that could be taken as representative, and neither of these contains any RCRRs. Table 2.2 is included for completeness, and shows the comparison where possible.

Sherlocc

In Sherlocc a *p*-value is obtained by fitting the average window scores to an extreme value distribution. It is difficult to see how this distribution could be constructed from the available data. It appears the process is actually more akin to selecting the bottom X % of scores. This distinction is key, because in the latter case there is no null hypothesis and the number of positives is pre-determined by the sample size alone. Also, the problem of gaps in the alignment is not addressed. Although the alignments in Pfam are carefully built and generally of a high quality, they still contain regions with a significant proportion of gaps, which could lead to a discrepancy between the level of conservation assumed by the method and that which is actually observed. Finally, although Pfam alignments are typically made non-redundant at the sequence level using a maximum sequence identity threshold

LED ID	RCRRs	E. coli ID	CRPs	Homologous family name
abH01.02	10	x	-	Mammalian carboxylesterases
abH08.14	2	x	-	Ccg1/TafII250-interacting factor B like
abH09.02	0	AAC76437	2	BioH protein like
abH12.01	3	x	-	Hydroxynitrile lyases
abH14.02	0	x	-	Gastric lipases
abH15.02	6	x	-	Burkholderia cepacia lipase like
abH17.01	3	x	-	Chloroflexus aurantiacus lipase like
abH19.01	4	x	-	Palmitoyl-protein thioesterase 1 like
abH23.01	0	x	-	Rhizomucor mihei lipase like
abH24.01	2	x	-	Pseudomonas lipases
abH26.01	0	x	-	Deacetylases
abH28	0	AAC74915	0	Prolyl endopeptidases
abH30.01	0	x	-	Cocaine esterases
abH31.02	0	x	-	Carboxymethylenebutenolidases
abH33.01	0	x	-	Antigen 85-C
abH34.02	7	x	-	Serine carboxypeptidase II like

Table 2.2: Unfortunately a comparison of the locations of RCRRs and CRPs is impossible because there is almost no overlap between the two datasets. The columns, from left to right, show the Lipase Engineering Database ID for the homologous family, the number of RCRRs in the alignment, the NCBI ID of a representative *Escherichia coli* K-12 protein if available, the number of CRPs in the alignment, and the name of the homlogous family. Only two families contain a protein that could be used for comparison, and neither of these contains any RCRRs.

of 80 %, multiple sequences from the same species are tolerated (Finn et al., 2010). Such paralogous sequences evolve in identical cytosolic environments and so cannot be considered independent examples of coding sequence evolution even when the protein sequences are divergent. The alignments used in this investigation were purpose-built and care was taken not to include paralogues.

The method presented here possesses a lower predicted false positive rate, higher sensitivity, and improved runtime compared with Sherlocc and RCRR, and as such can be considered at least a viable alternative.

2.6 Algorithm Behaviour

A striking pattern in the results of the algorithm is the very distinct peaks at the extrema of the p-value range (see Figure 2.7). Only 64 % of computed p-values lie between 0.001 and 0.99. This is an interesting result in itself, as it suggests the presence of strong opposing evolutionary pressures that exert their influence, or gain dominance, in a binary, mutually exclusive fashion. This goes against the traditional interpretation of variations in codon usage, which posits a single optimal coding pattern and uniform evolutionary pressures that are felt more keenly in highly-expressed genes. It has been noted previously that genes encoding aminoacyl tRNA synthetases (aaRSs) tend to use rare codons more frequently than the background. This is thought to improve the chances of the cell recovering from a state of amino acid starvation by ensuring that aaRSs can still be synthesised even when amino acid levels are low (Elf et al., 2003). The aaRS proteins also tend to avoid their own cognate tRNA for the same reason (Seligmann, 2012). The same pattern has been

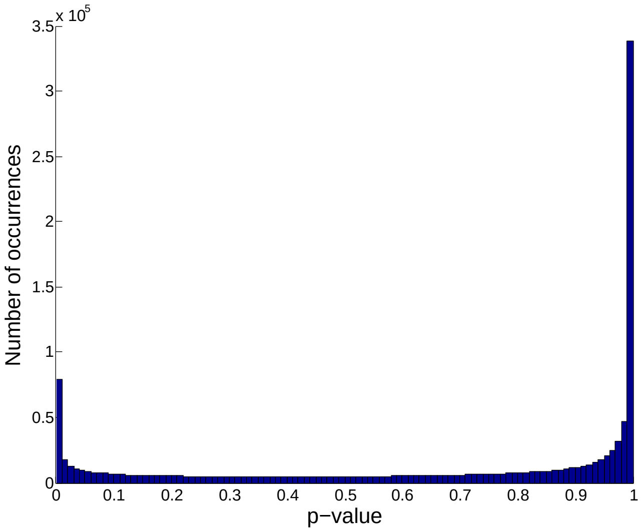


Figure 2.7: There are clear frequency peaks at the extremes of the p-value range. This suggests that in a large proportion of positions codon usage is either conserved high-frequency or conserved rare.

noted in genes that are up-regulated in response to stress (Gingold et al., 2012).

There is evidence that the cellular pool of tRNA is highly dynamic, adapting in response to growth rates and environmental conditions (Dong et al., 1996; Putzer and Laalami, 2003). The effect we are seeing could be the result of the same selective force, encouraging use of the codon corresponding to the most abundant tRNA, but adapted to two distinct sets of cellular conditions. Alternatively, it could be related to elongation rate. Since protein synthesis rates are the main determinant of prokaryotic fitness (Johansson et al., 2008), it makes sense that the majority of positions experience selection for codons that putatively enhance the rate of translation. It is plausible that the small number of key locations that require lower rates of translation, mediated by codon frequency, are under equally strong selection.

2.6.1 Amino Acid Composition

Even after correcting for amino acid composition by normalising the codon probabilities in our sampling technique, we find that some amino acids are enriched in CRPs and others are depleted with respect to the background occurrences of amino acids (see Figure 2.8). As discussed in Section 2.3.1, amino acids with greater degrees of bias in their synonymous sets contribute more strongly to the underlying distribution, so this is not unexpected. The most over-represented amino acid is lysine with an observed:expected ratio of 1.5. The most under-represented amino acids are methionine – many of which are excluded because of their N-terminal position – and aspartic acid, with ratios of 0.59 and 0.62 respectively.

The fact that CRP occurrence is not independent of amino acid composition is not a weakness – rather, it is a property of the measured signal – but it has the potential to confound the analysis, as amino acids have their own preferences that are determined by their physico-chemical properties. This is addressed in Section 3.4.1 in the next Chapter, which presents the results of the analysis of the data generated by the algorithm.

2.6.2 Experimental Evidence

There are two experimentally verified examples of beneficial rare codons in native *E. coli* genes. This section discusses those two examples and the findings of our algorithm in relation to them.

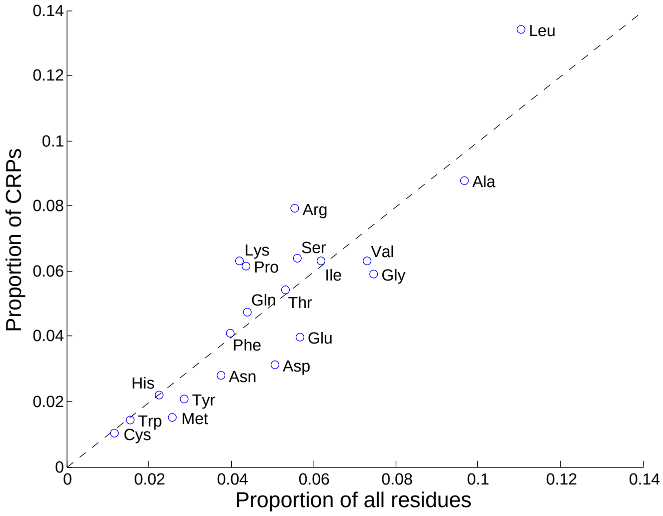


Figure 2.8: Frequency of amino acids in positions classified as CRPs against their frequency in the whole *E. coli* genome. The dashed line represents expected occurrences, if CRPs were completely independent of amino acids. Residues occurring above the line are over-represented in CRPs, and vice versa.

SufI

SufI is a 470 residue, three domain protein that is secreted via the Tat pathway. It is thought to play a structural role in cell division, but its precise function is unknown (Tarry et al., 2009).

Zhang et al. (2009) used an algorithm based on tRNA abundances in *E. coli* to predict regions of slow translation. They identified twenty codons encoded by rare tRNAs, in four different regions of the gene spanning all three domains. They mutated combinations of these locations, and found that just two common-for-rare codon substitutions in the latter-central region, at positions 244 and 252, were disruptive to proper folding of the protein. Similar results were achieved by over-expressing the rare tRNAs. They probed the mechanism of the disruption by examining the size and protease lability of fragments derived from cell-free systems, concluding that translational pauses were occurring near the sites they identified and that the change in the translation rate profile was altering the structure of folding intermediates. They also implemented a temperature-mediated global deceleration of translation, and were able to mitigate the deleterious effect of the mutations.

The results of our algorithm do not agree with any of the twenty the positions identified by Zhang et al. (2009), although several of them correspond to regions of low conservation where the alignment is poor. The requirement for high conservation could be considered a weakness of the method, but comes with improved accuracy in a trade-off that was considered worthwhile. Two positions in the earliest region, at residues 34 and 42, have p-values of order 10^{-5}.

CAT-3

Chloramphenicol acetyltransferase 3 (CAT-3) is a 213 residue, single domain protein that forms a functional trimer. CAT-3 confers resistance to the chloramphenicol antibiotic by covalently attaching an acetyl group that prevents it from binding to the ribosome. Komar et al. (1999) identified sixteen rare codons in a twenty-codon region near the midpoint of the protein. They found that substituting these rare codons for more common ones increased protein synthesis but reduced enzymatic functionality by 20% in cell free systems, implying that a portion of the protein had folded incorrectly.

Again, none of the positions mutated by Komar et al. (1999) met the threshold for identification as CRPs in our algorithm. We identify two positions with p-values in the order of 10^{-5} in the region specified in the paper, although not coinciding with the mutated positions. It is also worth noting that the modifications disrupt an internal Shine-Dalgarno-like sequence in the region, which has been suggested as an alternative explanation for the findings (Li et al., 2012).

2.6.3 Conclusions

It is disappointing that we do not find direct agreement between our method and the only two pieces of experimental evidence available. However, the lack of agreement does not invalidate our findings. Our method identifies position-specific conservation of rare coding. In some cases there may be flexibility in the precise location of rare codons within the gene. If rare codons are present in all homologues but are spread across a region rather than in a specific, conserved position, they may not produce CRPs.

Equally, because our method relies on homology to increase its statistical power it is not capable of identifying CRPs in regions of poorly conserved amino acid usage. It is feasible that a pause that is required for structural reasons may be instigated by different mechanisms in different homologue versions. For example, where one gene may use a sequence of rare codons to slow translation, another may have undergone an insertion mutation that introduces extra translational steps. This would still effect a relative translational delay between two structural positions either side of the region, but without slowing translation or using rare codons.

In both experimental investigations, a large number of residues are identified and mutated but only a small proportion of these are found to be relevant to translational mechanisms – the precision is very low. In CAT-III a very large window is mutated, without a great deal of explanation as to how it was selected. In SufI only two of the codons identified are reported to make a difference to the folding. We considered this trade-off worth making, because single-sequence methods are not reliable enough to be extended to general cases. The related computational methods, where the information is available, have similar levels of agreement with these experimental data (Widmann et al., 2008; Chartier et al., 2012).

Importantly, both studies used *in vitro* systems for the bulk of their experimental work. While such cell-free systems contain the essential machinery of translation, they are missing myriad components that are present in a living organism, such as chaperones and binding partners, and many regulatory mechanisms that control metabolism in response to environmental cues (Parry et al., 2014). This could well result in erroneous pausing due to the absence of co-factors that help to maintain and regulate translation in *in vivo*. Further, the cytosolic environment is so replete with constituents sized over several orders of magnitude that it inherently constrains the folding of the nascent chain in ways that remain poorly understood (Kim et al., 2013).

Although we do not see direct correspondence with experimental results, there are numerous possible explanations for the lack of agreement that do not invalidate our findings. The approach on the conservative side in terms of establishing statistical significance, and we can be confident that the rationale behind the algorithm is reasonable. Thus we can proceed to the analysis of results.

3 Analysis of Conserved Rare Positions in *Escherichia coli*

This chapter documents the large-scale analysis of codon usage data in *Escherichia coli* K-12 MG1655. Alignments of homologous proteins were built for 3886 coding regions in the *E. coli* K-12 genome, drawing from a database of about 2.3 million genes from 678 prokaryotic genomes. After the application of thresholds on alignment quality and window size, p-values were computed for 1,162,109 valid positions. 13,615 positions across 2336 alignments were identified as displaying conserved rare codon usage. The algorithm is described in detail in Chapter 2.

The aim of this part of the work was to identify a biological explanation for the presence and location of the conserved rare codon positions (CRPs). The stringent significance threshold that was applied, and the high degree of conservation required by the algorithm, make a strong argument for a location-specific selective pressure that defies the prevalent genomic tendencies. To elucidate this selective pressure, we are interested in finding data to support a translational mechanism by which CRPs might aid correct protein expression.

The investigation is predicated on the assumption that CRPs are genetic locations where a translational pause is necessary for proper protein expression, and hypotheses are constructed on this basis. Experimental evidence for this premise is admittedly sparse, and in reality a strong direct correlation between codon usage and translation rate probably cannot be relied upon. However, the same is true of any other currently known signal, including tRNA abundances (Bonekamp et al., 1989; Stadler and Fire, 2011). There is strong evidence, presented here and elsewhere, that rare codons provide some selective advantage, and the only known mechanism by which they could influence translation is through elongation rate. Codon usage is selected as a relatively convenient and accessible signal providing an abundance of data available for analysis. Although codon usage frequency may be a more abstract proxy for translation rate than tRNA abundances, it offers vastly more usable data and thus can add considerable statistical power to predictions. The measure designed herein is more sensitive and more specific than previous codon usage frequency measures, by virtue of the combination of the choice of window size and the use of homology, and more wide-reaching in its application to prokaryotic sequence data.

The investigation begins by mining for associations between CRPs and gene-level feature annotations such as functional groups, and also position-specific features, such as secondary structural elements and disordered regions. We then examine structural class and structural motifs. Finally, we examine in detail a small number of genes in which we find strong evidence of selectively adaptive rare codon enrichment.

A Note on Paralogues

Domain families often contain numerous closely-related proteins. Steps were taken to remove paralogues from alignments, so that the identification of CRPs is based only on correspondences between sequences from separate organisms. However, alignments based on paralogous genes are likely to contain overlapping samples of sequences from the database because the homology searches will return similar results, so treating them as independent could conflate some parts of the analysis. Paralogous redundancy was removed from the database for the latter part of the investigation in this chapter concerning structural motifs and domain families, but not the earlier part that examines functional annotations. This decision was made because functional conservation is lost relatively quickly in the evolutionary process, and gene duplication encourages diversification of sequence and function (Martínez-Núñez et al., 2010). Also, previous studies have shown that paralogues are not enriched in any particular functional group (Nembaware et al., 2002). Section 3.1.1 details how paralogues were identified and removed.

3.1 Materials and Methods

3.1.1 Removing Paralogues

Strong paralogues were identified as mutual hit pairs in the original BLAST searches – i.e. pairs of sequences where both sequences returned the other as a hit with an E-value of $< 10^{-5}$ in our database (see Section 2.5.1). These mutual hit pairs can be used to define edges in homology graphs of a set of interest, for example a structural family. Families were pruned by removing sequences until the homology graphs were completely disconnected – i.e. until there were no edges remaining. Sequences were pruned in a greedy fashion, removing sequences in descending order of their number of incident edges in the homology graph (i.e. the number of mutual hits remaining in the family). Pruning was repeated until the family contained no pairs of mutual hits. Multiple domains from the same protein were allowed, because although the domains probably arose from an historic duplication event the thresholds on alignment quality mean that the same multidomain pattern is present in the other sequences in the alignment. This means that the duplication event would very likely have occurred before speciation, so the same arguments about sequence divergence apply.

3.1.2 Assigning Structure to Sequence

Secondary Structure

Disorder annotations were taken from MobiDB (Di Domenico et al., 2012). The MobiDB database combines manually curated annotations, data mined from X-ray and NMR structures in the PDB, and the results of multiple disorder prediction algorithms. Steps are taken to resolve annotation conflicts, and a distinction is made between known and predicted disordered regions. MobiDB provides a webservice that returns serialised data on known and predicted disordered regions for UniProt entries, making residue assignment relatively easy.

Mapping sequence and structural data on a large scale is far from trivial. Proteins are often heavily modified from the wildtype for structural analysis, and many structures are missing coordinates for particular residues and/or contain multiple, sometimes homologous chains, so pairwise alignments between genomic and structural sequences can be very messy. For this study, simplified position-specific secondary structural features were extracted from the UniProt entries for the alignment seed sequences. UniProt maps the eight secondary structure classifications from DSSP onto three separate features: HELIX (G, H and I), STRAND (E and B) and TURN (T). The mapping process is complicated by multiple, sometimes conflicting annotations from various sources structures, and details of UniProt's process are unpublished (UniProt Consortium, personal communication). The mapping can be particularly problematic for helices, where the numerous different classes of helix result in some feature tables containing separately-listed helix features that are directly adjacent in the sequence. Such features were merged into single helical regions for the analysis. UniProt only takes annotations from known structures, so the population was limited to the approx. 1300 *E. coli* proteins that are represented in PDB. This was considered preferable to resorting to predicted secondary structure, which is inevitably imprecise (Deane and Saunders, 2011). Regions that did not belong to any of the other secondary structure classes and were not predicted (or known) as disordered were classified as coils.

The testing for enrichment in these regions was done as follows. The null hypothesis is that the presence of CRPs is independent with respect to the residue classification – in other words, that the structural classification is an unbiased sample of residues with respect to CRPs. The expected number of CRPs in a secondary structure class is equal to the number of CRPs per residue in the whole set of proteins with structure multiplied by the number of residues in the class. The probability of obtaining the observed number of CRPs in each secondary structural region was then calculated under this model using Fisher's exact test, which is equivalent to summing over the tail of a hypergeometric distribution. Regional sets of CRPs were tested separately against the structural classes. In these cases the sample and the population were limited to the span of the region – for example, in the case testing N-terminal CRPs in helices, the population of residues is the size of the region (38), multiplied by the number of proteins in the dataset with known structure (1298), and the sample is the number of helical residues in the first 38 residues of each protein. Although this constitutes a not-insignificant number of tests, it was not seen fit to adjust the *p*-values because the tests are not independent – since the classes are mutually exclusive, enrichment in one class comes at the expense of another.

Domains and Domain Linkers

Domain region annotations were taken from the Gene3D database of known and predicted domains and domain families (Lees et al., 2012). Gene3D uses hidden Markov model profiles generated from the domain families in the CATH database to assign domain predictions to sequences with unknown structure. These predictions are mapped to regions of UniProt sequences and conflicting regions are resolved. These mappings were used to define domains in proteins with good coverage of our dataset (3861 out of 3886 proteins contained at least one predicted domain), to classify proteins as multidomain, and to assign them to fold classes and domain families.

Gene3D predictions were also used to derive domain linker regions. Linker regions were identified as non-terminal regions in proteins with two or more predicted

Gene3D domains that were not within the extent of the predicted domains. Domain linker regions are therefore imprecise and do not necessarily correspond to regions of predicted disorder.

Structural Motifs

Structural motifs were identified using PROMOTIF (Hutchinson and Thornton, 1996), which analyses PDB structure files and outputs information on the tertiary configuration of structural elements. PROMOTIF identifies secondary structural elements (beta strands and helices) and their interactions in tertiary configurations (beta sheets and barrels, beta-alpha-beta units, beta hairpins), distinguishes between different classes of turn (beta and gamma), and gives the locations of point irregularities and modifications (beta bulges, disulphide bridges). Its precomputed output can be accessed via PDBSum (de Beer et al., 2014). The residue numbers given in PROMOTIF have to be mapped back to UniProt residue numbers, which can be done via the SIFTS database (Velankar et al., 2013). Motifs were assigned here by selecting a single representative structure and chain for each protein in the dataset with any known structure. 1187 structures were successfully parsed for motifs. Helix-turn-helix motifs are not explicitly analysed by PROMOTIF but can be inferred from the helix context.

3.1.3 Annotation Enrichment

The Database for Annotation, Visualization and Integrated Discovery (DAVID; Huang et al. 2009a,b) was used for assessing enrichment of functional and ontological annotations. DAVID traverses hierarchical annotation trees and adjusted p-value-like EASE scores (Hosack et al., 2003) for the enrichment of annotation classes in a given gene list against a specified background. The EASE scores are the upper bound of the Fisher's exact p-values under jackknifing – that is, the maximum value that can be obtained by removing any single gene. This accounts for the sensitivity of the p-value to sampling precision, giving a more robust measure in cases where sample sizes are small or particular classes are lightly populated. DAVID offers the ability to cluster sets of annotations from numerous databases according to their overlap, and uses a filtered version of the Gene Ontology hierarchy that attempts to remove the broader, redundant terms in favour of more specific annotations, based on the number of children belonging to a node.

3.1.4 Expression Levels

Highly expressed genes were identified using the Gene Expression Omnibus (GEO; Edgar et al. 2002). GEO Profiles can provide information on the ranking of genes by their expression levels across curated data sets. This can be used to identify genes that are consistently highly expressed regardless of experimental conditions, or genes that are highly expressed under at least some sets of conditions. GEO profile queries can be downloaded as text files and parsed for Gene ID information, which can be mapped to numerous other resource identifiers contained in UniProt entries. For this study, highly expressed genes were defined as those expressed in the top percentile of all experiments in at least one dataset.

3.1.5 Essential Genes

The Keio database (Baba et al., 2006) was used to identify essential genes in our dataset. The Keio project involved the construction of an exhaustive set of single-gene knockout mutants of *Escherchia coli* K-12. Some 303 genes that resisted all attempts to grow deletion mutant colonies were classified as essential. 267 of these were mapped back to our database via their unique gene name.

3.2 Strengthening the Argument for the Value of CRPs

Two possible arguments against the claims made here about the selective advantage of CRPs were identified. First, as is acknowledged earlier the lack of true independence between homologous sequences could be seen to undermine the statistical significance of the *p*-values used to identify CRPs. Because of this, the false discovery rate may be higher than was accounted for when setting the threshold for defining rareness. Second, a well-known result in the study of codon usage is that more highly-expressed genes use a higher proportion of common codons. This is taken as evidence that it is only in these genes that selection on synonymous codon usage is sufficiently potent to have a lasting effect on gene sequence. These two arguments are examined in this section: the first by looking at the effect of codon and amino conservation at CRPs, and the second by examining expression levels of genes with and without CRPs.

3.2.1 Conservation

As stated previously, the assumption of statistical independence between corresponding codons from homologous genes does not necessarily hold in all cases. The conservation of codon usage was examined and found not to be significantly influencing the identification of CRPs (see Section 2.5.4). It was also noted that positions with unusually high codon entropy are enriched in CRPs. This high variability in combination with conservation of rareness suggests a strong selective force at work.

High-entropy positions are over-represented in the set of CRPs compared with the underlying distribution of position entropies. In particular a cluster with high amino acid and relative codon entropy (i.e. codon entropy − amino acid entropy) was determined by eye as the region defined by amino acid entropy in the range $(0.44, 0.525)$ and relative codon entropy in the range $(0.1, 0.225)$. The cluster contained 344 CRPs from sixteen different proteins, an unusually high average number of CRPs per protein. It should be noted that this includes four homologous Rhs (rearrangement hotspot; Hill et al. 1994) proteins with large numbers of rares, whose alignments therefore contain similar sets of sequences. Even when all but one of these are removed the cluster is still apparent. Figure 3.1 shows the codon and amino acid entropy distributions of rare and non-rare positions. The density plot includes only one of the rhs genes.

The proteins represented in the cluster are listed in Table 3.1, along with details of their functions. They include proteins with known co-translational export mechanisms; and toxin-antitoxin systems that inhibit intracellular growth or confer antibiotic resistance, carrying a strong selective advantage.

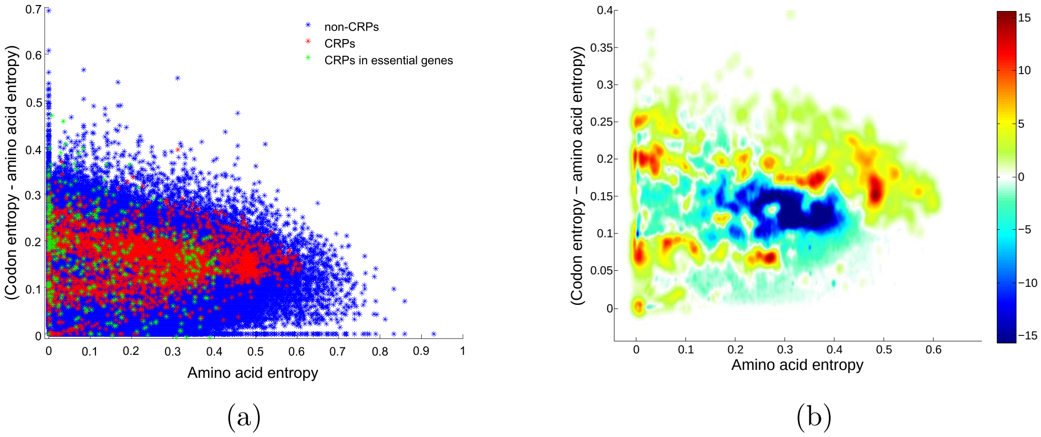


Figure 3.1: Shannon's entropy of codon and amino acid selection across alignments at CRPs and non-CRPs. (a) Scatter plot of codon vs. amino acid entropy at CRPs (red) and non-CRPs (blue). CRPs in essential genes are plotted separately in green (see Section 3.2.2). Note that the vertical axis shows the difference between codon and amino acid entropy, because there must be at least as many unique codons as there are unique amino acids. (b) The relative density of CRPs in entropy space. The normalised densities of CRPs and non-CRPs were estimated separately using Gaussian kernel density estimation. The differential density is shown in a heat map. The relative enrichment of high codon-entropy positions in the set of CRPs is clearly visible as a series of dark red patches towards the top of the shaded area. The dark red patch to the right of the figure contains the proteins described in Table 3.1. Note that zero-entropy positions were removed from the density estimation calculations. Also note that the plot shown includes only one of the rhs genes described in Table 3.1, to demonstrate that this cluster still exists if the others are discounted.

Table 3.1: Details of proteins containing CRPs found in the high-entropy cluster

CDS Acc	Gene Name and Description	# CRPs	Comments
AAC75027	dcm; DNA cytosine methyltransferase	1	
AAC74990	fliC; flagellar filament structural protein (flagellin)	1	Sole constituent of flagellar filament, which comprises up to 20,000 FliC subunits. Must be exported through the central channel of the flagella filament and therefore must be considerably unfolded. The narrow channel is thought to aid in preventing premature folding (Apel and Surette, 2008), and the presence of rare codons could also assist in this process by delaying translation.
AAC74580	hipA; inactivating GltX kinase facilitating persister formation; toxin of HipAB TA pair; autokinase	1	Toxic element of HipAB chromosomal toxin-antitoxin operon. The hipA gene is downstream of hipB. The HipA protein deactivates glutamate-tRNA ligase, preventing the charging of AA-tRNA(Glu). The toxicity of HipA is neutralised upon binding to an already-formed HipB dimer. Therefore slow translation of hipA may be critical for suppressing its toxic effects by allowing HipB dimers time to form. The HipAB TA module is associated with high persistance (Feng et al., 2013).
AAC77304	hsdS; specificity determinant for hsdM and hsdR	65	Part of the EcoKI DNA methyltransferase restriction/modification enzyme. HsdS forms a trimer with two copies of HsdM, which is encoded immediately upstream (overlapping by 1 nt). The hsdS gene has 203 CRPs out of 464 positions, a very high proportion of CRPs. In contrast the upstream hdsM gene has 17 out of 529.

CDS Acc	Gene Name and Description	# CRPs	Comments
AAC76617	rhsA	59	The rhs (rearrangement hotspot) genes are characterised by variable toxic C-terminal domains, which is the region where all of the CRPs in this group are found. Rhs proteins are excreted when environmental conditions are crowded and are toxic to neighbouring cells, inhibiting growth and thus offering a selective advantage to cells possessing the downstream immunity gene (Jackson et al., 2009; Zhang et al., 2012). Note that the CRPs in rhs genes make up a considerable portion of the cluster, and since the genes are homologous their alignments contain many of the same proteins. However, the cluster is still apparent when only one rhs gene is included in the analysis – see Figure 3.1.
AAT48186	rhsB	56	
AAC73794	rhsC	36	
AAC73599	rhsD	61	
AAC77307	mrr; methylated adenine and cytosine restriction protein	1	Laterally acquired type IV restriction endonuclease involved in stress response (Ghosh et al., 2014)
AAC75098	wzxB; predicted polisoprenol-linked O-antigen translocating flippase; lethality reduction protein	20	Membrane-embedded transporter protein that reduces the lethal effects of stress (Han et al., 2010).
AAC73999	ycaI; ComEC family inner membrane protein	2	Transmembrane pore involved in the uptake of exogenous single stranded DNA (Sun et al., 2009).
AAC75803	ygcB; cascade complex anti-viral R-loop helicase-annealase Cas3	27	Cas3 has two opposing functions: it either acts to anneal or unwind RNA-DNA R-loops (nucleic acid secondary where RNA hybridises with one strand of a separated stretch of DNA), depending on the levels of ATP.

CDS Acc	Gene Name and Description	# CRPs	Comments
AAC74229	cohE; e14 prophage; cI-like repressor protein phage e14	1	Numerous components of probable prophage elements, including repressors – which prevent the cell phage from entering the lytic phase – and the RecE homologous combination system.
AAC73638	intD; DLP12 prophage; putative phage integrase	7	
AAC74427	intR; Rac prophage; integrase	4	
AAC75673	yfjI; CP4-57 prophage; uncharacterized protein	2	

3.2.2 CRPs in Critical Genes

The argument has been made that only the most highly expressed genes are under sufficient evolutionary pressure to optimise their codon usage patterns (Ikemura, 1981a; Sharp et al., 1986, 2010; Klumpp et al., 2012). These models for codon sequence evolution posit that codon selection is a relatively marginal selective adaptation, and that weakly expressed genes are more weakly selected, so that mutational drift is sufficient to override selective codon adaptation. As discussed in Section 2.2, there is evidence to the contrary. Even weakly expressed genes display an observable difference in nucleotide composition compared with untranslated regions (Hershberg and Petrov, 2009, 2012), and many weakly expressed proteins can confer a significant evolutionary advantage in circumstances where they are expressed. It has been shown recently that synonymous mutations can be as strongly selected as non-synonymous mutations (Bailey et al., 2014), so the adaptive advantage to be gained from codon selection is clearly not marginal. Additionally, mutational drift does not account for the conserved patterns that our method highlights, where species consistently defy their genome-wide preferences at specific locations. However, there is evidence that nucleotide-specific biases are stronger and proportional use of high-frequency codons is higher in more highly expressed genes. Therefore, highly expressed genes that contained CRPs were sought.

Using the Gene Expression Omnibus (Edgar et al., 2002), 116 highly expressed genes (see Section 3.1.4) were identified. This set of genes was significantly enriched in CRPs in all regions, with the largest fold change (2.33) occurring in the internal (non-terminal) region where 45 genes had CRPs. There was a single pair of mutual hit paralogues in this set (see Section 3.1.1), and the significance was robust to its removal. This included numerous ribosomal constituents and translational components, proteins involved in stress- and misfolding response systems, transcription factors, and essential primary metabolic proteins involved in ATP and glucose metabolism.

There is also a heavy enrichment of CRPs in essential genes (see Section 3.1.5 for details of how essential genes were identified), with the highest fold change (2.11) again in the internal region. Essential genes are generally more evolutionarily constrained at the amino acid level on account of the often lethal effects of deleterious mutations, but CRPs can be found in them across the range of the entropy landscape, and especially in locations with high amino acid conservation (low entropy) and high codon divergence (see Figure 3.1a). This again suggests that CRPs in these genes can confer a selective advantage.

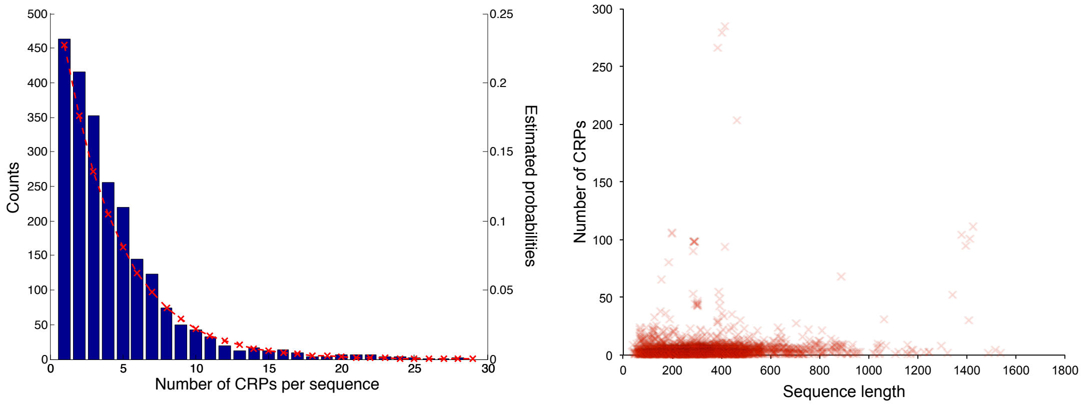
3.3 General Investigation

3.3.1 Overview

Distribution Between Sequences

Just under 1.2% of positions tested were classified as CRPs, and these were distributed between 60% of the genes examined. The number of CRPs per sequence appears to follow a geometric distribution (see Figure 3.2a). This is suggestive of random independent events. However, if CRPs occurred randomly at the nucleotide level there would be a strong correlation between the number of CRPs in a gene and the length of the sequence. In fact, there is virtually no correlation at all, indicating a gene-level effect that suggests a link with translation (see Figure 3.2b). Looking at single genes in isolation, one could conclude that the effect arises from a negative adaptive force against rare codons, where each successive rare window somehow damages the efficacy of a gene in an additive manner so that subsequent mutations producing rare windows prevail with diminishing probabilities. This aligns with the observation that longer genes use a higher proportion of common codons, perhaps to protect the metabolic investment of producing a large protein (see Section 1.2.4). However, the conservation of the pattern across an alignment is evidence for a positively adaptive, position-specific mechanism. A specific position may be critical with respect to another feature, or it may be that a number of rare windows are required anywhere in the gene but must be incorporated without disrupting other features, leaving a limited number of viable positions. The position must be important, but whether it is critical with respect to another feature, or must be incorporated into the gene without disrupting other sequence features, remains unclear.

Despite the lack of correlation between the number of CRPs and the length of a sequence, genes containing CRPs are on average about 30 residues longer than those not containing CRPs ($p < 10^{-5}$, Mann-Whitney U-test; see Figure 3.3). Sequence length is often taken as a measure of folding rate (De Sancho et al., 2009). The extended length of proteins with CRPs could therefore be taken as suggestive of some threshold in protein complexity, above which CRPs are more likely to be required to modulate translation. In search of further evidence for this hypothesis, the contact order of single-domain proteins with known structures was examined. Contact order is defined as the mean sequence separation between residues that are



(a) (b)

Figure 3.2: (a) Histogram of the number of CRPs per sequence. This is well approximated by a geometric distribution with estimated probability $\hat{p} = 0.23$ (goodness of fit $R^2 = 0.9843$) [Note: the figure shows the data up to 30 CRPs per sequence only] (b) Number of CRPs plotted versus sequence length. Points are plotted with a transparency of 0.75 to give an indication of density. If CRPs occurred randomly on the nucleotide level there would be strong correlation between sequence length and the number of CRPs per sequence. In fact the two variables are almost completely uncorrelated (Spearman's $\rho = 0.0842$). This indicates a gene-level process and is strong evidence for a link with translation.

less than 5 Å apart in the structure. It is a measure of the complexity of a protein structure and a better proxy for folding rate (Plaxco et al., 1998). There is no statistically-significant difference between proteins with and without CRPs, and no correlation between contact order and the number of CRPs in sequences that have them, indicating that folding rate alone cannot explain the role of CRPs.

Distribution Within Sequences

Examining the locations of conserved rare positions (CRPs) shows strong peaks towards the ends of proteins, especially the N-terminus (see Figure 3.4). The N-

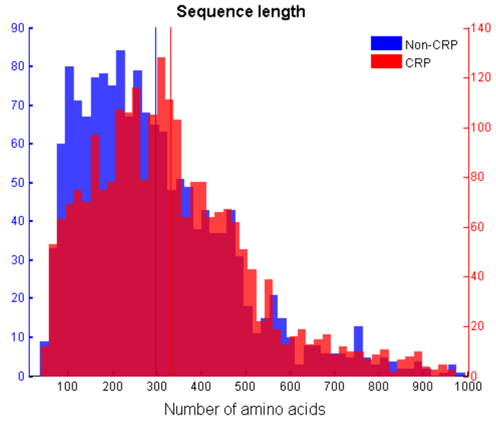
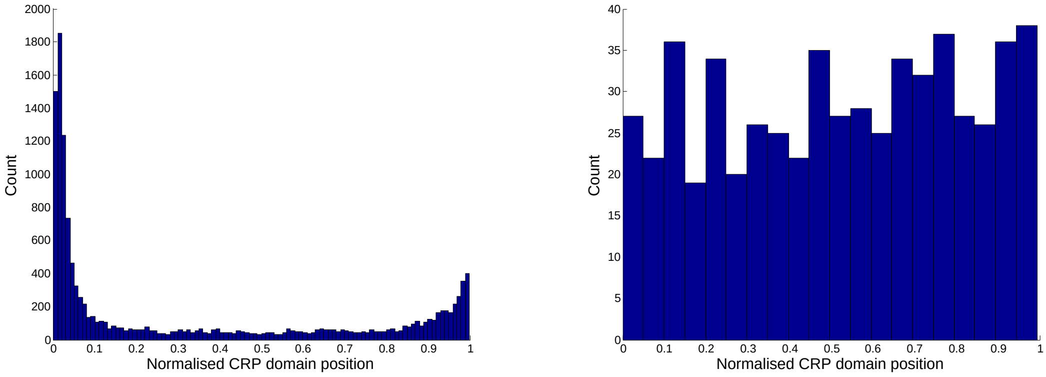


Figure 3.3: There is a statistically significant difference in the sequence lengths of proteins containing CRPs vs. those not containing CRPs.

terminal pattern is in agreement with numerous previous studies that attribute the effect to selection against strong mRNA secondary structure near the initiation region (Kudla et al., 2009; Goodman et al., 2013; Tsukuda and Miyazaki, 2013) or to the need to avoid ribosomal collisions during translation (Tuller et al., 2010a,b; Navon and Pilpel, 2011). The C-terminal peak is more novel but has been discussed in at least one previous study (Clarke and Clark, 2010), where it was suggested that a C-terminal pause may allow the nascent chain to take further advantage of the ribosomal environment for folding or promote interaction with cofactors or chaperones. Another possible explanation is that a substantial number of *E. coli* genes are in very close proximity to, or actually overlapping with, the Shine-Dalgarno sequence or even the coding sequence of the subsequent chromosomal gene (Eyre-Walker, 1996). This would lead to conflicting pressures related to the other gene rather than any feature of the gene in question; codons that bear a resemblance to parts of the Shine-Dalgarno sequence are often rare (Li et al., 2012), and the N-terminal mRNA secondary structure of the second gene would also become a constraint. This could be contributing to the C-terminal peak we observe if the operon or chromosome structure was broadly preserved between a number of species. This is difficult to check due to the limited availability of genomic structural data. The N-terminal peak in particular is associated with full protein sequences but not individual domains, supporting the explanations related to translation initiation or termination, or chromosomal gene position, rather than co-translational folding.

Goodman et al. (2013) also found that the GC content of a species' genome influenced the likelihood of finding rare codons in the N-terminal region. A-T hybridisation has lower free energy than G-C, and if the rest of the gene contains more GC-rich codons then hybridisation is also less likely if AT-rich codons are used in the N-terminal region. The alignment-averaged GC content is significantly higher in alignments containing N-terminal CRPs, suggesting that this is indeed a factor, but as Goodman et al. (2013) also pointed out, it does not explain the full distribution. The difference in GC content almost disappears when alignments containing CRPs in any position are considered against those without (see Figure 3.5).

The effects mentioned here do not fully account for previously documented links between expression and codon usage (Goodman et al., 2013), or for the conservation of rareness observed. Nonetheless, location was used to split the data into multiple sets to isolate known explanations from potential new discoveries. Goodman et al. (2013) identified codon usage in the region centred on the tenth residue as contributing most strongly to adaptation, whereas Kudla et al. (2009) suggested that the first 38 residues were responsible. Both of these regions are visibly separated from what appears to be the baseline CRP frequency in our data (see Figure 3.6). In this study the first and last 38 residues of proteins were treated as distinct regions and were examined separately and in combination with internal CRPs – those occurring at other locations.



(a)

(b)

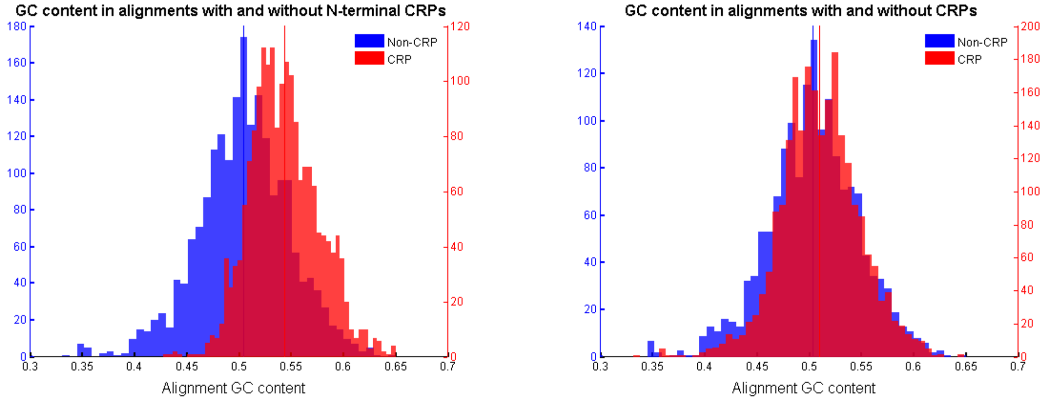
Figure 3.4: CRPs are heavily clustered at the termini, especially the N-terminus. (a) shows the distribution of CRPs normalised by sequence length in full protein sequences, (b) shows the distribution in individual, non-terminal domains (see Section 3.1.2). The N-terminal peak disappears when N-terminal domains are discounted, strongly suggesting a link with translational initiation. The C-terminal peak is also less pronounced but still present.

3.3.2 Gene Level Features

This section discusses gene level annotations that were found to be enriched in proteins with CRPs.

Annotation Enrichment

A striking feature of the set of proteins containing CRPs is the enrichment (1.6 fold, with a corrected p-value of 1.4×10^{-6}) of ribosomal constituents. Ribosomal components are likely under heavy selective pressure, due to their high expression



(a) (b)

Figure 3.5: Distributions of alignment-averaged GC content in alignments with and without CRPs (a) in the 38 residue N-terminal region and (b) anywhere in the sequence. The distributions of genes with and without CRPs are plotted on differently-scaled vertical axes, as indicated by the colour. GC content was calculated for alignments as the percentage of G or C nucleotides in all represented mRNA sequences after removal of gaps. The histogram bars are slightly transparent to show the shapes of both distributions. The coloured vertical lines indicate the means of the corresponding sample. In (a) the means are 0.543 in the group with CRPs versus 0.504 in those without; in (b) the means are 0.510 versus 0.501. The difference in means in the full set is entirely attributable to the difference in the means of the N-terminal samples.

levels and their vital role in the cell. Because of this, ribosomal genes have frequently been assumed to be translationally optimised and therefore exemplary in terms of codon usage (Sharp et al., 2005; Puigbò et al., 2007; Higgs and Ran, 2008; Suzuki et al., 2008; Wang et al., 2011). This finding is therefore of special significance and interest to the study of codon usage patterns.

There is no clear explanation relating the enrichment of CRPs in ribosomal genes to the elongation rate-related hypothesis. However, it does fit with the "two channel" hypothesis explained in Section 2.6 – that some tRNAs are reserved for urgent production of high-priority proteins. Since ribosomal genes are clearly of critical

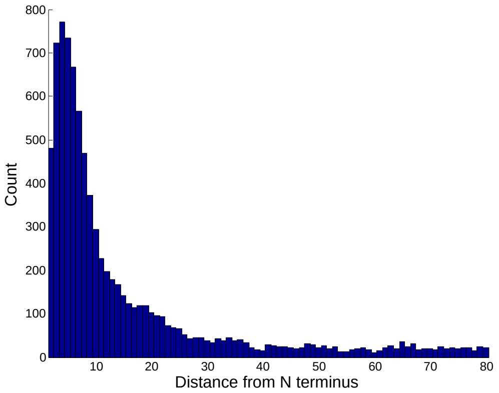


Figure 3.6: The peak in the occurrence of CRPs in the N-terminal region is strongest between the positions 5 and 10, plateauing around position 20. There is then a further drop towards the baseline frequency until position 38.

importance to the cell it would make sense for their production to be prioritised.

Interestingly, CRPs are also enriched in other components of the translational system and many processes involving DNA interaction. Several aminoacyl-tRNA synthetase, helicase, ligase and polymerase enzymes and nucleotide and ribonucleotide binding proteins are over-represented in the set of proteins with CRPs. DNA binding activity is strongly associated with the Helix-Turn-Helix (HTH) structural motif, which is also over-represented in the set. HTH motifs are very common in regulatory elements that bind specific nucleotide sequences, and are also involved in DNA and RNA metabolism (Aravind et al., 2005; Chakravartty and Cronan, 2013). HTH motifs are generally quite small (Pellegrini-Calace and Thornton, 2005) and can form fast-folding, stable, independent domains (Religa et al., 2007), so it would be surprising if they required translation rate modulation to fold under kinetic control. The effect could be due to an enrichment in disordered regions, which are also associated with DNA binding, or in domain boundaries, since the HTH motif is present in a large number of domain families (see Section 3.4.1).

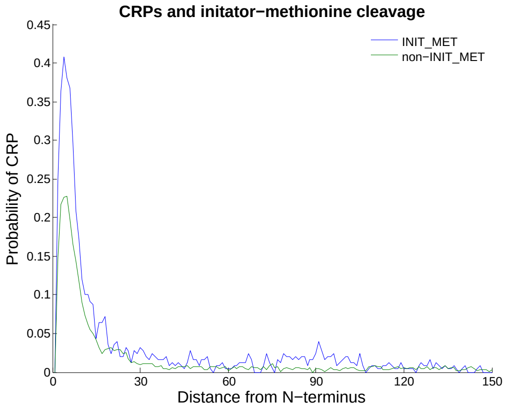


Figure 3.7: Probablity of finding a CRP with distance from the N-terminus, in proteins whose initiator methionine (iMet) is cleaved and those whose iMet is not cleaved. CRPs are enriched across the whole N-terminal region in proteins whose iMet is cleaved. This could suggest a role for CRPs in the co-translational cleavage mechanism.

CRPs are also heavily enriched in the N-terminal region of proteins whose N-terminal methionine is cleaved. N-terminal methionine (iMet) cleavage is a common modification that occurs co-translationally in a large percentage of proteins. It is a two stage process, requiring deformylation of the iMet and subsequent breakage of the peptide bond, and is often critical for function and structural stability (Liao et al., 2004). In *E. coli* there is a single enzyme, methionine aminopeptidase (MAP), that is responsible for all iMet cleavage. Cleavage of iMet by MAP is sensitive to the properties of the subsequent amino acid P1'. Bulky side-chains are thought to prevent cleavage from occurring at all, and among the smaller amino acids threonine and valine can hamper cleavage (Frottin et al., 2006). In particular a Glycine in position P1' is thought to readily enable cleavage.

If cleavage is hampered by a bulkier amino acid, the kinetics of cleavage could be aided by a translational pause, which could be mediated by the presence of CRPs.

Taking experimentally confirmed annotations from UniProt entries, the probabilities of finding CRPs in the N-terminal region of proteins that have their iMet cleaved were compared with those that have some experimental or structural annotation but not the iMet cleavage tag. Enrichment of CRPs was found across the whole N-terminal region, especially in the first 45 residues and extending less dramatically to approximately 100 residues (see Figure 3.7). Proteins were then classified according to their second amino acid, and the probabilities of finding N-terminal CRPs were calculated for each set. All of the proteins with valine in the P1' position, and nearly three quarters of those with threonine, were found to have N-terminal CRPs ($p = 8.78\times10^{-}6$, Fisher's exact test). There is also significant enrichment in proteins with serine and alanine P1', but not glycine, the P1' residue reported to be most amenable to iMet cleavage. This could well suggest a co-translational role for CRPs, where a translational pause kinetically aids iMet cleavage when the identity of the adjacent amino acid is inhibitory.

3.4 Correspondence With Structure

Numerous studies have failed to reach a consensus on the correspondence of the nucleotide composition and codon usage patterns in mRNA sequences with the secondary and tertiary structure of the proteins they encode, producing instead a bewildering set of overlapping and conflicting conclusions.

Specific codons have been determined to have varying preferences for secondary structural classes, and the N- and C-termini of structural elements. Some studies have found universal patterns relating to the third codon position (Adzhubei et al., 1996), while others have claimed that preferences exist only in higher organisms and not in prokaryotes or that only one or two codons per genome have varying preferences, and these differ between species (Tao and Dafu, 1998; Oresic and Shalloway, 1998; Gupta et al., 2000). Still other studies have suggested that the preferences of individual codons for or against secondary structural elements vary depending on the fold class of the protein as well as the species in which it is found (Gu et al., 2003).

Studies examining codon frequency in secondary structure are also divergent in their findings. Early evidence suggested an enrichment of rare codons in strands and a deficit in helices (Thanaraj and Argos, 1996a), while a more recent study drew almost the opposite conclusions – focussing on the transitions between helical and strand regions and coils, Saunders and Deane (2010) found an enrichment of rare codons at the transitions into helices and a deficit at the transitions into strands. Similar discrepancies exist in attempts to quantify codon usage in domain boundaries. Thanaraj and Argos (1996b) and more recently Makhoul and Trifonov (2002) and Zhang and Ignatova (2009) found a general enrichment of rare codons in, around, and downstream of domain boundaries, while Saunders and Deane (2010) found the opposite. Brunak and Engelbrecht (1996) are in the minority with their conclusion that, although nucleotide-level biases that correspond to structural regions do exist, they can be explained entirely by the preferences of the amino acids and there are no specific or frequency-related correlations between codons and secondary structure. The majority of experimental evidence supporting the existence of positively adaptive rare codons concerns domain boundary regions (Cortazzo et al., 2002; Angov et al., 2008; Zhang et al., 2009), but computational studies suggest that these advantages are not conserved at the fold level (Widmann et al., 2008; Saunders and Deane, 2010; Chartier et al., 2012).

As part of this study, the correspondence between CRPs and secondary structural elements, domain boundaries and fold classes was investigated. The results of this investigation are presented in this section.

3.4.1 Regional Enrichment

The investigation began by examining the presence of CRPs in classified structural regions. Residues in proteins with known structure were classified as helix, sheet, turn, coil, or disordered (the annotation and testing process is described in Section 3.1.2). The results are summarised in Table 3.2.

In accordance with Thanaraj and Argos (1996a), CRPs were found to be significantly depleted in helices in all regions of proteins, most significantly in the N-terminal region. The rates in strand residues are location-dependent; CRPs are enriched in strands appearing in the N-terminal region of proteins, but this preference reverses in the internal region and towards the C-terminus, where CRPs occur less than expected. CRPs are depleted in structural turns in the N-terminal and internal regions, but show no strong preference towards the C-terminus. In coil regions they are enriched at the N- and C-terminus, and show slight but not statistically significant enrichment in the internal region.

This aligns simplistically with what is known about how readily secondary structural units are acquired – helices are thought to fold quickly and stably, since most of their interactions are between sequential residues and propagation is thermody-

		Overall	N-term	Internal	C-term
Helix	fold change	0.5860	0.5921	0.5978	0.8601
	p-value	1.05e-123	5.41e-59	8.56e-31	7.96e-4
Strand		0.9317	1.1303	0.6304	0.8432
		4.33e-3	7.30e-5	4.96e-14	0.01027
Turn		0.6744	0.6511	0.4799	1.4163
		7.11e-6	2.74e-4	1.82e-4	0.0327
Coil		1.1662	1.2101	1.0545	1.1440
		1.72e-21	3.05e-22	0.0628	1.27e-3
Disorder (known)		1.9725	1.0604	2.3801	0.7577
		1.70e-26	0.204	1.97e-8	0.0294
Disorder (predicted)		2.9600	1.4439	2.6310	1.2497
		0.0	6.99e-44	1.23e-41	7.20e-6
Domain linker		N/A	N/A	0.9839	N/A
				0.399	

Table 3.2: Fold change against expectation and statistical significance of CRPs in the various residue classes and sequence regions. Statistically significant results (those with *p*-values $< 1\text{e-}3$) are coloured according to the $\log_2$ of the fold change, in red for depletion and in blue for enrichment.

namically favourable. In contrast beta-sheets are more complex and involve more long range contacts (Stanger et al., 2001). However, this explanation is not fully satisfactory as the nascent chain is not able to fold at the moment of translation. Rather, between 30 and 72 residues remain buried in the ribosomal exit tunnel where they are unable to attain tertiary structure; the variability comes from the ability of the nascent chain to form helices inside the exit tunnel, making the emerging structure more compact (Zhang et al., 2009; Fedyukina and Cavagnero, 2011). This effectively slows the emergence of the nascent chain with respect to translation rate. Therefore the comparative slowing of sheet forming residues could act to maintain a more consistent rate of extrusion and exposure to the cytosol over the course of normal translation.

As mentioned in Section 2.6.1, despite the fact that our measure corrects for amino acid usage frequencies, different amino acids do have differing propensities for classification as CRPs in accordance with the degrees of bias in their synonymous codon sets. This is not a departure from what was sought, but is a potential confounding factor in this sort of structural study because amino acids also have structural biases determined by their physical properties. To ensure that the variations in CRP levels in the different amino acid structures were not simply due to differences in amino acid structural preferences aligning with CRP tendencies, the proportion of CRPs in each amino acid in secondary structural regions was tested against the overall proportions. If the proportions were similar or appeared to be drawn from the same distribution, the variations in CRP levels in and out of secondary structures could be attributed to amino acid propensities and not because of differences in the pressures acting on codon usage. In all cases this null hypothesis can be rejected with very high confidence in favour of the explanation that codon usage differs in the various secondary structural classes (see Figure 3.8).

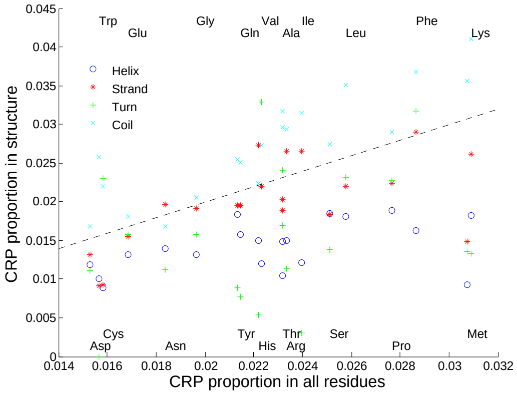


Figure 3.8: The proportion of amino acid residues involved in CRPs in the various structural classes, plotted against the proportion that are involved in CRPs across all residue types. Each series has a point for each amino acid. The text labels show the horizontal positions of the amino acids, representing their background CRP frequencies – the proportion of all occurrences that are present in a CRP window. The dashed line represents equilibrium, i.e. the likelihood of finding a CRP in a given amino acid does not depend on the type of secondary structure it belongs to. Backgrounds were measured in proteins with known structure only. The clearest pattern is in helices, were all amino acids have lower levels of CRPs compared with their overall levels. Amino acids in strands and turns also show a general reduction on average, whereas those in coils are generally enriched. All amino acids are heavily enriched in CRPs when found in disordered regions, either predicted or known – these data are not plotted. In all cases the deviation is significant under a Chi-squared test with $p < 0.001$.

On a residue level CRPs are enriched in disordered regions, both known and predicted, at the N-terminus and internally. At the C-terminus, there is actually a slight negative correlation between the occurrence of CRPs and known disorder, but the signal is positive in predicted disorder in this region, when more proteins are included in the population.

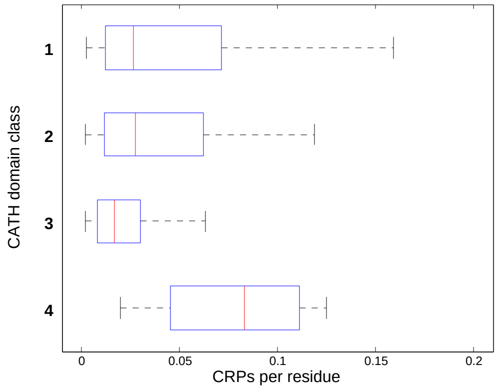


Figure 3.9: CRPs per residue in the various CATH domain classes. Class 1 domains are mainly alpha helical, class 2 are mainly beta strand, class 3 are mixed alpha/beta and class 4 have few secondary structures. Class 2 domains have slightly more CRPs than class 1, although there is little difference at this level. Mixed domains have fewer CRPs on average, and irregular domains have more (although the sample is small). Note that the plot only includes those domains containing at least one CRP.

Examining the numbers of CRPs per residue in known and predicted domains belonging to different CATH classes (see Section 3.1.2) supports the above findings but adds an interesting twist. As expected, the domain class with few secondary structures is relatively enriched in CRPs (although the sample size is small, with only 25 domains in the dataset from this class). In domains that have at least one CRP there is little difference between mainly alpha and mainly beta class domains, but mixed alpha/beta domains have fewer CRPs on average (see Figure 3.9). The overall preferences hold in all domain classes, but depletion of CRPs in helices is even stronger in mixed alpha/beta domains than in mainly alpha, with a fold change of about 0.52 compared with 0.56 in the internal region.

There is no significant difference between the level of enrichment in domain linker regions against the background. Domain linkers occur exclusively in the internal (i.e. non-terminal) regions of proteins. See Section 3.1.2 for information on how domain boundary annotations were derived.

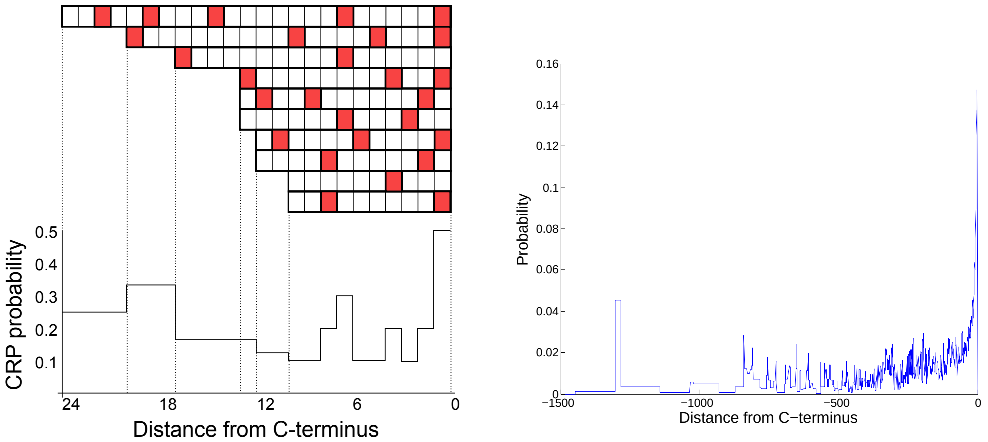
These findings are interesting, but there is no clear explanation relating direct positional enrichment to translation rate or co-translational processes because of the delay in cytosolic exposure following translation. Rather than examining the coding of specific regions, it would be interesting to look for enrichment in CRPs upstream or downstream of the features identified here. A statistical bootstrapping method was developed to examine the levels of CRPs over a range of distances up- and downstream of features and compare these to an expected level. This analysis is presented in the next section.

3.4.2 Relative Positional Enrichment

A bootstrapping method was developed to assess the significance of positional correspondences between CRPs and other features, without being influenced by the independent positional preferences of feature sets. For example, as noted in Section 3.3.1 above, there are accepted explanations for the abundance of CRPs at the termini that are largely unrelated to co-translational events. Helices and beta sheets occur more frequently in the internal regions of proteins because the termini are often unstructured. In combination these factors result in a correspondence that is unrelated to the interaction between the two features. It is necessary to determine how much of the positional correspondence between features and observed CRPs is due to co-dependence, rather than coincidence between independent positional preferences. This can be done by generating random CRP patterns that follow the overall positional distribution of CRPs in the actual data and comparing these to the features.

For each set of features, probability distributions were built to describe the likelihood of finding a CRP at a given distance from the C-terminus in sequences containing the feature. Because they are unlikely to be involved in co-translational events, CRPs in the N-terminal 38 residues were ignored. The distributions were built by sorting the sequence lengths in descending order, anchoring them at the C-terminus, and looking at segments defined by the gaps between successive lengths. Probabilities were calculated by dividing the number of CRPs at each position by the number of sequences represented in that segment. If a segment did not have at least one rare codon in each position then the probabilities were averaged over the whole segment. Distributions fitted in this way are termed "rare profiles". Figure 3.10a illustrates the process of constructing a rare profile for a set of proteins, and Figure 3.10b shows the resulting distribution of CRPs in sequences containing helices as an example.

Having constructed a rare profile for a set of sequences containing both CRPs and a feature of interest, all the distances between the actual CRPs and the start and end positions of features are counted. New, random sets of CRPs are then generated using the rare profile fitted to the set, and the distance counting is repeated. This was performed one million times for each feature, and the distance counts were averaged over the repeats. By comparing the observed counts to the averaged counts, the relative enrichment in CRPs can be calculated at each position – that is, the enrichment compared with what would be expected if the CRPs and features appeared independently. Relative enrichment was calculated at each

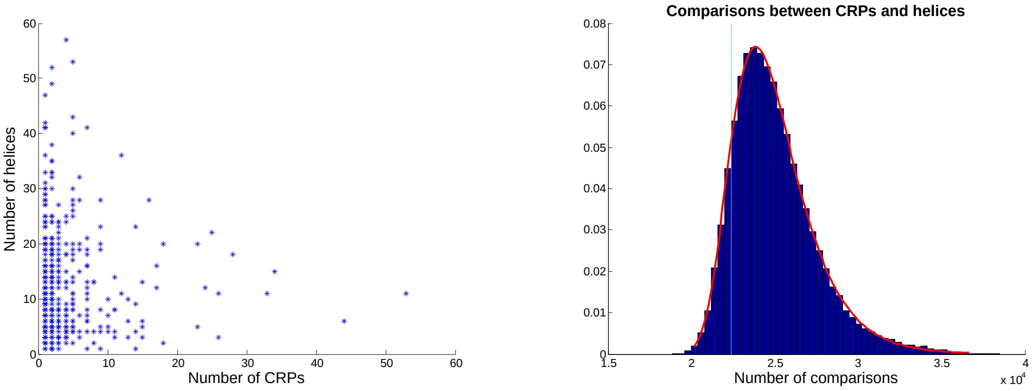


(a) (b)

Figure 3.10: Calculating the rare profiles of a set of proteins. (a) shows an illustration of the method. The rows represent protein sequences aligned at the C-terminus, with CRPs shown as red boxes. Distances from the C-terminus are divided into regions formed naturally by the ordered lengths of proteins. Probabilities are averaged over the sparsely populated regions at high distances. Closer to the C-terminus, where there is at least one CRP at each distance, the probabilities are position-specific. These profiles are used to generate random CRPs patterns whose positions are compared to the positions of feature sets. This allows effective correction for inherent positional biases. (b) shows the actual probability distribution used to generate random rare profiles for sequences containing helices. The distribution is coarse at high distances where there are few proteins, but becomes more precise closer to the C-terminus.

position as $\frac{n_e - n_d}{n_e + n_d}$, where n_e is the number of random patterns with fewer CRPs than the observed pattern (indicating enrichment in the observed pattern) and n_d is the number with more (indicating depletion). The relative enrichment score at a given position can range from -1, when the observed pattern has fewer CRPs at that position than occurred in any of the randomly generated profiles, to 1 when the observed pattern has more CRPs than any of the random profiles.

Applying this analysis to the secondary structure classes, additional evidence



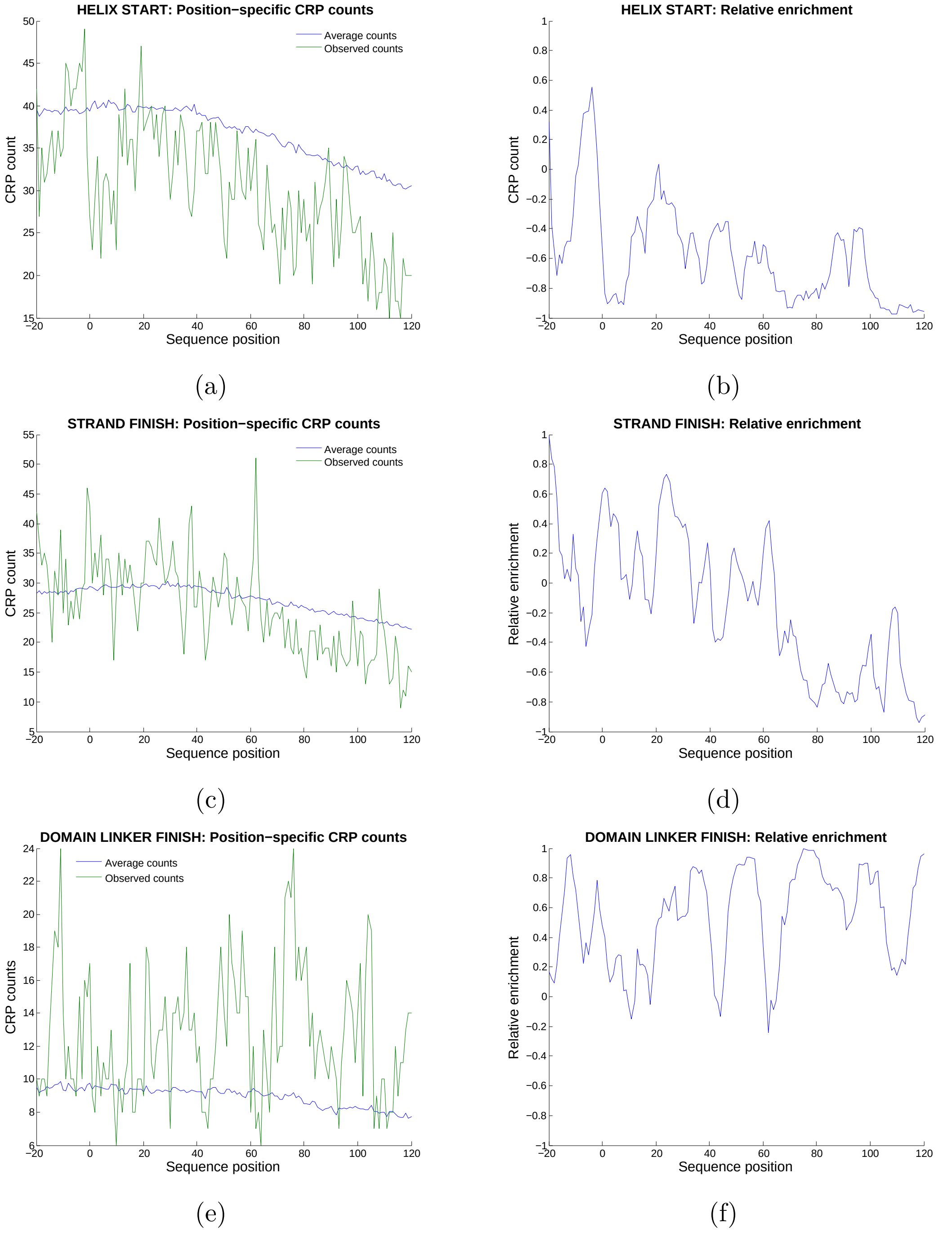
(a) (b)

Figure 3.11: (a) Scatter plot of the number of helices against the number of non-N-terminal CRPs in sequences that have both. There is a slight negative correlation between the two values, especially in sequences with larger numbers of rares. The number of comparisons is the sum of the products of the number of helices and CRPs in each sequence. When the random profiles are generated based on the overall positional distributions, this results in more comparisons on average than are observed in the original sequences. (b) The number of comparisons made between CRPs and helices. The distribution is built by randomly re-pairing the observed numbers of CRPs and helices, and is well described by a generalised extreme value distribution (shown in red). The vertical cyan line shows the actual number of comparisons in the observed data.

was found for the depletion of CRPs in helices, where a depleted region can be seen downstream of the starts and upstream of the finishes corresponding to the average helix length in the dataset (9.73 residues; see Figure 3.12a and b). An overall depletion in the region downstream of helices can also be observed. This arises partly from a slight negative correlation between the number of CRPs and the number of helices in a sequence, especially in those sequences with large numbers of CRPs, which reduces the overall number of comparisons made (see Figure 3.11). It could also suggest a disincentive to use rare codons after a helix has occurred in a protein.

In strands, the slight enrichment within strand regions can again be observed, but a stronger depletion far downstream, above 60-80 residues, is also apparent (see Figure 3.12c and d). As mentioned above, this distance is significant with respect to translation as it corresponds approximately to the number of residues of the nascent chain that remain buried in the ribosome exit tunnel between the P-site and the cytosol, although there is no obvious explanation in this case. There are no strong positional enrichment patterns with respect to turn or coil regions.

Examining domain boundaries, there is no significant enrichment difference in linker regions against the background level of CRPs. There does appear to be a peak shortly upstream, and most strongly about 80 residues downstream of the finishes of boundary regions (see Figure 3.12e and f). Again, this latter peak corresponds to the number of residues that remain buried in the ribosomal tunnel after translation. This could well be indicative of a pause to allow N-terminal domains to fold before more C-terminal portions are translated. Looking at a larger region, it appears that there is a slight enrichment in CRPs across the whole downstream region relative to domain boundaries. Examining the numbers of rares per domain in N-terminal and non-N-terminal domains, there is a small but significant difference (1.2757 vs. 1.1725, $p < 0.0005$ Wilcoxon's Rank Sum test). This suggests that non-N-terminal domains are more likely to contain CRPs, contrasting with the results of Chartier et al. (2012).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.12: The positional enrichment of CRPs relative to selected structural features: the starts of helices in (a) and (b), the finishes of strands in (c) and (d), and the finishes of domain linker regions in (e) and

(f). The horizontal axes show the sequence position, where negative is upstream (N-terminal) of the helix start. (a) The observed counts at each position and the average counts generated from random profiles based on the positional distribution of CRPs in helical proteins. (b) The relative enrichment of CRPs at each position. Relative enrichment scores were smoothed using a moving average window of size 5. The depletion of CRPs within and far downstream of helices is clearly visible in both plots. (c) and (d) show the observed and average counts and the relative enrichment with respect to the finish positions of strand elements. The slight enrichment within strands is visible, but the most noticeable feature is the depletion downstream, starting between 60 and 80 residues. (e) and (f) show the same data with respect to the finishes of domain linker regions. The enriched peak at 80 residues downstream is the strongest feature, suggesting a possible link with separate co-translational folding of domains.

3.4.3 Tertiary Structure

Protein secondary structure classifies the hydrogen bonding state of the backbone of amino acid residues (Branden and Tooze, 1999). Secondary structural elements in functional protein structures combine into more complex motifs, which themselves compose independently folding domains. The size and complexity of motifs and domains containing the same secondary structural elements can vary dramatically. This prompted an investigation into whether CRPs might play a role in mediating the folding of specific tertiary structures that are perhaps more complex than most, or that form most efficiently when allowed to follow a certain kinetic pathway.

A number of structural motifs were assigned using the PROMOTIF output via PDBSum (see Section 3.1.2). Four tertiary structural motifs were selected initially: beta sheets (superstructures of beta strands; sheet and barrel configurations were considered together), beta hairpins (consecutive anti-parallel hydrogen-bonded beta strands), beta-alpha-beta units (two parallel hydrogen-bonded beta strands connected by a helix-containing loop), and helix-turn-helix motifs (interacting helices separated by a turn, often involved in DNA binding and recognition). The rare profiling method described in Section 3.4.2 was applied to the C-terminal points of these motifs. In every case the samples were relatively small, but even so there was no significant signal over the sequence range of -20 to +120 residues.

As mentioned in Section 3.4.1, there is a slight difference in the number of CRPs per residue in different domain classes, although the overall preferences of secondary structural classes hold. The CATH domain classification of our dataset was analysed at the Superfamily level (see Section 3.1.2) in order to determine whether any particular structural family was heavily enriched in CRPs.

1047 domain superfamilies are represented in our dataset. As in the entire CATH database, the sizes are heavily skewed; 487 superfamilies contain only one domain, whereas the most heavily represented superfamily contains 284. After removing paralogues (see Section 3.1.1), there were 673 superfamilies with only one domain and the largest contained 66. The proportion of sequences in each family that had CRPs in the internal or C-terminal region was examined. 170 superfamilies have domains containing CRPs. Of the 43 families containing more than two domains with CRPs, there are only three in which more than half of the domains contain CRPs, and even these are not heavily populated – they contain 3, 4, and 5 domains, and in each case there is one domain with no CRPs at all. In all three families there is only partial correspondence between the locations of the CRPs (see Figure 3.13).

This strongly suggests that CRPs are not conserved in relation to tertiary structural features. Under the assumption that the folding pathway is conserved at the

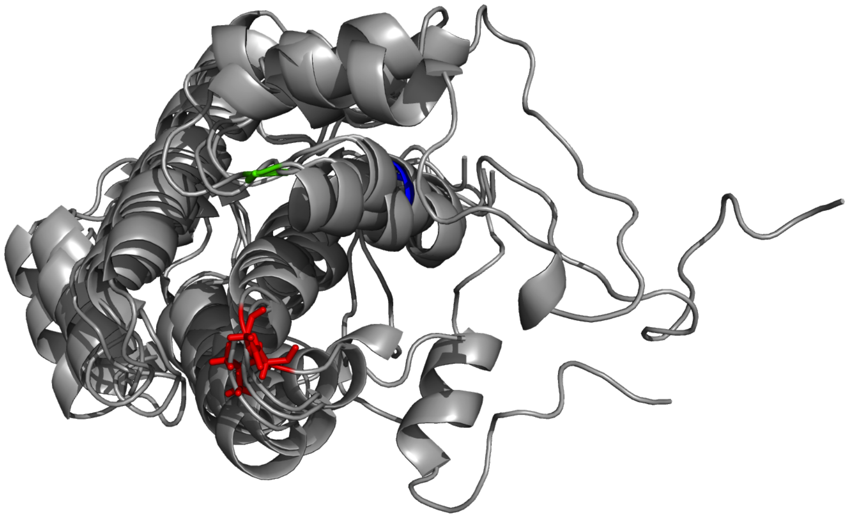


Figure 3.13: Superposition of four non-paralogous structures in CATH superfamily 1.10.340.30. The alignment shows the four domains from the CATH family in our database. Three out of four domains contain CRPs, but their positions do not correspond. The CRPs are coloured differently in each sequence, in red, blue, and green.

superfamily level, one could conclude that CRPs are unlikely to be directly involved in intra-domain co-translational folding. However, the specifics of folding pathways remain mysterious, and it is quite possible that small variations in the amino acid sequence could influence the specific pathway. Codon usage could modulating the elongation rate through a more convoluted route than can be described by pure frequency, or could be affecting the folding pathway through a different mechanism altogether. Until more detailed investigation of folding pathways is available alternative explanations for the influence of codon usage on structure, such as mRNA secondary structure or RNA interference, look more likely.

3.5 Conclusions

This chapter presents evidence for position-specific evolutionary pressures that shape the synonymous coding-sequences of genes in an organism-dependent manner. The intention was to identify co-translational processes in which CRPs may play a role. Although there are clear patterns in their locations within genes, and strong evidence of varying preferences in secondary structural elements, there does not appear to be a signal relating to protein folding at the domain level. There is a slight suggestion that the region downstream of domain boundaries is enriched in CRPs, which could indicate a pausing mechanism to allow proper folding of the preceding domain, but it is difficult to draw firm conclusions in this regard.

The strongest signal relates to the coding in regions that adopt helical structure. These regions consistently contain significantly fewer CRPs than any other class of residue. It is possible that this pattern arises from a pressure to maintain a consistent rate of extrusion. The ability of the nascent chain to adopt helical structure while still inside the exit tunnel would result in an effective slowing of the rate of exposure as the chain becomes more compact along the axis of the tunnel.

Additionally, there are numerous CRPs to be found in locations with disproportionately high variability in codon selection compared with amino acid selection across the alignment. Although there is no apparent unifying principle that governs their occurrence, the identification of CRPs in essential and highly expressed genes, including ribosomal constituents, is solid evidence for some conferred selective advantage.

4 Experimental Work

This chapter documents the experimental work that was undertaken in an effort to verify the indications of the algorithm. Because of time pressures, candidate genes highlighted by the still-ongoing computational work were selected for experimentation. These genes were to be cloned, modified as indicated by the intermediate computational results, and tested for altered activity.

Unfortunately, a mistake was discovered in the preliminary computational work while the experimental work was still ongoing. The initial construction of the sequence database contained a number of organisms represented by multiple strains. This redundancy introduces near-duplicate gene sequences with very similar codon usage statistics, which distorts the significance values calculated by the algorithm, so the redundant strains must be removed (see Section 2.5.1). The resulting changes to the database structure eliminated the positions that had been identified as CRPs in the selected candidate genes, rendering the experiments useless. At this stage in the project time constraints precluded restarting the experimental process with new candidates, so the experimental work presented here did not reach a satisfactory conclusion. This chapter documents the work that was undertaking up until the discovery of the computational error.

4.1 Materials and Methods

4.1.1 DNA Synthesis and Manipulation

All DNA and primers were synthesised by DNA2.0 (`https://www.dna20.com`) to order. Restriction digests were performed using New England Biolabs' (NEB's) *Nde*I and *Hin*dIII restriction enzymes. These were performed in NEBuffer 2. Ligations were performed with the NEB Quick Ligation kit, according to the manufacturer's protocol.

Gene variants were made using site directed mutagenesis. This was performed with Agilent QuikChange Lightning Site Directed Mutagenesis kits as per the kit protocol, in a thermal cycler with the lid pre-heated to 90 °C. Primers were designed using Agilent Technologies' QuikChange Primer Design web tool (`http://www.genomics.agilent.com/primerDesignProgram.jsp`; see Section 4.2.4).

4.1.2 Competent Cells

Two commercial strains of *E. coli* obtained from Agilent Technologies were used: XL10-Gold Ultracompetent Cells (catalogue #200314) for cloning of plasmid DNA produced by ligation or mutagenesis, and BL21(DE3)pLysS (catalogue #200132) for protein expression.

Chemically competent cells were produced from the *aphA* Keio (Baba et al., 2006) strain JW4015-1 using the following protocol: an overnight culture of the strain was grown in 5 ml kanamycin-selective Nutrient Broth (see Section 4.1.10). The following morning the overnight culture was added to 20 ml Nutrient Broth containing 20 mM $MgCl_2$, pre-warmed to 37 °C, and grown for one hour in a shaker-incubator at 37 °C, 225 rpm. The culture was removed and chilled in an ice bath, then centrifuged at 8000 g, 4 °C for ten minutes. The culture was returned to an ice bath, the supernatant discarded and the pellet resuspended in 2 ml of just-thawed 75 mM $CaCl_2$ in 15% glycerol. The suspension was divided into 200 µl aliquots in 0.5 ml tubes pre-chilled in dry ice. These were flash frozen in liquid nitrogen before storage at −80 °C for subsequent use.

4.1.3 Transformations

A different protocol was followed when working with the induced-competent cells derived from the *aphA* Keio strain as opposed to commercial competent cells. The commercial cells were transformed using a "standard" method involving a 42 °C heat shock, described as Protocol 1 below, whereas the Keio cells were transformed using the quick method detailed in Protocol 2, in which cells are spread directly onto preheated plates (Pope and Kent, 1996). The second protocol was found to produce high transformation yields using the ampicillin- and kanamycin-resistant Keio cells.

Protocol 1: Standard method

DNA was prepared in a solution of sterile water to a final concentration of 50 ng µl^{-1}. 2 µl DNA were added to a 15 µl aliquot of competent cells suspended in glycerol, defrosted on ice from −80 °C. This was mixed by flicking, then returned to ice. The mixture was exposed to a 30 second heat shock in a water bath at 42 °C, then returned to ice again. 0.5 ml Terrific Broth (TB

– see Section 4.1.10) at room temperature was added, mixed by inverting and returned to ice for 5 minutes. The mixture was then incubated for 1 hour at 37 °C, 225 rpm in a shaker-incubator. Following incubation, 0.1-0.5 ml cell mixture was spread on kanamycin-selective agar plates and incubated overnight at 37 °C. Plates were sealed with Parafilm and stored at 4 °C until needed.

Protocol 2: Quick method

DNA was again prepared in a solution of sterile water at the same final concentration and added in the same volume to an equivalent aliquot of cells. The mixture was gently mixed and incubated on ice for 5 minutes, then spread directly onto selective agar plates pre-heated to 37 °C and incubated at the same temperature over night. Plates were sealed with Parafilm and stored at 4 °C until needed.

4.1.4 Plasmid Preparation

To prepare cultures, colonies were picked from plates and used to inoculate 5 ml lysogeny broth (LB – see Section 4.1.10) with 50 µg ml^{-1} kanamycin in 50 ml Falcon tubes. Tubes were laid horizontally and incubated overnight in a shaker incubator at 37 °C, 225 rpm. Plasmid extraction then proceeded with a miniprep using QIAGEN QiaPrep kits according to the manufacturer's protocol.

4.1.5 DNA Gel Electrophoresis

DNA gel electrophoresis was performed in 100 ml of 1% agarose gel in Tris-borate-EDTE buffer containing 5 µl ethidium bromide, with an electric potential of 120 V for 45 minutes. NEB Quick-Load DNA Marker (catalogue number NEB #N0303) was used to indicate fragment sizes. Gel extractions were done with QIAGEN's QIAquick and MinElute spin kits.

4.1.6 Protein Expression

Protein expression was done in the BL21(DE3)pLysS strain of *E. coli*, with genes incorporated into the pET29-a vector. Shaker-incubators were set to 37 °C, 250 rpm unless otherwise specified.

To express the genes, colonies were picked from plates and used to grow starter cultures of 10 ml TB with kanamycin, in 50 ml Falcon tubes placed horizontally in a shaker-incubator overnight (16 hours). 100 ml TB with kanamycin in a 500 ml shaker flask was inoculated with 4 ml of the overnight cultures and placed in a shaker-incubator at 37 °C, 250 rpm for 3 hours. Expression was induced with IPTG at a final concentration of 0.5 mM. Cultures were left to express in the shaker-incubator under the same conditions for 3 hours, then divided into 50 ml aliquots and centrifuged at 15 000 G for 10 minutes. The resulting pellets were stored at −20 °C until required.

4.1.7 Protein purification

Protein samples were purified prior to electrophoresis using nickel affinity chromatography. Cell pellets (see Section 4.1.6) were resuspended in binding buffer (20 mM imidazole, 0.1 M HEPES, 300 mM potassoim chloride) and lysed by sonication (10 µm amplitude, 20 seconds on, 30 seconds off for 20 cycles). The lysate was placed in a centrifuge and spun at 15 000 g for 15 minutes. The supernatant containing the soluble proteins from the cell lysate was extracted and placed in a settled column of nickel sepharose suspension. This was washed with imidazole in increasing concentrations (5 mM, 20 mM, and three times with 500 mM) to elute the bound protein. The elution samples were collected and separated by size using SDS-PAGE (see Section 4.1.8).

4.1.8 Protein Gel Electrophoresis

Protein gel electrophoresis was used for checking soluble expression of the protein products. Polyacrylamide gels were made using National Diagnostics ProtoGel 15% resolving gel and 4% stacking gel, made according to the supplied protocol. Samples were mixed in equal volumes with Loading Buffer (National Diagnostics, catalogue number EC-886) and heated to 95 °C for 10 minutes, then cooled to room temperature before loading. Separation was done under a constant current of 35 mA for 40 minutes. Life Technologies PageRuler Unstained Protein Ladder (10 - 200 kDa) was used for sizing proteins.

4.1.9 Assays

Spectrophotometric assays were the most appropriate option given the resources give the resources and expertise available in the research environment.

aphA – class B acid phosphatase

The acid phosphatase protein cleaves a phosphate group from para-nitrophenyl phosphate, producing para-nitrophenol (pNP). This activity can be measured by monitoring the electromagnetic absorption spectrum; pNP absorbs highly at wavelengths in the range 390-420 nm.

Pellets from 50 ml aliquots of expressed cell culture (see Section 4.1.6 for details of expression) were resuspended in 10 ml of assay buffer consisting of 20 mM sodium acetate, 1 mM $MgCl_2$ at pH 6.5 with 1 mM DTT. The tube containing the resuspended cells was placed in an ice bath and the cells were lysed by sonication (10 µm amplitude, 20 seconds on, 30 seconds off for 20 cycles). The lysate was then placed in a centrifuge, pre-chilled to 4 °C, and spun at 15 000 g for 15 minutes. The supernatant containing the soluble proteins from the cell lysate was extracted and further diluted two-fold in the assay buffer.

Phosphatase activity was monitored using a stopped assay. A 25 ml volume of diluted supernatant from the previous step was warmed to 37 °C before the addition of 1 mM pNPP. The reaction was incubated at 37 °C. 250 µl samples were taken at successive time points and quenched in 750 µl 2M NaOH to halt phosphatase activity and allow measurement of pNP levels. Absorption was measured at 406 nm, the observed peak absorbance of the standard samples, and the final concentration of phosphate-cleaved pNP was determined from the calibration curve shown in Figure 4.3.

4.1.10 Media Recipes

Lysogeny Broth

Lysogeny broth was made by dissolving 20 g Sigma-Aldrich LB Broth powder (catalogue number L3022) in 1 l distilled water. The solution was autoclaved at 121 °C for 15 minutes.

Terrific Broth

Terrific broth (TB or TB+) was made by dissolving 47.6 g Sigma-Aldrich Terrific Broth powder (catalogue number T0918) and 8 ml glycerol in 1 l distilled water. The solution was autoclaved at 121 °C for 15 minutes.

A modified Terrific Broth, Gly- (Glu+) TB or TB^{-}, was also used. This was made by dissolving 47.6 g Sigma-Aldrich Terrific Broth powder (catalogue number T0918) and 8 g glucose in 1 l distilled water. The solution was autoclaved at 121 °C for 15 minutes.

Nutrient Broth

Nutrient broth was made by dissolving 8 g Acumedia Nutrient Broth powder (catalogue number 7146A) in 1 l distilled water. The solution was autoclaved at 121 °C for 15 minutes.

4.2 Results

4.2.1 Experimental Approach

A general approach for the experimental verification of the importance of conserved rare codons was conceived:

1. From the set of analysed alignments (see Section 2.5), identify an *Escherichia coli* gene containing conserved rare codons and suitable for experimental analysis (as defined in Section 4.2.2 below).
2. Build a library of variants of this gene, including the wild-type and a set of variants with high-frequency codons substituted into the identified rare regions.
3. Insert the variants into high-expression plasmid vectors, over-express them in *E. coli* and measure the activity and expressions level of each variant.
4. Look for correspondence between the measured activity and the presence or absence of the identified rare codons.

The expectation was that substitution of synonymous common codons for conserved rare codons in the identified regions would have a deleterious effect on protein function, while possibly simultaneously increasing expression levels due to the elimination of translational pauses. The approach has several advantages. Careful design of the library of variants would provide extra information about the aptitude of the algorithm in identifying functionally relevant rare codons and elucidate any combinatorial effects. The wild type is assumed to be evolutionarily optimised for the cellular environment in *E. coli*, so expressing in the same species takes advantage of this. The use of plasmids makes transformation convenient, and over-expression should mean that activity of the synthetic protein is significantly greater than that of any native proteins, providing a healthy dynamic range across which to measure any reduction in activity arising from the substitution of common codons for rare ones. Monitoring the development of activity after induction could provide evidence that the mechanism of change is related to translation rate.

4.2.2 Candidate Selection

Although alignments were generated and analysed for every gene in the *E. coli* K-12 MG1655 genome (see Section 2.5.1), a large number of these are not easily amenable to experimental analysis. The search for candidates was restricted to genes that met the following three criteria: firstly, being of known structure, with at least one corresponding entry in the protein data bank (PDB, Berman et al. (2000)); secondly, having only a single structural domain, determined from the domain entries for the PDB structure in the CATH database (Sillitoe et al., 2013); and thirdly, having an Enzyme Commission (EC) code for functional classification. The first two requirements were intended to facilitate structural analysis, eliminate domain linker regions, and yield a protein that would fold efficiently in the best case. The third gives a rough indication of how readily the activity of the protein assayed. These restrictions left a shortlist of 262 potential candidate genes.

The shortlist was further reduced in a few steps. First and foremost, a gene with a conserved rare codon region showing a low p-value was sought. At this stage in the analysis this included all genes in the shortlist, so genes were ranked in ascending order according to their minimum average codon usage frequency across a window of three codons – this is approximately the test statistic used by the algorithm. The literature was then consulted to assess the feasibility of assaying the highest-ranked enzymes. The equipment and expertise available in the laboratory in which the research took place favoured spectrophotometric assays.

Quantitative measures of the codon usage in alignments for enzymes that passed the stringent constraint of assayability were examined in detail. The p-values for windows in the alignment, the overall average codon usage, and the proportion of aligned sequences at each position were considered. The codon usage in the seed sequence was also closely examined to ensure that the experimental target shared the general pattern of rareness across the alignment. This involved developing a special metric, which is described in Section 4.2.2.

Seed Sequence Codon Frequency Analysis

To check whether the coding in the seed sequence over a particular window could be said to be rarer than expected by chance, the sum of the frequencies observed in the window was compared with the sum of the averages of the synonymous codon sets. Using the notation introduced in Chapter 2 (see Table 2.1 for a quick reference guide), the expectation $\bar{f}_i$ and variance s_i^2 of codon frequency for a single amino acid a_i can be written

$$\bar{f_i} = \sum_{k \in 1:n(a_i)} f_k \times \Pr(Fr(a_i) = f_k)$$

$$s_i^2 = \sum_{k \in 1:n(a_i)} (f_k - \bar{f_i})^2 \times \Pr(Fr(a_i) = f_k)$$

Parameters of a distribution for the sum of the codon frequencies in a window can be derived from the expectations and variances of the composite amino acids:

$$\bar{F} = \sum_{i \in 1:z} \bar{f_i}$$

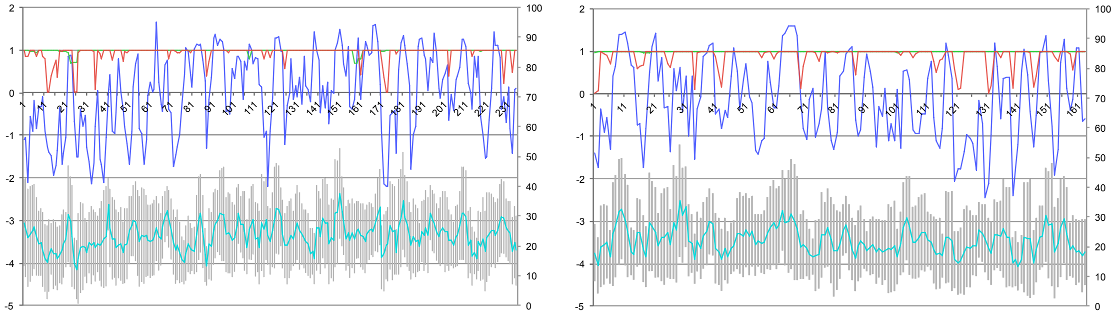
$$S^2 = \sum_{i \in 1:z} s_i^2$$

Dividing the difference between the observed sum F_o and the expectation $\bar{F}$ by the expected standard deviation of the window S yields a z-score, where a negative z-score indicates rarer than expected coding:

$$z = \frac{F_o - \bar{F}}{S}$$

This z-score was used as a relatively simple confirmation that the seed sequence was included in the conservation pattern of the alignment.

This analysis identified two genes which were taken forward for further experimental investigation: *aphA*, encoding a class-B acid phosphatase; and *ubiC*, which produces the chorismate pyruvate-lyase enzyme. Figure 4.1 shows plots of the statis-



aphA *ubiC*

Figure 4.1: Codon usage statistics plotted against residue number in *aphA* and *ubiC*. Regions of interest are those that show a concurrent drop in p-value, average frequency, and seed z-score (see Section 4.2.2), together with high coverage. Details of the regions identified are given in Table 4.1 and 4.2. Regions with low seed z-scores but comparatively high p-values were also selected to compare the efficacy of the alignment model against the single-sequence z-score. The grey bars around the window-average frequencies show the standard deviations of frequencies within the window.

tics that were examined in the process.

4.2.3 Gene Design

Synthetic constructs of the two wild type genes were obtained, with some minor modifications to aid expression. A sequence of six histidine residues (a His-tag), encoded with alternating CAT and CAC codons, was added to the 3' end of the constructs to give the option of protein purification via nickel affinity chromatography. Two successive stop codons, TGA and TAA, were also added to ensure reliable termination under strong promotion. Genes were designed with flanking restriction sites to allow double digests. The intragenic region was scanned for restriction sites and determined to be compatible with numerous restriction enzymes. The sites chosen were for the NdeI enzyme at the 5' end and HinDIII at the 3' end. These were

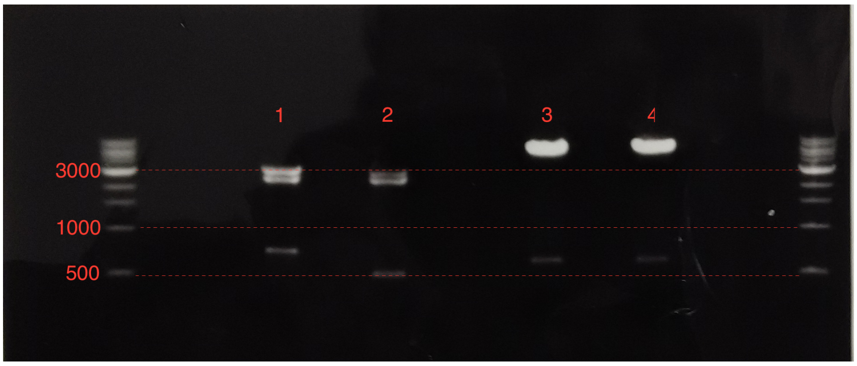


Figure 4.2: The first stage of the experimental work was transferring the gene inserts via restriction and ligation reactions to the pET29-a plasmid for expression. The image shows a UV-exposed agarose gel containing double restriction digests after electrophoresis. Lanes 1 and 2 show the synthetic constructs of wild type *aphA* (711 bp) and *ubiC* (606 bp) being removed from the pJ201 plasmids in which they were delivered. The double bands in the kilobase region in these lanes probably arise from the restriction digest being only partially successful – the samples contain a portion of plasmid still bound to the insert. However, the lower bands containing the inserts are dense enough for extraction and usage. Lanes 3 and 4 contain pET29-a with an unknown insert, present in a sample in the laboratory. This was prepared for insertion of the synthetic *aphA* and *ubiC* cassettes. (Note: image has been cropped)

both present in the polylinker regions of the chosen plasmid, with NdeI upstream.

The synthetic genes were delivered in the pJ201 plasmid. Genes were transferred via restriction and ligation reactions to the pET29-a plasmid for expression in strain BL21(DE3)pLysS. The pJ201 and pET29-a plasmids contain kanamycin resistance markers, so kanamycin was used as the selective agent at 50 µg ml^{-1} when working with them. The pET29-a plasmid contains the *lac* repressor and operator and is inducible by IPTG, and it uses the T7 promoter region, which is targeted by the T7 RNA polymerase present in BL21(DE3) (but not K-12). Figure 4.2 shows an image of a DNA gel electrophoresis procedure used to separate the restricted inserts from the delivery plasmids, and prepare a sample of expression plasmid for re-insertion.

4.2.4 Variant Design

A library of variants was designed to reveal the effects of synonymous mutations to the rare codon regions identified. As a control, variants were also designed to include mutations to regions that were rare in the seed sequence, but where the rareness was not conserved over the alignment.

The designed library of variants was produced from the synthetic wild-type constructs using site-directed mutagenesis. Because of the impact of additional mutations on primer binding energies it is not recommended to perform more than seven nucleotide changes with a single primer, so some regions had to be modified in multiple steps. These regions were split into a core section, including the extremum of the *p*-value signal, and a peripheral section covering flanking rare codons in the seed sequence. This meant that some variants required a series of mutagenesis reactions. Modifications were verified by sequencing after all reactions.

An initial oversight in the design of some primers was identified. Primers were designed to perform the required modifications on the wild-type sequence. In some serial reactions on closely-located regions, later reactions would reverse the modifications made in earlier ones because the primers were designed to bind to the wild type sequence and not the variant. Primers were redesigned to work in series and the reactions repeated and again verified by sequencing.

To reduce the potential for introducing confounding variables, all variants were checked for hybridisation with the *E. coli* anti-Shine-Dalgarno sequence. This is the tail of the 16S rRNA, GAUCACCUCCUUA (5'-3'). Hybridisation energies were checked using the free2bind package (Starmer et al., 2006; `http://www.unc.edu/`

***aphA* variants**

				Variant						
	Index	Residue	Codon change	A	B	C	D	E	F	G
RARE	25	Ala	GCC>GCG		•	•	•			
	26	Ser	TCA>AGC	•	•	•	•			
	27	Ser	TCT>AGC	•	•	•	•			
	28	Pro	CCT>CCG	•	•	•	•			
	29	Ser	TCA>AGC		•	•	•			
	33	Pro	CCT>CCG			•	•	•		
	34	Gly	GGG>GGC			•	•	•		
	35	Thr	ACT>ACC			•	•	•		
	37	Asn	GTT>GTG			•	•	•		
	172	Pro	CCA>CCG				•		•	
	173	Gly	GGG>GGC				•		•	
	174	Gln	CAA>CAG				•		•	
	175	Asn	AAT>AAC				•		•	
	176	Thr	ACA>ACC				•		•	
CONTROL	117	Glu	GAG>GAA							•
	118	Val	GTC>GTG							•
	119	Ala	GCT>GCG							•

Table 4.1: Codon changes in the *aphA* gene variant library

~starmer/free2bind/) and found not to deviate significantly from the wildtype.

Tables 4.1 and 4.2 show the variants that were designed and what modifications they included.

4.2.5 Assay Development

The *aphA* gene produces a protein, AphA, E.C. code 3.1.3.2, with phosphatase activity on a broad range of organic substrates in the pH range 4.0-7.0 (Thaller et al., 1997). A relatively simple assay for phosphatase activity can be constructed

***ubiC* variants**

				Variant									
	Index	Residue	Codon change	A	B	C	D	E	F	G	H	I	J
RARE	34	Ser	TCC>AGC	●	●								
	120	Phe	TTC>TTT				●		●	●			
	121	Thr	ACA>ACC				●		●	●			
	122	Ser	TCA>AGC		●	●	●	●	●	●			
	123	Ser	TCG>AGC				●		●	●			
	124	Thr	ACA>ACC				●		●	●			
	125	Leu	TTA>CTG							●	●		
	127	Arg	CGG>CGC							●	●		
	128	Asp	GAC>GAT					●	●	●	●	●	
	131	Glu	GAG>GAA					●	●	●	●	●	
	132	Ile	ATA>ATT					●	●	●	●	●	
	134	Arg	CGT>CGC					●	●	●	●	●	
CONTROL	54	Gly	GGG>GGC										●
	56	Phe	GTC>GTG										●
	57	Val	GAG>GAA										●

Table 4.2: Codon changes in *ubiC* gene variant library

using para-nitrophenyl phosphate (pNPP) as a substrate. While pNPP is clear in solution, cleavage of the phosphate group yields para-nitrophenol (pNP) which has a strong yellow colour at basic pH that can be measured in the wavelength range 390-420 nm Passariello et al. (2006).

A review of the literature suggested that it may be possible to run a more convenient continuous assay, conducting the reaction in a cuvette and constantly monitoring the change in absorption. AphA requires an acidic pH to function as a phosphatase, but the absorbance of pNP is strongly pH dependent. Absorbance is low at low pH, but above the $\mathrm{pK_a}$ of around 7.2, as the equilibrium of proton dissociation from the hydroxyl group shifts higher, absorbance increases sharply. These two requirements are incompatible, as the low absorbance of pNP at the working pH would have drastically reduced the sensitivity of the assay, so this approach was ruled out after some preliminary work in favour of a stopped assay using a method based on Passariello et al. (2006).

The strategy determined was to use raw cell lysate from cell cultures containing the expressed protein variants. After expression (see Section 4.1.6), cultures would be resuspended in assay buffer (see Section 4.1.9), lysed by sonication, and centrifuged to separate the soluble fraction of protein. This supernatant would contain many cellular proteins in addition to the synthetic variant, but if the use of a high copy number and the powerful T7 promoter should lead to high protein yields. A volume of supernatant would then be mixed with 1 mM pNPP substrate and incubated at 37 °C. Small portions of the supernatant-substrate mix would be extracted at time points and quenched using high-concentration sodium hydroxide. This also serves to lower the pH. The quenched samples could then be measured for absorbance against a background of cell lysate and sodium hyroxide without reaction substrate, to determine the amount of cleavage activity that had occurred in the incubation time.

AphA-mediated phosphate cleavage of pNPP has a k_{cat} of $156\,\mathrm{s}^{-1}$, and a K_{M} of 169 µM. A substrate concentration of 1 mM was selected in order to be well in excess of the K_{M}, which should yield a reasonably constant reaction rate over short time periods. Only a modest yield (around $6\,\mathrm{mg\,ml}^{-1}$) would be required to achieve an easily reaction over a timescale of a few minutes.

A calibration curve (see Figure 4.3) was derived to suit the intended assay conditions by adding 250 µl samples of pNP in assay buffer (see Figure 4.3) at various concentrations to 750 µl 2M NaOH. A check for base-mediated cleavage of phosphate in conditions equivalent to the stopped assay (pH 13.78) revealed no significant change in absorption over 30 minutes.

A series of experiments were conducted to attempt to determine an appropriate timescale for the assay reactions. Colonies of expression strains transformed with plasmids containing wild type *aphA* were cultured, induced to express as described in Section 4.1.6, and assayed as described in Section 4.1.9. Cells containing the same plasmid with no insert were treated in the same way as a control. This revealed a very high level of background activity in the cells containing the empty plasmid, presumably arising from the chromosomal *aphA* and possibly other acid phosphatases (Dassa et al., 1991). In a test of phosphatase activity of the supernatant extracted from cultures containing wild type *aphA* against cultures containing pET29-a with no insert, the supernatant from the cultures with no insert actually showed more activity over ten minutes (see Figure 4.4).

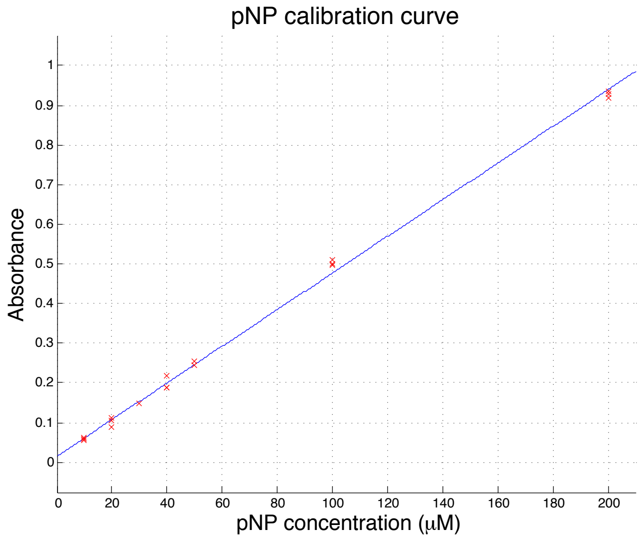


Figure 4.3: Calibration curve for determination of pNP concentration in experimental samples. 250 µl samples of pNP in assay buffer were prepared at the correct concentration, then added to 750 µl 2M NaOH in a 1 ml optical cuvette for absorbance measurements to obtain a sample pH above 13.5, matching what could be expected from the assay protocol and well above the pK_a of pNP. Absorbance was measured at 400 nm, which was found to be the peak response frequency. Measurements were taken at 10, 20, 30, 40, 50, 100 and 200 µM in triplicate. The response is confirmed to be linear in this range. Regression line: slope = 0.0046 μM^{-1} (std. err. = 4.97e-5), intercept = 0.0152 (std. err. = 4.42e-3), mean squared error = 1.97e-4

The background activity would have drastically reduced the dynamic range of the assay. Rossolini et al. (1994) found that acid phosphatase is not expressed in cells using glucose as a carbon source, so repeat tests were conducted using a modified Terrific Broth recipe, TB^-, in which the glycerol was replaced with glucose (see Figure 4.5, Section 4.1.10 for media recipes). This was found to slightly reduce the background activity, but probably not sufficiently to allow sensitive investigation of the effects of genetic variants. Dassa et al. (1991) report that *E. coli* produces acid phosphatase in response to oxygen deprivation. The experiment was repeated in

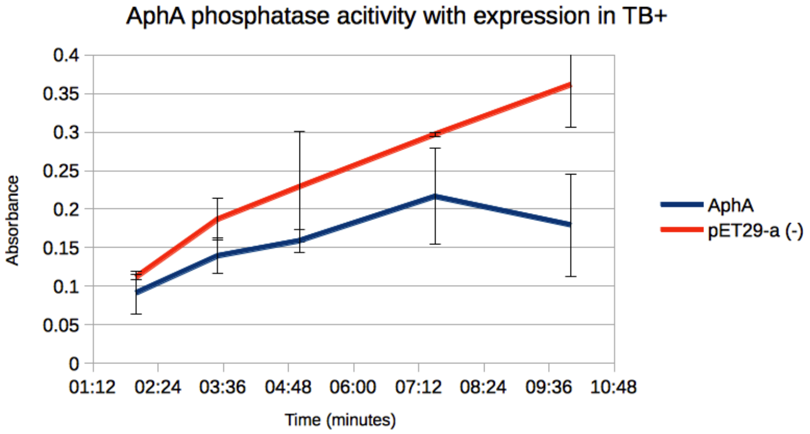


Figure 4.4: Comparison of phosphatase activity in the supernatants extracted from cell cultures with the pET29-a plasmid containing the wild type *aphA* gene or no insert. See Section 4.1.9 for full details of the assay method. 250 µl samples were taken from the reaction at timestops of 2:00, 3:30, 5:00, 7:30 and 10:00 miniutes. Each line shows the average of two replicates, and the error bars indicate the actual recorded values. Although all samples showed significant activity over the recorded time period, the replicates were noisy. More alarmingly, the cultures with no active insert being expressed displayed more phosphatase activity than those containing the *aphA* gene.

smooth and baffled flasks to increase oxygen levels, but no difference was found.

The possibility of using a multifactorial design-of-experiments approach to find growth conditions that resulted in reduced background activity was considered. However, this would have meant deliberately perturbing the environment away from the very conditions in which the enzyme had likely evolved, somewhat undermining the reasoning behind the experiment. Purification of the protein using the His-tag (see Section 4.1.1) prior to assaying was considered. However, the effect we are seeking to measure could well arise from only slight alterations in the structure compared to the native state. The energetic perturbations to the structure resulting from binding to the nickel column, as well as the extra time between expression

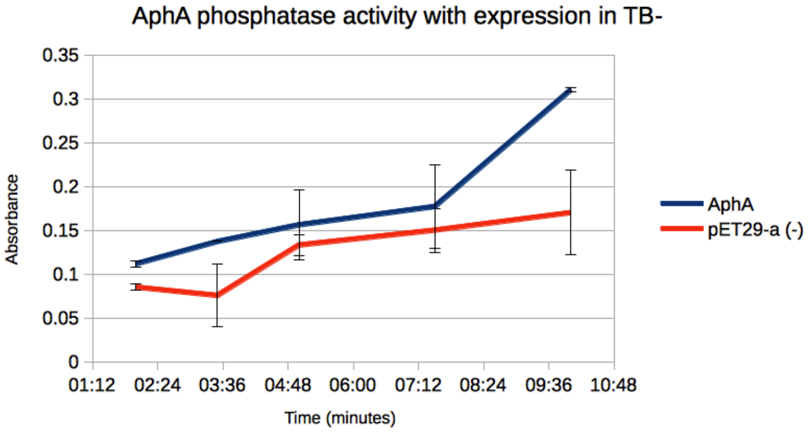


Figure 4.5: Comparison of phosphatase activity in the supernatants extracted from cell cultures with the pET29-a plasmid containing the wild type *aphA* gene or no insert, expressed in modified Terrific Broth TB- containing glucose instead of glycerol. See Section 4.1.9 for full details of the assay method, and Section 4.1.10 for details of the media recipe. 250 µl samples were again taken from the reaction at timestops of 2:00, 3:30, 5:00, 7:30 and 10:00 miniutes. The lines each show the average of two replicates of the two culture types, and the error bars indicate the actual recorded values. Overall activity was slightly reduced compared with expression in TB+, and the cultures containing the insert displayed more activity than those without, but the difference was not considered great enough to give a viable dynamic range.

and assaying, might be enough to reverse the structural alterations arising from the mutations to the coding sequence. Therefore the decision was made to attempt expression in a knockout strain deficient in chromosomal *aphA*.

4.2.6 Keio Knockout Strains

The Keio collection (Baba et al., 2006) consists of a library of 3985 mutant strains of *E. coli* K-12 in which single, non-essential genes have been deleted and replaced with a kanamycin resistance marker. This provides an extremely useful and flexible framework for detailed genetic analysis of the species. By expressing variants of the candidate genes in the relevant Keio strain, it should be possible to eliminate background activity arising from the native, chromosomal copy of the gene under investigation. To this end a culture of the *aphA* Keio strain JW4015-1 was obtained.

All Keio strains are delivered with kanamycin resistance cassettes flanked by FLP recombination target sites in place of the deleted gene. Rather than remove the kanamycin cassette and risk contaminating the culture, it was considered prudent to retain it and switch the variants to a plasmid containing an alternative resistance marker. An alternative plasmid would have been needed in any case because pET29-a is under the control of the T7 phage promoter and polymerase, which is not present in *E. coli* K-12.

The pJ444-01 vector was selected as a suitable substitute, as it contains an ampicillin resistance marker and is uses the T5 promoter, which is recognised by the native *E. coli* RNA polymerase and so can be expressed in most strains. Doubly-selective media containing both ampicillin and kanamycin 50 $\mu g\,ml^{-1}$ was used when working with the pJ444-01 plasmid. The library of variants was transferred to the pJ444-01 vector via restriction-ligation reactions.

Chemically competent *aphA*-deficient Keio cells were prepared according to the protocol described in Section 4.1.2, and transformed with the library of variants in pJ444-01 using the quick transformation method described in Section 4.1.3. The transformed cultures were grown in doubly-selective media containing both kanamycin and ampicillin.

An alternative approach of measuring activity of the *aphA* variants indirectly through organism fitness, as assayed by growth rate, was considered but deemed unlikely to produce results as differences in the growth rates of the Keio strain that is completely deficient in *aphA* against the wild-type are not distinguishable from biological noise (Baba et al., 2006, 2007). Early in the process of expression testing in the keio strains it became clear that time pressures would not allow completion of the experimental work, and unfortunately it had to be abandoned.

4.3 Discussion

The experimental work conducted as part of this thesis did not reach a satisfactory conclusion. A key mistake made in the undertaking of the experimental work presented here was the failure to properly verify the expression of the plasmid constructs before concluding that the experiments were not viable in a standard *E. coli* expression strain and moving onto a Keio variant. However, the results would almost certainly have been improved using the Keio variant as the measurable background activity could be expected to have been eliminated, so this decision was not without base. The work yielded a great deal of experience with broadly applicable molecular biology techniques, and the general approach was sound. The comparison of differences in activity between proteins produced by synonymous genetic variants remains an interesting avenue of investigation in the exploration of the effect of codon usage.

5 Conclusions and Future Work

Codon usage reflects the combination of manifold selective forces acting on the extremely complex and vital process of protein metabolism. The data is available in some form for any fully or partially sequenced genome across all kingdoms of life, and is relatively amenable to analysis. For these reasons codon usage has received a lot of attention over several decades of research. However, it remains an abstraction of the actual biological process of translation. The nature of the impact of changes to codon usage and its precise relationship with protein expression is not only qualitatively uncertain, it may depend upon entirely unknown mechanisms. The broadly accepted trend is that, in general, most genes with a high proportion of common codons are expressed at higher levels. There appear to be cases where the typical preferences of an organism are selectively defied, and reversing this is to the detriment of protein function. These cases are currently indistinguishable from instances of non-preferred coding that are either neutral or slightly deleterious.

The aims of this thesis were three-fold:

1. Develop a novel method for identifying selectively adaptive rare codons with a sound statistical basis
2. Analyse the findings of the method and search for biological correlates as a

form of validation

3. Seek experimental confirmation of the computational method

The first two aims were met. The alignment-based approach offers clear advantages over single-sequence approaches in the form of reduced noise. In combination with the window and gap handling methods developed here, this affords the algorithm the ability to determine a local measure that is robust to alignment quality. Use of the convolution as the test statistic, instead of a figure derived from the geometric mean of frequencies (as used in tAI and RCRR; see Section 1.4), in effect places a more stringent requirement on conservation of rare codon usage. This stringency makes the algorithm more robust to close homologues and cases where a rare coding is only possessed by a set of outliers. The very stringent significance threshold applied to determine rareness produces a very low expected false positive rate, which gives confidence in the findings. For these reasons, the algorithm presented here could be fairly described as being among the best available methods for pure codon usage analysis.

The algorithm was used to generate significance measures of the rareness of coding over a dataset covering the majority of the *Escherichia coli* genome. Analysing these data provided some validation of the algorithm in the form of agreement with the most well-established codon usage patterns relating to the terminal regions of genes, particularly the N-terminus (Tuller et al., 2010a; Clarke and Clark, 2010; Goodman et al., 2013). The focus was then shifted to co-translational and structural aspects. The algorithm identified a novel association with N-terminal methionine cleavage and a striking depletion of rare codons in helical regions. The novel hypothesis proposed here is that this latter finding may be related to a requirement for the maintenance of a steady rate of extrusion of the nascent chain into the cytosol. There is no significant association with the transitions between secondary structural features or with domain boundaries themselves, although there is a slight enrichment of rare codons in non-N-terminal domains and at a biologically relevant offset downstream of domain boundaries. However, there is no evidence for the conservation of rare positions at the level of structural domain families. The algorithm also indicates that ribosomal and translational components are enriched in rare codons, a surprising result that contributes to the debate on the use of such genes as exemplary of optimal codon usage (Wang et al., 2011; Hershberg and Petrov, 2012).

5.1 Limitations and Potential Improvements

The algorithm, while currently state-of-the-art, could certainly be improved. Explicit computation of window distributions in place of Monte-Carlo simulations would enhance the accuracy of significance scores. Moreover, a thorough and systematic exploration of the effect of varying the model parameters would be of interest. Adjusting the window size and significance threshold in combination and monitoring the significance and strength of the correlation with biological signals might help to elucidate optimum values beyond estimates from the literature. However, such approaches remain difficult to verify while there is a dearth of suitable experimental examples. The development of more advanced and accurate methods will become easier as relevant experimental data accumulates.

Another avenue for improvement to the algorithm would be accounting for the shared evolution between similar genomes in the statistical significance measure.

Such an extension to the probability calculations would be challenging, because it would require some measure of the conditional probabilities of codon usage between all combinations of organisms. These probabilities could perhaps be estimated using a sample of representative alignments, but not without the potential for introducing extra biases, and still not without considerable computational expense.

It is unfortunate that the experimental work presented in this thesis did not reach a satisfactory conclusion, especially because lack of experimental evidence is the major obstacle to progress in the field. Even negative results would have been of utility in informing future algorithmic approaches. Experimental investigations in iterative combination with computational work could yet shed light on this enigmatic problem.

Alternative experimental approaches might also be worth exploring. The approach adopted here – over-expression of a modified version of a gene in comparison with the wild-type – is promising and especially relevant for gene design where over-expression is often the goal. However, it has its limitations with respect to biological understanding because over-expression is inherently unnatural. Some possible alternative approaches are suggested in the next section.

5.2 Related Future Work

The focus of this project is the signal of codon usage, which as established is limited in its ability to capture the complexity of translation and distil it into an estimate of elongation rate. Future computational work that focusses on establishing evidence for the hypothesis of elongation rate-mediated cotranslational folding should look to build more sophisticated models of elongation rate. Careful experimental work to deconvolute the relative contributions of various sequence properties would facilitate the development of such models. One avenue is large-scale comparative assessment of the translational efficiency and functionality of synonymous transcripts. Effectively sampling sequence space for peptides of significant length is virtually impossible, because it is so large and the fitness landscape has many local minima, but for small numbers of residues it is feasible (Chevance et al., 2014). Although there are many factors this reductionist approach could bear fruit. The applicability of such reductive results to full-scale proteins should then be tested with two avenues of investigation; measurement of the translational efficiency and functional efficacy of specially designed synthetic constructs; and genomic analysis, to detect natural occurrences against expectation of features identified as critical.

The information thus generated would inform more specialised approaches. Knowledge of the relative and combinatorial contribution to elongation rate of the various sequence properties would allow the construction of species-specific models of elongation rate. To fully leverage this information for the detection of evolutionary signals of cotranslational folding would require a specialised alignment method that attempts to align proteins based on the expected translational time signature rather than on sequence divergence. This would be possible using the same dynamic programming principles applied in traditional sequence alignment, but with a scoring system that aimed to minimise, for example, the variance in predicted elongation time between aligned residues. In proteins whose cotranslational folding depends on elongation rate, this might elucidate the folding pathway by highlighting cases where insertions fulfil the same translational delay that could be provided by other sequence properties.

Experimental protocols to investigate the effect of small changes to the elongation rate on the functionality of specific proteins are challenging to develop. The requirement for ready assayability of protein function greatly constrains the choice of target. Further, in preparing and purifying the sample there is a risk that the protein structure could be perturbed through binding to column media or simply through thermodynamic changes over time, thus obliterating or introducing noise to the very effect that the experiment is seeking to detect. Over-expression negates the need for purification, which mitigates this risk, but as mentioned above, the paradigm is removed from the biological reality of translation. An alternative approach would be to indirectly measure the efficacy of the protein product by monitoring the relative fitness of modified populations against the wild-type, either in isolation or in direct competition. Fitness can be measured using growth rate, or maximal optical density of a liquid culture. Detecting the small impact on fitness that would likely be conferred by changes to single genes could be challenging, since most genes can be deleted altogether without altering maximal optical density over growth periods of about one day (Baba et al., 2006). Growing modified strains in competition would reduce problems arising from extrinsic noise, as would longer experimental time scales. Another strategy would be to deliberately select growth conditions in which the target protein made the biggest contribution to the chosen fitness measure. These conditions could be found *a priori* using a multifactorial experimental design approach. This would allow the experiments to mirror the process of evolution.

This work has focussed entirely on prokaryotes, but the ideal gene design algorithm would be applicable to expression of any protein in any organism. Prokaryotes are often used in bioreactors for their flexibility, robustness, high growth density, and rapid metabolisms, and so are of great interest as target organisms; and biomedical components are often derived from human metabolites. The ability to optimise a eukaryotic gene for expression in prokaryotes would be of great utility, but the differences between the domains pose problems. Translation rates are globally different (about half the speed in eukaryotes), so even if elongation rate is modulated by coding sequence in eukaryotes, and this is detectable, the same protein sequence may need entirely different modulation when expressed in a prokaryotic system, just to follow the same folding pathway. Also, eukaryotic organisms possess many more cofactors and chaperones that may obviate the need for elongation rate modulation in specific proteins. The ability to make predictions of protein folding pathways under the two distinct sets of conditions, and how the elongation rate might influence the dominant pathway, is beyond the scope of current full-scale protein structure predictors and molecular dynamic models, but not beyond the realms of what may be possible in the near future with greater computational power enhancing the resolution and timespan of simulations, not to mention the advent of new techniques.

5.3 Conclusions

The hypothesis that cotranslational folding is modulated by elongation rate is still in its infancy and the body of evidence, although growing, is constrained to a small number of isolated examples. There could well be uncontrolled variables, such as RNA interference (Makarova et al., 2006), interaction with unknown cofactors, tRNA depletion or interaction, ribosome interactions, or something else as-yet undiscovered. It is possible that the selected threshold used in this study was too stringent to allow detection of a cotranslational signal, but it is probably fair to assume that a vital signal would be detectable with very high significance. Therefore, it is safe to conclude on the strength of this research that if the phenomenon of elongation rate-modulated cotranslational folding is important to protein production, is widespread, and is conserved across relatively close homologues, then codon usage is not a sufficiently powerful signal to detect it.

However, there is still much work to be done in the field. Incorporating new signals into the development of algorithms and combining with an increased body of experimental evidence will inform future gene design projects and studies of synonymous variant-linked pathologies, working towards a full understanding of the many layers of information carried in the genetic code.

Bibliography

Adzhubei, A. A., Adzhubei, I. A., Krasheninnikov, I. A., and Neidle, S. (1996). Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS letters*, 399(1-2):78–82.

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., and Marx, C. J. (2013). Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Molecular biology and evolution*, 30(3):549–60.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3):927–35.

Akashi, H. (2003). Translational selection and yeast proteome evolution. *Genetics*, 164(4):1291–303.

Akashi, H. and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3695–700.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2008). *Molecular Biology of the Cell.* Garland Science, 5th edition.

Allert, M., Cox, J. C., and Hellinga, H. W. (2010). Multifactorial determinants of protein expression in prokaryotic open reading frames. *Journal of molecular biology*, 402(5):905–18.

Andersson, S. G. and Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2):198–210.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230.

Angov, E., Hillier, C. J., Kincaid, R. L., and Lyon, J. A. (2008). Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host. *PLOS ONE*, 3(5).

Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and William M Gelbart (2000). *An Introduction to Genetic Analysis*. W. H. Freeman, 7th edition.

Apel, D. and Surette, M. G. (2008). Bringing order to a complex molecular machine: the assembly of the bacterial flagella. *Biochimica et biophysica acta*, 1778(9):1851–8.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., and Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS microbiology reviews*, 29(2):231–62.

Arslan, M. A., Chikina, M., Csermely, P., and Soti, C. (2011). Misfolded proteins inhibit proliferation and promote stress-induced death in SV40-transformed mammalian cells. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 26.

Artsimovitch, I. and Landick, R. (2002). The Transcriptional Regulator RfaH Stimulates RNA Chain Synthesis after Recruitment to Elongation Complexes by the Exposed Nontemplate DNA Strand. *Cell*, 109:193–203.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. a., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2.

Baba, T., Huan, H.-c., Datsenko, K., Wanner, B. L., and Mori, H. (2007). The Applications of Systematic In-Frame, Single-Gene Knockout Mutant Collection of Escherichia coli K-12 Tomoya. In Osterman, A. L. and Gerdes, S. Y., editors, *Microbial Gene Essentiality: Protocols and Bioinformatics*, volume 416, pages 183–194. SpringerProtocols.

Bailey, S. F., Hinz, A., and Kassen, R. (2014). Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature Communications*, 5(6):1–7.

Baker, D. (1998). Metastable states and folding free energy barriers. *Nature Structural Biology*, 5(12):1021–4.

Baskakov, I. V., Legname, G., Prusiner, S. B., and Cohen, F. E. (2001). Folding of prion protein to its native alpha-helical conformation is under kinetic control. *The Journal of Biological Chemistry*, 276(23):19687–90.

Beletskii, A., Grigoriev, A., Joyce, S., and Bhagwat, A. S. (2000). Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *Journal of Molecular Biology*, 300(5):1057–65.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, 36(Database issue):D25–30.

Berg, O. G. and Silva, P. J. (1997). Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Research*, 25(7):1397–404.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–42.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504.

Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S., and Puglisi, J. D. (2004). tRNA selection and kinetic proofreading in translation. *Nature Structural & Molecular Biology*, 11(10):1008–14.

Blumenthal, T. (1998). Gene clusters and polycistronic transcription in eukaryotes. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 20(6):480–7.

Bonekamp, F., Dalbø ge, H., Christensen, T., and Jensen, K. F. (1989). Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli. *Journal of bacteriology*, 171(11):5812–6.

Boycheva, S., Chkodrov, G., and Ivanov, I. (2003). Codon pairs in the genome of Escherichia coli. *Bioinformatics*, 19(8):987–998.

Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure.* Garland Publishing.

Brunak, S. r. and Engelbrecht, J. (1996). Protein Structure and the Sequential Structure of mRNA : α-Helix and β-Sheet Signals at the Nucleotide Level. *PROTEINS: Structure, Function, and Genetics*, 252:237–252.

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C. M., and Stefani, M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880):507–11.

Bulmer, M. (1991). The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics*, 129:897–907.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(421).

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

Chakravartty, V. and Cronan, J. E. (2013). The wing of a winged helix-turn-helix transcription factor organizes the active site of BirA, a bifunctional repressor/ligase. *The Journal of Biological Chemistry*, 288(50):36029–39.

Charneski, C. A. and Hurst, L. D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLOS Biology*, 11(3):e1001508.

Charneski, C. A. and Hurst, L. D. (2014). Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp. *Molecular Biology and Evolution*, 31(1):70–84.

Chartier, M., Gaudreault, F., and Najmanovich, R. (2012). Large scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, pages 1–8.

Chen, D. and Texada, D. E. (2006). Low-usage codons and rare codons of Escherichia coli. *Gene Therapy and Molecular Biology*, 10:1–12.

Chen, J., Petrov, A., Tsai, A., O'Leary, S. E., and Puglisi, J. D. (2013). Coordinated conformational and compositional dynamics drive ribosome translocation. *Nature Structural & Molecular Biology*, 20(6):718–27.

Cheng, L. and Goldman, E. (2001). Absence of effect of varying Thr-Leu codon pairs on protein synthesis in a T7 system. *Biochemistry*, 40(20):6102–6.

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

Chu, D., Barnes, D. J., and von der Haar, T. (2011). The role of tRNA and ribosome competition in coupling the expression of different mRNAs in Saccharomyces cerevisiae. *Nucleic Acids Research*, pages 1–10.

Clarke, T. F. and Clark, P. L. (2008). Rare codons cluster. *PLOS ONE*, 3(10):e3412.

Clarke, T. F. and Clark, P. L. (2010). Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC Genomics*, 11(118).

Cortazzo, P., Cerveñansky, C., Marin, M., Reiss, C., Ehrlich, R., and Deana, A. (2002). Silent mutations affect in vivo protein folding in Escherichia coli. *Biochemical and Biophysical Research Communications*, 293(1):537–541.

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

Dassa, J., Fsihi, H., Marck, C., Dion, M., Kieffer-Bontemps, M., and Boquet, P. L. (1991). A new oxygen-regulated operon in Escherichia coli comprises the genes for a putative third cytochrome oxidase and for pH 2.5 acid phosphatase (appA). *Molecular & General Genetics*, 229(3):341–52.

de Beer, T. a. P., Berka, K., Thornton, J. M., and Laskowski, R. a. (2014). PDBsum additions. *Nucleic acids research*, 42(Database issue):D292–6.

De Sancho, D., Doshi, U., and Muñoz, V. (2009). Protein folding rates and stability: how much is there beyond size? *Journal of the American Chemical Society*, 131(6):2074–5.

Deane, C. M. and Saunders, R. (2011). The imprint of codons on protein structure. *Biotechnology journal*, 6(6):641–9.

Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M., and Yusupova, G. (2012). A new understanding of the decoding principle on the ribosome. *Nature*, pages 1–5.

Di Domenico, T., Walsh, I., Martin, A. J. M., and Tosatto, S. C. E. (2012). MobiDB:

a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, 28(15):2080–1.

Dong, H., Nilsson, L., and Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *Journal of Molecular Biology*, 260(5):649–63.

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

Drummond, D. and Wilke, C. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–352.

Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature*, 287.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, 12(6):640–9.

Duret, L. and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4482–7.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–10.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–7.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science (New York, N.Y.)*, 300(5626):1718–22.

Evans, D. and Leemis, L. (2004). Algorithms for computing the distributions of sums of discrete random variables. *Mathematical and Computer Modelling*, 40(13):1429–1452.

Eyre-Walker, A. (1996). The close proximity of Escherichia coli genes: Consequences for stop codon and synonymous codon use. *Journal of Molecular Evolution*, 42(2):73–78.

Fedorov, A. and Baldwin, T. (1997). Cotranslational protein folding. *Journal of Biological Chemistry*, (2).

Fedyukina, D. V. and Cavagnero, S. (2011). Protein folding at the exit tunnel. *Annual Review of Biophysics*, 40:337–59.

Feng, J., Kessler, D. A., Ben-jacob, E., and Levine, H. (2013). Growth feedback as a basis for persister bistability. *Proceedings of the National Academy of Sciences*, 111(1):544–549.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–22.

Fleishman, S. and Baker, D. (2012). Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell*, 149(2):262–273.

Fluitt, A., Pienaar, E., and Viljoen, H. (2007). Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Computational Biology and Chemistry*, 31(5-6):335–46.

Francino, M. P. and Ochman, H. (2001). Deamination as the Basis of Strand-Asymmetric Evolution in Transcribed Escherichia coli Sequences. *Molecular Biology and Evolution*, 18(6):1147–1150.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., and Meinnel, T. (2006). The proteomics of N-terminal methionine cleavage. *Molecular & Cellular Proteomics*, 5(12):2336–49.

Galtier, N. and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–7.

Ghosh, A., Passaris, I., Tesfazgi Mebrhatu, M., Rocha, S., Vanoirbeek, K., Hofkens, J., and Aertsen, A. (2014). Cellular localization and dynamics of the Mrr type IV restriction endonuclease of Escherichia coli. *Nucleic Acids Research*, 42(6):3908–18.

Gilchrist, M. A. and Wagner, A. (2006). A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology*, 239(4):417–34.

Gilis, D., Massar, S., Cerf, N. J., and Rooman, M. (2001). Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biology*, 2(11).

Gingold, H., Dahan, O., and Pilpel, Y. (2012). Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic acids research*, 40(20):10053–10063.

Gloge, F., Becker, A. H., Kramer, G., and Bukau, B. (2014). Co-translational mechanisms of protein maturation. *Current Opinion in Structural Biology*, 24:24–33.

Goetz, R. M. and Fuglsang, A. (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. *Biochemical and Biophysical Research Communications*, 327(1):4–7.

Goodarzi, H., Nejad, H. A., and Torabi, N. (2004). On the optimality of the genetic code, with the consideration of termination codons. *BioSystems*, 77(1-3):163–73.

Goodman, D. B., Church, G. M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(6157):475–9.

Govindarajan, S. and Goldstein, R. A. (1998). On the thermodynamic hypothesis of protein folding. *Proceedings of the National Academy of Sciences*, 95(10):5545–9.

Greenbaum, B. D., Cocco, S., Levine, A. J., and Monasson, R. (2014). Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proceedings of the National Academy of Sciences*, 111(13):5054–9.

Grosjean, H., de Crécy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Letters*, 584(2):252–64.

Gu, W., Li, M., Xu, Y., Wang, T., Ko, J.-H., and Zhou, T. (2014). The impact of RNA structure on coding sequence evolution in both bacteria and eukaryotes. *BMC evolutionary biology*, 14(1):87.

Gu, W., Zhou, T., Ma, J., Sun, X., and Lu, Z. (2003). Folding type specific secondary structure propensities of synonymous codons. *IEEE Transactions on Nanobioscience*, 2(3):150–157.

Gupta, S. K., Majumdar, S., Bhattacharya, T. K., and Ghosh, T. C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochemical and Biophysical Research Communications*, 269(3):692–6.

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology*, 22(7):346–53.

Gutman, G. a. and Hatfield, G. W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 86(10):3699–703.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–30.

Han, X., Dorsey-Oresto, A., Malik, M., Wang, J.-Y., Drlica, K., Zhao, X., and Lu,

T. (2010). Escherichia coli genes that reduce the lethal effects of stress. *BMC Microbiology*, 10(35).

Harr, B., Todorova, J., and Schlötterer, C. (2002). Mismatch repair-driven mutational bias in D. melanogaster. *Molecular Cell*, 10:199–205.

Harrison, R. J. and Charlesworth, B. (2011). Biased gene conversion affects patterns of codon usage and amino acid usage in the Saccharomyces sensu stricto group of yeasts. *Molecular Biology and Evolution*, 28(1):117–29.

Hartl, D. L. and Jones, E. W. (1998). *Genetics: Principles and Analysis*. Jones and Bartlett Publishers, 4th edition.

Hendriks, G., Calléja, F., Besaratinia, A., Vrieling, H., Pfeifer, G. P., Mullenders, L. H. F., Jansen, J. G., and de Wind, N. (2010). Transcription-dependent cytosine deamination is a novel mechanism in ultraviolet light-induced mutagenesis. *Current Biology*, 20(2):170–5.

Hershberg, R. and Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, 42(iv):287–99.

Hershberg, R. and Petrov, D. A. (2009). General rules for optimal codon choice. *PLOS Genetics*, 5(7).

Hershberg, R. and Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLOS Genetics*, 6(9).

Hershberg, R. and Petrov, D. A. (2012). On the limitations of using ribosomal genes as references for the study of codon usage: a rebuttal. *PLOS ONE*, 7(12).

Higgs, P. G. and Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular Biology and Evolution*, 25(11):2279–91.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLOS Genetics*, 6(9).

Hill, C., Sandt, C., and Vlazny, D. (1994). Rhs elements of Escherichia coli: a family of genetic composites each encoding a large mosaic protein. *Molecular Microbiology*, 12(6):865–871.

Hosack, D. A., Jr, G. D., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10).

Huang, C.-J., Lin, H., and Yang, X. (2012). Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. *Journal of Industrial Microbiology & Biotechnology*.

Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. a. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

Huard, F. P. E., Deane, C. M., and Wood, G. R. (2006). Modelling sequential protein folding under kinetic control. *Bioinformatics (Oxford, England)*, 22(14):e203–10.

Hutchinson, E. and Thornton, J. (1996). PROMOTIF - A program to identify and analyze structural motifs in proteins. *Protein Science*, 5:212–220.

Ikemura, T. (1981a). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, 146(3):1–21.

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

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. *Journal of Molecular Biology*, 158(4):573–97.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2(1):13–34.

Irwin, B., Heck, J., and Hatfield, G. (1995). Codon Pair Utilization Biases Influence Translational Elongation Step Times. *Journal of Biological Chemistry*, 270(39):22801–06.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., Ho, P. Y., Kakazu, Y., Sugawara, K., Igarashi, S., Harada, S., Masuda, T., Sugiyama, N., Togashi, T., Hasegawa, M., Takai, Y., Yugi, K., Arakawa, K., Iwata, N., Toya, Y., Nakayama, Y., Nishioka, T., Shimizu, K., Mori, H., and Tomita, M. (2007). Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science*, 316(5824):593–7.

Itzkovitz, S. and Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, pages 405–412.

Jackson, A. P., Thomas, G. H., Parkhill, J., and Thomson, N. R. (2009). Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. *BMC Genomics*, 10:584.

Jana, S. and Deb, J. K. (2005). Strategies for efficient production of heterologous proteins in Escherichia coli. *Applied Microbiology and Biotechnology*, 67(3):289–98.

Johansson, M., Bouakaz, E., Lovmar, M., and Ehrenberg, M. (2008). The kinetics of ribosomal peptidyl transfer revisited. *Molecular cell*, 30(5):589–98.

Jones, M., Wagner, R., and Radman, M. (1987). Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. *Genetics*, 115:605–610.

Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1):143–55.

Khade, P., Shi, X., and Joseph, S. (2013). Steric complementarity in the decoding center is important for tRNA selection by the ribosome. *Journal of molecular biology*, 425(20):3778–3789.

Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., and Hartl, F. U. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annual Review of Biochemistry*, 82:323–55.

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(5811):525–8.

Klumpp, S., Dong, J., and Hwa, T. (2012). On ribosome load, codon bias and protein abundance. *PLOS ONE*, 7(11).

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

Kramer, G., Boehringer, D., Ban, N., and Bukau, B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature Structural & Molecular Biology*, 16(6):589–97.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science*, 324:255–258.

Kunkel, T. A. and Erie, D. A. (2005). DNA mismatch repair. *Annual Review of Biochemistry*, 74:681–710.

Larson, M. H., Mooney, R. a., Peters, J. M., Windgassen, T., Nayak, D., Gross, C. a., Block, S. M., Greenleaf, W. J., Landick, R., and Weissman, J. S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science*, 344(6187):1042–7.

Lavner, Y. and Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, 345(1):127–38.

Lawrence, J. G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology*, 2(5):519–23.

Le Roch, K. G., Johnson, J. R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S. F., Williamson, K. C., Holder, A. a., Carucci, D. J., Yates, J. R., and Winzeler, E. a. (2004). Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Research*, 14(11):2308–18.

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., and Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research*, 40:D465–71.

Levinthal, C. (1968). Are there pathways for protein folding. *Journal de Chimie Physique*, 65(1):44.

Li, G.-W., Oh, E., and Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484:538–41.

Liao, Y., Jeng, J., and Wang, C. (2004). Removal of N-terminal methionine from recombinant proteins by engineered E. coli methionine aminopeptidase. *Protein Science*, 13:1802–1810.

Liljenström, H. and von Heijne, G. (1987). Translation rate modification by preferential codon usage: intragenic position effects. *Journal of Theoretical Biology*, 124(1):43–55.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362:709–715.

Lu, A. L. and Chang, D. Y. (1988). Repair of single base-pair transversion mismatches of Escherichia coli in vitro: correction of certain A/G mismatches is independent of dam methylation and host mutHLS gene functions. *Genetics*, 118(4):593–600.

Lu, J. and Deutsch, C. (2008). Electrostatics in the ribosomal tunnel modulate chain elongation rates. *Journal of Molecular Biology*, 384(1):73–86.

Makarova, K. S., Grishin, N. V., Shabalina, S. a., Wolf, Y. I., and Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*, 1(7).

Makhoul, C. H. and Trifonov, E. N. (2002). Distribution of rare triplets along mRNA and their relation to protein folding. *Journal of Biomolecular Structure & Dynamics*, 20(3):413–20.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, 19(6):330–338.

Martínez-Núñez, M. a., Pérez-Rueda, E., Gutiérrez-Ríos, R. M., and Merino, E.

(2010). New insights into the regulatory networks of paralogous genes in bacteria. *Microbiology*, 156:14–22.

McVean, G. A. T. and Hurst, G. D. D. (2000). Evolutionary lability of context-dependent codon bias in bacteria. *Journal of Molecular Evolution*, 50:264–275.

Meyerovich, M., Mamou, G., and Ben-Yehuda, S. (2010). Visualizing high error levels during gene expression in living bacterial cells. *Proceedings of the National Academy of Sciences*, 107(25):11543–8.

Michel, A. M. and Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdisciplinary Reviews. RNA*, 4(5):473–90.

Modrich, P. and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annual Review of Biochemistry*, 65:101–33.

Nakamura, Y., Gojobori, T., and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1):292.

Nakken, S., Rø dland, E. a., and Hovig, E. (2010). Impact of DNA physical properties on local sequence bias of human mutation. *Human Mutation*, 31(12):1316–25.

Navon, S. and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biology*, 12(R12).

Nei, M. and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97(1):145–63.

Nembaware, V., Crum, K., Kelso, J., and Seoighe, C. (2002). Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Research*, pages 1370–6.

Nicola, A. V., Chen, W., and Helenius, A. (1999). Co-translational folding of an alphavirus capsid protein in the cytosol of living cells. *Nature Cell Biology*, 1(6):341–5.

Novoa, E. M., Pavon-Eternod, M., Pan, T., and Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell*, 149(1):202–13.

O'Brien, E. P., Vendruscolo, M., and Dobson, C. M. (2012). Prediction of variable translation rate effects on cotranslational protein folding. *Nature Communications*, 3(868).

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. *Nature Communications*, 5(2988).

Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. a., Kramer, G., Weissman, J. S., and Bukau, B. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, 147(6):1295–308.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246:96–98.

Oresic, M. and Shalloway, D. (1998). Specific Correlations between Relative Synonymous Codon Usage and Protein Secondary Structure. *Journal of Molecular Biology*, 281:31–48.

Pan, T., Artsimovitch, I., Fang, X. W., Landick, R., and Sosnick, T. R. (1999). Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proceedings of the National Academy of Sciences*, 96(17):9545–50.

Parmley, J. L. and Huynen, M. a. (2009). Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLOS Genetics*, 5(7).

Parry, B. R., Surovtsev, I. V., Cabeen, M. T., O'Hern, C. S., Dufresne, E. R., and Jacobs-Wagner, C. (2014). The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. *Cell*, 156(1-2):183–94.

Passariello, C., Forleo, C., Micheli, V., Schippa, S., Leone, R., Mangani, S., Thaller, M. C., and Rossolini, G. M. (2006). Biochemical characterization of the class B acid phosphatase (AphA) of Escherichia coli MG1655. *Biochimica et Biophysica Acta*, 1764:13–19.

Pechmann, S. and Frydman, J. (2012). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology*, 20(2):237–243.

Pedersen, S. (1984). Escherichia coli ribosomes translate in vivo with variable rate. *The EMBO journal*, 3(12):2895–8.

Pellegrini-Calace, M. and Thornton, J. M. (2005). Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. *Nucleic Acids Research*, 33(7):2129–40.

Percudani, R., Pavesi, A., and Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *Journal of Molecular Biology*, 268(2):322–30.

Plaxco, K. W., Simons, K. T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology*, 277(4):985–94.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42.

Pope, B. and Kent, H. M. (1996). High efficiency 5 min transformation of Escherichia coli. *Nucleic Acids Research*, 24(3):536–7.

Power, P. M., Jones, R. a., Beacham, I. R., Bucholtz, C., and Jennings, M. P. (2004). Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of Escherichia coli. *Biochemical and Biophysical Research Communications*, 322(3):1038–44.

Price, W. N., Handelman, S. K., Everett, J. K., Tong, S. N., Bracic, A., Luff, J. D., Naumov, V., Acton, T., Manor, P., Xiao, R., Rost, B., Montelione, G. T., and Hunt, J. F. (2011). Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli. *Microbial Informatics and Experimentation*, 1(6).

Puigbò, P., Guzmán, E., Romeu, A., and Garcia-Vallvé, S. (2007). OPTIMIZER:

a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Research*, 35:W126–31.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, 40:D290–301.

Purvis, I. J., Bettany, A. J. E., Santiago, T. C., Coggins, J. R., Duncan, K., Eason, R., and Brown, A. J. P. (1987). The Efficiency of Folding of Some Proteins is Increased by Controlled Rates of Translation in Vivo: A Hypothesis. *Journal of Molecular Biology*, 193:413–417.

Putzer, H. and Laalami, S. (2003). Regulation of the expression of aminoacyl-tRNA synthetases and translation factors. In *Translation Mechanisms*, chapter 24, pages 388–415. Landes Bioscience.

Qu, X., Wen, J.-D., Lancaster, L., Noller, H. F., Bustamante, C., and Tinoco, I. (2011). The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature*, 475(7354):118–21.

Ran, W. and Higgs, P. G. (2010). The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Molecular Biology and Evolution*, 27(9):2129–40.

Religa, T. L., Johnson, C. M., Vu, D. M., Brewer, S. H., Dyer, R. B., and Fersht, A. R. (2007). The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. *Proceedings of the National Academy of Sciences*, 104(22):9272–7.

Rocha, E. P. C. (2004). Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 14:2279–86.

Rogozin, I. B., Makarova, K. S., Natale, D. a., Spiridonov, A. N., Tatusov, R. L., Wolf, Y. I., Yin, J., and Koonin, E. V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Research*, 30(19):4264–71.

Rossolini, G. M., Thaller, M. C., Pezzi, R., and Satta, G. (1994). Identification of an Escherichia coli periplasmic acid phosphatase containing of a 27 kDa-polypeptide component. *FEMS microbiology letters*, 118(1-2):167–73.

Russell, J. and Cook, G. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological Reviews*, 59(1):48–62.

Sabi, R. and Tuller, T. (2014). Modelling the Efficiency of Codon-tRNA Interactions Based on Codon Usage Bias. *DNA Research*, pages 1–15.

Sander, I. M., Chaney, J. L., and Clark, P. L. (2014). Expanding Anfinsen's Principle: Contributions of Synonymous Codon Selection to Rational Protein Design. *Journal of the American Chemical Society*.

Saunders, R. and Deane, C. M. (2010). Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Research*.

Schmid, P. and Flegel, W. a. (2011). Codon usage in vertebrates is associated with a low risk of acquiring nonsense mutations. *Journal of Translational Medicine*, 9(1):87.

Seligmann, H. (2012). Positive and Negative Cognate Amino Acid Bias Affects Compositions of Aminoacyl-tRNA Synthetases and Reflects Functional Constraints on Protein Structure. *BIO*, 2:11–26.

Shah, P. and Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences*, 108(25):10231–6.

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 Research*, 33(4):1141–53.

Sharp, P. M., Emery, L. R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London B, Biological sciences*, 365(1544):1203–12.

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 Research*, 15(3):1281–1295.

Sharp, P. M., Tuohy, T. M. F., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis dearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, 14(13):5125–5143.

Siller, E., DeZwaan, D. C., Anderson, J. F., Freeman, B. C., and Barral, J. M. (2010). Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *Journal of Molecular Biology*, 396(5):1310–8.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees,

J. G., Lewis, T. E., Studer, R. a., Rentzsch, R., Yeats, C., Thornton, J. M., and Orengo, C. a. (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic acids research*, 41(Database issue):D490–8.

Skewes, A. D. and Welch, R. D. (2013). A Markovian analysis of bacterial genome sequence constraints. *PeerJ*, 1:e127.

Sohl, J., Jaswal, S., and Agard, D. (1998). Unfolded Conformations of α-lytic Protease are More Stable than its Native State. *Nature*, 395(October):817–819.

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. *Journal of Molecular Biology*, 222(2):265–80.

Spencer, P. S., Siller, E., Anderson, J. F., and Barral, J. M. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*, 422:328–335.

Stadler, M. and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, 17(12):2063–2073.

Stanger, H. E., Syud, F. a., Espinosa, J. F., Giriat, I., Muir, T., and Gellman, S. H. (2001). Length-dependent stability and strand length limits in antiparallel beta-sheet secondary structure. *Proceedings of the National Academy of Sciences*, 98(21):12015–20.

Starmer, J., Stomp, A., Vouk, M., and Bitzer, D. (2006). Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLOS Computational Biology*, 2(5):e57.

Stoletzki, N. and Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular Biology and Evolution*, 24(2):374–81.

Sun, D., Zhang, X., Wang, L., Prudhomme, M., Xie, Z., Martin, B., and Claverys, J.-P. (2009). Transforming DNA uptake gene orthologs do not mediate spontaneous plasmid transformation in Escherichia coli. *Journal of Bacteriology*, 191(3):713–9.

Suzuki, H., Brown, C. J., Forney, L. J., and Top, E. M. (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Research*, 15(6):357–65.

Tao, X. and Dafu, D. (1998). The relationship between synonymous codon usage and protein structure. *FEBS Letters*, 434:93–96.

Tarry, M., Arends, S. J. R., Roversi, P., Piette, E., Sargent, F., Berks, B. C., Weiss, D. S., and Lea, S. M. (2009). The Escherichia coli cell division protein and model Tat substrate SufI (FtsP) localizes to the septal ring and has a multicopper oxidase-like structure. *Journal of Molecular Biology*, 386(2):504–19.

Thaller, M. C., Schippa, S., Bonci, A., Cresti, S., and Rossolini, G. M. (1997). Identification of the gene (aphA) encoding the class B acid phosphatase/phosphotransferase of Escherichia coli MG1655 and characterization of its product. *FEMS Microbiology Letters*, 146:191–198.

Thanaraj, T. A. and Argos, P. (1996a). Protein secondary structural types are differentially coded on messenger RNA. *Protein science*, 5(10):1973–83.

Thanaraj, T. A. and Argos, P. (1996b). Ribosome-mediated translational pause and protein domain organization. *Protein Science*, 5:1594–612.

Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–22.

Trotta, E. (2011). The 3-base periodicity and codon usage of coding sequences are correlated with gene expression at the level of transcription elongation. *PLOS ONE*, 6(6):e21590.

Tsai, C.-J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., and Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of Molecular Biology*, 383(2):281–91.

Tsukuda, M. and Miyazaki, K. (2013). Directed evolution study unveiling key sequence factors that affect translation efficiency in Escherichia coli. *Journal of Bioscience and Bioengineering*, 116(5):540–5.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J. M., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S. cerevisiae. *PLOS Computational Biology*, 3(12):e248.

Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences*, 107(8):3645–50.

Uemura, S., Aitken, C. E., Korlach, J., Flusberg, B. a., Turner, S. W., and Puglisi, J. D. (2010). Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature*, 464(7291):1012–7.

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

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41:D483–9.

Voigts-Hoffmann, F., Klinge, S., and Ban, N. (2012). Structural insights into eukaryotic ribosomes and the initiation of translation. *Current Opinion in Structural Biology*, 22(6):768–77.

Wang, B., Shao, Z.-Q., Xu, Y., Liu, J., Liu, Y., Hang, Y.-Y., and Chen, J.-Q. (2011). Optimal Codon Identities in Bacteria: Implications from the Conflicting Results of Two Different Methods. *PLOS ONE*, 6(7):e22714.

Weixlbaumer, A., Leon, K., Landick, R., and Darst, S. (2013). Structural basis of transcriptional pausing in bacteria. *Cell*, 152(3):431–441.

Welch, M., Villalobos, A., Gustafsson, C., and Minshull, J. (2009). You're one in a googol: optimizing genes for protein expression. *Journal of the Royal Society Interface*, 6:467–476.

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(7187):598–603.

Widmann, M., Clairo, M., Dippon, J., and Pleiss, J. (2008). Analysis of the distribution of functionally relevant rare codons. *BMC Genomics*, 9:207.

Wilson, D. N. and Beckmann, R. (2011). The ribosomal tunnel as a functional environment for nascent polypeptide folding and translational stalling. *Current Opinion in Structural Biology*, 21(2):274–82.

Wolffe, A. P. and Matzke, M. A. (1999). Epigenetics: regulation through repression. *Science*, 286(5439):481–6.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87:23–29.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–74.

Yamamoto, K., Kusano, K., Takahashi, N. K., Yoshikura, H., and Kobayashi, I. (1992). Gene conversion in the Escherichia coli RecF pathway: a successive half crossing-over model. *Molecular & General Genetics*, 234(1):1–13.

Yokoyama, S. and Nishimura, S. (1995). Modified Nucleosides and Codon Recognition. In *tRNA: Structure, Biosynthesis and Function*, chapter 12, pages 207–223. ASM Press.

Zaborske, J. M., Narasimhan, J., Jiang, L., Wek, S. A., Dittmar, K. A., Freimoser, F., Pan, T., and Wek, R. C. (2009). Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. *The Journal of Biological Chemistry*, 284(37):25254–67.

Zalucki, Y. M., Beacham, I. R., and Jennings, M. P. (2009). Biased codon usage in signal peptides: a role in protein export. *Trends in Microbiology*, 17(4):146–50.

Zhang, D., de Souza, R., and Anantharaman, V. (2012). Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative. *Biology Direct*, 7(18).

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology*, 16(3):274–80.

Zhang, G. and Ignatova, Z. (2009). Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLOS ONE*, 4(4):e5036.

Zouridis, H. and Hatzimanikatis, V. (2008). Effects of codon distributions and tRNA competition on protein translation. *Biophysical Journal*, 95(3):1018–33.