The Dissertation Committee for Efrain Siller certifies that this is the approved version of the following dissertation:

STUDIES ON TRANSLATION SPEED AND PROTEIN FOLDING EFFICIENCY IN BACTERIA AND EUKARYOTES

Committee:

José M. Barral, M.D., Ph.D., Supervisor

Darren F. Boehning, Ph.D., Chair

Hiram F. Gilbert, Ph.D.

Robert O. Fox, Ph.D.

Andres F. Oberhauser, Ph.D.

Henry F. Epstein, M.D.

Dean, Graduate School

STUDIES ON TRANSLATION SPEED AND PROTEIN FOLDING EFFICIENCY IN BACTERIA AND EUKARYOTES

by

Efrain Siller, MD

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas Medical Branch
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch
August 2011

Dedication

To my wife, Randi.

Acknowledgements

First and foremost I want to thank Dr. José Barral, he took a big risk when he accepted me in his lab. I came to his lab with no prior research experience. I am forever grateful to him. I owe him my research skills and scientific mind. Additionally, I would like to acknowledge Paige Spencer and Jason Chandler for their assistance through my PhD studies. I especially want to thank, Dr. John Anderson, my colleague and friend, he was a driving force behind this project and an extraordinary friend.

My committee members, Dr. Henry F. Epstein, Dr. Andres F. Oberhauser, Dr. Hiram F. Gilbert, Dr. Robert O. Fox and Dr. Darren F. Boehning provided invaluable guidance throughout this work for which I am very grateful. I would like to especially thank Dr. Darren F. Boehning for being a second mentor to me. He was deeply involved in the details of my project from the beginning of my education at UTMB.

Financial support for this study was supplied by the Welsh Foundation.

STUDIES ON TRANSLATION SPEED AND PROTEIN FOLDING EFFICIENCY IN BACTERIA AND EUKARYOTES

Publication No.____________

Efrain Siller, M.D., Ph.D.

The University of Texas Medical Branch, 2011

Supervisor: José M. Barral

The mechanisms for *de novo* protein folding differ significantly between bacteria and eukaryotes, as evidenced by the often observed poor yields of native eukaryotic proteins upon recombinant production in bacterial systems. Polypeptide synthesis rates are faster in bacteria than in eukaryotes, but the effects of general variations in translation rates on protein folding efficiency have remained largely unexplored. By employing *E. coli* cells with mutant ribosomes whose translation speed can be modulated. In this work, it is showed that reducing polypeptide elongation rates leads to enhanced folding of diverse proteins of eukaryotic origin. These results suggest that in eukaryotes, protein folding necessitates slow translation rates.

The degeneracy of the genetic code allows most amino acids to be encoded by multiple codons. The distribution of these so-called synonymous codons among protein coding sequences is not entirely random and multiple theories have arisen to explain the biological significance of such non-uniform codon selection. Most ideas revolve the notion that certain codons allow for faster or more efficient translation, whereas the presence of others result in slower translation rates. The presence of these different types of codons along a message is postulated in turn to confer variable rates of emergence of the nascent polypeptide from the ribosome, which may influence its capacity to fold towards the native state, among other properties. A metric to predict organism-specific polypeptide elongation rates of any mRNA based on whether each codon is decoded by tRNAs capable of Watson-Crick, non-Watson-Crick or both types of interactions was developed. By pulse-chase analyses in living *E. coli* cells it was demonstrated that sequence engineering based on these concepts predictably modulate translation rates due to changes in polypeptide elongation and show that such alterations significantly impact the folding of proteins of eukaryotic origin. Finally, this work shows that sequence harmonization based on expression host tRNA pools designed to mimic ribosome movement of the original organism can significantly increase the folding of the encoded polypeptide.

Table of Contents

List of Figures

List of tables

Chapter 1. Introduction

Proteins are the macromolecules primarily responsible for expressing genetic information.

Proteins participate in virtually all living processes.

Proteins are involved in essentially everything that symbolizes life. Throughout evolution, the cell has harnessed the ability of protein molecules to acquire extremely diverse shapes to perform countless biochemical pathways and to serve as versatile structural scaffolds. Under most conditions where life has arisen, most chemical reactions would occur too slowly if it were not for enzymes, and the vast majority of enzyme-catalyzed reactions are performed by proteins. Proteins can relay information within cells, where complex signaling cascades depend on their abilities to change shapes and form specific complexes. In multi-cellular organisms, entirely new functions have arisen that critically depend on proteins. For instance, proteins can relay information among cells,

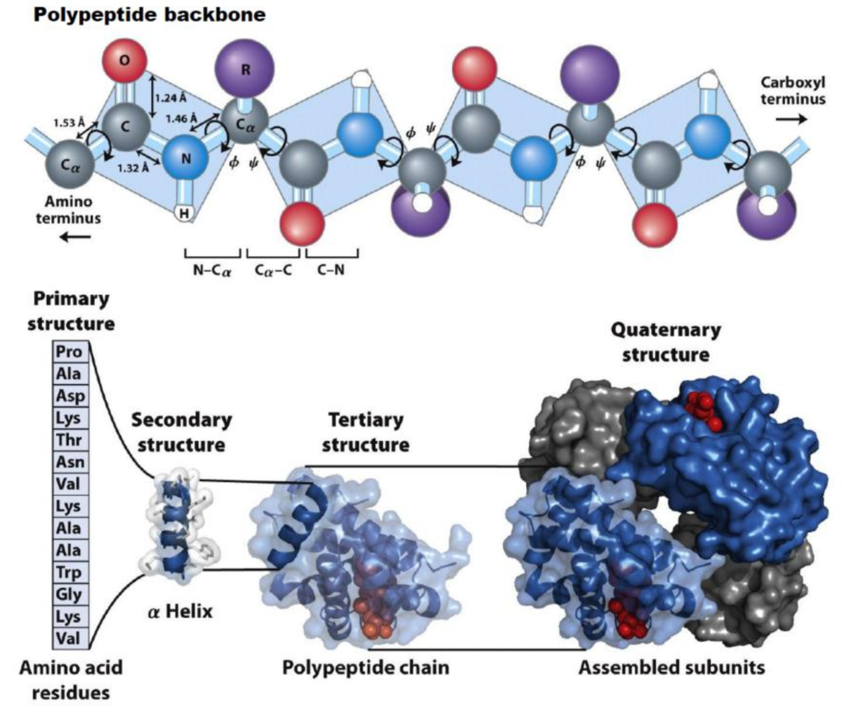


Figure 1. Protein structure levels.
Permission pending. Lehninger Principles of Biochemistry. 5th Ed. 2009

as is the case with hormones, where they can pass a signal between cells that might not be close to each other. Proteins are the primary components of most tissues, such as muscle, which is mainly composed of actin and myosin, or they can provide structural support, as is the case with bones, whose organic matrix is composed of fibrillar proteins, including collagen. Furthermore, proteins play very important roles in transport of oxygen to tissues by hemoglobin, and in the immune system where antibodies can recognize and destroy foreign entities such as bacteria and viruses.

Protein molecules are polymers of 20 amino acids.

Although proteins can carry out innumerable functions, in most cases they are constituted by only 20 amino acids, each with particular chemical characteristics, conferred to by their different side chains (Fig. 1). Proteins are generally linear polymers composed by different combinations of those 20 amino acids, bound together by covalent bonds of the amide type, referred to as *peptide bonds*. This linear arrangement of amino acid residues along a protein chain is known as its *primary structure*. The backbone of the linear polypeptide consists of three main atoms *per* amino acid: a nitrogen atom forming an amide (N), the α carbon (Cα) and a carbonyl (C) (Fig. 2). The spatial relations among these atoms along a polypeptide can be described by their relative rotational (or dihedral) angles, referred to as φ, ψ and (Fig. 1). Hydrogen bonding interactions between the amide hydrogen and the carbonyl oxygen along the backbone can result in acquisition of relatively regular local conformations, known as

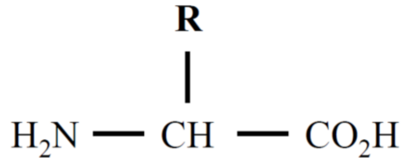


Figure 2. Chemical structure of an amino acid. R, side chain; H_2N, amine; CO_2H, carboxyl

secondary structure, which include the α helix and the β strand, among others (Fig. 1). More extensive interactions among the side chains and the backbone, which involve hydrogen bonds, van der Waals forces, hydrophobic effects and salt bridges allow polypeptides to acquire and maintain specific three-dimensional conformations, known as their *tertiary structure* (Fig. 1). When the functional form of a protein consists of more than one independently synthesized polypeptide chain, it is said to have acquired *quaternary structure* (Fig. 1).

How proteins are synthesized: information transfer and genetic code degeneracy.

How are these linear polymers of amino acid residues put together in the cell? Protein synthesis is a template-mediated process. In essence, the amino acid sequence of a protein is determined by the sequence of nucleotides present in a particular segment of nucleic acids (traditionally known as a *gene*), which constitute the inheritable (or genetic) material of the cell. In most organisms, the genetic material exists in the form of DNA (retroviruses are an exception, since their genetic material is composed of RNA). In order to direct protein synthesis, the genetic information in DNA needs to be *transcribed* into *messenger RNA* (mRNA). *Transcription* is a highly regulated process that allows the cell to express certain regions of the genetic material at specific times and in order to respond to particular conditions. It is carried out by protein complexes that bind to regulatory regions along DNA and may include numerous steps (such as splicing, addition of non-templated nucleotides at both ends, *etc.*), depending on the organism. Since the main objectives of this study involve processes that occur once an mRNA molecule has been synthesized, transcription will not be discussed in greater detail.

Ribosomes are the machinery responsible for *translating* the nucleotide sequence present mRNA into polypeptide sequence. Thus, they have two crucial functions: (1) to decode the information present in mRNA and, based on this, (2) catalyze the formation of the peptide bond between the correct amino acids. All ribosomes contain two main subunits, termed *small* and *large* (Fig. 3) and both subunits are composed of *ribosomal RNA* (rRNA) and protein components.

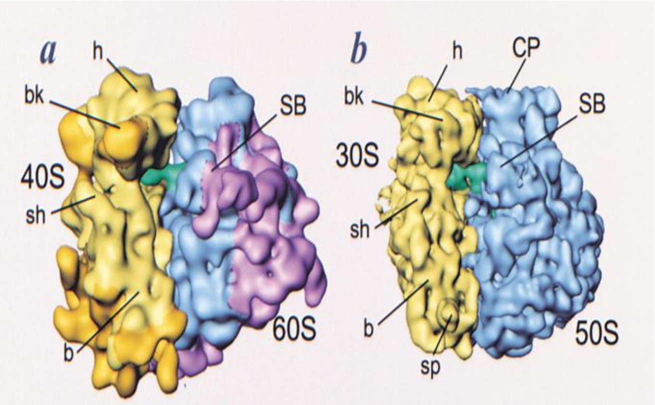
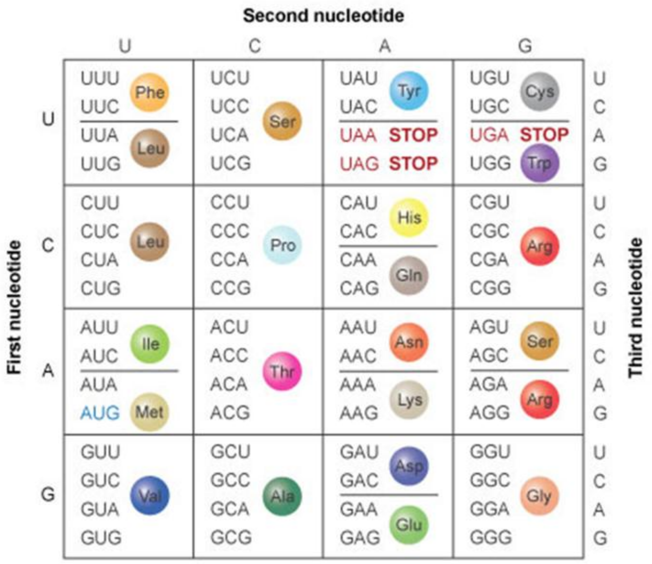


Figure 3. Comparison of the 80S Ribosome from S. cerevisiae (a) with the E. coli 70S Ribosome (b). The small subunits are in yellow, the large subunits in blue, and the P site-bound tRNA in green.b, body; bk, beak; h, head; sh, shoulder; sp, spur; CP, central protuberance; SB, stalk base.

Reprinted from Cell, 107, 3, Christian M.T. Spahn, Roland Beckmann, Narayanan Eswar, Pawel A. Penczek, Andrej Sali, Günter Blobel and Joachim Frank, Structure of the 80S Ribosome from Saccharomyces cerevisiae—tRNA-Ribosome and Subunit-Subunit Interactions, Pages No. 373-386, Copyright (2001), with permission from Elsevier.

At the early stages molecular biology, it was known that proteins were made up of 20 different amino acids, while nucleic acids contained only for different nucleotides (adenine [A], cytosine [C], guanine [G], and uracil [U] in RNA, and DNA with thymine [T] instead of uracil). George Gamow postulated that the *genetic code* had to be based on units of at least three nucleotides (4^3 = 64), as two-nucleotide units would be insufficient to encode all 20 amino acids (4^2 = 16). This was proven by experiments performed by Francis H. Crick, Sydney Brenner and colleagues, in which the mutagen proflavin (which usually results in deletions or insertions of single nucleotides) was used to show that a T4 bacteriophage gene was rendered non-functional by mutations of one, two or four nucleotides, but would become functional again when the total number of mutations equaled three nucleotides (Crick et al, 1961). These three-nucleotide coding units were termed *codons*.

The next quest was the actual deciphering the genetic code: which nucleotide combinations encode which amino acids? Marshall W. Nirenberg and colleagues were the first to elucidate part of the genetic code (Matthaei et al, 1962). By utilizing a cell free system and a poly-U RNA, they discovered that a polypeptide was synthesized that contained only the amino acid phenylalanine. Thus, the codon UUU was specific for phenylalanine. Later, Severo Ochoa used a poly-C RNA and found that it coded for proline, while poly-A RNA coded for lysine (Gardner et al, 1962; Wahba et al, 1963). Using polymers of different nucleotides, the rest of the codons were deciphered, mainly by Nirenberg and H. Gobind Khorana. Importantly, three of the 64 codons were found to function as *stop* codons, used to terminate translation (Table 1). Thus, the genetic code is said to be *degenerate* (or redundant), in that that an amino acid can be encoded by more than one codon. Yet, the code displays no ambiguity: each codon specifies a particular amino acid (or a stop signal).



	Second nucleotide				
First nucleotide	U	C	A	G	Third nucleotide
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA STOP	UGA STOP	A
	UUG	UCG	UAG STOP	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

Table 1. The Genetic Code.
©Copyright 2009, Nature Education.

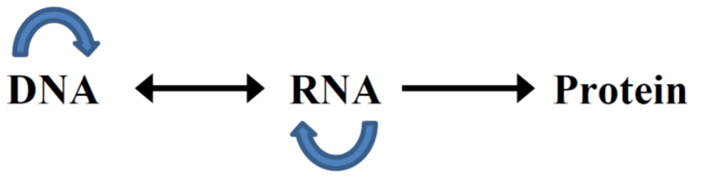


Figure 4. The Central Dogma of Molecular Biology.

While working on the mechanisms behind protein synthesis, Crick developed the concept known as The Central Dogma of Molecular Biology, which states “once sequential information has passed into protein it cannot get out again” (Fig. 4) (Crick, 1970). Because of the degeneracy of the code, the *originally exact* nucleotide sequence encoding a particular protein cannot be known from its amino acid sequence alone. Additionally, he hypothesized the existence of an “adaptor molecule” that was needed as a bridge between two alphabets, the nucleotides from RNA and the amino acids (Crick, 1958). This molecule was later discovered by Robert W. Holley in 1965 and termed it *transfer RNA* (tRNA) (Holley et al, 1965). tRNAs are generally 70 to 90 nucleotides in length and contain a crucial sequence, *the anticodon*, that pairs with a cognate *codon* present in mRNA. At its 3’ end, the tRNA has a covalently attached amino acid, which is transferred to the growing polypeptide by the ribosome as it reads the mRNA (Fig. 5). The overall accuracy of protein synthesis depends on two processes. (a) codon-anticodon recognition and, (b) correct tRNA charging by aminoacyl-tRNA synthetases

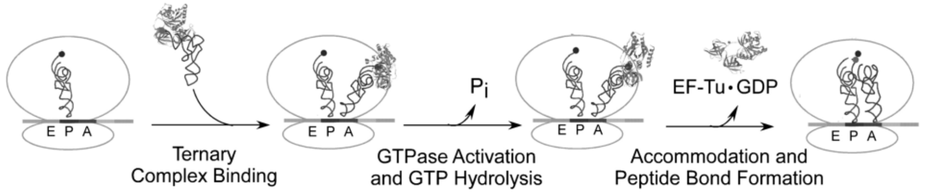

E P A
Ternary Complex Binding
E P A
P_i
GTPase Activation and GTP Hydrolysis
E P A
EF-Tu•GDP
Accommodation and Peptide Bond Formation
E P A

Figure 5. tRNA decoding and peptide bond formation.
“Reprinted from Molecular Cell, Vol 31, Issue 1, Ledoux, S. and Uhlenbeck, OC. Different aa-tRNAs Are Selected Uniformly on the Ribosome, 114-123., Copyright (2008), with permission from Elsevier.

(AARS). AARSs have an error rate of 1 in 10^4, accurate aminoacylation of tRNAs is in fact a critical element of the decoding process, as the ribosome will not correct mis-charged tRNAS (Ibba & Soll, 2000).

tRNAs are encoded by their corresponding *tRNA genes* (and thus constitute examples of genetic information not expressed as protein, similarly to rRNA). Remarkably, not a single genome sequenced to date (Chan & Lowe, 2009) contains tRNA genes for all 61 anticodons. Yet, in every organism, all 61 codons are indeed utilized to encode amino acids. How, then, is decoding accomplished for codons for which there is no strictly-matching anticodon-bearing tRNAs? Crick postulated the “wobble hypothesis” to answer this apparent discrepancy (Crick, 1966). The first two bases of a codon bind to the second and third bases of the anticodon in the tRNA by so-called Watson-Crick base pairing (A:U and C:G). However, the third base of the codon can bind via additional, non-Watson-Crick pairing to certain tRNAs (Crick, 1966). A variety of post-transcriptional chemical modifications of the first (or wobble) position in tRNAs (which pairs with the third base in the codon), such as deamination of adenine to inosine, allow three or even four degenerate codons to be recognized (Agris et al, 2007).

Effect of codon:anticodon interactions in polypeptide elongation rates.

It has been demonstrated that ribosome movement along an mRNA molecule is not uniform during translation and may be influenced by the presence of particular kinds of codons. In *E. coli*, codons decoded by tRNAs that utilize wobble have been experimentally demonstrated to be translated more slowly than synonymous ones that do form strict Watson-Crick base pairings in all three positions (Curran & Yarus, 1989; Sorensen & Pedersen, 1991). Moreover, codons that can be read by more than one tRNA species (Watson-Crick *plus* wobble) are translated faster than those that can only be decoded by a single tRNA species (*via* Watson-Crick only) (Higgs & Ran, 2008). Thus, it is likely that the rate at which a particular codon is translated depends on (1) tRNA concentration and (2) the chemical nature of the codon-anticodon interaction at the wobble position. This is in agreement with the notion that the rate-limiting step in polypeptide elongation is the binding of the charged tRNA complex to the cognate codon on the mRNA (Johansson et al, 2008).

The fate of the polypeptide upon its emergence from the ribosome.

Once amino acid residues are covalently brought together by the ribosome, their different moieties interact with each other *via* various forces that allow the nascent polypeptide to adopt three-dimensional shapes or *folds* in the aqueous environment of the cell. As briefly mentioned above, protein function is critically dependent on such shapes (or *conformations*) that polypeptide chains finally acquire, known as their *native state(s)*. The process (or processes) by which polypeptide chains acquire such structures is known as *protein folding*. Most of the current knowledge about how proteins fold has been obtained by *in vitro* refolding experiments, where full-length proteins are allowed to refold after being removed from solutions containing agents that disrupt the forces that stabilize protein structure, known as *denaturants* (such as urea or guanidinium). Pioneering experiments by Christian Anfinsen and colleagues showed that all the information required for a protein to acquire its native state is encoded in its amino acid sequence (Anfinsen, 1973). That is, a particular linear arrangement of amino acid residues *somehow* dictates the particular shape (or shapes) that the protein is to acquire. In essence, this is known as *the protein folding problem*. Now, why is it a problem? First, unlike the straightforward information transfer that occurs during protein synthesis (it is possible to exactly predict the amino acid sequence based on the mRNA sequence), it remains very challenging to predict what kind of fold is encoded by what sequence of amino acids, especially for larger and more complex conformations. In other words, although advances have been made (Bradley et al, 2005; Dimaio et al, 2011; Shirts & Pande, 2000; Snow et al, 2005; Thompson & Baker, 2011) (http://folding.stanford.edu/ (Shirts & Pande, 2000)) a fully dependable *folding code* capable of predicting three-dimensional structures has not emerged. Furthermore, examples from nature and the laboratory have demonstrated that proteins with remarkably similar sequences (>90% sequence identity) can adopt drastically different conformations (Alexander et al, 2007; Alexander et al, 2009; Alexander et al, 2005; Dalal & Regan, 2000). A different aspect of the protein folding problem pertains to the time necessary for a polypeptide to acquire it native state. As indicated by Cyrus Levinthal (Levinthal, 1969), polypeptides have quite large degrees of spatial freedom, due to the ability of their constituent atoms to rotate around their bonds. For example, assuming that there are only three possible conformations *per* peptide unit (and there are many more), a 100 amino acid protein could in principle populate 3^{198} different conformations. If each conformation could be sampled within the time required for a bond rotation (10^{-13} s), it would take 10^{27} years to complete the search (the age of the Universe is estimated to be ~ 1.4 x 10^{10} years). Strikingly, it is known that it only takes from milliseconds to a few seconds for most proteins to fold. This paradox suggests that each protein must follow a *preferred* or *sequential* folding pathway, where transition folding states arise and thus guide and significantly speed up the folding process (Fig. 6).

De novo protein folding in the cell differs from *in vitro* protein refolding in various ways (Bukau et al, 2006; Frydman, 2001; Hartl & Hayer-Hartl, 2002). *In vivo*, proteins emerge gradually from the ribosome as they are being synthesized. Thus, the full-length protein chain is not immediately available for folding, as it is when diluted out of denaturant. The vectorial nature of ribosomal protein synthesis imparts additional constraints on the folding process: the N-terminus of the protein is always exposed to solvent (or to interacting proteins, see below) before its more C-terminal elements. The rate of proteins synthesis is generally significantly slower (seconds to minutes) than measured rates of folding (nanoseconds to seconds). Furthermore, in contrast to the optimal conditions prepared for *in vitro* refolding of particular proteins, protein folding in the cytosol occurs under

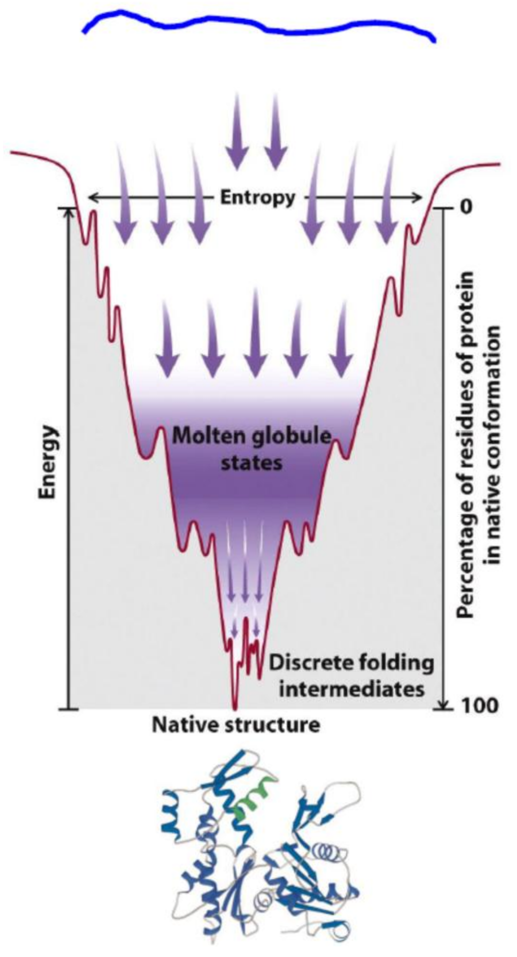

Entropy
0
Energy
Percentage of residues of protein in native conformation
Molten globule states
Discrete folding intermediates
100
Native structure

Figure 6. Protein folding landscape. Permission pending. Lehninger Principles of Biochemistry. 5th Ed. 2009

significant macromolecular crowding, and at fixed temperature and ionic strength (Ellis & Minton, 2006).

Certain proteins can readily attain native-like structures as they are being synthesized, whereas others, due to constraints mentioned above, need assistance from a class of proteins known as molecular chaperones (Barral et al, 2004; Hartl et al, 2011; Hartl & Hayer-Hartl, 2009). Chaperones that assist protein folding during translation bind to aggregation-prone, unstructured regions of the nascent polypeptide, thereby preventing intra- and inter-molecular misfolding. The iterative nature of chaperone binding results in release of these segments once additional polypeptide chain is available for folding. Thus, by maintaining the more N-terminal regions of a polypeptide relatively unstructured and competent for subsequent folding, chaperones offer an alternative to overcome the “N- to C-terminal emergence problem” of protein synthesis. If the protein folds successfully, its hydrophobic regions are usually buried and do not re-bind to the chaperone. Failure to fold usually results in re-binding to the chaperone or transfer to a “downstream” chaperone (Bracher et al, 2011). Proteins that are incapable of reaching their native state are either degraded and/or form misfolded species that undergo aggregation (Anderson et al, 2011).

Differences in protein folding mechanisms between bacteria and eukaryotes.

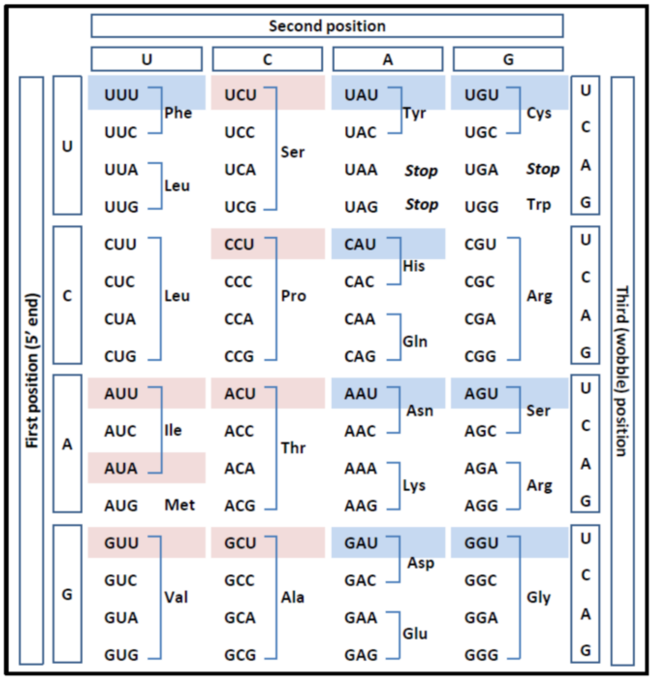
Translation speed and molecular chaperone content differ between bacteria and eukaryotes.

Since the process of polypeptide synthesis is directional, proteins emerge gradually from ribosomes as incomplete chains that are susceptible to misfolding.

Indeed, when proteins of eukaryotic origin are synthesized in bacterial expression systems, they are often incapable of acquiring their native state (Agashe et al, 2004; Chang et al, 2005). Eukaryotic protein misfolding upon recombinant production in bacteria has allowed the examination of conditions that contribute to *de novo* protein folding (Agashe et al, 2004; Chang et al, 2005; Stemp et al, 2005). A number of differences between the protein biosynthetic machines of bacteria and eukaryotes could be responsible for this phenomenon. In bacteria, folding of nascent chains can be delayed relative to their synthesis ("post-translational" folding), a process that may promote misfolding of certain eukaryotic recombinant proteins. In contrast, the eukaryotic cytosol appears to be highly capable of efficiently folding protein domains as they emerge from the ribosome ("co-translational" folding). Organism-specific chaperones have been demonstrated to support each of these distinct folding regimes, including Trigger Factor (TF) in bacteria and the ribosome-associated complex (RAC) in yeast. In addition to their different chaperone complements, a major difference between bacteria and eukaryotes is their translation speed. In *E. coli*, polypeptide elongation rates vary from ~12 amino acids per second (aa/s) during slow growth to ~22 aa/s during fast growth. In contrast, elongation rates in eukaryotes are thought to be considerably slower (~3 – 8 aa/s). Thus, the folding mechanisms of eukaryotic proteins evolved in the context of slower synthesis rates than those present in bacteria. It is possible that polypeptides of eukaryotic origin are less capable of attaining their folded state when emerging from the ribosome at unusually faster rates. Although ribosomal pausing at rare codons along mRNAs encoding particular proteins has been suggested to affect their activities, the effect of general variations in polypeptide synthesis rates on protein folding efficiency has remained largely unexplored as well as the effect of translation speed on de novo folding.

Differences in isoacceptor tRNA gene content between bacteria and eukaryotes

The increasing number of organisms whose genomes have been completely (or nearly completely) sequenced, coupled to computational algorithms designed to identify and curate tRNA genes have allowed a hitherto unparalleled analysis of the patterns of tRNA gene content across evolution (Chan & Lowe, 2009). Strikingly, our lab has found that, although there is conservation throughout evolution among tRNA gene families (Higgs & Ran, 2008), clear differences in tRNA gene content exist between bacteria and eukaryotes (Fig. 7). For example, bacterial genomes tend not to contain genes for tRNAs with adenine at the wobble position for alanine, glycine, proline, threonine, valine and isoleucine. These codons would be predicted to be translated slowly, since decoding depends on non-matching tRNAs bearing a guanine (instead of the strictly complementary uracil) at their first position (Fig. 8); inosine is not an option, since it is biochemically derived from adenosine).



	Second position								
First position (5' end)	U		C		A		G		Third (wobble) position
U	🟦 UUU	Phe	🟥 UCU	Ser	🟦 UAU	Tyr	🟦 UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	*Stop*	UGA	*Stop*	A
	UUG		UCG		UAG	*Stop*	UGG	Trp	G
C	CUU	Leu	🟥 CCU	Pro	🟦 CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	🟥 AUU	Ile	🟥 ACU	Thr	🟦 AAU	Asn	🟦 AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	🟥 AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	🟥 GUU	Val	🟥 GCU	Ala	🟦 GAU	Asp	🟦 GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 7. Differences in tRNA gene content between bacteria and eukaryotes. Codons boxed in blue denote tRNA genes often absent in bacteria and eukaryotes, while codons boxed in pink denote genes mostly absent only in bacteria. Data obtained from (Chan & Lowe, 2009), Image from Dr. Barral.

Throughout this dissertation, experiments have been designed to utilize the differences between bacteria and eukaryotes in terms of their protein biogenesis machineries in order to gain insight into the relationships between coding sequences, translation rates and *de novo* protein folding efficiencies.

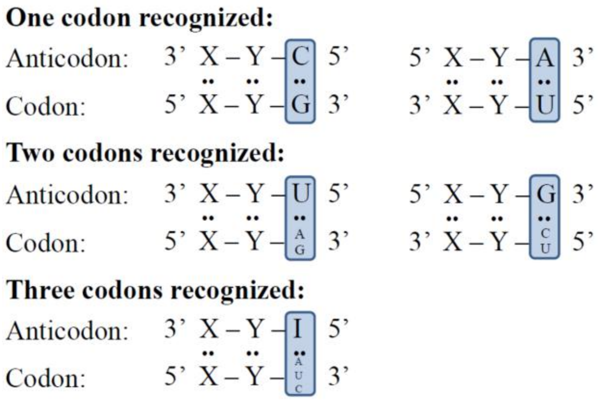


Figure 8. Wooble Hypothesis. The nature of the nucleotide at the wobble position, dictates the number of codons that can be recognized.

In the next chapters of this work, I will first explore whether a global reduction in bacterial translation speed could enhance eukaryotic protein folding efficiency. Secondly, I will examine whether polypeptide elongation rates and folding information are encoded in the genome as well.

Chapter 2: Slowing Bacterial Translation Speed Enhances Eukaryotic Protein Folding Efficiency[1]

Introduction

Misfolding of eukaryotic proteins upon recombinant production in bacteria has placed great limitations on their biochemical and structural analyses and their therapeutic utilization (Baneyx & Mujacic, 2004; Pavlou & Reichert, 2004). In bacteria, folding of polypeptide nascent chains can be delayed relative to their synthesis ("post-translational" folding), a process that may promote misfolding of certain recombinant proteins (Agashe et al, 2004; Netzer & Hartl, 1997). In contrast, the eukaryotic cytosol appears to be highly capable of efficiently folding protein domains as they emerge from the ribosome ("co-translational" folding) (Agashe et al, 2004; Chang et al, 2005; Netzer & Hartl, 1997). Kingdom-specific molecular chaperones have been demonstrated to support each of these distinct folding regimes, including Trigger factor (TF) in bacteria (Agashe et al, 2004; Kaiser et al, 2006) and the ribosome-associated complex in fungi (Gautschi et al, 2002). In addition to their different chaperone complements, a major difference between bacteria and eukaryotes is their translation speed. In *E. coli*, polypeptide elongation rates vary from ~10 amino acids *per* second (aa/s) during slow growth to ~20 aa/s during fast growth (Liang et al, 2000; Pedersen, 1984). In contrast, elongation rates in eukaryotes are thought to be fairly constant and considerably slower (3 – 8 aa/s) (Mathews et al, 2000). Although ribosomal pausing at rare codons along

[1] "Modified from Journal of Molecular Biology, Vol. 396, Issue 5, Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM., Slowing bacterial translation speed enhances eukaryotic protein folding efficiency, 1310 – 1318. Copyright (2010), with permission from Elsevier"

mRNAs encoding particular proteins has been shown to affect their activities (Kimchi-Sarfaty et al, 2007; Zhang et al, 2009), the effect of general variations in polypeptide synthesis rates on protein folding efficiency has remained largely unexplored. In this work, we aimed to study the impact of global changes in protein synthesis rates on *de novo* protein folding by utilizing streptomycin (Sm) pseudo-dependent (Sm^P) ribosomes of *E. coli* (Zengel et al, 1977), whose polypeptide elongation rates can be modulated by varying the concentration of Sm present in the growth medium.

Results

Polypeptide elongation rates can be modulated in SmP bacteria with no detrimental effects on the folding of endogenous proteins or activation of the stress response.

Sm^P ribosomes contain mutations in protein S12 of their decoding center (see Methods) (Kurland et al, 1996). In the absence of Sm, bacteria harboring these ribosomes display a “hyper-accurate” phenotype, with considerable reduction in translation rates (~5 aa/s) and a ~20-fold increase in accuracy of amino acid incorporation compared to wild type (Ruusala et al, 1984). Addition of Sm relieves this phenotype in a concentration-dependent fashion, restoring translation speed to nearly wild type levels, as reflected by restoration of growth rates (which correlate directly with protein synthesis speed; Fig. 9a) (Ruusala et al, 1984), albeit with a concomitant ~7-fold increase in misincorporation rates compared to wild type (Ruusala et al, 1984).

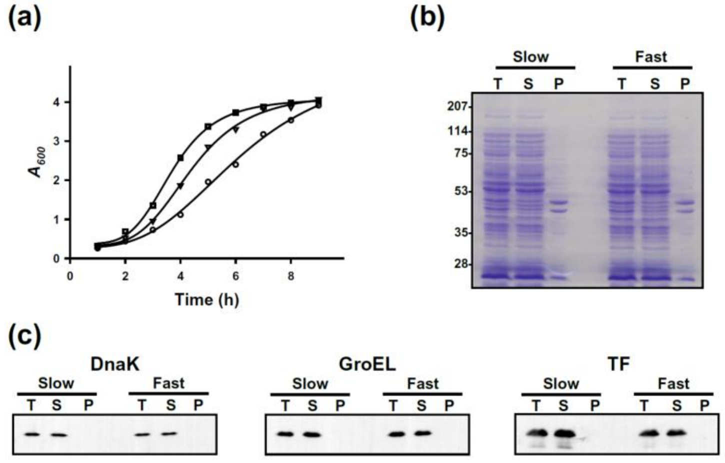


Figure 9. Protein synthesis rates can be modulated in SmP bacteria with no major effects on endogenous protein misfolding or upregulation of molecular chaperones. (a) Growth of Sm^P bacteria in the absence (circles) or increasing concentrations of Sm (triangles, 200 µg/ml; squares, 500 µg/ml). (b) Coomassie brilliant blue SDS-PAGE of protein content present in total (T), supernatant (S) and pellet (P) fractions of Sm^P bacteria grown in the absence (slow) or presence of Sm (fast; 500 µg/ml), harvested at equivalent A_{600} values and lysed under native conditions. (c) Immunoblots with antibodies against DnaK, GroEL and TF (as indicated) of total (T), supernatant (S) and pellet (P) fractions of Sm^P bacteria grown in the absence (slow) or presence of Sm (fast; 500 µg/ml), harvested at equivalent A_{600} values and lysed under native conditions.

Utilization of Sm^P ribosomes allowed us to focus on the effects of decreased polypeptide elongation rates on protein folding, since for every comparison between bacteria harboring slow (without Sm) and fast (with Sm) ribosomes, all other experimental parameters were identical and constant. We also wished to ascertain that a general and constant decrease in translation speed did not lead to upregulation of molecular chaperones, due for example to misfolding and aggregation of certain endogenous *E. coli* proteins. Thus, we began our analysis by comparing the levels of aggregated proteins present in the Sm^P strain grown under different concentrations of Sm (Fig. 9b). We observed no major differences in the patterns or levels of proteins present in the insoluble fraction of lysates prepared under native conditions (Chang et al, 2005). We next analyzed whether slow translation led to activation of the bacterial stress response and accumulation of molecular chaperones (Bukau et al, 2000). We performed immunoblot analyses of the steady state levels of the major chaperone systems known to influence nascent protein folding in *E. coli* (Fig. 9c). We observed no major differences between cultures grown in the absence or presence of Sm in the accumulation of the Hsp70 homolog DnaK, the chaperonin GroEL or TF. Thus, we concluded that a general reduction in translation speed in *E. coli* does not result in major alterations in the folding efficiency of its endogenous proteins.

Slow translation speed enhances the de novo folding of firefly luciferase.

Having established that there were no major differences in the molecular chaperone capacity of Sm^P bacteria under different antibiotic concentrations, we wished to assess whether alterations in polypeptide elongation rates *per se* influenced the folding efficiency of firefly luciferase (FL, 64 kDa) (Conti et al, 1996), a model protein whose *in vivo* folding and *in vitro* refolding requirements have been extensively characterized (Agashe et al, 2004; Frydman et al, 1999; Schroder et al, 1993). Production of FL in *E. coli* is characterized by very poor folding yields, with the majority present as aggregated, inactive material (Agashe et al, 2004). In contrast, heterologous production of FL in the yeast *S. cerevisiae*, whose ribosomes are slower than those of *E. coli* (Mathews et al, 2000), results in nearly 100% of the protein as a soluble, active species (Agashe et al, 2004). Remarkably, the bacterial chaperone system DnaK/DnaJ/GrpE is highly capable of assisting the *in vitro* refolding of FL upon dilution from denaturant, as evidenced by the high yields (~90%) of native luciferase obtained in its presence compared to its spontaneous refolding (<10%) (Schroder et al, 1993). These and other results have suggested that the misfolding of luciferase during its *de novo* production in bacteria occurs, at least partially, from a co-translational misfolding event that the DnaK/DnaJ/GrpE system is incapable of resolving (Agashe et al, 2004; Frydman et al, 1999). Thus, we wished to determine whether production of FL by bacteria with slower translation rates, more closely resembling those of eukaryotes, led to beneficial effects on its folding yield. Faster ribosomes synthesize more FL chains than slower ones in the same amount of time. In order to assess FL solubility at equivalent levels of accumulation between bacteria containing slow (in the absence of Sm) and fast (in the presence of Sm) ribosomes, we set up the following experiment (Figs. 10a,b) (see Methods). A starter culture of Sm^P bacteria transformed with a plasmid encoding FL under control of an arabinose promoter (grown in the absence of Sm) was diluted into equal volumes of growth medium containing arabinose. One volume contained Sm (fast translation) and the other one did not (slow translation). Aliquots of each culture were harvested at regular intervals (15 min for “low accumulation” and 1 h for “high accumulation”). The content of total FL chains produced was monitored by SDS-PAGE followed by densitometry of the band corresponding to FL in immunoblots (for “low accumulation”) or Coomassie blue-stained gels (for “high accumulation”). For solubility assessment, cell pellets containing equivalent amounts of total FL protein were lysed under native conditions (Chang et al, 2005) and separated into supernatant and pellet fractions by centrifugation. We consistently found that, when translated by slow ribosomes, a larger fraction of luciferase was present in the supernatant, and the pellet contained less aggregated material (Figs. 10a,b). This effect did not depend on the concentration of recombinant protein produced, since very similar results were obtained with shorter or longer induction times (Figs. 10a,b). We next examined whether this increase in solubility corresponded to increased enzymatic activity, which would confirm that a greater fraction of FL indeed reached its native state under slow translation conditions. Equal volumes from the total and supernatant fractions used for solubility determination were assayed for luciferase activity (see

Methods). Fractions containing FL synthesized by slow ribosomes displayed ~2-fold greater activity than those from faster ribosomes, and this effect was also independent of the levels of FL accumulation (Figs. 10a,b).

The absence of molecular chaperone induction and the higher activity of FL produced in Sm^P bacteria in the absence of Sm (irrespective of levels of accumulation) suggest that the folding of FL nascent chains is enhanced by slower polypeptide elongation rates. An alternative explanation would be that, in the absence of Sm, misfolded FL chains are more efficiently degraded and thus lesser amounts of insoluble material accumulate in this condition than in the presence of the antibiotic. To rule out this possibility, we set up an experiment to compare the *total* activity of FL produced in each condition within a defined time period, irrespective of the *number* of FL nascent chains produced. Recombinant expression of FL was initiated in cultures with and without Sm and allowed to proceed for the same amounts of time. Equal culture volumes were harvested and the yield of total enzymatic activity was determined for each. We observed higher accumulation of active enzyme in the culture grown without Sm, in spite of the fact that this culture had less cells than the one with Sm (Fig. 10c). Since slower ribosomes could not have synthesized a greater number of total nascent chains than the faster ones within the same period of time, a higher fraction of total nascent chains must have folded correctly in the culture synthesizing FL more slowly. Thus, enhanced degradation alone cannot account for our findings.

A different scenario that could also explain our findings involves the higher rate of amino acid misincorporation by Sm^P ribosomes in the presence of Sm (Ruusala et al, 1984), which could render nascent chains simply incapable of folding. If this were the case, misfolded FL produced by error-prone ribosomes should be less capable of refolding to the native state after denaturation. To test this possibility, FL isolated from inclusion bodies from bacteria grown in the presence and in the absence of Sm, was denatured in urea and allowed to refold into buffer supplemented with the DnaK/DnaJ/GrpE chaperone system. The kinetics and refolding yields of FL translated by slow or fast ribosomes were essentially identical (Fig. 10d). Thus, we concluded that amino acid misincorporation cannot solely account for the increased misfolding of FL produced in the presence of Sm. Taken together, the above findings strongly suggest that a reduction in translation rates *per se* leads to significantly higher yields of native FL upon recombinant production in *E. coli*.

The folding of diverse aggregation-prone eukaryotic proteins is promoted by decreased bacterial translation rates.

Having established that FL synthesized at slower speeds was more capable of acquiring its native state, we wished to determine whether this effect was generally applicable to eukaryotic proteins prone to aggregation when synthesized in bacterial systems. The green fluorescent protein from *Aequorea victoria* (GFP, 27 kDa) (Tsien, 1998) is a single domain protein composed mostly of beta strands that does not depend on assistance from molecular chaperones for *in vitro* refolding (Fukuda et al, 2000). However, it displays considerable misfolding and aggregation upon recombinant production in *E. coli* (Chang et al, 2005; Fukuda et al, 2000). A variant of GFP selected for efficient maturation in bacteria at 37 °C, the so-called “Cycle3” mutant (Crameri et al, 1996) (utilized in this study), displays soluble yields of <50% (Chang et al, 2005; Fukuda et al, 2000). In order to analyze the behavior of GFP under fast and slow translation conditions, we set up experiments similar to those described above for FL. We observed that, when synthesized by slower ribosomes, the fraction of fluorescent GFP, indicative of correct acquisition of its native state, was ~2-fold higher than when translated at faster rates (Fig. 11a). A similar, yet more modest, behavior was observed for its solubility.

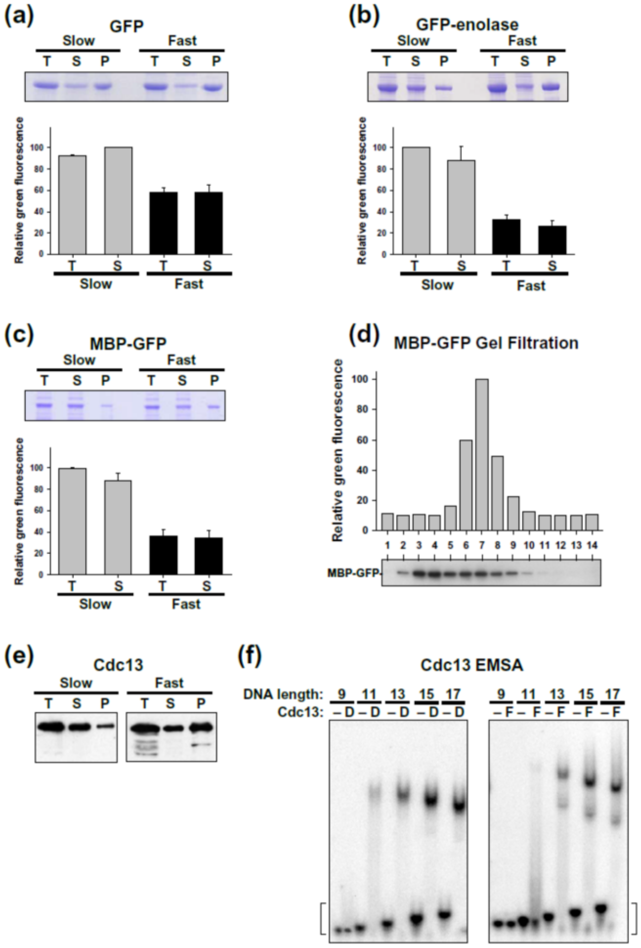


Figure 11. Production of native aggregation-prone eukaryotic proteins is promoted by slower bacterial translation rates. (a, b and c) Assessment of solubility (top panels) and quantification of fluorescence (bottom panels) of GFP (a), GFP-enolase (b) and MBP-GFP (c) fusion proteins produced in Sm^P bacteria in the absence (slow) or presence of Sm (fast; 500 µg/ml). Solubility was determined by SDS-PAGE after production of total lysates (T) under native conditions followed by fractionation into supernatant (S) and pellet (P) fractions as in 2a. Fluorescence emission was determined for total (T) and supernatant fractions (S) as described(Chang et al, 2005). (d) Gel filtration behavior of soluble MBP-GFP produced in Sm^P bacteria grown in the presence of Sm (500 µg/ml). Eluted fractions were assessed for fluorescence emission (top panel) and total recombinant protein content by immunoblotting (bottom panel). (e) Solubility of Cdc13 produced in Sm^P bacteria in the absence (slow) or presence of Sm (fast; 500 µg/ml) determined by immunoblotting with an anti-His$_6$ tag antibody after production of total lysates (T) under native conditions followed by fractionation into supernatant (S) and pellet (P) fractions. (f) Electromobility shift assay (EMSA) of purified full-length (F; 100 nM) Cdc13 or its DNA binding domain (D; 100 nM). The single-stranded DNA length required for binding was determined using telomeric oligonucleotides (50 pM) with the indicated lengths. The brackets mark the approximate position of free probe.

GFP and its derivatives have been extensively utilized as reporter domains in a wide variety of fusion proteins (Giepmans et al, 2006). It has been shown that the GFP moiety can impose significant constraints on *de novo* folding of fusion proteins in bacteria, probably as a result of intra-molecular interference with the folding of adjacent domains (Chang et al, 2005). In order to assess whether interference of GFP on its fusion partner could be ameliorated by decreased translation speed, we conducted similar experiments with the previously characterized model fusion proteins GFP-enolase (74 kDa) and maltose binding protein (MBP)-GFP (70 kDa). Both enolase and MBP are produced as soluble, native species even when recombinantly expressed to very high levels in bacteria (Chang et al, 2005). In contrast, when fused to GFP, the resulting proteins are present mainly in the insoluble fraction, and their fluorescence emission is drastically reduced (Chang et al, 2005). We observed that both GFP fusion proteins displayed ~3-fold higher fluorescence emission when produced by the Sm^P strain under slower translation conditions (Figs. 11b,c). Similar to GFP alone, the amount of recombinant proteins present in the soluble fraction was only moderately enhanced. We next performed an experiment to explain the observed discrepancy between the considerably higher changes in fluorescence *versus* solubility observed for GFP and its fusion proteins. We hypothesized that some amount of the recombinant proteins produced could be misfolded (and thus non-fluorescent), yet have remained in the soluble fraction under our centrifugation conditions (22,000 *g* for 10 min). To test this possibility, we performed a gel filtration experiment on the supernatant fraction of MBP-GFP synthesized by Sm^P bacteria in the presence of Sm (Fig. 11d). We analyzed each of the eluted fractions for native protein content by fluorescence emission and total protein content by immunoblotting. We found that the native MBP-GFP peak accounted for only a small fraction the total MBP-GFP content present in the supernatant, the majority of which eluted at earlier fractions, corresponding to higher apparent molecular weights. These results confirm our idea that the supernatants of our GFP fusions contain misfolded recombinant proteins that do not sediment under our experimental conditions. Thus, experiments involving biochemical activities, rather than solubility, more accurately reflect the folding behaviors of the proteins being studied.

We next investigated whether this approach could be successfully applied to large, multi-domain eukaryotic proteins previously shown to be inefficiently folded to their native state upon recombinant production in bacteria. We selected the telomere-binding protein Cdc13 from *Saccharomyces cerevisiae* (105 kDa), a protein essential for telomere maintenance that protects chromosome ends from damage and recruits the telomerase complex (Nugent et al, 1996; Pennock et al, 2001). Cdc13 contains three distinct regions: an N-terminal telomerase recruitment domain, a central DNA binding domain and a C-terminal capping region (Chandra et al, 2001; Wang et al, 2000). The central DNA binding domain of Cdc13 has been expressed as a soluble and active species in *E. coli*, which has facilitated its biochemical and structural analysis (Mitton-Fry et al, 2004). However, these analyses for the full-length protein have been hindered by the inability of wild type bacteria to yield native material. Similar to our results with FL and the GFP fusion proteins, we observed that most of the full-length Cdc13 protein was present in the soluble fraction when synthesized by Sm^P ribosomes under slow translation conditions (Fig. 11e). To assess whether full-length Cdc13 produced in this manner was indeed native, we purified it from the soluble fraction (see Supplementary Information) and compared its DNA binding activity to that of the central DNA binding domain alone by electromobility shift assays (EMSA) (Toogun et al, 2007) (Fig. 11f). We found that purified full-length Cdc13 displayed comparable DNA binding affinity and selectivity (*i.e.* 11-base length requirement) to the well characterized properties of the central DNA binding domain (Hughes et al, 2000) (Fig. 11f). Furthermore, examination of various Cdc13 derivatives produced in the Sm^P strain, including full-length, shows that the additional domains of this protein are produced in native form, as the N-terminal domain stimulates telomerase activity and the C-terminal domain associates with additional telomere proteins to cap telomeric DNA *in vitro* (DeZwaan et al, 2009).

Discussion

In this study, we set out to investigate whether the fast polypeptide elongation rates of the bacterial ribosome could be responsible, at least partially, for the poor capacity of *E. coli* to fold proteins of eukaryotic origin, normally translated at the considerably slower speed of the eukaryotic ribosome. We have found that indeed, decreasing bacterial polypeptide elongation rates to rates similar to those of eukaryotes promotes the folding of a diverse set of heterologous proteins. How do slower translation rates favor correct eukaryotic protein folding? The folding pathways of eukaryotic proteins evolved in context of slower translation rates (Agashe et al, 2004; Bremer & Dennis, 1996; Chang et al, 2005). However, when eukaryotic proteins emerge from the bacterial ribosome at faster rates, longer nascent chains are exposed to greater conformational possibilities, some of which may result in misfolded species that were never originally selected against. Slower emergence from the ribosome may thus temporally restrict the number of incorrect conformations an elongating nascent chain can adopt. Additionally, the chaperone complement of the bacterial cytosol may be incompatible with the folding regimes of certain eukaryotic proteins (Hartl & Hayer-Hartl, 2002).

Diverse methodologies have been previously attempted to increase the yield of correctly folded aggregation-prone eukaryotic proteins produced in bacterial systems, with varying degrees of success. A widely utilized approach is to decrease temperature during the induction period, which has proven to be beneficial for a diverse set of proteins (Baneyx & Mujacic, 2004; Schein, 1989). However, the folding yield of a considerable number of proteins appears not to improve even at induction temperatures as low as 18 °C (such as Cdc13, discussed above). These different behaviors may be explained, at least partially, by the varying extent with which decreased temperature affects cellular processes and parameters (Bremer & Dennis, 1996), including protein synthesis and folding rates as well as the hydrophobic interaction. For proteins that misfold even at reduced temperatures, the beneficial effects of lowering temperature (*e.g.* slower translation rates and decreased hydrophobic interactions) might be masked by adverse effects (*e.g.* a strong reduction in intrinsic folding rates and the biochemical activities of molecular chaperones). Since the strategy outlined in this study targets translation speed exclusively, all other cellular processes are maintained constant. Thus, a nascent chain emerging slowly from the ribosome may benefit from unaltered folding rates and full chaperone assistance. In cases where the adverse effects of low temperature have no major repercussions on the folding properties of the protein being produced, reducing translation rate may further enhance its folding yields. Similarly, utilization of Sm^{P} ribosomes may be beneficial for the production of proteins whose *de novo* folding regimes depend upon over-expression or supplementation with molecular chaperones (Stemp et al, 2005), a strategy that by itself has so far proven only of limited success.

In summary, we have found that decreasing translation rates in bacteria *per se* promotes the folding of a diverse set of proteins of eukaryotic origin. We found that reduced protein synthesis rates led in no case to detrimental effects on the folding of recombinant proteins. Slower translation did not result in endogenous protein misfolding or activation of the bacterial stress response. We believe that our findings provide a general strategy for the production of recombinant proteins that does not rely on individual manipulation of coding sequences or introduction of specific accessory factors.

Materials and Methods

Strains and growth conditions.

The *E. coli* Sm^P strain utilized here was CH184 (a W3110 derivative), a kind gift from Prof. D. Hughes (Uppsala University). It contains two mutations in the *rpsL* gene (see below), C256A and C272A, resulting in R86S and P91Q substitutions in protein S12, corresponding to the *rpsL1204* in Ref. 14 and thus corresponds to strain SM3 in Ref. 13. For recombinant protein production (see below), this strain or a λDE3-lysogenized derivative (Novagen) were transformed with the following arabinose-controlled promoter-based plasmids (Guzman et al, 1995): pBAD-Luc (encoding FL with a C-terminal c-myc-His$_6$ epitope tag) (Agashe et al, 2004), pBAD-GFP$_{uv}$ (encoding the Cycle3 variant of GFP (Crameri et al, 1996)), pBAD-GFP-Eno and pBAD-MBP-GFP (encoding fusion proteins of GFP fused to *E. coli* enolase or MBP *via* a 16 amino acid flexible linker) (Chang et al, 2005); or T7-driven promoter-based plasmids pET28-Cdc13 (encoding amino acids 1 – 924 of Cdc13 with an N-terminal His$_6$ tag) or pET6H-Cdc13-DBD (encoding amino acids 451 – 694 of Cdc13 with an N-terminal His$_6$ tag). Cells were grown in LB broth at 37 °C with 250 rpm orbital shaking in volumes that occupied at most one fourth of the total vessel volume, in the presence of ampicillin (100 µg/ml).

For fast translation rates, Sm was added to a final concentration of 500 µg/ml, unless indicated otherwise. Sm^P cells consistently grew faster in the presence of Sm. For the experiments in Fig. 9, volumes of cells containing equivalent A_{600} values were harvested by centrifugation and lysed by spheroplasting (Ausubel et al, 2003) under native conditions (Chang et al, 2005). Similar amounts of total protein in the resulting lysates were verified by the Bradford assay (BioRad).

Recombinant protein production.

Starter cultures were grown as described above, diluted into two equal volumes (for experiments with and without addition of Sm) and protein induction was carried out when cell density reached A_{600} = 0.8 with 0.2% (w/v) arabinose or 1 mM IPTG and harvested at either 15 min (for low protein accumulation experiments) or 1 h (for high protein accumulation experiments) intervals. For experiments in Figs. 10a, 10b and 11, total amounts of recombinant protein produced during each interval were assessed by examining equivalent amounts of cells (equal A_{600} values), which were subsequently lysed, ran on SDS-PAGE and either immunoblotted (for low protein accumulation experiments) or Coomassie brilliant blue-stained (for high protein accumulation experiments). Aliquots harvested at time points containing equivalent levels of each recombinant protein produced in the presence and absence of Sm (as assessed by band densitometry) were then lysed under native conditions as described (Chang et al, 2005) and their solubility and activity or fluorescence emission assessed (see below). For the experiments in Fig. 10c, equal culture volumes were harvested at the time points indicated, lysed under native conditions and total luciferase activity was determined (see below).

rpsL sequencing.

The entire *rpsL* gene from strain CH184 was amplified by PCR with *Pfu* turbo DNA polymerase (Stratagene) with oligonucleotide primers RPSLup (5’ CAG ACT TAC GGT TAA G 3’) and RPSLdn (5’ CAG GAT TGT CCA AAA C 3’) and sequenced with an ABI Prism 3730 capillary sequencer (Sequiserve).

Determination of protein solubility.

Cells were harvested by centrifugation and spheroplasts were prepared as described(Ausubel et al, 2003). Spheroplasts were lysed by dilution into an equal volume of native lysis buffer (5 mM $MgSO_4$, 0.2% (v/v) Triton X-100 (Sigma), Complete EDTA-free protease inhibitors (Roche), 100 units/ml Benzonase (Roche), 50 mM Tris-HCl, pH 7.5). Aliquots were centrifuged into supernatant and pellet fractions (20,000*g* for 10 min) and analyzed by SDS-PAGE followed by either staining with Coomassie brilliant blue or immunoblotting with the anti-DnaK 8E2/2 monoclonal antibody (Stressgen), the anti-GroEL 9A1/2 monoclonal antibody (Stressgen), an anti-TF polyclonal antibody (a kind gift from Dr. P. Genevaux, Cologne), the anti-c-myc 9E10 monoclonal antibody (Roche), the anti-His_6 tag monoclonal antibody HIS6.H8 (Abcam) or the anti-GFP JL8 monoclonal antibody (Clontech).

Determination of luciferase activity and green fluorescence.

Total and supernatant fractions from cells expressing FL, GFP and GFP fusion proteins were prepared as above and equivalent dilutions to those used for solubility assessment were utilized. FL activity was determined using the Luciferase Assay System (Promega) in a Sirius luminometer (Berthold) as described (Agashe et al, 2004). Green fluorescence was measured in a Fluorolog 3 fluorescence spectrometer (Horiba/Jobin Yvon) with excitation at 398 nm and emission at 508 nm as described (Chang et al, 2005).

Gel filtration experiments.

A supernatant fraction from Sm^P bacteria grown in the presence of Sm was prepared by native lysis as described above. It was applied to a Superdex 75 column (GE) pre-equilibrated in PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4). Fractions were collected and equivalent volumes were immediately assayed for fluorescence emission (as above) or ran on SDS-PAGE and immunoblotted with the anti-GFP antibody (as above).

Electromobility shift assays.

Indicated Cdc13 protein versions utilized in the electromobility shift assays (Toogun et al, 2007) were in TMG-30 buffer supplemented with 200 µg/ml bovine serum albumin, 200 µg/ml poly[d(I-C)] and end-radiolabeled oligonucleotide. The single-stranded telomeric oligonucleotides were 5’ GTG GGT GTG 3’, 5’ GTG GGT GTG TG

3’, 5’ GTG GGT GTG TGT G 3’, 5’ GTG GGT GTG TGT GTG 3’ and 5’ GTG GGT GTG TGT GTG GG 3’. Following a 20 min incubation at 22 °C, the samples were resolved on a 6% native polyacrylamide gel in 1X GTG buffer (29 mM Taurine, 0.7 mM EDTA, 90 mM Tris), which was subsequently dried. The products were visualized with a Phosphoimager instrument (Molecular Dynamics).

In vitro refolding assays.

10 µM of FL from inclusion bodies of the Sm^P strain grown in the presence (500 µg/ml) or absence of Sm were denatured in denaturation buffer (6 M Gdm-HCl, 5 mM DTT, 5 mM magnesium acetate, 50 mM potassium acetate, 25 mM HEPES-KOH, pH 7.4) at 25 °C for 30 min. Refolding was started by 100-fold dilution into refolding buffer (5 mM ATP, 10 mg/ml bovine serum albumin, 1 mM DTT, 5 mM magnesium acetate, 50 mM potassium acetate, 25 mM HEPES-KOH, pH 7.4) and allowed to proceed at 25 °C in the presence of DnaK (10 µM), DnaJ (4 µM) and GrpE (6 µM) (Agashe et al, 2004). Luciferase activity was determined as above.

Supplementary Information

Purification of full-length and DNA binding domain of Cdc13. 6 liters of LB medium were seeded with SmP transformants of the plasmids described above from an overnight culture and grown at 18 °C to A595 = 0.1 and induced with IPTG after the A595 = 0.3. The cultures were clarified by centrifugation, resuspended in ice-cold 1X Talon binding buffer, 1% NP-40 and 0.5 µg/ml lysozyme and flash frozen in liquid nitrogen. Recombinant protein present in the soluble fraction was isolated using metal affinity chromatography (Talon resin, Clontech). Eluted protein was diluted 2:1 with TEN0G Buffer (0.1 mM EDTA, 10% glycerol, 20 mM Tris, pH 6.9), applied to a MonoQ column (GE), which was washed with TEN0.10G buffer (as above, with 100 mM NaCl final). Full-length Cdc13 was eluted with a 100 – 300 mM NaCl gradient, while the DNA binding domain was eluted with a 100 – 500 mM NaCl gradient. Eluted proteins were concentrated to ~ 200 µl volumes using a micro-concentration device and applied to a Superdex 200 gel filtration column (GE) equilibrated in TMG30 buffer (1.1 mM MgCl2, 0.1 mM EDTA, 1.5 mM DTT, 10% glycerol, 0.1% Triton X-100, 30 mM sodium acetate, 20 mM Tris-HCl, pH 7.0). Eluted proteins were concentrated to ~100 µl volumes, flash frozen and stored at – 20 °C (Fig. 12).

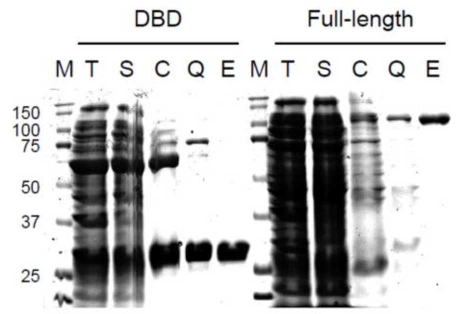


Figure 12. Supplementary Figure
Purification of full-length and DNA binding domain of Cdc13. Both variants of Cdc13 were produced as described in Supplementary Methods. Total lysates (T) were cleared by centrifugation and the supernatant fraction (S) were subjected to cobalt-affinity (C), MonoQ ion exchange (Q) and Superdex 200 size exclusion (E) resins. Aliquots of each step and chromatographic fraction and the final purified proteins (0.5 µg) were co-resolved by SDS-PAGE and the proteins were visualized by staining with Coomassie brilliant blue.

Chapter 3. Polypeptide elongation rates and folding efficiencies can be predictably manipulated by synonymous codon substitutions

Introduction

Proteins are made up of amino acids which are encoded in DNA by tri-nucleotide codons. There are 64 codons, 61 of which encode 20 amino acids in most organisms. Since as many as six synonymous codons can code for a single amino acid, the genetic code is said to be “degenerate” or “redundant.” Furthermore, all 61 codons can be translated by as few as 32 tRNA anticodons. This is because a single tRNA can decode more than one codon through non-cognate pairing, or wobble pairing, in the third (wobble) position of the codon (Crick, 1966). Wobble pairing refers to any base pairing that is not standard Watson-Crick pairing (G::C and A::U).

There is evidence that these two methods of decoding differ in translation rate with the former being translated more slowly (Curran & Yarus, 1989; Sorensen & Pedersen, 1991). When the translation rates of two synonymous codons for glutamate were measured, Sorensen, (Sorensen & Pedersen, 1991) found that the Watson-Crick read codon was decoded 3.4 fold faster than the Wobble decoded codon even though these codons were decoded by the same tRNA. These findings, coupled with the knowledge that the rate limiting step of ribosomal elongation is the arrival of the correct tRNA to the waiting codon (Johansson et al, 2008; Varenne et al, 1984) , indicate the translation speed of a particular codon likely depends on tRNA concentration and the chemical nature of the codon-anticodon interaction at the wobble position. Indeed, it is largely accepted that tRNA concentration plays an important role in determining translation speed and has even been found to be “the best and most informative estimator of codon translation speed" as the result of a large scale analysis (Saunders & Deane, 2010).

Interestingly, the use of synonymous codons is nonrandom, and this nonrandom distribution believed to greatly influence translation rate. Codon usage varies among and within organisms with highly expressed genes showing bias toward certain synonymous codons (Grantham et al, 1980). These codons are referred to as "frequent", "common", "optimal", and "favorable" because they are used most frequently within highly expressed genes, and because they have been found to correlate somewhat with tRNA content, they are considered to be translated at faster rates (Ikemura, 1981) On the other hand, codons that are used less frequently within highly expressed genes are termed "rare", "non-optimal", "unfavorable", and "slow", assuming there is less tRNA available to decode them. Indeed, codon bias is utilized quite regularly as a predictor of translation speed (Clarke & Clark, 2008; dos Reis et al, 2004; Sharp & Li, 1987) Using codon bias to define which codons are "slow" and which are "fast" is certainly contentious since there are several instances in which the most frequently used codons have no cognate tRNA genes and must rely on the slower Wobble decoding (Chan & Lowe, 2009) http://gtrnadb.ucsc.edu/); (Fig. 18). Furthermore, the findings of other groups (Bonekamp et al, 1989; Saunders & Deane, 2010) have shed doubt on the presence of a correlation between codon bias and tRNA concentration.

It has long been proposed that translation speed affects protein folding (Purvis et al, 1987), and recent evidence has provided substantial support for this hypothesis (Kimchi-Sarfaty et al, 2007; Siller et al, 2010; Zhang et al, 2010). Notably, Siller *et al.*

demonstrated an increase in folding efficiency of certain proteins produced in *E. coli* as the result slower ribosomal elongation rates. However, much of the evidence linking mRNA encoded speed information to protein secondary and tertiary structure uses translation speed predictions based on codon bias (once referred to as "one of the most controversial areas in molecular evolution") (dos Reis et al, 2004; Gupta et al, 2000; Thanaraj & Argos, 1996a; Xie & Ding, 1998). Indeed, there is discordance within this group of studies (Brunak & Engelbrecht, 1996; Thanaraj & Argos, 1996b; Zhou et al, 2009), some of which is likely due to predicting translation speed from codon bias rather than tRNA content. Others have recently shown that tRNA concentration is critical for predicting translation speed and elucidating its effects on protein folding albeit *in silico* (Saunders & Deane, 2010; Tuller et al, 2010; Zhang et al, 2009).

These studies all support the notion that translation speed is encoded in mRNA and that it affects the co-translational folding of the nascent polypeptide. However, there is a lack of experimental evidence demonstrating that codon composition does influence polypeptide elongation rate and/or folding. Here, we utilize pulse-chase methods to measure *in vivo* translation rates of mRNA recoded based on the two most common predictors of translation speed: tRNA concentration and codon bias. We show that the former is the superior translation speed predictor because it accelerates translation rate beyond that of the codon bias recoded mRNA. Furthermore, rate acceleration decreases the folding efficiency of the recoded gene products. This prompted us to construct an mRNA by harmonizing the natural variations in speed (observed in our translation speed profiles) along the mRNA sequence with the tRNA pools in the host, *E. coli,* which resulted in increased folding efficiency of the recoded gene product.

Results

Patterns of tRNA gene content differ significantly among the three domains of life.

To gain insight into the extent by which different organisms utilize different sets of tRNAs during protein synthesis, we conducted an analysis of the current version of the Genomic tRNA Database (GtRNAdb) (Chan & Lowe, 2009), a manually curated database documenting the predicted number of genes for each tRNA isoacceptor for a large number of organisms whose genomes have been sequenced. Upon examination of organism-specific tRNA gene content for all genera available in the database (35 archaea, 223 bacteria and 35 eukaryotes), we observed striking differences in the pattern of tRNA genes present in the genomes of organisms belonging to each domain of life (Fig. 13). The distribution of tRNA genes for most synonymous codons within a Domain tends to be rather constant, but clear differences arise when comparisons are made across the three Domains. For instance, in the case of isoleucine (encoded by AUU, AUC and AUA), most bacteria have tRNA genes that decode AUC, and none that decode AUU. In eukaryotes, the situation is reversed: most have tRNA genes for AUU and only very few have genes for AUC (only a small fraction of organisms in all three domains have tRNA genes that decode AUA). In other cases, the tRNA gene is present in a considerable fraction of eukaryotic genomes, yet completely absent in bacterial genomes (for example, GUU, CCU, CUU, UCU, ACU and GCU). In yet other cases, all three domains appear to contain mostly the same tRNA genes for a particular isoacceptor, especially for amino acids with only two synonymous codons (for example, UAC, CAC, AAC, GAC and UGC), but even in these cases, a minor fraction of eukaryotic genomes may contain “rare tRNA genes” that decode the other isoacceptor (UAU, CAU, AAU, GAU and UGU), whereas *no* bacterial genomes appear to possess them.

It is important to emphasize at this point that absence of tRNA genes that decode a particular codon in a given organism is not correlated with the absence (or underrepresentation) of that codon in the protein coding sequences of that organism. In other words, all cellular organisms utilize all 61 codons to encode proteins, even though certain tRNA genes are missing in every genome analyzed to date. In fact, we have observed in our analyses that, in every organism, some of the most frequent codons have no matching tRNA isoacceptor genes. When such a codon for which there is no tRNA isoacceptor gene(s) is encountered by the ribosome, it is decoded by a tRNA that base pairs to it *via* non-Watson-Crick interactions (*i.e.*, by wobble). For certain codons, it has been shown experimentally that such non-Watson-Crick codon-anticodon interactions result in decreased elongation rates compared to decoding *via* strict Watson-Crick binding (Curran & Yarus, 1989; Sorensen & Pedersen, 1991).

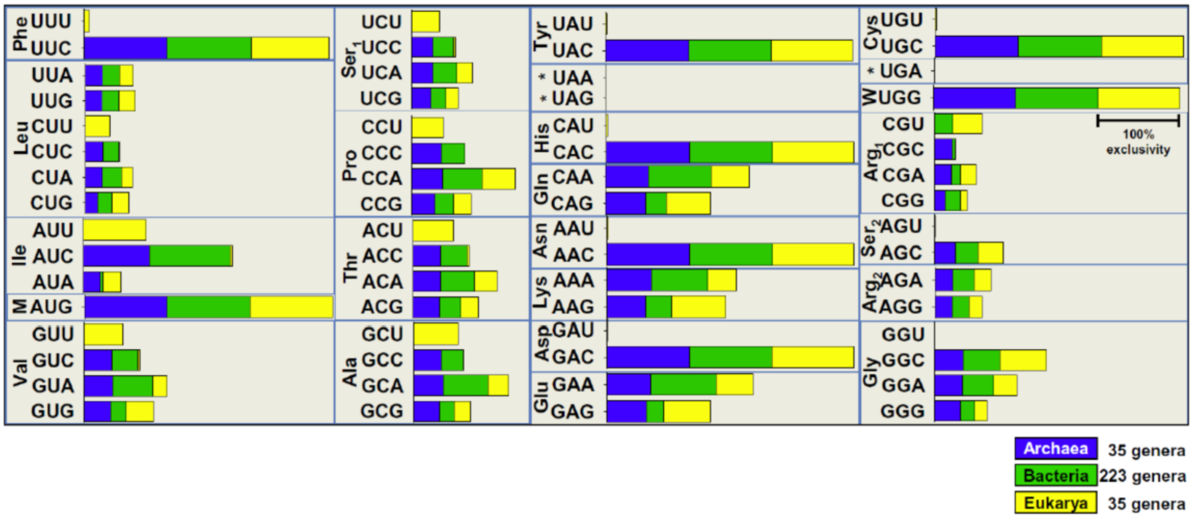


Figure 13. The distribution of genes encoding tRNAs of different decoding capacities vary among archaea, bacteria and eukarya. Predicted gene content for tRNAs capable of decoding the standard genetic code according to GtRNAdb (Chan & Lowe, 2009) is plotted for each codon in histogram form (as indicated) by each domain of life in different colors (as indicated). The length of each box represents the extent to which genes for tRNAs capable of decoding the corresponding codon are present in a domain. For example, for Ala, no eukaryotic genera examined contain tRNA genes capable of decoding GCC, whereas ~60%, ~25% and ~15% of them contain tRNA genes to decode GCU, GCA and GCG, respectively. For Met or Trp, 100% of genera examined in each domain are predicted to contain a single species of tRNA genes to decode these codons (and thus the length of these bars corresponds to "100% exclusivity").

Since both eukaryotes and bacteria lack tRNAs for a significant number of codons (Fig. 13) (and the abundances of those present vary substantially (Chan & Lowe, 2009; Sorensen & Pedersen, 1991), it is likely that the non-uniform movement of the ribosome along an mRNA will be considerably influenced by the pool of available tRNAs in each organism. Thus, it may be possible that, although bacterial ribosomes are entirely capable of decoding the genetic information of eukaryotes, their different patterns of tRNA availability may lead to differences in the rates at which various segments of the polypeptide emerge from the ribosome. Such variations in the rates of appearance of segments of the polypeptide that are critical for folding may contribute to the often observed misfolding of eukaryotic proteins upon production in bacteria. For example, a subtle increase in the concentration of a partially folded intermediate during translation of its polypeptide sequence may exceed the critical concentration of the intermediate and lead to its nucleation-dependent aggregation, thus forming intracellular aggregates. In order to explore these differences, we sought to develop a formula that would allow us to predict the relative polypeptide elongation rates along a given mRNA on any expression host whose genome has been annotated with respect to tRNA gene content.

Prediction of relative polypeptide elongation rates based on expression host tRNA availability

We wished to develop a metric to assess the influence of the different patterns of tRNA availability of bacteria and eukaryotes on the relative rates of emergence of the nascent polypeptide. Our metric (see Materials and Methods, p. 55) generates a relative speed value for each codon along an mRNA molecule based on whether the cognate Watson-Crick tRNA isoacceptor is present for that codon in the expression host, whether non-Watson-Crick tRNA isoacceptors capable of decoding that codon are present as well as the number of tRNA genes that fulfill one and/or both of the above conditions. As mentioned previously, it has been shown that codons differing only in the wobble position are translated by the same tRNA species at different rates. We utilized the experimentally determined translation rates of 31 individual codons (Curran & Yarus, 1989; Sorensen & Pedersen, 1991) and current knowledge of the tRNAs responsible for decoding them to calculate a general ratio of wobble-based decoding to Watson-Crick-based decoding (see Materials and Methods, p.55). The relative speed values thus obtained for each codon are then averaged over a sliding window of 30 codons (which corresponds to the number of amino acid residues the ribosomal exit tunnel can accommodate (Ban et al, 2000; Harms et al, 2001). These values, which we have termed “translation speed index” are plotted against codon position to generate a profile that depicts the predicted variations in polypeptide elongation rates based on tRNA availability of a given expression host (Fig. 14). Regions of high relative speed value (or translation speed index) predict a faster polypeptide elongation rate compared to regions of lower translation speed value. Due to the similarities in tRNA gene content among eukaryotes, such profiles should be similar for a eukaryotic coding sequence translated by another eukaryote but different when translated by a bacterium. When we examine the sequence that encodes the enzyme luciferase from the firefly *Photinus pyralis*, a model protein whose folding behavior has been previously characterized in our laboratory (Agashe et al, 2004; Kaiser et al, 2006; Siller et al, 2010) we see that the predicted differential translation speed profiles are indeed similar between an insect (*D. melanogaster*) and a yeast (*S. cerevisiae*), but different in the bacterium *E. coli* (Fig. 14). Accordingly, we have found that luciferase folds well when recombinantly produced in yeast, but not in bacteria (Agashe et al, 2004; Siller et al, 2010). Interestingly, the predicted region of

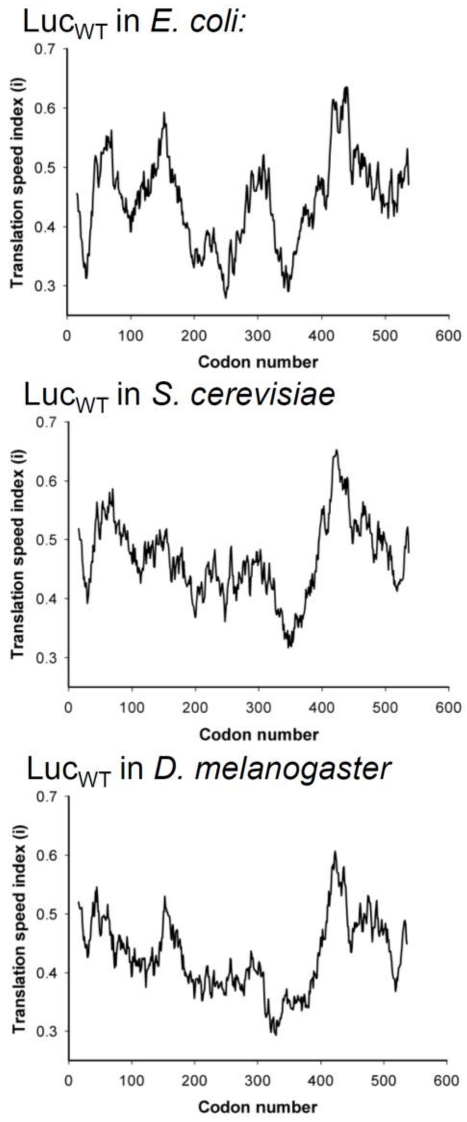


Figure 14. Relative polypeptide elongation rates can be predicted for any mRNA based on the tRNA gene content of the expression host. Plots depicting predicted translation speed indices (see main text and Materials and methods), calculated for the sequence encoding firefly luciferase, utilizing the tRNA gene content of the organisms indicated obtained from GtRNAdb (Chan & Lowe, 2009). Regions with high *i* values are predicted to be translated rapidly, whereas regions with lower *i* values are predicted to be translated more slowly.

fastest translation speed for luciferase in the eukaryotic profiles correlates well with the presence the C-terminal domain (residues 437 – 544), a topologically independent structural domain of the enzyme (Conti et al, 1996).

Polypeptide elongation rates can be predictably accelerated by manipulating wobble base composition.

We next sought to determine whether the predicted different rates by which individual codons are decoded (depending on whether they are read by tRNAs capable of binding *via* Watson-Crick *vs.* non-Watson-Crick interactions) were of sufficient magnitude to affect *overall* ribosome movement along an mRNA molecule *in vivo*. We began by asking whether a sequence whose codons were decoded exclusively by tRNAs pairing *via* Watson-Crick interactions would be translated at observably faster rates than the original wild type sequence. In both cases, the actual amino acid sequences emerging from the ribosome are identical. We employed DNA synthesis to build a bacterial expression construct for the model protein firefly luciferase in which every amino acid is encoded by a synonymous codon read by the tRNA species with the highest number of tRNA genes in *E. coli* (Fig. 15). This synthetic construct (termed Luc$_{\text{fast}}$) and the original luciferase sequence (Luc$_{\text{WT}}$) were placed under control of identical regulatory sequences (T7-driven promoter and terminator) for expression in *E. coli* and their respective mRNAs accumulated to similar levels (Fig. 16).

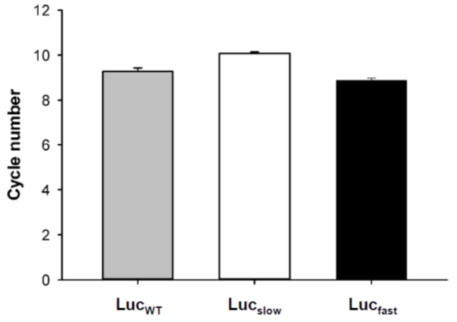


Figure 16. Steady-state accumulation of mRNA synthesized from the wild type and sequence-engineered constructs. Histogram depicting the results of a quantitative reverse transcriptase PCR reaction to evaluate the levels of accumulation of mRNA produced from each of the indicated constructs. Error bars represent standard errors of the mean.

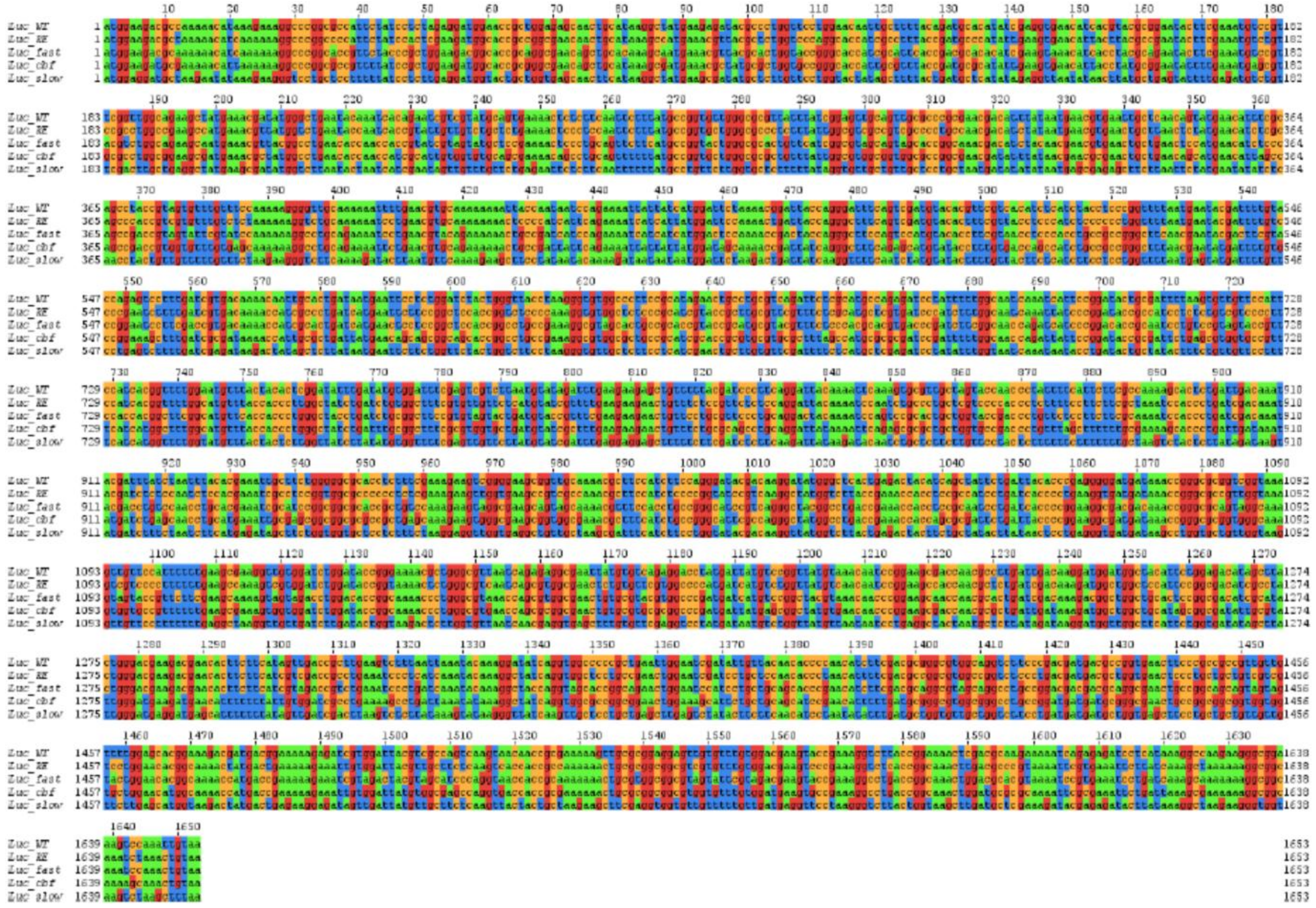


Figure 15. Coding sequences of the firefly luciferase constructs utilized in this study. Multiple sequence alignment of the nucleotide sequences of the wild type and the various sequence-engineered constructs of firefly luciferase synthesized for this study. Each nucleotide has been placed in a box of different color to facilitate visual inspection of similarities and differences across sequences.

We then proceeded to determine their polypeptide elongation rates by performing pulse-chase analyses in live *E. coli* cells (Materials and Methods). We found that, indeed, luciferase protein synthesis was clearly accelerated in cells harboring the Luc_{fast} construct compared to those harboring the Luc_{WT} plasmid (Fig. 17a). In order to obtain a quantitative idea of the magnitude of the observed rate acceleration, we generated simulated curves of the calculated rates of appearance of methionine incorporated into full length firefly luciferase at various predicted average polypeptide elongation rates (Fig. 17b) (Sorensen & Pedersen, 1991) (Materials and Methods).

As can be observed, the rate of full length protein appearance from Luc$_{WT}$ most closely fits the theoretical curve corresponding to 10 amino acids *per* second (aa/s), whereas that produced from Luc$_{fast}$ clearly approaches 20 aa/s.

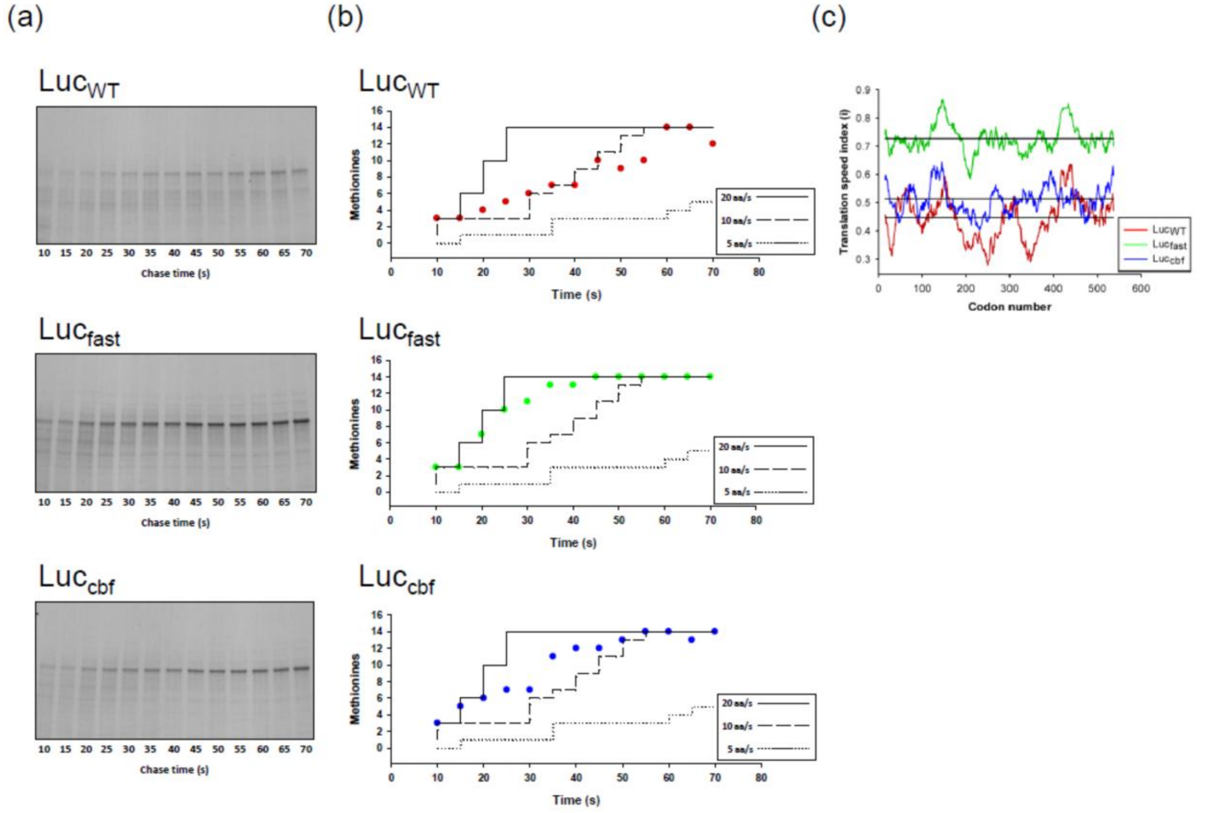
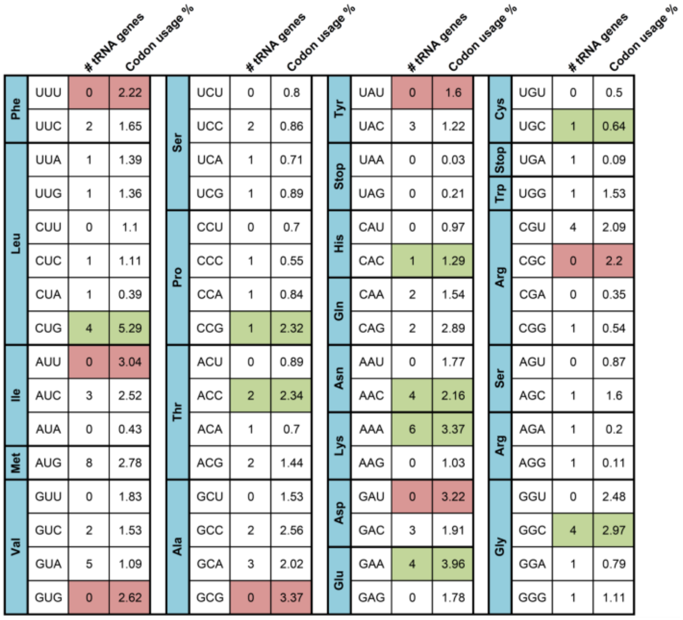


Figure 17. Translation rates can be accelerated by engineering a sequence to contain only codons predicted to be decoded by abundant tRNAs. (**a**) Autoradiograms of SDS-PAGE gels from pulse-chase experiments of live *E. coli* cells synthesizing recombinant firefly luciferase from the indicated sequence-engineered constructs (see main text and Materials and methods). (**b**) Plots depicting the appearance of incorporated [^{35}S]methionine in full length firefly luciferase produced from the indicated constructs (colored dots; values obtained by denistometric analysis of the data in **a**) and curves for the theoretical appearance of methionine with three calculated average translation rates, as indicated (full, broken and dotted lines). (**c**) Plot of the predicted polypeptide elongation rates for luciferase synthesized from the constructs indicated, calculated as in Figure 2 (see main text and Materials and methods). Straight broken lines represent the average predicted translation rates over the entire sequence (*avg.*), as indicated.

Previously, predictions of the speed at which codons are translated have been based on their frequency of occurrence in a given set of coding sequences in a given organism (Gupta et al, 2000; Sharp & Li, 1987; Thanaraj & Argos, 1996a; Thanaraj & Argos, 1996b). In this so-called *biased codon usage* (or codon bias) (Comeron & Aguade, 1998; Grantham et al, 1980; Ikemura, 1985; Lynn et al, 2002), frequent codons have traditionally been considered fast, while rare ones have been predicted to be translated more slowly. We next considered whether sequence engineering based on this metric would also lead to rate acceleration. We designed a luciferase sequence composed entirely of the most frequently used codons in *E. coli*, regardless of the number of tRNA genes associated with those codons (termed Luc$_{cbf}$) (Chan & Lowe, 2009; Grantham et al, 1980; Sharp & Li, 1987). Pulse-chase analysis revealed that protein production rates from the Luc$_{cbf}$ plasmid was intermediate between those of Luc$_{fast}$ and Luc$_{WT}$ (Fig. 17a). Since a considerable fraction of codons predicted to be translated fast by codon usage bias criteria correspond to the codons for which the highest number of tRNA genes exist in *E. coli* (Fig. 18), it is not surprising that the luciferase produced from Luc$_{cbf}$ accumulated with faster rates than that produced from Luc$_{WT}$ and probably occurred as a result of the over representation of those codons. Indeed, predictions based on our metric suggested that Luc$_{cbf}$ would be translated with rates intermediate between those of Luc$_{WT}$ and Luc$_{fast}$ (Fig. 17c).



		# tRNA genes	Codon usage %			# tRNA genes	Codon usage %			# tRNA genes	Codon usage %			# tRNA genes	Codon usage %
Phe	UUU	0	2.22	Ser	UCU	0	0.8	Tyr	UAU	0	1.6	Cys	UGU	0	0.5
	UUC	2	1.65		UCC	2	0.86		UAC	3	1.22		UGC	1	0.64
Leu	UUA	1	1.39		UCA	1	0.71	Stop	UAA	0	0.03	Stop	UGA	1	0.09
	UUG	1	1.36		UCG	1	0.89		UAG	0	0.21	Trp	UGG	1	1.53
	CUU	0	1.1	Pro	CCU	0	0.7	His	CAU	0	0.97	Arg	CGU	4	2.09
	CUC	1	1.11		CCC	1	0.55		CAC	1	1.29		CGC	0	2.2
	CUA	1	0.39		CCA	1	0.84	Gln	CAA	2	1.54		CGA	0	0.35
	CUG	4	5.29		CCG	1	2.32		CAG	2	2.89		CGG	1	0.54
Ile	AUU	0	3.04	Thr	ACU	0	0.89	Asn	AAU	0	1.77	Ser	AGU	0	0.87
	AUC	3	2.52		ACC	2	2.34		AAC	4	2.16		AGC	1	1.6
	AUA	0	0.43		ACA	1	0.7	Lys	AAA	6	3.37	Arg	AGA	1	0.2
Met	AUG	8	2.78		ACG	2	1.44		AAG	0	1.03		AGG	1	0.11
Val	GUU	0	1.83	Ala	GCU	0	1.53	Asp	GAU	0	3.22	Gly	GGU	0	2.48
	GUC	2	1.53		GCC	2	2.56		GAC	3	1.91		GGC	4	2.97
	GUA	5	1.09		GCA	3	2.02	Glu	GAA	4	3.96		GGA	1	0.79
	GUG	0	2.62		GCG	0	3.37		GAG	0	1.78		GGG	1	1.11

Figure 18. tRNA gene content and biased codon usage values for E. coli. Table of data obtained from (Chan & Lowe, 2009) depicting the number of tRNAs capable of decoding each codon as well as the codon usage frequency of each codon for the bacterium *E. coli*. Boxes shaded in green indicate instances where the most frequent codon coincides with the highest number of tRNA genes for that codon. Boxes shaded in red indicate instances where the most frequent codon has no tRNA genes available for strict Watson-Crick decoding.

Next, we wished to assess whether we could employ reasoning similar to the above to engineer a sequence that would be translated more slowly. Thus, we synthesized a luciferase construct composed of codons relying solely on non-Watson-Crick decoding tRNAs for their translation (except for codons encoding Met, Trp and Gln, see figure 18), which we termed Luc_{slow}. This construct was placed under regulatory sequences identical to those described above. Although we could detect luciferase activity in cells harboring this plasmid, we could not accurately measure accumulation of full length protein in our pulse-chase analyses, precluding determination of polypeptide elongation rates for protein produced from this construct (Fig. 19). It is probable that such pronounced frequency of codons relying on wobble tRNA interactions for decoding led to marked ribosomal stalling, which resulted in sequestration of ribosomes and the consequent activation of cellular mechanisms to rescue such ribosomes (Liu et al, 2010; Seidman et al, 2011), leading to very little production of full length protein. Regardless, we believe that our results, taken together, suggest that predictions based on tRNA availability (based for example on the presence and number of tRNA genes) rather than biased codon usage might yield more accurate results regarding the translation rates of synonymous codons, at least in *E. coli*.

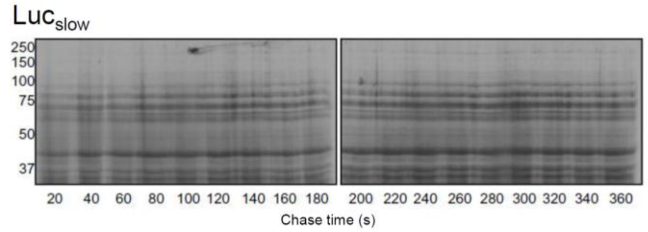


Figure 19. Autoradiogram of an SDS-PAGE from a pulse-chase experiment with the Lucsslow construct. The experiment was carried out as described in the main text, except that aliquots were taken for considerably longer times, as indicated.

Translation initiation does not play a significant role in sequence-based acceleration.

It is well known that nucleotide composition can influence mRNA secondary structure (Mathews et al, 2007; Zuker, 2003) and that secondary structural elements in regions at and/or near the ribosomal binding and translation initiation sites can significantly affect translation initiation rates (Kudla et al, 2009; Plotkin & Kudla, 2011). Although all our constructs contained identical ribosomal binding sites and their mRNA stabilities around critical translation initiation sites were similar, we nevertheless wished to ensure that changes in translation initiation were not responsible for our observed effects on translation acceleration. We engineered a set of sequences in which the first 50 codons were identical among themselves (derived from the Luc_{WT} sequence, to yield $\text{Luc}_{\text{WT-fast}}$ and $\text{Luc}_{\text{WT-cbf}}$) and conducted experiments in the same manner as before. As can be observed (Fig. 20b), $\text{Luc}_{\text{WT-fast}}$ and the $\text{Luc}_{\text{WT-cbf}}$ lead to production of

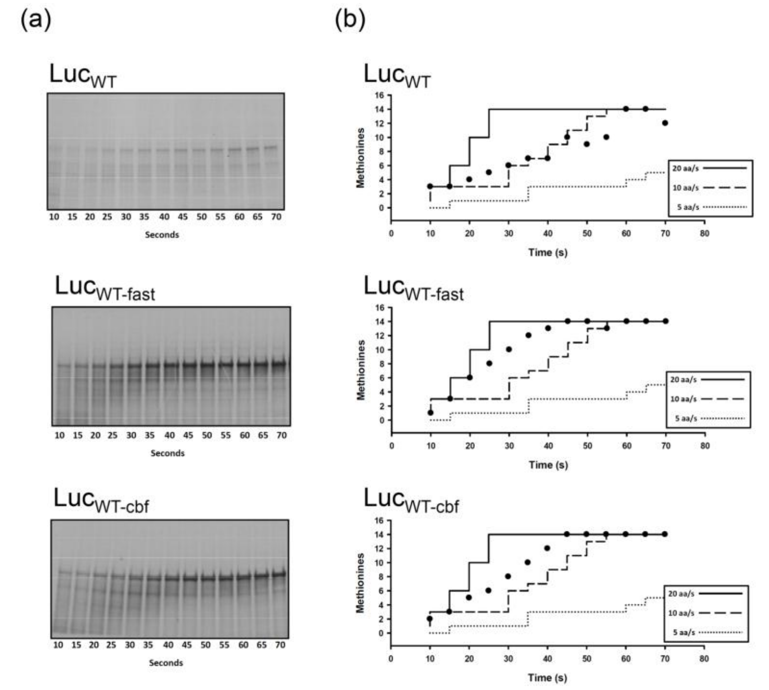


Figure 20. Sequence-based translation acceleration is not due to changes in initiation rates. (**a**) Autoradiograms of SDS-PAGE gels from pulse-chase experiments of live *E. coli* cells synthesizing recombinant firefly luciferase from the indicated sequence-engineered constructs (see main text and Materials and methods). (**b**) Plots depicting the appearance of incorporated [^{35}S]methionine in full length firefly luciferase produced from the indicated constructs (filled dots; values obtained by denistometric analysis of the data in **a**) and curves for the theoretical appearance of methionine with three calculated average translation rates, as indicated (full, broken and dotted lines).

full length proteins with accelerated rates similar to those of their Luc_{fast} and Luc_{cbf} counterparts (Fig. 17a). Thus, the presence of wild type translation initiation sites does not affect the overall effects on rate acceleration conferred to by the rest of the sequences. We believe that changes in mRNA secondary structure throughout the sequence are unlikely to mediate the observed effects, as the Luc_{fast} construct actually contains a higher GC content (54%) than Luc_{WT} (45%). Thus, even though Luc_{fast} might contain more stable secondary structural elements (which could provide an impediment to ribosomal movement (Qu et al, 2011), we nevertheless observe a significant rate acceleration, which argues for the robustness of this sequence manipulation and suggests it is due primarily to an effect on polypeptide elongation.

Acceleration of translation rates by synonymous codon substitutions impacts the folding of the encoded polypeptide.

We have previously utilized *E. coli* strains harboring mutant ribosomes that can translate at variable rates, depending whether the antibiotic streptomycin is absent (~5aa/s) or present (~11 aa/s) (Ruusala et al, 1984; Zengel et al, 1977) to investigate the effects of translation rates on protein folding efficiencies (Siller et al, 2010). We have shown that the folding efficiency of firefly luciferase (and several other recombinant proteins of eukaryotic origin) increases about two-fold when produced by ribosomes translating at slower rates (Siller et al, 2010). Our coding sequence-based manipulations described above now allowed us to test whether further increases in polypeptide elongation rates beyond those observed under wild type conditions had an impact on folding efficiencies.

To elucidate the effect of rate acceleration on folding efficiency, we expressed our set of sequence-engineered luciferase constructs, determined the accumulation total (folded and misfolded) recombinant protein produced, and measured luciferase activity for each, as an indication of acquisition of the native state (Fig. 21) (Materials and Methods). It is important to note here that the amino acid sequences among all these sequence-engineered constructs are predicted to be identical, as all manipulations involved synonymous substitutions. As can be observed, at very similar levels of total recombinant protein accumulation, the protein produced from the Luc$_{fast}$ construct displayed remarkably lower levels of activity compared to that produced from the Luc$_{WT}$ plasmid (Fig. 21a). When specific activities are obtained by dividing total luciferase activity over total amount of protein production, it can be seen that acceleration results in 10-fold decrease in folding efficiency (Fig. 21b). Consistent with the results presented above,

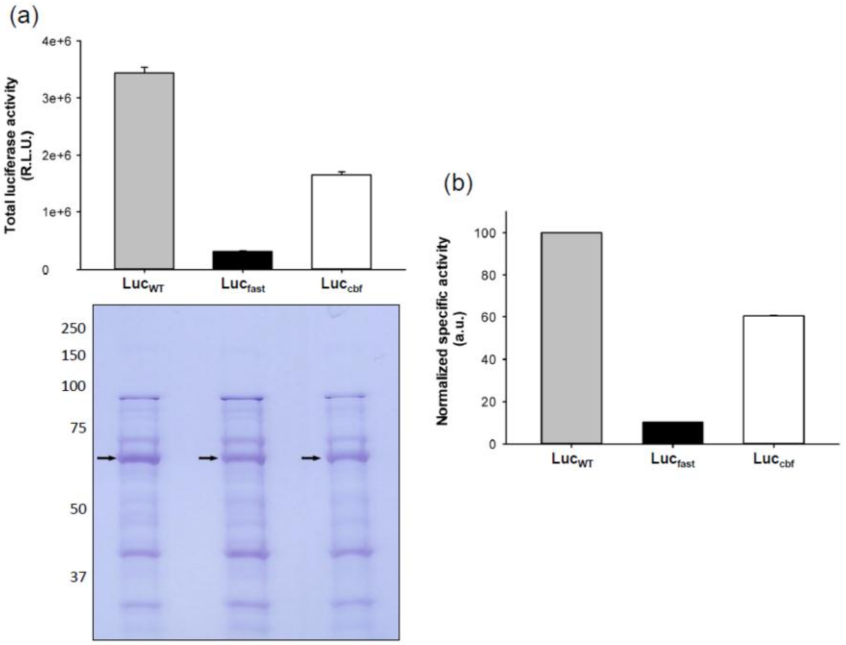


Figure 21. Sequence-based Acceleration of translation rates affects the folding of the encoded polypeptide. (**a**) Histogram depicting total firefly luciferase activity (top panel) and SDS-PAGE documenting total full-length recombinant protein (bottom panel) produced in *E. coli* from the indicated sequence-engineered constructs. (**b**) Histogram depicting the specific activities of the luciferase protein produced from the indicated constructs, obtained by dividing the values in the top panel of figure **a** over the amount of full-length protein shown in the top panel (measured by densitometric analysis) and setting the value of the protein from the wild-type sequence to 100%. R.L.U.: relative light units. a.u.: arbitrary units. Error bars represent standard errors of the mean.

the luciferase translated from the Luc_{cbf} construct exhibited an intermediate degree of folding efficiency. Thus, it appears that, at least for firefly luciferase, increments in overall polypeptide elongation rates correlate with decrements in folding efficiency, consistent with our previous results (Siller et al, 2010).

Translation speed harmonization leads to enhanced folding efficiency of identical polypeptides.

Our synonymous codon substitutions have allowed us to uncover principles by which sequence composition can affect polypeptide elongation rates and folding. However, *in vivo*, messages are not uniformly slow or fast (Varenne et al, 1984), (Fig. 14). Rather, the ribosome alternatively accelerates and decelerates as it moves along a given mRNA, which is presumably reflected by the peaks and valleys in our profiles (Fig. 14), corresponding to the presence of clusters of faster and slower codons, respectively. We next asked whether recreating these naturally occurring variations in translation speed by implementing the synonymous sequence-based manipulations described above might enhance the folding efficiency of eukaryotic proteins produced in *E. coli*. Because, as shown above (Fig. 13) and previously (Grosjean et al, 2010) tRNA gene content differs between bacteria and eukaryotes, a codon with abundant tRNA content in the firefly (predicted to be a “fast” codon by our metric) may correspond to a codon lacking strict Watson-Crick-decoding tRNA genes in *E. coli* (*i.e.*, predicted to be a “slow” codon in *E. coli*). We thus employed a *harmonization* strategy to recode the firefly luciferase sequence in which codons predicted to be translated fast in the fruit fly *D. melanogaster* (a species evolutionarily close to the firefly whose entire tRNA gene set is known) were matched with synonymous codons predicted to be translated fast in *E. coli*. Conversely, codons predicted to be translated by non-Watson-Crick tRNAs in the fruit fly were matched by codons with no matching tRNA genes in *E. coli* (Materials and Methods; supplementary information). The resulting construct (Luc$_{RE}$) was utilized to produce recombinant luciferase in *E. coli* cells. Although the predicted average translation speed of the Luc$_{RE}$ sequence is very similar to that of Luc$_{WT}$, the luciferase protein produced from the former folded with higher efficiency (Fig. 22). As can be observed, even under strong induction conditions, cell viability was unaffected and total luciferase activity was consistently higher upon induction (Fig. 21). Significantly, at equivalent amounts of total recombinant protein, luciferase produced from Luc$_{RE}$ consistently displayed about a three-fold higher activity. Thus, it appears that subtle manipulations of ribosome movement along a recombinant mRNA molecule that mimic its movement in the original host appear to be a robust method to improve the folding efficiency of certain

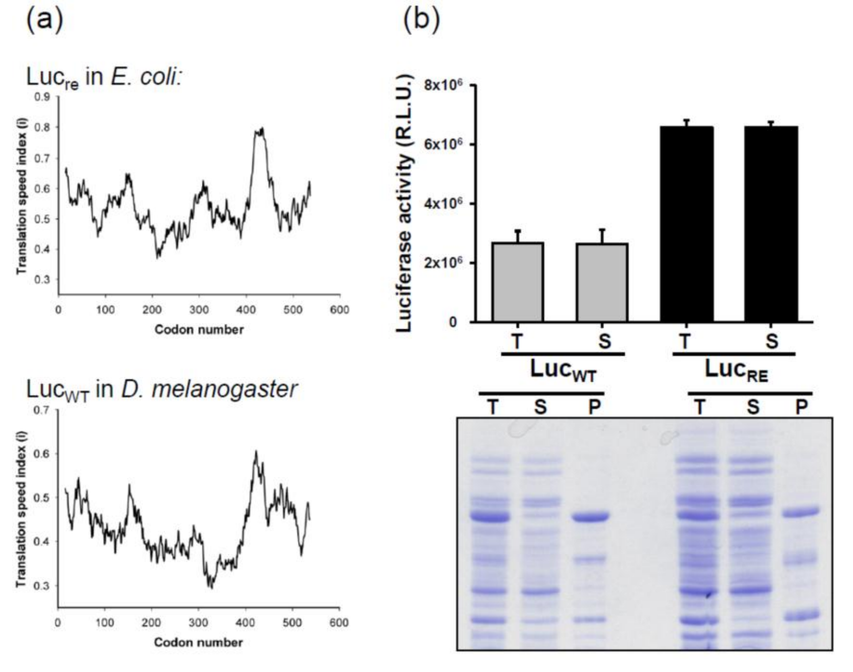


Figure 22. Sequence harmonization based on tRNA gene content between the original and the expression host increases folding efficiency. (**a**) Plots depicting predicted translation speed indices (see main text and Materials and methods), calculated for firefly luciferase translated from a harmonized redesigned sequence (Luc$_{RE}$) upon expression in E. coli and for the original (Luc$_{WT}$) sequence expressed in *D. melanogaster* (see main text and Materials and methods). (**b**) Histogram depicting firefly luciferase activity (top panel) and SDS-PAGE documenting total (T), soluble (S) and insoluble (P) recombinant protein (bottom panel) produced in *E. coli* from the indicated sequence-engineered constructs. R.L.U.: relative light units. Bars represent standard errors of the mean.

eukaryotic proteins.

Discussion

Multiple previous studies have explored the notion that the observed non-uniform distribution of codons among protein coding sequences is associated in some manner with folding events of the encoded polypeptide (Adzhubei et al, 1996; Makhoul & Trifonov, 2002; Zhou et al, 2009). These studies have reported a wide variety of conclusions, probably due in part to the fact that predictions of the speed at which particular codons are translated have been based on different sets of data, including statistical analyses of codon frequencies among highly expressed genes, experimentally determined concentrations of decoding tRNAs, *etc.* (Sharp & Li, 1987; Zhang et al, 2009). Pioneering work on determining translation rates in vivo demonstrated that codons that are decoded by tRNAs capable of making Watson-Crick interactions are generally translated faster than those depending on tRNAs only binding to the codon *via* non-Watson-Crick interactions (Curran & Yarus, 1989; Sorensen & Pedersen, 1991). In the present study, we analyzed a large set of data containing predicted numbers of tRNA genes from fully sequenced genomes across the domains of life and observed striking differences in the distribution of tRNA genes capable of decoding codons from synonymous groups between bacteria and eukaryotes. As it is well known that proteins of eukaryotic origin often misfold upon recombinant production in bacteria, we decided to investigate whether these differences in tRNA pools across the domains of life could perhaps contribute to this phenomenon. Thus, we began by developing a strategy to predict translation rates based on whether a codon along an mRNA in a given organism is decoded by tRNAs capable of Watson-Crick, non-Watson-

Crick or both types of interactions. Our calculations took into consideration the number of tRNA genes capable of decoding a given set of codons in each particular organism, which has been shown to correlate well with actual tRNA concentrations (Kanaya et al, 1999). Based on these considerations, we were able to predictable engineer sequences to be translated faster, by selecting only codons decoded by Watson-Crick type interactions and demonstrated that these sequences are indeed translated approximately twice as fast as the non-engineered sequence by performing pulse-chase analyses in living *E. coli* cells. In agreement with previous results from our laboratory (Siller et al, 2010), sequences encoding proteins of eukaryotic origin that were translated at faster rates led to the production of polypeptides that folded with decreased efficiencies. We utilized the knowledge of the available pool of tRNAs in *E. coli* (based on its tRNA gene content) to redesign the sequence of the model protein firefly luciferase (which folds poorly in bacteria) in an attempt to compel the *E. coli* ribosome to move at similar segmental rates along the mRNA that the original eukaryotic ribosome would in the original host. This sequence led to a significant increase in the folding efficiency of this model protein, as reflected by its increased solubility and specific activity upon production in *E. coli*.

How can subtle differences polypeptide elongation rates impact the folding of the polypeptide emerging from the ribosome? Although 2-3 fold differences in the rates of ordinary reactions might not be generally considered significant from a chemical kinetics point of view, a 2-3 fold difference in the rate of synthesis of a protein may have profound biological consequences. For example, a subtle increase in the concentration of a partially folded, aggregation-prone polypeptide intermediate during translation may exceed the critical concentration of the intermediate and lead to its nucleation-dependent aggregation, thus forming intracellular aggregates. In essence, our findings that variations in translation rates impact protein folding support the notion that not all proteins fold globally, but rather follow particular pathways throughout the available structural space, influenced by the speed at which they emerge vectorially from the ribosome. This idea may find applications in a variety of fields and settings, including improvements in the production of recalcitrant proteins for vaccine development, recombinant pharmaceuticals and structure-determination studies. Moreover, these results may provide further insight into how so-called “silent” polymorphisms may result in human disease (Plotkin & Kudla, 2011) and on how physiological and disease-related variations in tRNA concentrations impact cellular proteostasis, critical in a wide variety of oncologic, cardiovascular and neurodegenerative disorders (Dittmar et al, 2006; Pavon-Eternod et al, 2009; Prudencio et al, 2010).

Materials and Methods

Prediction of polypeptide elongation rates. In order to assign a *translation speed index* (*i*) to each of the 61 codons in a given organism, the following rules were assigned regarding the nature codon($N_1N_2N_3$):anticodon($N_{34}N_{35}N_{36}$) interactions (where $N_1N_2N_3$ represents each codon along the 5' → 3' direction in an mRNA and $N_{34}N_{35}N_{36}$ represents the 5 → 3' anticodon loop of the decoding tRNA): (1) Watson-Crick interactions are allowed to occur between $N_1N_2G_3$:$C_{34}N_{35}N_{36}$, $N_1N_2C_3$:$G_{34}N_{35}N_{36}$, $N_1N_2A_3$:$U_{34}N_{35}N_{36}$, $N_1N_2U_3$:$A_{34}N_{35}N_{36}$ and $N_1N_2C_3$:$I_{34}N_{35}N_{36}$ (where I_{34} represents inosine, derived from post-transcriptional deamination of some A_{34}-bearing tRNAs); (2) non-Watson-Crick interactions are allowed to occur between $N_1N_2G_3$:$U_{34}N_{35}N_{36}$, $N_1N_2U_3$:$G_{34}N_{35}N_{36}$, $N_1N_2U_3$:$I_{34}N_{35}N_{36}$, and $N_1N_2A_3$:$I_{34}N_{35}N_{36}$. Inosination was assumed to occur for all A_{34}-bearing tRNAs in eukaryotes and for A_{34}-bearing tRNAs that decode Arg codons in bacteria (Grosjean et al, 2010). Since a $U_{34}A_{35}C_{36}$-bearing species of tRNA is generally utilized to decode AUA codons in bacteria, it was assumed that a $U_{34}A_{35}C_{36}$-bearing tRNAs would partition equally for decoding AUG and AUA codons. In order to obtain normalized values for tRNA gene abundances across organisms for each codon, we divided the number of tRNA genes for every codon by the total number of tRNA genes in the respective synonymous codon group. These values (termed $NNN_{\%}$ for each codon) were then utilized, according to the above assumptions, to calculate a translation speed index (*i*) for each codon (termed NNN_i) in a given organism according to the following formulas (where *w* is a "penalizing" factor for non-Watson-Crick interactions; in this study, *w* = 3 for all such interactions, as these have been experimentally shown to result in ~3-fold slower polypeptide elongation rates

(Sorensen & Pedersen, 1991): (1) For all bacterial codons (except those for Ile, Met and Arg): NNU_i = $NNU_{\%}$+$NNC_{\%}/w$; NNC_i = $NNC_{\%}$; NNA_i = $NNA_{\%}$; NNG_i = $NNA_{\%}/w$. (2) For bacterial Ile: AUU_i = $AUU_{\%}$ + $AUC_{\%}/w$; AUC_i = $AUC_{\%}$ and AUA_i = $AUG_{\%}/w$*2. (3) For bacterial Met: AUG_i = $AUG_{\%}/2$. (4) For bacterial Arg: treat as a eukaryotic Arg. (5) For eukaryotic two-codon groups and both similar codons of six-codon gropus: NNU_i = $NNU_{\%}$ + $NNC_{\%}/w$; NNC_i = $NNC_{\%}$ + $NNU_{\%}$; NNA_i = $NNA_{\%}$; NNG_i = $NNG_{\%}$ + $NNA_{\%}/w$. (6) For eukaryotic four-codon groups, the four similar codons of six-codon groups and Ile: NNU_i = $NNU_{\%}/w$ + $NNC_{\%}/w$; NNC_i = $NNC_{\%}$ + $NNU_{\%}$; NNA_i = $NNA_{\%}$ + $NNU_{\%}/w$; NNG_i = $NNG_{\%}$ + $NNA_{\%}/w$. Once these values were obtained for each organism, they were assigned to the corresponding codons of any protein coding sequence. From the start of the coding sequence, *i* values of 30 consecutive codons were added and the average value plotted at position number 15. The same operation was performed repeatedly by sliding the window of 30 values one codon position at a time, until of the coding sequence was reached. The resulting *i* values were plotted as a function of codon position.

Coding sequence engineering. Luciferase mRNAs were engineered as follows: for sequences to be translated slowly (Luc_{slow}), codons that lack isoacceptor tRNA genes in *E. coli* were selected for each amino acid, with the exception of methionine and tryptophan. If genes for all the anticodons of a particular amino acid are present, then the codon with the least amount of available anticodon interactions at the wobble position was selected. For sequences to be translated faster (Luc_{fast}), codons with the highest number of isoacceptor tRNA genes were selected. In cases were more than one codon had the highest number of tRNA genes, the codon with the most amount of available anticodon interactions at the wobble position was selected. Similarly, Luc_{cbf} was designed to harbor codons for that are the most frequently used in *E. coli*. Sequences for $Luc_{WT\text{-}fast}$ and $Luc_{WT\text{-}cbf}$ contain the nucleotides 1-50 from Luc_{WT} and the remaining nucleotides from Luc_{fast} and Luc_{cbf} respectively.

Strains and growth conditions. The *E. coli* utilized here was BL21 (New England Biolabs). For recombinant protein production (see below), this strain was transformed with the following β-gal-controlled promoter-based plasmids: $pLuc_{WT}$ (encoding Luc_{WT} with a C-terminal c-myc-His_6 epitope tag) (Agashe et al, 2004), $pLuc_{slow}$, $pLuc_{fast}$, $pLuc_{cbf}$, $pLuc_{WT\text{-}fast}$ and $pLuc_{WT\text{-}cbf}$. For activity measurements, cells were grown in LB broth at 37 °C with 250 rpm orbital shaking in volumes that occupied at most one fourth of the total vessel volume, in the presence of ampicillin (100 µg/ml). For the experiments in Fig. 5, volumes of cells containing equivalent A_{600} values were harvested by centrifugation and lysed by spheroplasting (Ausubel et al, 2003) under native conditions (Chang et al, 2005). Similar amounts of total protein in the resulting lysates were verified by the Bradford assay (BioRad). For pulse-chase analysis, cells were grown in a methionine free defined medium (Teknova) at 37 °C with 250 rpm orbital shaking in volumes that occupied at most one fourth of the total vessel volume, in the presence of ampicillin (100 µg/ml).

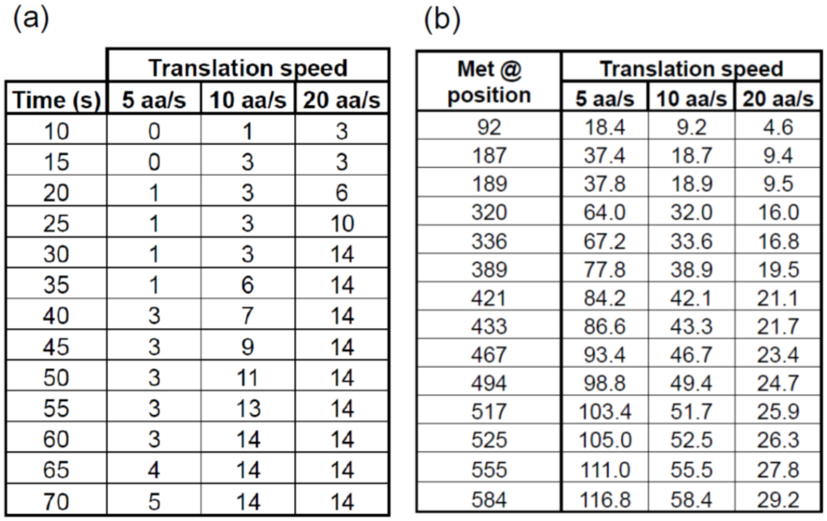
Recombinant protein production. Starter cultures were grown overnight as described above and diluted the next day. Protein expression was induced at A_{600} = 0.4 with 1 mM IPTG and harvested at 10 min intervals for activity measurements and at 5 second intervals for pulse-chase analysis. Total amounts of recombinant protein produced during each interval were assessed by examining equivalent amounts of cells (equal A_{600} values), which were subsequently lysed, ran on SDS-PAGE and Coomassie brilliant blue-stained. Aliquots harvested at time points containing equivalent levels of each recombinant protein produced were then lysed under native conditions as described (Chang et al, 2005) and their solubility and activity assessed (see below).

Determination of protein solubility. Cells were harvested by centrifugation and spheroplasts were prepared as described (Ausubel et al, 2003). Spheroplasts were lysed by dilution into an equal volume of native lysis buffer (5 mM $MgSO_4$, 0.2% (v/v) Triton X-100 (Sigma), Complete EDTA-free protease inhibitors (Roche), 100 units/ml Benzonase (Roche), 50 mM Tris-HCl, pH 7.5). Aliquots were centrifuged into supernatant and pellet fractions (20,000*g* for 10 min) and analyzed by SDS-PAGE followed by staining with Coomassie brilliant blue.

Determination of luciferase activity. Lysates from cells expressing Luc_{WT}, Luc_{fast}, Luc_{cbf} were prepared as above and equivalent dilutions to those used for solubility assessment were utilized. Luciferase activity was determined using the Luciferase Assay System (Promega) in a Sirius luminometer (Berthold) as described (Agashe et al, 2004).

Pulse-chase analysis. Pulse-chase experiments were performed as described (Sorensen & Pedersen, 1998). Cells expressing the desired construct were grown and protein expression was induced as described above. At time 0, (30 minutes post-induction), ^{35}S-Met was added to the culture, 10 seconds later excess unlabeled Met was added. Aliquots were taken every 5 seconds and placed in ice-cold tubes containing chloramphenicol. Cells were harvested and lysates were run on SDS-PAGE followed by autoradiography.

Predicted average polypeptide elongation rates were performed as described (Sorensen & Pedersen, 1998). In our constructs, there are 14 methionine residues at positions 92, 187, 189, 320, 336, 389, 421, 433, 467, 495, 518, 526, 555 and 584 from the C' terminus. The theoretical appearance of radiolabeled methionines was calculated for translation speeds of 5, 10 and 20 aa/s (Table 2).



(a)

	Translation speed		
Time (s)	5 aa/s	10 aa/s	20 aa/s
10	0	1	3
15	0	3	3
20	1	3	6
25	1	3	10
30	1	3	14
35	1	6	14
40	3	7	14
45	3	9	14
50	3	11	14
55	3	13	14
60	3	14	14
65	4	14	14
70	5	14	14

(b)

Met @ position	Translation speed		
	5 aa/s	10 aa/s	20 aa/s
92	18.4	9.2	4.6
187	37.4	18.7	9.4
189	37.8	18.9	9.5
320	64.0	32.0	16.0
336	67.2	33.6	16.8
389	77.8	38.9	19.5
421	84.2	42.1	21.1
433	86.6	43.3	21.7
467	93.4	46.7	23.4
494	98.8	49.4	24.7
517	103.4	51.7	25.9
525	105.0	52.5	26.3
555	111.0	55.5	27.8
584	116.8	58.4	29.2

Table 2. Prediction of appearance of methionines. (a) projected appearance of methionines according to translation speed. (b) calculated time required (s) to synthesize methionines based on met position and translation rates.

Chapter 4. Conclusions and perspectives.

Conclusions

There are other factors that could potentially affect overall polypeptide elongation rates, such as programmed stalling of the ribosome (Vazquez-Laslop et al, 2010) and signal recognition particles (Nagai et al, 2003), among others. However, in this study we decided to concentrate in the effects of the in translation speed by the ribosome's decoding mechanisms.

In chapter 2 of this work, our laboratory was able to show that the speed at which a polypeptide comes out of the ribosome plays an important role in its fate. We observed that proteins of eukaryotic origin are more likely to effectively acquire their three dimensional structure when translation proceeds at a slow rate in bacteria. This provides an insight into the evolutionary processes of both bacteria and eukaryotes. As mentioned before, the translation speed in bacteria is considerably faster than that of eukaryotes. Not surprisingly, proteins in bacteria tend to be shorter in length and simpler in domain composition compared to eukaryotes (Netzer & Hartl, 1997). This significant difference is likely due to the translation speeds that proteins are translated at, a more reasonable explanation is that in order to evolve and form more complex proteins that have more functions, at some point in time, translation had to be slowed down.

How could this have been achieved? One possible answer is that organisms devised a mechanism of incorporating translation speed information in their genome. In chapter 3, we discuss that ribosome movement along an mRNA is not uniform. This, due to the chemical nature of each codon and the concentration of their isoacceptor tRNA. We showed that codons that are decoded by abundant tRNAs are translated faster than those that rely on wobble-decoding. Thus, is very likely that higher organisms were able to evolve in part by using the degeneracy of the genetic code to modulate their translation speed.

Codon usage bias and translation speed.

Codons that are used more often in proteins of a particular organism are commonly referred to as “frequent” codons. Additionally, this usually also implies that frequent codons are translated faster than “rare” ones. However, it has been shown in the literature (Bremer & Dennis, 1996; Kudla et al, 2009; Zhang et al, 2009) as well as in this work, that there is not a direct correlation between codon frequency and translation speed.

There are cases where the most “frequent” codon lacks genes for the cognate isoacceptor tRNA. The opposite is also true, there are “rare” codons that have several genes encoding the isoacceptor tRNA. For example, bacteria lack isoacceptor tRNA genes for the most frequent codons for phenylalanine (UUU), isoleucine (AUU), valine (GUG), alanine (GCG), tyrosine (UAU), aspartic acid (GAU) and arginine (CGC). The same situation is found in humans, were genes for the most frequent codons for serine (UCC), proline (CCC), threonine (ACC) and alanine (GCC) are not present in the human genome. Therefore, utilizing the so-called frequent codons to “recode” a protein in order to achieve an increase in translation speed and/or yield of active protein will not result in a significant increase of speed and/or yield. As was shown in chapter 3 of this study (Fig. 17 and 21).

Thus, the preferred method to increase the yield of active recombinant protein is to match the speed profile of that protein in the expression host to that of its native host based on tRNA gene content and decoding characteristics (Fig. 22).

Perspectives

There are still several unanswered questions. What role does the differences in molecular chaperones across species play in translation speed? Do silent mutations that cause disease alter translation in a way that leads to misfolding and aggregation? Similarly, can variations on tRNA abundance lead to disease?

Molecular chaperones, translation speed and folding efficiency.

As mentioned earlier, there are a number of differences in the translational machinery between bacteria and eukaryotes. Differences left unexplored are molecular chaperones. In bacteria, folding of nascent chains occurs mostly post-translationally, while in eukaryotes, co-translational folding is favored. In bacteria, TF binds nascent chains as they come out of the ribosome and has been shown to remain bound to the nascent polypeptide once it has left the ribosome (Kaiser et al, 2006). Thus, promoting post-translational folding. Post-translational folding may hinder the capability of proteins of eukaryotic origin to fold in bacteria, since eukaryotic proteins evolved using a co-translational pathway. However, TF might be detrimental to the effective folding of proteins of eukaryotic origin when expressed in bacteria.

It would be very interesting to explore if the expression of eukaryotic proteins in bacteria translated at slower rates, either by the utilization of the SmP ribosomes, by “re-coding” of the mRNA or both, the deletion of TF would further enhance the folding yield.

Silent mutations and disease.

Diseases such as amyotrophic lateral sclerosis (ALS), some types of Alzheimer’s and Parkinson’s disease are thought to be caused by protein misfolding and aggregation (Prudencio et al, 2010; Seetharaman et al, 2009). The mutations in these diseases, particularly ALS, are of the so-called silent mutation type, those that alter the DNA but not the amino acid sequence. Perhaps, these thought to be innocuous type of mutations, alter the translation speed at the local level and thus, promote misfolding. In fact, it has been shown that in some cases, silent mutations indeed alter protein function (Hamano et al, 2007).

Future studies, should investigate whether these mutations lead to differences in translation speed of those pseudo-mutated proteins and thus, in the ability of acquiring their native structure.

Variations in tRNA concentration and disease.

A study showed that in humans, the expression of tRNA genes is not uniform among tissues. Indeed, the expression levels can vary up to tenfold (Dittmar et al, 2006). This suggests that tRNA expression and thus, translation speed might play an important role in the function of certain tissues. Furthermore, tRNA concentration may influence cellular differentiation during development. Additionally, a tenfold increase in tRNA expression has been observed in breast cancer patients (Pavon-Eternod et al, 2009), which further confirms that tRNA expression levels might modulate the function of proteins.

Translation speed and folding ability of these altered proteins should be assessed. This could potentially be done by artificially synthesizing the sequences harboring those mutations and follow that experimental procedures discussed in chapter 3 of this work.

References

Adzhubei AA, Adzhubei IA, Krasheninnikov IA, Neidle S (1996) Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett* **399:** 78-82

Agashe VR, Guha S, Chang HC, Genevaux P, Hayer-Hartl M, Stemp M, Georgopoulos C, Hartl FU, Barral JM (2004) Function of trigger factor and DnaK in multidomain protein folding: increase in yield at the expense of folding speed. *Cell* **117:** 199-209

Agris PF, Vendeix FA, Graham WD (2007) tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol* **366:** 1-13

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A* **104:** 11963-11968

Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A* **106:** 21149-21154

Alexander PA, Rozak DA, Orban J, Bryan PN (2005) Directed evolution of highly homologous proteins with different folds by phage display: implications for the protein folding code. *Biochemistry* **44:** 14045-14054

Anderson JF, Siller E, Barral JM (2011) Disorders of protein biogenesis and stability. *Protein Pept Lett* **18:** 110-121

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181:** 223-230

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (2003) *Current Protocols in Molecular Biology*, New York, NY: John Wiley & Sons, Inc.

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289:** 905-920

Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in Escherichia coli. *Nat Biotechnol* **22:** 1399-1408

Barral JM, Broadley SA, Schaffar G, Hartl FU (2004) Roles of molecular chaperones in protein misfolding diseases. *Semin Cell Dev Biol* **15:** 17-29

Bonekamp F, Dalboge H, Christensen T, Jensen KF (1989) Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli. *J Bacteriol* **171:** 5812-5816

Bracher A, Starling-Windhof A, Hartl FU, Hayer-Hartl M (2011) Crystal structure of a chaperone-bound assembly intermediate of form I Rubisco. *Nat Struct Mol Biol*

Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* **309:** 1868-1871

Bremer H, Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Neidhart FC (ed), pp 1553-1569. Washington, DC.: ASM Press

Brunak S, Engelbrecht J (1996) Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. *Proteins* **25:** 237-252

Bukau B, Deuerling E, Pfund C, Craig EA (2000) Getting newly synthesized proteins into shape. *Cell* **101:** 119-122

Bukau B, Weissman J, Horwich A (2006) Molecular chaperones and protein quality control. *Cell* **125:** 443-451

Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37:** D93-97

Chandra A, Hughes TR, Nugent CI, Lundblad V (2001) Cdc13 both positively and negatively regulates telomere replication. *Genes Dev* **15:** 404-414

Chang HC, Kaiser CM, Hartl FU, Barral JM (2005) De novo folding of GFP fusion proteins: high efficiency in eukaryotes but not in bacteria. *J Mol Biol* **353:** 397-409

Clarke TFt, Clark PL (2008) Rare codons cluster. *PLoS One* **3:** e3412

Comeron JM, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. *J Mol Evol* **47:** 268-274

Conti E, Franks NP, Brick P (1996) Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4:** 287-298

Crameri A, Whitehorn EA, Tate E, Stemmer WP (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol* **14:** 315-319

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

Crick FH (1958) On protein synthesis. *Symp Soc Exp Biol* **12:** 138-163

Crick FH (1966) Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19:** 548-555

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* **192:** 1227-1232

Curran JF, Yarus M (1989) Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol* **209:** 65-77

Dalal S, Regan L (2000) Understanding the sequence determinants of conformational switching using protein design. *Protein Sci* **9:** 1651-1659

DeZwaan DC, Toogun OA, Echtenkamp FJ, Freeman BC (2009) The Hsp82 molecular chaperone promotes a switch between unextendable and extendable telomere states. *Nat Struct Mol Biol* **16:** 711-716

Dimaio F, Leaver-Fay A, Bradley P, Baker D, Andre I (2011) Modeling symmetric macromolecular structures in rosetta3. *PLoS One* **6:** e20450

Dittmar KA, Goodenbour JM, Pan T (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet* **2:** e221

dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. In *Nucleic Acids Res* Vol. 32, pp 5036-5044. England

Ellis RJ, Minton AP (2006) Protein aggregation in crowded environments. *Biol Chem* **387:** 485-497

Frydman J (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. *Annu Rev Biochem* **70:** 603-647

Frydman J, Erdjument-Bromage H, Tempst P, Hartl FU (1999) Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. *Nat Struct Biol* **6:** 697-705

Fukuda H, Arai M, Kuwajima K (2000) Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry* **39:** 12025-12032

Gardner RS, Wahba AJ, Basilio C, Miller RS, Lengyel P, Speyer JF (1962) Synthetic polynucleotides and the amino acid code. VII. *Proc Natl Acad Sci U S A* **48:** 2087-2094

Gautschi M, Mun A, Ross S, Rospert S (2002) A functional chaperone triad on the yeast ribosome. *Proc Natl Acad Sci U S A* **99:** 4209-4214

Giepmans BN, Adams SR, Ellisman MH, Tsien RY (2006) The fluorescent toolbox for assessing protein location and function. *Science* **312:** 217-224

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8:** r49-r62

Grosjean H, de Crecy-Lagard V, Marck C (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett* **584:** 252-264

Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC (2000) Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem Biophys Res Commun* **269:** 692-696

Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* **177:** 4121-4130

Hamano T, Matsuo K, Hibi Y, Victoriano AF, Takahashi N, Mabuchi Y, Soji T, Irie S, Sawanpanyalert P, Yanai H, Hara T, Yamazaki S, Yamamoto N, Okamoto T (2007) A single-nucleotide synonymous mutation in the gag gene controlling human immunodeficiency virus type 1 virion production. *J Virol* **81:** 1528-1533

Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107:** 679-688

Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. *Nature* **475:** 324-332

Hartl FU, Hayer-Hartl M (2002) Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295:** 1852-1858

Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* **16:** 574-581

Higgs PG, Ran W (2008) Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol* **25:** 2279-2291

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A (1965) Structure of a ribonucleic acid. *Science* **147:** 1462-1465

Hughes TR, Weilbaecher RG, Walterscheid M, Lundblad V (2000) Identification of the single-strand telomeric DNA binding domain of the Saccharomyces cerevisiae Cdc13 protein. *Proc Natl Acad Sci U S A* **97:** 6457-6462

Ibba M, Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* **69:** 617-650

Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* **146:** 1-21

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2:** 13-34

Johansson M, Bouakaz E, Lovmar M, Ehrenberg M (2008) The kinetics of ribosomal peptidyl transfer revisited. *Mol Cell* **30:** 589-598

Kaiser CM, Chang HC, Agashe VR, Lakshmipathy SK, Etchells SA, Hayer-Hartl M, Hartl FU, Barral JM (2006) Real-time observation of trigger factor function on translating ribosomes. *Nature* **444:** 455-460

Kanaya S, Yamada Y, Kudo Y, 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:** 143-155

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315:** 525-528

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324:** 255-258

Kurland CG, Hughes D, Ehrenberg M (1996) Limitations of translational accuracy. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Neidhart FC (ed), pp 979-1004. Washington, DC.: ASM Press

Levinthal C (1969) How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems:Proceedings of the University of Illinois Bulletin*, Urbana, IL. Vol. 67, pp 22–24.

Liang ST, Xu YC, Dennis P, Bremer H (2000) mRNA composition and control of bacterial gene expression. *J Bacteriol* **182:** 3037-3044

Liu Y, Wu N, Dong J, Gao Y, Zhang X, Shao N, Yang G (2010) SsrA (tmRNA) acts as an antisense RNA to regulate Staphylococcus aureus pigment synthesis by base pairing with crtMN mRNA. *FEBS Lett* **584:** 4325-4329

Lynn DJ, Singer GA, Hickey DA (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res* **30:** 4272-4277

Makhoul CH, Trifonov EN (2002) Distribution of rare triplets along mRNA and their relation to protein folding. *J Biomol Struct Dyn* **20:** 413-420

Mathews DH, Turner DH, Zuker M (2007) RNA secondary structure prediction. In *Curr Protoc Nucleic Acid Chem* Vol. Chapter 11, 2008/04/23 edn, p Unit 11 12. Rochester, NY: University of Rochester

Mathews MB, Sonenberg N, Hershey JWB (2000) Origins and principles of translational control. In *Translational control of gene expression*, Sonenberg N, Hershey JWB, Mathews MB (eds), pp 1-31. Cold Spring Harbor, New York, NY: Cold Spring Harbor Laboratory Press

Matthaei JH, Jones OW, Martin RG, Nirenberg MW (1962) Characteristics and composition of RNA coding units. *Proc Natl Acad Sci U S A* **48:** 666-677

Mitton-Fry RM, Anderson EM, Theobald DL, Glustrom LW, Wuttke DS (2004) Structural basis for telomeric single-stranded DNA recognition by yeast Cdc13. *J Mol Biol* **338:** 241-255

Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L (2003) Structure, function and evolution of the signal recognition particle. *EMBO J* **22:** 3479-3485

Netzer WJ, Hartl FU (1997) Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* **388:** 343-349

Nugent CI, Hughes TR, Lue NF, Lundblad V (1996) Cdc13p: a single-strand telomeric DNA-binding protein with a dual role in yeast telomere maintenance. *Science* **274:** 249-252

Pavlou AK, Reichert JM (2004) Recombinant protein therapeutics--success rates, market trends and values to 2010. *Nat Biotechnol* **22:** 1513-1519

Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T (2009) tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res* **37:** 7268-7280

Pedersen S (1984) Escherichia coli ribosomes translate in vivo with variable rate. *Embo J* **3:** 2895-2898

Pennock E, Buckley K, Lundblad V (2001) Cdc13 delivers separate complexes to the telomere for end protection and replication. *Cell* **104:** 387-396

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12:** 32-42

Prudencio M, Durazo A, Whitelegge JP, Borchelt DR (2010) An examination of wild-type SOD1 in modulating the toxicity and aggregation of ALS-associated mutant SOD1. *Hum Mol Genet* **19:** 4774-4789

Purvis IJ, Bettany AJ, Santiago TC, Coggins JR, Duncan K, Eason R, Brown AJ (1987) The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. In *J Mol Biol* Vol. 193, pp 413-417. England

Qu X, Wen JD, Lancaster L, Noller HF, Bustamante C, Tinoco I, Jr. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475:** 118-121

Ruusala T, Andersson D, Ehrenberg M, Kurland CG (1984) Hyper-accurate ribosomes inhibit growth. *Embo J* **3:** 2575-2580

Saunders R, Deane CM (2010) Synonymous codon usage influences the local protein structure observed. In *Nucleic Acids Res* Vol. 38, pp 6719-6728.

Schein CH (1989) Production of soluble recombinant proteins in bacteria. *Biotechnology* **7:** 1141-1149

Schroder H, Langer T, Hartl FU, Bukau B (1993) DnaK, DnaJ and GrpE form a cellular chaperone machinery capable of repairing heat-induced protein damage. *Embo J* **12:** 4137-4144

Seetharaman SV, Prudencio M, Karch C, Holloway SP, Borchelt DR, Hart PJ (2009) Immature copper-zinc superoxide dismutase and familial amyotrophic lateral sclerosis. *Exp Biol Med* **234:** 1140-1154

Seidman JS, Janssen BD, Hayes CS (2011) Alternative fates of paused ribosomes during translation termination. *J Biol Chem*

Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-1295

Shirts M, Pande VS (2000) COMPUTING: Screen Savers of the World Unite! *Science* **290:** 1903-1904

Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J Mol Biol* **396:** 1310-1318

Snow CD, Sorin EJ, Rhee YM, Pande VS (2005) How well can simulation predict protein folding kinetics and thermodynamics? *Annu Rev Biophys Biomol Struct* **34:** 43-69

Sorensen MA, Pedersen S (1991) Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol* **222:** 265-280

Sorensen MA, Pedersen S (1998) Determination of the peptide elongation rate in vivo. *Methods Mol Biol* **77:** 129-142

Stemp MJ, Guha S, Hartl FU, Barral JM (2005) Efficient production of native actin upon translation in a bacterial lysate supplemented with the eukaryotic chaperonin TRiC. *Biol Chem* **386:** 753-757

Thanaraj TA, Argos P (1996a) Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci* **5:** 1973-1983

Thanaraj TA, Argos P (1996b) Ribosome-mediated translational pause and protein domain organization. *Protein Sci* **5:** 1594-1612

Thompson J, Baker D (2011) Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins* **79:** 2380-2388

Toogun OA, Zeiger W, Freeman BC (2007) The p23 molecular chaperone promotes functional telomerase complexes through DNA dissociation. *Proc Natl Acad Sci U S A* **104:** 5765-5770

Tsien RY (1998) The green fluorescent protein. *Annu Rev Biochem* **67:** 509-544

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

Varenne S, Buc J, Lloubes R, Lazdunski C (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* **180:** 549-576

Vazquez-Laslop N, Ramu H, Klepacki D, Kannan K, Mankin AS (2010) The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J* **29:** 3108-3117

Wahba AJ, Gardner RS, Basilio C, Miller RS, Speyer JF, Lengyel P (1963) Synthetic polynucleotides and the amino acid code. VIII. *Proc Natl Acad Sci U S A* **49:** 116-122

Wang MJ, Lin YC, Pang TL, Lee JM, Chou CC, Lin JJ (2000) Telomere-binding and Stn1p-interacting activities are required for the essential function of Saccharomyces cerevisiae Cdc13p. *Nucleic Acids Res* **28:** 4733-4741

Xie T, Ding D (1998) The relationship between synonymous codon usage and protein structure. *FEBS Lett* **434:** 93-96

Zengel JM, Young R, Dennis PP, Nomura M (1977) Role of ribosomal protein S12 in peptide chain elongation: analysis of pleiotropic, streptomycin-resistant mutants of Escherichia coli. *J Bacteriol* **129:** 1320-1329

Zhang F, Saha S, Shabalina SA, Kashina A (2010) Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. In *Science* Vol. 329, pp 1534-1537. United States

Zhang G, Hubalewska M, Ignatova Z (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol* **16:** 274-280

Zhou T, Weems M, Wilke CO (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. In *Mol Biol Evol* Vol. 26, pp 1571-1580. United States

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406-3415