Top Med Chem
DOI: 10.1007/7355_2016_2

The Art of Gene Redesign and Recombinant Protein Production: Approaches and Perspectives

Anton A. Komar

Abstract In recent years, the demand for recombinant proteins for use in research laboratories or in medical settings has increased dramatically. Although a wide variety of recombinant protein expression systems and gene redesign approaches are available, obtaining active, correctly folded recombinant proteins in sufficient amounts remains a challenge in many cases. One of the main approaches to gene redesign with the potential to increase protein production involves introduction of synonymous codon substitutions in mRNAs aimed at increasing the rate/efficiency of translation. However, a number of recent studies have shown that synonymous codon substitutions can also negatively impact mRNA biogenesis, mRNA decoding, as well as protein folding and function. Maximizing the speed and output of translation may put conflicting demands on the protein synthesis machinery resulting in reduced accuracy of the decoding process and/or improper protein folding. An improved understanding of the impact of synonymous codon substitutions on mRNA/protein biogenesis and function is critically important for the development of safer and more effective recombinant protein therapeutics. This review discusses the most common approaches to gene redesign that involve synonymous codon substitutions and provides recommendations for their optimal use in light of recent developments in the field regarding the impact of synonymous codon usage on various aspects of protein production and function.

Keywords Codon usage, Gene redesign, Mistranslation, mRNA turnover, Protein folding, Protein synthesis, Rare synonymous codons, Recombinant protein therapeutics, Synonymous codons

A.A. Komar (✉)
Center for Gene Regulation in Health and Disease and Department of Biological, Geological and Environmental Sciences, Cleveland State University, Cleveland, OH 44115, USA

DAPCEL, Inc., Cleveland, OH 44106, USA
e-mail: a.komar@csuohio.edu

Contents

1 Introduction

Production of soluble and functionally active proteins in heterologous and homologous host organisms is the cornerstone of many modern biotechnology applications. In recent years, the demand for recombinant proteins used in research laboratories or in medical settings (e.g., for therapeutic applications) has increased dramatically. Specifically, the protein therapeutic market was valued in excess of $85 billion in 2010 and is predicted to double by the end of 2018, reaching up to $165 billion, as new products (especially therapeutic monoclonal antibodies) become available (http://www.researchandmarkets.com/reports/2729030/global_protein_therapeutics_market_outlook_2018). Despite the strong existing and potential significance of efficient recombinant protein production for both research applications and development of novel therapeutics, obtaining soluble, active recombinant proteins in sufficient amounts remains challenging in many cases.

A wide variety of recombinant protein expression systems are well established. These include, but are not limited to, various cellular systems, such as bacterial, yeast, insect and mammalian systems [1–7], and cell-free in vitro systems [8, 9]. The urgent need for robust and highly scalable protein manufacturing systems has further led to the development of in vivo plant- and animal-based systems [10–13]. All of these systems have their own advantages and disadvantages [14]. The choice of system to use for a particular application depends on the specific requirements for the final recombinant protein product (e.g., requirements for proper protein processing and/or co- and posttranslational protein folding and modifications) [14]. In most cases, use of a recombinant protein expression system that closely resembles the protein's natural in vivo expression system/environment is highly desirable, but this is obviously not always achievable [14]. For example, toxicity of the final product may not allow enhanced expression of a protein in a homologous, or even heterologous, cellular system(s) [15, 16]. In such cases, cell-free protein synthesis systems on a larger scale, particularly with continuous action, may offer an alternative solution [8, 9, 15, 17, 18]. In addition, expression of unmodified natural genes in a homologous environment frequently does not support levels of protein expression sufficient for large-scale protein production. The key to solving this problem lies in development of gene redesign approaches that result in robust expression of functionally active proteins both inside and outside their natural (homologous) cellular environments.

One of the main approaches to gene redesign facilitating protein production in heterologous and homologous organisms [19–21] takes advantage of the degeneracy of the genetic code (meaning a given amino acid may be encoded by more than one “synonymous” codon). Synonymous codons are present at different frequencies in different organisms and are decoded at different rates [22–24]. Therefore, substitution of synonymous codons in a gene can dramatically affect the rate/efficiency of synthesis of the encoded protein without altering its amino acid sequence [19–21]. In a given organism, frequently used codons are typically translated more rapidly than infrequently used ones due to the fact that tRNAs corresponding to the frequently used codons are relatively more abundant [25–31]. Many synonymous codons that are frequently used in eukaryotes (especially mammals) are utilized with low frequency in prokaryotes [22–24] such as the bacteria *Escherichia coli*, one of the most common hosts for heterologous protein production [14]. The impact of these differences on recombinant protein production is now well appreciated, and it has been clearly demonstrated that the level of protein expression in heterologous and homologous organisms can be increased through suitable selection of synonymous (frequent) codons along target mRNAs [19–21].

In addition to the effects of differential codon usage, the secondary structure of messenger RNAs (mRNA) has been recognized as a factor that can have a negative impact on translation and reduce protein yields by slowing or blocking translation initiation and/or the movement of ribosomes along the mRNA [32–39].

Several other considerations important for recombinant protein production (e.g., choice of appropriate vector/promoter system(s), means of gene delivery, etc.) are outside the scope of this short review.

Approaches involving substitution of the majority of infrequently used codons with synonymous frequently used ones, often combined with elimination of extreme GC content that could contribute to formation of stable mRNA secondary structures, have been widely used by many biotechnology companies and research groups for optimization of heterologous gene/protein expression, but with mixed results ([19, 40] and references therein). Use of gene sequences optimized through the abovementioned approaches often yielded large amounts of recombinant proteins [19, 40]; however, in many cases, the products formed biologically inactive insoluble aggregates which had to be refolded (whenever it was possible) in order to regain similarity in structure and biological activity with native analogues [19, 28]. Moreover, even when proteins expressed in heterologous or homologous hosts remained soluble, they were not necessarily natively folded [41].

These and other experiments brought about awareness of the scientific community to the impact of synonymous codon usage on not only the efficiency of translation but also on other aspects of gene function, particularly, protein folding. The significance of synonymous codon usage on protein folding was highlighted by findings showing that multiple and, more surprisingly, single synonymous substitutions/mutations can affect proteins' activity [42–44], interactions with drugs and inhibitors [43], phosphorylation profiles [45], sensitivity to limited proteolysis [43, 45, 46], spectroscopic properties [47], and aggregation propensity [47–49] and ultimately change protein structure [50].

Many recent studies have shown that synonymous substitutions or naturally occurring synonymous mutations are not neutral and may affect gene function by multiple mechanisms [51, 52], including but not limited to those mentioned above, as well as mechanisms exerting effects on mRNA splicing and/or mRNA stability [53, 54]. Synonymous codon choice has been also suggested to affect efficient interaction of nascent polypeptides with the signal recognition particle [55]. Changes in codon context caused by synonymous mutations may also induce mistranslation leading to protein misfolding [56].

While in many instances complete understanding of the exact effects caused by synonymous substitutions and/or mutations is still lacking, it nevertheless seems possible to use existing knowledge for the development of some common rules to gene design and redesign that should increase the chances of getting the desired levels and activity of the expressed recombinant proteins and reduce protein misfolding and aggregation.

This review discusses the most common approaches to gene redesign that involve synonymous codon substitutions and contains a set of recommendations for optimizing protein synthesis and folding through this approach. These recommendations take into account recent developments in the field highlighting the impact of synonymous codon usage on protein production and function.

2 Synonymous Gene Exploration in Protein Production and Folding

Designing an optimal gene for recombinant protein production requires choosing from an enormous number of possible DNA/RNA sequences. It is a combinatorial problem, giving approximately 3^N variants for a sequence with N codons. However, as discussed below, this number can be substantially reduced by taking into account a set of critical considerations.

In general, two global gene design/redesign approaches predominate (1) de novo gene design based on reverse translation from an amino acid sequence to DNA/RNA and (2) gene redesign based on recoding of a natural DNA/RNA sequence. Numerous online/web-based and stand-alone platforms are available for use in one or both of these approaches. These include, for example, Codon Optimization OnLine (COOL) [57], DNA Works [58], D-Tailor [59], EuGene [60], GeneDesign [61], Gene Designer 2.0 [62], Jcat [63], mRNA Optimiser [64], OPTIMIZER [65], Synthetic Gene Designer [66], TmPrime [67], Visual Gene Developer [68], and others (for a review see [69]). The majority of available tools, however, start with a natural DNA/RNA sequence and employ either codon or RNA structure optimization algorithms (or both) to maximize gene expression; only TmPrime [67] is a “pure” de novo back-translation tool. GeneDesign [61] and OPTIMIZER [65] offer both possibilities – de novo back-translation from protein to DNA/RNA sequence and recoding of the natural DNA/RNA sequence.

Most of the abovementioned platforms customize codon usage by setting codon frequency percentage [70] and/or Codon Adaptation Index (CAI) [71] thresholds and then substituting rare synonymous codons with frequent ones along the entire open reading frame (ORF) of a gene to achieve the desired threshold level(s). Substitutions are selected based on known organism-specific codon biases [22–24, 68]. The COOL [57], D-Tailor [59], EuGene [60], OPTIMIZER [65], and Visual Gene Developer [68] tools also take into account the RNA structure and/or GC/AT content, aiming to reduce obstacles related to formation of stable RNA structures. mRNA Optimizer [64] and TmPrime [65] focus solely on mRNA secondary structure optimization to avoid stable secondary structures by means of maximizing the minimum free energy (MFE) of the nucleotide sequences without changing the final resulting amino acid sequence.

As mentioned above, all currently available algorithms (with the exception of TmPrime [65]) typically start from the original/natural coding sequence and then evolve the sequence through iterations of synonymous codon changes that would increase/maximize the MFE and/or codon usage frequency/CAI or both to achieve the desired outcome. However, none of the abovementioned tools typically considers the impact of synonymous codon usage on protein folding (rather than simply on translation efficiency). They also fail to take into account some other important considerations that can affect mRNA translatability and stability and, therefore, preclude efficient expression of correctly folded and functional proteins. Below, I examine some of these considerations that may facilitate gene design and redesign toward optimized expression of active, correctly folded proteins.

2.1 Codon Usage at ORF (Open Reading Frame) 5′ Termini

The occurrence of synonymous codons in protein-coding open reading frames (ORFs) of genes is not random, thus revealing the existence of evolutionary pressure on codon choice [23, 24, 28, 72–74]. Clustering of synonymous codons has been observed at specific conserved locations in mRNAs indicating that there are forces that influence the selection of these codons at specific locations within mRNA sequences [28, 33, 37, 38, 55, 75, 76]. Strategic placement of specific synonymous codons, particularly those that are rare, in gene ORFs suggests a functional role conserved in evolution rather than random chance. Therefore, the randomized and/or global substitution of rare synonymous codons with frequent ones that is offered by the majority of tools aimed at simply increasing CAI/codon usage frequency and/or MFE (see above) might not be beneficial for the production of a functional protein.

An example of nonrandom synonymous codon usage within ORFs is the observed enrichment of rare codons at the 5′ termini of genes in *E. coli* and many other prokaryotes, as well as in genes of some eukaryotes such as the yeast *Saccharomyces cerevisiae* [75, 76]. The clustering of rare codons at 5′ gene termini (typically at codon positions 1 to ~20 [33, 37, 38, 76]) clearly indicates an influence of evolutionary pressure on their selection. This particular aspect of natural codon usage may be explained by fact that rare codons in many bacteria are largely AT-rich [70]. Thus, their clustering at 5′ORF termini leads to reduced secondary structure in that region of the mRNA and, consequently, enhanced protein expression (it is known that mRNA secondary structure at 5′ ORF termini negatively affects protein expression by limiting access of the ribosomes to the ribosome binding site (RBS) on the mRNA [33, 37, 38, 55, 75]).

It should be noted, however, that the enrichment of rare codons at 5′ ORF termini has been mostly found in bacteria with genomes with overall GC content of at least 50% [77]. Recent work showed that, in general, AT-rich codons as opposed to rare codons are preferentially located at 5′ ORF termini in prokaryotes [33, 34, 37, 38, 54]. This further implicates secondary structure as the driving force for specific codon selection at 5′ ORF termini in bacteria [33, 38, 54]. Interestingly, the higher the GC content of a genome, the more mRNA stability is reduced at the region near the start codon [78].

It should be also noted that despite differences in translation apparatus and the mechanism of protein synthesis between prokaryotes and eukaryotes, many eukaryotic ORFeomes also are characterized by reduced 5′-terminal mRNA secondary structure near the start codon [78]. This indicates that reduced 5′-terminal ORF mRNA secondary structure may have been evolutionary selected in all organisms. In eukaryotes, this can be expected to facilitate start-codon recognition by the scanning ribosome [78].

Could there be additional reasons for preferential use of rare codons at the 5′ ORF termini of some natural genes, including those in *E. coli*? It has been suggested that clustering of rare codons at 5′ ORF termini may in certain cases allow slow co-translational formation of the N-terminal folding nucleus of the protein, thus facilitating overall correct protein folding in the cell [28].

Interestingly, strong enrichment of rare codons at 5′ gene termini has been preferentially observed (with very high statistical significance ($P < 0.0001$)) in genes/ORFs encoding secretory proteins [76]. It has been suggested that for genes encoding secretory proteins with N-terminal signal sequences, 5′ rare codon clusters could have a functional role related to secretion, by transiently slowing down translation prior to membrane localization of the nascent chain(s) [79]. It has been experimentally shown in yeast that local slowdown of translation caused by presence of rare codons (located ~35–40 codons downstream of signal sequences or transmembrane segments) promotes nascent-chain recognition by signal recognition particle (SRP), which assists in protein translocation across membranes [55]. Similarly, strategically located Shine-Dalgarno-like elements were identified in ORFeomes of *E. coli* secretory proteins; these elements serve to transiently slow down translation elongation in order to allow efficient integration of the transmembrane helix of many membrane proteins [80].

Therefore, based on the considerations described above, carefully planned placement of rare/non-optimal (or AT-rich) codons in the 5′ ORF termini of mRNAs, especially for those encoding secretory and transmembrane proteins, may represent an important strategy for successful gene design and redesign enhancing proper protein production, secretion, and folding.

2.2 Conserved Rare Codon Clusters Within Gene ORFs

It is widely believed that the major influence of codon usage is on global and local translation rate. As mentioned above, frequently used codons are translated more rapidly than infrequently used ones [25–31]. However, which codons are more rare or frequent varies by organism [22–25, 70]. Surprisingly, across all organisms, rare codons appear to occur in clusters, rather than being randomly scattered across genes [28, 75]. Although there is a general tendency for rare codons to cluster at the 5′ termini of ORFs (see above), such clustering is also observed within ORFs [28, 75, 81]. These clusters are not confined to the 5′ end of ORFs or to ORFs of genes/proteins that are expressed at a low level (as might be expected if rare codons are thought of as simply correlating with reduced translation rate). Rather, they are found to occur equally in genes for all types of proteins, including abundant/highly expressed proteins [75, 81].

Analyses of ORFeomes from prokaryotic and eukaryotic organisms revealed that rare codon clustering (1) is not limited to a particular set of genes or genotype, (2) does not depend on and is not related to the overall GC content of the organism's genome, and (3) is significantly more abundant than would be expected based on random selection [75, 81]. Furthermore, for some protein families, the locations of rare codon-rich regions within mRNAs are highly conserved across homologs in different organisms; this is observed, for example, in families of cytochromes c, globins, gamma-B crystallins [28], ocular lacritins [82], and chloramphenicol acetyltransferases [28, 83].

Enrichment of rare codon clusters at specific locations in a broad range of genes and organisms suggests that evolutionary selection determines such clustering and that it must have some functional significance [28, 75, 81–83]. One hypothesis links the location of rare codon clusters to the process of protein folding in the cell [84, 85]. This proposes that sequential folding events that occur during co-translational folding of proteins might be separated by rare codon clusters, with such clusters serving to reduce the speed of translation at these positions and thereby facilitating proper folding through temporal separation of folding events on the ribosome [28, 74, 86–91]. This is consistent with the finding that there seems to be a certain hierarchy in the location of rare codon-rich regions along mRNAs. Frequently, but not always, the rarest codons seem to encode boundaries of relatively large structural units (e.g., protein domains), whereas less rare codons encode boundaries of smaller units (e.g., protein motifs and subdomains) [28]. This might reflect the need to provide a more substantial translational delay for independent co-translation folding of larger units in comparison with smaller ones [28].

In summary, while there is a substantial body of literature underlining the overall negative effects of rare codons on levels of protein production (see [19] for a review), it is becoming increasingly clear that strategic placement of conserved rare codons clusters can have positive effects on protein biogenesis (particularly proper folding) and function. Some biotech companies, such as DAPCEL, Inc., are already using this knowledge to enhance protein production and facilitate correct co-translational protein folding.

2.3 Codon Usage at ORF (Open Reading Frame) 3′ Termini

Enrichment of rare codons at the 3′ terminus of *E. coli* ORFs (and ORFs of 11 other prokaryotes) has also been observed [76]. While significant enrichment of rare codons at the 5′ termini of genes in *E. coli* can be explained as a mechanism that facilitates interaction between ribosomes and ribosome binding sites on mRNAs (see above; [33, 37, 38, 55, 75]), the observed incidence (albeit less pronounced) of increased rare codon abundance at the 3′ termini of *E. coli* ORFs is not that easy to explain. It is possible that rare codon clusters at 3′ ORF termini could be required for more robust termination of translation and/or for reducing the rate of protein folding before release from the ribosome [76]. Queuing of ribosomes at the 3′ termini of ORFs due to presence of rare codons may also protect mRNAs from degradation. An improved understanding of the impact of codon usage at 3′ ORF termini is required before this feature can be rationally exploited in gene design and redesign strategies and/or interpretation of in vivo folding pathways.

3 Synonymous Codons and mRNA Stability

mRNA turnover plays a critical role in regulating gene expression. mRNAs with longer half-lives generally produce more protein than those with shorter half-lives simply because they are available to be translated for a longer period of time. A link between codon usage and mRNA turnover rate has been long recognized in both prokaryotes and eukaryotes [92–94], but has not been well understood until recently [53, 54]. Previously, it was generally believed that more thermodynamically stable mRNAs would also be more resistant to degradation. However, recent work showed that, at least in yeast, so-called codon optimality [53] rather than mRNA thermodynamic stability has a broad and powerful influence on in vivo mRNA degradation rates. Codon optimality is a scale that reflects the balance between the supply of specific charged tRNA molecules and the demand for their use by translating ribosomes, thus representing a measure of translation efficiency [53]. Optimal codons (typically, these are frequent codons) are decoded faster. In the yeast study, it was found that many stable/long-lived mRNAs harbor optimal codons within their ORFs, while many unstable/short-lived mRNAs harbor non-optimal codons [53]. Moreover, it was found that substitution of optimal codons with synonymous, non-optimal codons results in dramatic destabilization of the mRNA and vice versa [53]. Interestingly, very similar results were obtained in *E. coli* [54]. These findings suggest that transcript-specific translation elongation rate is an important determinant of mRNA stability and that more rapidly translated mRNAs (at least in yeast and *E. coli*) are likely to be more stable and, thus, produce more protein. This new information presents an opportunity to upscale protein production in yeast and *E. coli* via reassignment of codon optimality in an mRNA to increase its stability and, thus, its capacity to produce protein. Whether the same paradigm exists in higher eukaryotic organisms remains to be determined. However, this approach should be applied with caution since assignment of codons that are optimal for translation rate and mRNA stability could lead to incorrect protein folding.

4 Synonymous Codons and Mistranslation/Frameshifting

Another aspect of mRNA biology that can be impacted by synonymous codon usage is the accuracy with which they are translated. Clearly, mRNAs must be translated accurately in order for fully functional proteins to be produced. Estimates of missense error rates (referred to as miscoding or mistranslation) during protein synthesis from natural mRNAs vary from 10^{-3} to 10^{-4} per codon ([95–98] and references therein). Mistranslation is the incorporation of an amino acid that is different from the one encoded by a specific codon in the mRNA. Recent research has enhanced our understanding of mistranslation mechanisms and how it is controlled [95–98]. While it is generally believed that synonymous codon changes should be silent (not changing the amino acid that is incorporated), that is not always the case [95–98]. Moreover, certain codons are mistranslated more frequently than others [95, 98]. This is apparently due to the fact that translation speed and mistranslation rate are carefully balanced during protein synthesis and situations maximizing translation speed place demands on the translational machinery that reduces accuracy [95–98]. In general, translation has multiple layers of proofreading; however, most errors occur during decoding, which takes place on the ribosome [96, 98]. The frequency of miscoding of different codons varies over a nearly 20-fold range ([95] and references therein). Mispairing at the wobble position and scarce availability of cognate competitor tRNAs appear to play major roles in mistranslation [95–98]. For example, the frequency of miscoding of the AAU (Asn) codon in *E. coli* leading to incorporation of Lys (encoded by AAG and AAA) instead of Asn is about fourfold higher than that for the AAC (Asn) codon [95]. It should be noted, however, that the AAU codon is used more frequently than the AAC codon (codon usage frequency per 1,000 codons is 29.32 for AAU vs. 20.26 for AAC [70]); thus, substitution of AAC with AAU with the intention of maximizing codon frequency/CAI could result in increased levels of miscoding, which in turn could lead to loss of protein activity due to misfolding [56] or absence of a functionally important amino acid.

While, as described above, there is considerable evidence linking codon usage and missense errors, little is known about the relationship between codon usage and frameshifting errors. Programmed ribosomal frameshifting is utilized by many viruses and bacteria to increase the information content of their genomes; through frameshifting, multiple proteins can be produced from a single span of sequence [99, 100]. Signals in mRNAs have been identified that cause frameshifting by one base in the 5′ (−1) or 3′ (+1) direction [99, 100]. While beneficial in some cases for bacteria and viruses as mentioned above, unintended frameshifting during translation is clearly not desirable. Frameshifting errors can lead to premature termination of translation or generate abnormal proteins with toxic effects on the cell [56]. Attempts have been made to develop computational tools to assess whether codon usage can be optimized to minimize the frequency of frameshifting errors [101]. The results of this work indicate that natural synonymous codon usage is biased toward specific patterns correlated with avoidance of mistranslation and frameshifting-induced protein misfolding [101]. Overall, an understanding of the impact of codon usage on mistranslation and frameshifting errors may be helpful in minimizing the risk of producing subpopulation of proteins with different amino acid sequences when undertaking recombinant protein production from a redesigned gene.

5 The Impact of Single Synonymous Codon Substitutions

Gene redesign usually involves numerous substitutions of synonymous codons. However, recent studies have shown that some specific single synonymous mutations are deleterious for proper protein expression and, moreover, organism health ([51, 52] for a review). The majority of identified deleterious single synonymous mutations exert effects on mRNA splicing (in eukaryotes), but there are also quite a few that may alter protein folding and, as a consequence, protein activity and/or resistance to degradation [51, 52]. These single synonymous mutations can produce disease in the expressing organism, and their inadvertent introduction into genes of therapeutic proteins may produce undesirable effects. It should be noted that the exact mechanisms underlying the effects of many synonymous mutations linked to disease are not yet well understood [51, 52]. One of the major challenges in the field is to understand why some disease-causing synonymous mutations are more deleterious than others and to predict the likely effects of a single mutation.

Evaluation of mRNA stability of fragments of genes of several proteins carrying neutral vs. disease-associated mutations and synonymous vs. non-synonymous mutations revealed that deleterious synonymous mutations tend to occur in mRNA regions with higher MFE levels and often lead to a reduction in MFE [102–105]. It is not yet clear how broadly applicable this situation originally identified for “disease-associated” mutations in the *F8* and *F9* genes encoding blood-coagulation factors VIII and IX, respectively, might be [102, 105]. Mutations in the *F8* and *F9* genes lead to blood clotting disorders known as hemophilia A and B [102, 105]. While further investigation into the deleterious effects of specific synonymous mutations is required, it is clear that known disease-associated mutations should be avoided in gene redesign efforts.

6 Concluding Remarks and Future Perspectives

Gene design and redesign approaches target protein-coding genes and aim to introduce predefined features of interest into the final protein product. These approaches frequently involve changes in synonymous codon usage intended to improve protein production in homologous and/or heterologous hosts without compromising the integrity of the encoded protein. Optimization of gene design and protein production is of strong significance due to the high, and continually increasing, demand for recombinant proteins for use in research and in therapeutic applications. Advances in DNA synthesis have enabled construction of numerous gene variants and facilitated our understanding of the impact of codon usage on gene function. Additional knowledge came from genome-wide studies aimed at uncovering the impact of synonymous mutations on gene function and phenotype and understanding their association with various diseases.

It has become clear that synonymous codon usage and synonymous mutations do not only alter the speed of protein synthesis but affect many critical aspects of mRNA and protein biogenesis (ranging from mRNA stability to protein mistranslation and folding), thus ultimately changing the phenotype associated with the protein. Importantly, it was revealed that even a single synonymous mutation may be deleterious to protein function. While complete understanding of the effects caused by multiple and single synonymous mutations remains lacking, it is possible, as done in this review, to use existing knowledge to develop some common rules to gene design and redesign that should increase the probability of achieving the desired quantity and activity of an expressed recombinant protein.

A combination of evolutionary, computational, and synthetic biology should ultimately enable (1) full genome-based understanding of the impact of individual synonymous mutations on gene function, mRNA biogenesis, protein production, and protein folding; (2) efficient manufacturing of safer, more effective, and even potentially individualized protein therapeutics; and (3) improved understanding of evolutionary processes.

7 Notes

1. Carefully planned placement of rare/non-optimal (or AT-rich) codons in the 5′ termini of mRNA ORFs, especially those encoding secretory and transmembrane proteins, may represent an important strategy for successful gene design and redesign enhancing proper protein production, secretion, and folding.
2. Enrichment of rare codon clusters at specific locations in a broad range of genes implies that they have functional significance. Therefore, strategic placement of evolutionarily conserved rare codon clusters within ORFs may facilitate correct protein folding.
3. Use of optimal synonymous codons during gene design and redesign may lead to substantial stabilization of the mRNA and enhancement of protein production (at least in yeast and *E. coli*).
4. Mistranslation as a result of synonymous codon changes may lead to incorrect protein folding; this should be taken into consideration when planning production of recombinant proteins.
5. Although a variety of methods are available for gene redesign, approaches that take into account the effect(s) of synonymous codon substitutions on translation efficiency, protein folding, and protein activity will allow the most productive manufacturing of safer and more effective protein therapeutics.

Acknowledgments I apologize to those whose work or original publications could not be cited in this article because of space limitations. I thank Patricia Stanhope Baker for help with manuscript preparation. This work was supported in part by grants to A.A.K. from the Human Frontier Science Program (grant # RGP0024/2010), AHA (grant # 13GRNT17070025), and NIH (grant # 1R15HL121779).

References

1. Schmidt FR (2004) Recombinant expression systems in the pharmaceutical industry. Appl Microbiol Biotechnol 65:363–372
2. Berlec A, Strukelj B (2013) Current state and recent advances in biopharmaceutical production in *Escherichia coli*, yeasts and mammalian cells. J Ind Microbiol Biotechnol 40:257–274
3. Khan KH (2013) Gene expression in Mammalian cells and its applications. Adv Pharm Bull 3:257–263
4. Assenberg R, Wan PT, Geisse S, Mayr LM (2013) Advances in recombinant protein expression for use in pharmaceutical research. Curr Opin Struct Biol 23:393–402
5. Young CL, Robinson AS (2014) Protein folding and secretion: mechanistic insights advancing recombinant protein production in *S. cerevisiae*. Curr Opin Biotechnol 30:168–177
6. Sugiki T, Fujiwara T, Kojima C (2014) Latest approaches for efficient protein production in drug discovery. Expert Opin Drug Discov 9:1189–1204
7. van Oers MM, Pijlman GP, Vlak JM (2015) Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. J Gen Virol 96:6–23
8. Carlson ED, Gan R, Hodgman CE, Jewett MC (2012) Cell-free protein synthesis: applications come of age. Biotechnol Adv 30:1185–1194

9. Whittaker JW (2013) Cell-free protein synthesis: the state of the art. Biotechnol Lett 35:143–152
10. Yusibov V, Streatfield SJ, Kushnir N (2011) Clinical development of plant-produced recombinant pharmaceuticals: vaccines, antibodies and beyond. Hum Vaccin 7:313–321
11. Abiri R, Valdiani A, Maziah M, Shaharuddin NA, Sahebi M, Yusof ZY, Atabaki N, Talei D (2015) A critical review of the concept of transgenic plants: insights into pharmaceutical biotechnology and molecular farming. Curr Issues Mol Biol 18:21–42
12. Houdebine LM (2000) Transgenic animal bioreactors. Transgenic Res 9:305–320
13. Bösze Z, Baranyi M, Whitelaw CB (2008) Producing recombinant human milk proteins in the milk of livestock species. Adv Exp Med Biol 606:357–393
14. Demain AL, Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. Biotechnol Adv 27:297–306
15. Klammt C, Schwarz D, Löhr F, Schneider B, Dötsch V, Bernhard F (2006) Cell-free expression as an emerging technique for the large scale production of integral membrane protein. FEBS J 273:4141–4153
16. Saïda F (2007) Overview on the expression of toxic gene products in *Escherichia coli*. Curr Protoc Protein Sci 50:1–5
17. Ryabova LA, Morozov IY, Spirin AS (1998) Continuous-flow cell-free translation, transcription-translation, and replication-translation systems. Methods Mol Biol 77:179–193
18. Murray CJ, Baliga R (2013) Cell-free translation of peptides and proteins: from high throughput screening to clinical production. Curr Opin Chem Biol 17:420–426
19. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22:346–353
20. Elena C, Ravasi P, Castelli ME, Peirú S, Menzella HG (2014) Expression of codon optimized genes in microbial systems: current industrial applications and perspectives. Front Microbiol 5:21
21. Quax TE, Claassens NJ, Söll D, van der Oost J (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161
22. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1998) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–18211
23. Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287–299
24. Sharp PM, Emery LR, Zeng K (2010) Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci 365:1203–1212
25. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34
26. Buchan JR, Stansfield I (2007) Halting a cellular production line: responses to ribosomal pausing during translation. Biol Cell 99:475–487
27. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223
28. Komar AA (2009) A pause for thought along the co-translational folding pathway. Trends Biochem Sci 34:16–24
29. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15:205–213
30. Dana A, Tuller T (2014) The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res 42:9171–9181
31. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B (2014) Measurement of average decoding rates of the 61 sense codons in vivo. Elife 3, eLife.03735
32. Hatfield GW, Roth DA (2007) Optimizing scaleup yield for protein production: computationally optimized DNA assembly (CODA) and translation engineering. Biotechnol Annu Rev 13:27–42

33. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255–258
34. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107:3645–3650
35. Kim HJ, Lee SJ, Kim HJ (2010) Optimizing the secondary structure of human papillomavirus type 16 L1 mRNA enhances L1 protein expression in *Saccharomyces cerevisiae*. J Biotechnol 150:31–36
36. Castillo-Méndez MA, Jacinto-Loeza E, Olivares-Trejo JJ, Guarneros-Pena G, Hernandez-Sanchez J (2012) Adenine-containing codons enhance protein synthesis by promoting mRNA binding to ribosomal 30S subunits provided that specific tRNAs are not exhausted. Biochimie 94:662–672
37. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science 342:475–479
38. Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N (2013) Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9:675
39. Li GW (2015) How do bacteria tune translation efficiency? Curr Opin Microbiol 24:66–71
40. Wu G, Zheng Y, Qureshi I, Zin HT, Beck T, Bulka B, Freeland SJ (2007) SGDB: a database of synthetic genes re-designed for optimizing protein over-expression. Nucleic Acids Res 35: D76–D79
41. de Marco A, Vigh L, Diamant S, Goloubinoff P (2005) Native folding of aggregation-prone recombinant proteins in *Escherichia coli* by osmolytes, plasmid- or benzyl alcohol-overexpressed molecular chaperones. Cell Stress Chaperones 10:329–339
42. Komar AA, Lesnik T, Reiss C (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462:387–391
43. 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
44. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y (2015) Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell 59:744–754
45. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, Sachs MS, Liu Y (2013) Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 495:111–115
46. Zhang G, Hubalewska M, Ignatova Z (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol 16:274–280
47. Sander IM, Chaney JL, Clark PL (2014) Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. J Am Chem 136:858–861
48. Hu S, Wang M, Cai G, He M (2013) Genetic code-guided protein synthesis and folding in *Escherichia coli*. J Biol Chem 288:30855–30861
49. Kim SJ, Yoon JS, Shishido H, Yang Z, Rooney LA, Barral JM, Skach WR (2015) Protein folding. Translational tuning optimizes nascent protein folding in cells. Science 348:444–448
50. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA (2016) Synonymous codons direct cotranslational folding toward different protein conformations. Mol Cell 61:341–351. http://www.sciencedirect.com/science/article/pii/S1097276516000095
51. Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12:683–691
52. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C (2014) Exposing synonymous mutations. Trends Genet 30:308–321
53. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J (2015) Codon optimality is a major determinant of mRNA stability. Cell 160:1111–1124

54. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, Luff JD, Valecha M, Everett JK, Acton TB, Xiao R, Montelione GT, Aalberts DP, Hunt JF (2016) Codon influence on protein expression in *E. coli* correlates with mRNA levels. Nature 529:358–363
55. Pechmann S, Chartron JW, Frydman J (2014) Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat Struct Mol Biol 21:1100–1105
56. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352
57. Chin JX, Chung BK-S, Lee D-Y (2014) Codon optimization on-line (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics 30:2210–2212
58. Hoover DM, Lubkowski J (2002) DNA Works: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res 30, e43
59. Guimaraes JC, Rocha M, Arkin AP, Cambray G (2014) D-Tailor: automated analysis and design of DNA sequences. Bioinformatics 30:1087–1094
60. Gaspar P, Oliveira JL, Frommlet J, Santos MAS, Moura G (2012) EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics 28:2683–2684
61. Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD (2006) GeneDesign: rapid, automated design of multikilobase synthetic genes. Genome Res 16:550–556
62. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S (2006) Gene designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformat 7:285
63. Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, Jahn D (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res 33:W526–W531
64. Gaspar P, Moura G, Santos MAS, Oliveira JL (2013) mRNA secondary structure optimization using a correlated stem-loop prediction. Nucleic Acids Res 41, e73
65. Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S (2007) Optimizer: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 35:W126–W131
66. Wu G, Bashir-Bello N, Freeland S (2005) The synthetic gene designer: a flexible web platform to explore sequence space of synthetic genes for heterolo-gous expression. In: 2005 I.E. computational systems bioinformatics conference, workshops and poster abstracts, 2005 Aug 8–11. Stanford University, California, pp 258–259
67. Li MH, Bode M, Huang MC, Cheong WC, Lim LS (2012) *De novo* gene synthesis design using TmPrime software. Methods Mol Biol 852:225–234
68. Jung S-K, McDonald K (2011) Visual gene developer: a fully programmable bioinformatics software for synthetic gene optimization. BMC Bioinformat 12:340
69. Gould N, Hendy O, Papamichail D (2014) Computational tools and algorithms for designing customized synthetic genes. Front Bioeng Biotechnol 2:41
70. Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res 28:292
71. 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
72. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42
73. Pechmann S, Frydman J (2011) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20:237–243
74. Chaney JL, Clark PL (2015) Roles for synonymous codon usage in protein biogenesis. Annu Rev Biophys 44:143–166
75. Clarke TF 4th, Clark PL (2008) Rare codons cluster. PLoS One 3, e3412
76. Clarke TF 4th, Clark PL (2010) Increased incidence of rare codon clusters at 5′ and 3′ gene termini: implications for function. BMC Genomics 11:118
77. Allert M, Cox JC, Hellinga HW (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames. J Mol Biol 402:905–918

78. Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 6, e1000664
79. Zalucki YM, Beacham IR, Jennings MP (2009) Biased codon usage in signal peptides: a role in protein export. Trends Microbiol 17:146–150
80. Fluman N, Navon S, Bibi E, Pilpel Y (2014) mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. Elife 3:eLife.03440
81. Chartier M, Gaudreault F, Najmanovich R (2012) Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. Bioinformatics 28:1438–1445
82. McKown RL, Raab RW, Kachelries P, Caldwell S, Laurie GW (2013) Conserved regional 3′ grouping of rare codons in the coding sequence of ocular prosecretory mitogen lacritin. Invest Ophthalmol Vis Sci 54:1979–1987
83. Widmann M, Clairo M, Dippon J, Pleiss J (2008) Analysis of the distribution of functionally relevant rare codons. BMC Genomics 9:207
84. 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. J Mol Biol 193:413–417
85. Krasheninnikov IA, Komar AA, Adzhubeĭ IA (1988) Role of the rare codon clusters in defining the boundaries of polypeptide chain regions with identical secondary structures in the process of co-translational folding of proteins. Dokl Akad Nauk SSSR 303:995–999
86. Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R (2008) Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J Mol Biol 383:281–291
87. Kramer G, Boehringer D, Ban N, Bukau B (2009) The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol 16:589–597
88. Zhang G, Ignatova Z (2011) Folding at the birth of the nascent chain: coordinating translation with co-translational folding. Curr Opin Struct Biol 21:25–31
89. Waudby CA, Launay H, Cabrita LD, Christodoulou J (2013) Protein folding on the ribosome studied using NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 74:57–75
90. O'Brien EP, Ciryam P, Vendruscolo M, Dobson CM (2014) Understanding the influence of codon translation rates on cotranslational protein folding. Acc Chem Res 47:1536–1544
91. Gloge F, Becker AH, Kramer G, Bukau B (2014) Co-translational mechanisms of protein maturation. Curr Opin Struct Biol 24:24–33
92. Hoekema A, Kastelein RA, Vasser M, de Boer HA (1987) Codon replacement in the PGK1 gene of *Saccharomyces cerevisiae*: experimental approach to study the role of biased codon usage in gene expression. Mol Cell Biol 7:2914–2924
93. Caponigro G, Muhlrad D, Parker R (1993) A small segment of the MAT alpha 1 transcript promotes mRNA decay in *Saccharomyces cerevisiae*: a stimulatory role for rare codons. Mol Cell Biol 13:5141–5148
94. Deana A, Ehrlich R, Reiss C (1996) Synonymous codon selection controls in vivo turnover and amount of mRNA in *Escherichia coli* bla and ompA genes. J Bacteriol 178:2718–2720
95. Kramer EB, Farabaugh PJ (2007) The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. RNA 13:87–96
96. Zaher HS, Green R (2009) Fidelity at the molecular level: lessons from protein synthesis. Cell 136:746–762
97. Kramer EB, Vallabhaneni H, Mayer LM, Farabaugh PJ (2010) A comprehensive analysis of translational missense errors in the yeast *Saccharomyces cerevisiae*. RNA 16:1797–1808
98. Ribas de Pouplana L, Santos MA, Zhu JH, Farabaugh PJ, Javid B (2014) Protein mistranslation: friend or foe? Trends Biochem Sci 39:355–362
99. Dinman JD (2012) Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip Rev RNA 3:661–673

100. Caliskan N, Peske F, Rodnina MV (2015) Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. Trends Biochem Sci 40:265–274
101. Huang Y, Koonin EV, Lipman DJ, Przytycka TM (2009) Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. Nucleic Acids Res 37:6799–6810
102. Hamasaki-Katagiri N, Salari R, Simhadri VL, Tseng SC, Needlman E, Edwards NC, Sauna ZE, Grigoryan V, Komar AA, Przytycka TM, Kimchi-Sarfaty C (2012) Analysis of *F9* point mutations and their correlation to severity of haemophilia B disease. Haemophilia 18:933–940
103. Edwards NC, Hing ZA, Perry A, Blaisdell A, Kopelman DB, Fathke R, Plum W, Newell J, Allen CE, Shapiro SGA, Okunji C, Kosti I, Shomron N, Grigoryan V, Przytycka TM, Sauna ZE, Salari R, Mandel-Gutfreund Y, Komar AA, Kimchi-Sarfaty C (2012) Characterization of coding synonymous and non-synonymous variants in ADAMTS13 using ex vivo and in silico approaches. PLoS One **7**:e38864
104. Salari R, Kimchi-Sarfaty C, Gottesman MM, Przytycka TM (2013) Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. Nucleic Acids Res 41:44–53
105. Hamasaki-Katagiri N, Salari R, Wu A, Qi Y, Schiller T, Filiberto AC, Schisterman EF, Komar AA, Przytycka TM, Kimchi-Sarfaty C (2013) A gene-specific method for predicting hemophilia-causing point mutations. J Mol Biol 425:4023–4033