CellPress

A critical analysis of codon optimization in human therapeutics

Vincent P. Mauro and Stephen A. Chappell

Department of Cell and Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

Codon optimization describes gene engineering approaches that use synonymous codon changes to increase protein production. Applications for codon optimization include recombinant protein drugs and nucleic acid therapies, including gene therapy, mRNA therapy, and DNA/RNA vaccines. However, recent reports indicate that codon optimization can affect protein conformation and function, increase immunogenicity, and reduce efficacy. We critically review this subject, identifying additional potential hazards including some unique to nucleic acid therapies. This analysis highlights the evolved complexity of codon usage and challenges the scientific bases for codon optimization. Consequently, codon optimization may not provide the optimal strategy for increasing protein production and may decrease the safety and efficacy of biotech therapeutics. We suggest that the use of this approach is reconsidered, particularly for *in vivo* applications.

Optimizing codon usage for increased protein expression

The polypeptide chain(s) of most proteins can be encoded by a seemingly infinite number of mRNA sequences owing to the degenerate nature of the genetic code (see Glossary) [1]. Interestingly, mRNAs encoding the same polypeptide via different codon assignments can vary dramatically in the amount of protein expressed [2,3]. The attempt to produce more protein by altering codon assignments has led to the broad use of codon-optimized mRNAs for the bioproduction of protein pharmaceuticals and nucleic acid therapies. However, considerable evidence demonstrates that synonymous codon choices in natural mRNAs have evolved in response to diverse selective pressures at both the RNA and protein levels [4]. In addition, various studies have shown that synonymous codon changes can have unanticipated effects. Synonymous codon changes may affect protein conformation and stability, change sites of post-translational modifications, and alter protein function [5–9]. Moreover, synonymous mutations have been linked to numerous diseases [4,10–13]. Some potential risks associated with the use of codon-optimized mRNAs for producing recombinant protein drugs have been discussed recently [11,12,14,15]. These risks include the production of anti-drug antibodies which can reduce drug efficacy and cause allergic reactions.

In this article we critically review the scientific bases for codon optimization and identify additional risks. These include two potentially serious side-effects that pose unique risks for applications in nucleic acid therapies: (i) the production of novel peptides from alternative out-of-frame open reading frames (ORFs); and (ii) altered sites of post-transcriptional nucleotide modifications that can lead to the production of novel protein variants and ensembles. Understanding the potential risks of codon optimization so that they can be minimized or eliminated is crucial as nucleic acid therapies begin to gain traction. We suggest that the use of these approaches for human therapeutics should be carefully considered to avoid introducing unnecessary problems.

The genetic code, tRNAs, and wobble

The genetic code is degenerate because most amino acids are encoded by multiple synonymous codons (Figure 1). However, cells and organelles do not express 61 different tRNAs and vary dramatically in the relative expression of individual tRNAs [16,17]. For instance, in humans around 500 tRNA genes correspond to 48 codons; there are no tRNA genes for the remaining 13 codons [17]. Interestingly, an overlapping but different set of tRNA genes is missing in Chinese hamster ovary (CHO) cells, a cell line that is often used to produce therapeutic proteins. Despite the absence of these tRNA genes, mRNAs use the full complement of codons and synonymous codon usage is not affected by the absence of a cognate tRNA. For example, two codons encode aspartic acid (D) and have similar codon-usage even though there is no tRNA gene corresponding to the GAU codon. This is possible because of 'wobble', which enables both codons to be decoded by the same tRNA.

Wobble involves tRNAs with U or G in position 34, which base-pair to the third base in the codon (Figure 2). Some tRNAs with U in position 34 are capable of U-A and U-G base-pairing. Likewise, some tRNAs with G in position 34 are capable of G-C and G-U base-pairing. In addition, for some tRNAs adenine in position 34 is deaminated to inosine (I), which can base-pair to U, C, and A. This occurs for one tRNA in prokaryotes and 7–8 tRNAs in eukaryotes [18,19]. Modifications at different positions in tRNAs can restrict wobble in some cases and expand it in others [20]. Experimentally, it has been demonstrated that 25 cognate tRNAs comprise a minimum set that can sustain protein synthesis by the use of extended wobble interactions referred to as 'superwobbling'

Corresponding author: Mauro, V.P. (vmauro@scripps.edu).

Keywords: codon optimization; gene therapy; mRNA therapy; vaccine; A-to-I editing; tRNA wobble.

1471-4914/

 http://dx.doi.org/10.1016/j.molmed.2014.09.003

Glossary

Cistron: historically refers to a gene. In more contemporary usage, a cistron refers to the nucleic acid sequence that encodes a polypeptide chain.
Codon optimization: refers to experimental approaches designed to improve the codon composition of a recombinant gene based on various criteria without altering the amino acid sequence. This is possible because most amino acids are encoded by more than one codon. Most codon-optimization approaches avoid the use of rare codons. However, different approaches vary in the extent of other features considered. Features include mRNA elements that can inhibit expression, for example mRNA instability elements, the nucleotide context of the initiation codon, mRNA secondary structures, sequence repeats, nucleotide composition, internal ribosome entry sites, promoter sequences, and putative splice donor and acceptor sites [33,37,38,90]. In addition, some programs consider protein structural information, intragenic poly(A) sites, stop codons in alternative reading frames, and dinucleotides that are targets for RNase cleavage, mutation, and methylation-dependent gene silencing [38,90,91]. Moreover, some approaches have features that facilitate cloning, for example by adding or removing restriction sites [31,34,92]. Some approaches also allow oligonucleotides to be designed and optimized for gene synthesis using different strategies [31,32,34,42,91,93]. In some cases, gene synthesis is the primary consideration and the ability to alter codon usage provides flexibility for good oligonucleotide design [94,95].
Codon-optimization applications: include optimizing mRNAs for expression in different organisms by using organism-specific codon-usage frequencies (e.g., [36,37,93,96]), designing RNAi resistant genes (e.g., for gene rescue experiments [96]), and embedding genetic watermarks into genes [97]. Specific codon-optimization approaches have also been developed for DNA vaccine and gene therapy applications [98,99].
Codon usage: refers to the non-random use of codons in mRNAs. Codon usage in many organisms has been quantified using various calculations, including the frequency of use of optimal codons [28], the codon bias index [29], relative synonymous codon usage [30], the codon adaptation index [100], and the effective number of codons [101].
Genetic code: refers to the nucleotide triplets termed codons that specify specific amino acids. Codons comprise the coding sequences of genes and are recognized at the mRNA level during the process of translation. The genetic code consists of 64 trinucleotide codons: 61 triplets specify 20 amino acids and three serve as stop codons. Only two amino acids, methionine (Met, M) and tryptophan (Trp, W) are encoded by single codons; other amino acids are encoded by 2, 3, 4, or 6 synonymous codons.
Major histocompatibility complex I (MHC): in vertebrates, MHC molecules bind to peptides that are typically derived from endogenous proteins and present them at the cell surface where they serve as ligands for antigen receptors of cytotoxic ($CD8^+$) T cells. These MHC/peptide complexes allow the immune system to distinguish normal, healthy cells from those harboring pathogenic infection, or having undergone tumorigenic transformation. Peptides were originally believed to originate from the proteolytic cleavage of mature, functional proteins [102]. It was subsequently proposed that these peptides may also result from defective ribosomal products, prematurely terminated proteins, and misfolded polypeptides [103–105]. However, it is now evident that many peptides are derived from newly synthesized proteins [106]. In addition, there is evidence that many of these peptides are translated during the pioneer round of mRNA translation [107,108]. Moreover, some peptides are encoded in alternative reading frames [74]. Interestingly, many of the epitopes that originate from alternative reading frames initiate translation at non-AUG codons. In the best studied case, a CUG start codon was shown to initiate translation using an elongator leucine tRNA ($tRNA^{Leu}CAG$) as opposed to the initiator methionine-tRNA ($tRNA^{Met}CAU$) [109].
Nucleic acid therapies: refers to approaches that use DNA or RNA to mediate a therapeutic effect. Nucleic acid therapies include gene therapy, mRNA therapy, DNA vaccines, and RNA vaccines.
Ribosomal tethering and clustering: a hypothesis which involves direct binding of the initiator Met-tRNA to an accessible initiation codon. This binding occurs while the initiator Met-tRNA is associated with a ribosomal subunit that is either bound to the mRNA (tethered) or localized by more transient interactions (clustered). According to this model, alternative initiation is an inevitable consequence of initiation.
Ribosome profiling: a technique whereby the positions of translating ribosomes (ribosome footprints) can be mapped onto mRNAs at single-nucleotide resolution. This technique has also been used to map ribosomal complexes at translation initiation sites by first treating cells with a drug (Harringtonine or Lactinomycin) that blocks translation initiation and freezes initiation complexes at the start-site [110].
RNA editing: a process that describes various post-transcriptional modifications that alter specific nucleotides in RNA molecules. For many organisms these modifications increase transcriptome complexity and contribute to a higher level of protein diversity than is indicated by the number of genes and alternative splicing variants. A-to-I editing involves the selective deamination to inosine of particular adenosines contained within imperfect, double-stranded regions of RNA. This process is catalyzed by ADARs (adenosine deaminases acting on RNA) (reviewed in [111]). A-to-I editing occurs predominantly in tissues derived from the nervous system [112]; this editing occurs most frequently within non-coding sequences and particularly in the RNA structures formed by inverted Alu repeats (reviewed in [113]). However, A-to-I editing also occurs in other types of repetitive sequences, including microRNAs, and tRNAs [85,111]. A-to-I editing within intronic sequences of pre-mRNAs may serve to modulate alternatively spliced variants by altering splice acceptor or donor sites, or by introducing new splice sites. Recent findings in *Drosophila* demonstrate that the extent of editing is determined cotranscriptionally [114], and supports a close relationship between splicing and editing [115].
Scanning hypothesis: a hypothesis proposed by Marilyn Kozak which suggests that, during translation initiation, the small ribosomal subunit scans from the mRNA cap-structure found at the 5′ ends of mRNA to the initiation codon [116]. This model proposes 5′ to 3′ linear ribosomal movement in which the 5′ leader is inspected nucleotide by nucleotide until the initiation codon is identified. At this point, scanning stops, the large ribosomal subunit attaches, and peptide synthesis begins. Kozak later extended and modified the model to include leaky scanning and reinitiation to accommodate examples of translation initiation that are inconsistent with the original model. In addition, the model has recently been modified by others to allow scanning from an internal mRNA recruitment site, scanning across the base of some stem-loop structures, and bidirectional movement resulting from backward scanning or diffusion.
Superwobble: this type of wobble pairing, also referred to as four-way wobbling or hyperwobbling, occurs between an unmodified uridine in the tRNA at position 34 and the third nucleotide of the codon.
Synonymous codon: refers to groups of codons that encode the same amino acid. A mutation that changes a codon to a synonymous codon is termed a silent mutation because the amino acid sequence is unaltered. However, the term silent mutation may be a misnomer as numerous diseases are associated with synonymous codon changes.
tRNA channeling: suggests that the translation machinery is organized during protein synthesis to facilitate charging, use, and recharging of tRNAs without their diffusion into the cytosol. This model is based on a series of studies from Deutscher in the 1990s [117,118], and is consistent with various recent experimental observations [52,119–121].
Wobble: Crick's wobble hypothesis [122] suggests that standard base-pairing is used for the first two nucleotides of a codon, but that the stringency of base-pairing is relaxed in the third position.

[21,22]. Superwobbling can explain how translation can occur with fewer tRNAs than are predicted by the wobble hypothesis. Although superwobbling has only been demonstrated in chloroplasts to date, its occurrence in mammals is plausible.

The history, scientific basis, and art of codon optimization

Degeneracy in the genetic code enabled the first recombinant peptide, a mammalian somatostatin, to be expressed in *E. coli* without knowing the mRNA sequence for the peptide [23]. A coding sequence was obtained by reverse translating the amino acid sequence. Codon usage was biased with consideration of the effects of various gene sequences on translation and transcription, as well as to facilitate gene synthesis.

When the first gene sequences were determined, it was noted that codons are used in a non-random manner [24–26]. For highly expressed genes in *E. coli* and yeast, the non-random use of synonymous codons was found to be correlated with tRNA abundance [27–30]. The observation that some highly expressed genes preferentially use a subset of codons suggested that codon bias and protein expression are causally linked, and that it might be possible to enhance expression by mimicking the pattern of codon bias of highly expressed mRNAs. This prospect led to the development of numerous codon-optimization programs and commercial services. These approaches differ in how codon bias is measured, the number of variables