Critical reflections on synthetic gene design for recombinant protein expression
Annabel HA Parret¹, Hüseyin Besir² and Rob Meijers ¹

Gene synthesis enables the exploitation of the degeneracy of the genetic code to boost expression of recombinant protein targets for structural studies. This has created new opportunities to obtain structural information on proteins that are normally present in low abundance. Unfortunately, synthetic gene expression occasionally leads to insoluble or misfolded proteins. This could be remedied by recent insights in the effect of codon usage on translation initiation and elongation. In this review, we discuss the interplay between optimal gene and vector design to enhance expression in a particular host and highlight the benefits and potential pitfalls associated with protein expression from synthetic genes.

Addresses
¹ European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany
² Protein Expression and Purification Core Facility, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Corresponding author: Meijers, Rob (r.meijers@embl-hamburg.de)

Current Opinion in Structural Biology 2016, 38:155–162
This review comes from a themed issue on New constructs and expression of proteins
Edited by Rob Meijers and Anastassia Perrakis
For a complete overview see the Issue and the Editorial
Available online 21st July 2016
http://dx.doi.org/10.1016/j.sbi.2016.07.004
0959–440X/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction
The efficient overexpression of proteins is a prerequisite for successful structure determination. Structural genomics studies on a wide range of soluble protein targets over more than a decade have shown that more than 40% of the constructs failed due to limited expression (Figure 1). The high failure rate is a result of the use of standardized approaches, and can serve as a baseline for a structure determination project on a specific target. To alleviate these ominous statistics, a number of approaches have been adopted which address recombinant protein expression. The use of parallel cloning strategies employing different construct lengths and purification tags can increase success rate to a certain extent [1]. Alternatively, larger amounts of protein can be obtained by scaling up expression using large-scale fermentation [2]. However, large-scale fermentation often requires optimization to address problems of plasmid loss, oxygen deprivation and pH fluctuation [3].

A more cost-effective approach is the optimization of the protein source, either by opting for a protein species of high natural abundance, or by using an efficient recombinant expression system. Traditionally, selecting protein homologues from species that can be purified from natural source has been rather successful [4–8]. Nevertheless, there is an increasing need to obtain the structure of a protein from a particular species, in order to assess species-specific functional aspects. Structures of human or murine proteins involved in the immune or nervous system can help to relate the effect of specific point mutants or posttranslational modifications. Similarly, structural analysis of virulent proteins from bacterial pathogens can reveal species-specific details important for drug design. In this review, we focus on the opportunities that gene synthesis provides to optimize recombinant protein expression of specific protein targets.

Application of synthetic genes for protein production
The use of synthetic genes for the production of proteins for structural characterization is widespread. In general, gene synthesis may expedite the production of a series of constructs with engineered restriction sites. It is often preferred to the use of cDNA, which can contain several isoforms or splice variants. For target proteins identified from metagenomes or extremophiles, gene synthesis may provide the only route to protein expression, because cDNA is not available. Protein expression from codon-optimized genes has been especially successful for proteins involved in cell signalling whose native codon sequence can be strongly regulated, such as Irisin [9], Parkin [10] and Netrin [11]. Codon optimization has also uncovered hidden levels of gene regulation that are encoded in codon bias. Codon optimization of the FRQ protein involved in the regulation of circadian rhythms resulted in a loss of function, although expression levels were increased [12]. Codon optimization also revealed a hidden quality control mechanism encoded in the gene of the cystic fibrosis transmembrane conductance regulator (CFTR), an anion channel membrane protein [13]. Similarly, optimization of the CTP1L endolysin gene led to the discovery of a secondary translation site that produced a truncated protein. The truncated protein forms a 1:1 complex with the full length protein,
and lysis activity is substantially reduced when the secondary translation site is removed by codon optimization [14]. Toll-like receptors (TLRs) contain leucine-rich domains, and it was shown through targeted leucine codon optimization that transcription of TLRs is regulated by leucine codons to balance the abundance of certain receptors [15]. It is clear that gene synthesis may provide even unexpected benefits, but there are also numerous reports where it leads to reduced protein yield [16,17].

Adapting codon usage of target genes using codon optimization algorithms
Exploiting gene synthesis for transgene expression enables researchers to influence the multitude of variables controlling protein yield. The biased usage of synonymous codons has a strong effect on protein expression [18**]. Species-specific differences in this so-called codon bias can result in tRNA depletion, which slows down protein translation. This effect can be in part compensated by using an expression host that expresses rare tRNAs [16], however this is often insufficient [19].

An extensive collection of software tools (Table 1) are available to analyze codon usage and identify unwanted features that have been identified to hamper expression (Table 2). Gene design tools enable researchers to optimize the sequence of their gene of interest (GOI) [20]. A common strategy to manipulate codon bias to maximize protein translation is based on the ‘codon adaptation index’ (CAI) [21]. CAI is a species-specific index for codon frequency based on a set of highly expressing genes. Individual genes can then be analyzed, comparing the actual codons used to a fully optimized gene that contains only the most frequently used codons. It should be noted that CAI maximization does not necessarily correlate with high protein yield [22–25]. In general, the best results using CAI optimization are obtained when a subset of less frequent codons is replaced, after which a secondary list of criteria should be optimized involving DNA and RNA sequence elements that can negatively influence expression of the GOI (Table 2). It appears some features of codon bias are universal for all expression systems, whereas others are specific for prokaryotic or eukaryotic expression.

Codon optimization algorithms will generally allow adjustment of several of these undesirable sequence properties simultaneously [20,26–28], although the weight given to each sequence parameter varies between different algorithms. It is important to note that proprietary CAI-based codon optimization algorithms adjust a subset of rare codons in a random manner. As a result, gene optimization is ambiguous, since the replacement of rare codons is not based on empirical data. Species-specific gene optimization routinely involves the use of the species-relevant CAI index, taking into account the codon frequency of the particular species. Other factors that affect for example eukaryotic protein expression are most often not considered, because the influence on expression levels is poorly understood.

Recent insights into the relationship between mRNA, protein expression and protein folding
It is obvious that codon-mediated translational control is not yet well understood, and further insight will benefit gene design for protein expression. Recent studies indicate that rare codons may play different roles, such as safeguarding mRNA structure [29], depending on their positioning within the gene. A large scale study on bacterial genes expression in *Escherichia coli* under a common promoter system revealed that mRNA stability correlates with expression levels [30*]. However, there is a difference in the effect between the ‘head’ of the gene (covering approximately 16 codons) and the rest of the gene. In the ‘head’, the mRNA structure is more important, whereas the rest of the gene is rather affected by codon-mediated translation elongation efficiency. This phenomenon is likely a universal property for both prokaryotic and eukaryotic protein expression [31]. Cell-free expression studies have shown that universal translation initiation tags can be introduced upstream of the GOI that boost expression both
in bacterial and eukaryotic systems [32,33]. To boost protein expression, it may be sufficient to modify the codon usage at the ‘head’ of the gene in order to optimize the mRNA structure for efficient translation initiation, and leave the codon sequence of the rest of the gene intact.

In prokaryotes, translation of a transcript is initiated before the transcription process is complete, due to the recruitment of ribosomes to the newly synthesized mRNA. In genes which are not highly expressed, the translation efficiency is not necessarily correlated with protein synthesis rates, as ribosome profiling studies have shown [34–36]. These studies point to another level of regulation, where codon usage affects protein folding and post-translational modifications. A bioinformatics analysis in E. coli revealed that putative translational attenuation sites are widespread, and were identified in about 60% of protein coding sequences [37]. In this study, Li et al. proposed that these internal Shine–Dalgarno sequences are the main determinants of translation elongation rates in bacteria allowing for translational pausing where necessary.

Bioinformatic analyses indicate that rare codon clusters are found in a wide variety of genes, especially in the 5′ and 3′ termini [38–40]. Such clusters are thought to slow down the ribosome during translation elongation in order to allow co-translational protein folding [41–43]. Slow elongation is especially important in case of membrane protein targeting [44,45,46]. As a consequence, modifying the codons involved in the regulation of translation speed can culminate in protein misfolding. In one particular case, the replacement of frequently used codons by rare codons at domain boundaries resulted in an improved solubility of an enzyme produced in E. coli due to slower translation rate at these sites [47]. To circumvent deleterious effects of codon usage during co-translational protein folding, Angov et al. proposed a so-called codon harmonization strategy where rare codons are positioned where domain boundaries were predicted [48,49]. When compared to conventional codon optimization, this strategy resulted in improved expression level and soluble protein yield when overexpressing optimized genes from the eukaryotic parasite Plasmodium falciparum.
in *E. coli*. All these studies indicate that in gene synthesis, it may be important to conserve codon usage in some parts of the gene, to maintain the regulatory elements that affect folding and post-translational modifications.

### The interface between the synthetic gene and the expression vector

It is important to keep in mind that the mRNA produced to translate into protein is often not limited to the synthetized gene, but contains fragments derived from the expression vector. In fact, the challenge to optimize recombinant protein expression involves three intersecting areas. First, the choice of expression host, second, the selection of the expression vector [50] and third, the GOI (Figure 2). Much research has been done to optimize expression vectors adapted to the expression host. However, the adaptation of the GOI towards the expression vector is often neglected. Most bacterial vector systems are used as plug-and-play, where the GOI is inserted downstream of a vector-encoded ribosome binding site (RBS). As part of the 5' untranslated region (5' UTR), the RBS controls the rate of ribosome recruitment as well as translation initiation.

### Table 2

<table>
<thead>
<tr>
<th>Codon sequence features potentially influencing protein expression levels</th>
<th>Possible effect</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G + C content</td>
<td>mRNA secondary structure formation; can slow down or inhibit translation</td>
<td>[15,22,88,89]</td>
</tr>
<tr>
<td>Very high G + C content (&gt;70%)</td>
<td>Can slow down transcription elongation</td>
<td></td>
</tr>
<tr>
<td>Very low G + C content (&lt;30%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mRNA secondary structure</strong></td>
<td>Can promote mRNA stability</td>
<td>[31,53,56,90-92]</td>
</tr>
<tr>
<td>Near RBS</td>
<td>Reduction of translational initiation</td>
<td></td>
</tr>
<tr>
<td><strong>Repetitive sequences</strong></td>
<td>Inhibition of translation through formation of mRNA stem–loops</td>
<td>[16]</td>
</tr>
<tr>
<td><strong>Rare codons</strong></td>
<td>Can result in translational pausing, tRNA depletion resulting in low protein yield, mRNA degradation</td>
<td>[27,39,93,94]</td>
</tr>
<tr>
<td>Global</td>
<td>Promotes accurate co-translational folding</td>
<td>[40-42,48]</td>
</tr>
<tr>
<td>At protein domain boundaries</td>
<td>Shorter protein (if in frame); mixture of two proteins; incorrect protein if out of frame</td>
<td>[14]</td>
</tr>
<tr>
<td><strong>Alternative start codons</strong></td>
<td>Recombination hotspots; DNA instability</td>
<td>[95]</td>
</tr>
<tr>
<td>Chi-site stretches</td>
<td>Leaderless transcripts; shorter or incorrect protein</td>
<td>[96-98]</td>
</tr>
<tr>
<td>Alternative transcriptional initiation sites</td>
<td>Can effect mRNA stability</td>
<td>[31,99]</td>
</tr>
<tr>
<td>RNase E sites</td>
<td>Can cause translational pausing (ribosome stalling)</td>
<td>[37]</td>
</tr>
<tr>
<td>SD-like RBS sequences</td>
<td>Can result in incorrect transcription initiation</td>
<td>[88,100]</td>
</tr>
<tr>
<td><strong>Eukaryotic features</strong></td>
<td>Can result in incorrect transcription initiation</td>
<td></td>
</tr>
<tr>
<td>High CpG content</td>
<td>Missplicing of mRNA; incorrect protein</td>
<td>[101]</td>
</tr>
<tr>
<td>Internal TATA-boxes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptic splice sites</td>
<td>Can induce premature termination of translation</td>
<td>[102,103]</td>
</tr>
<tr>
<td>Potential polyA sites</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Optimization triangle illustrating the key determinants in recombinant protein expression. The inset illustrates the anatomy of a typical bacterial overexpression vector belonging to the popular pET-series in the region surrounding the insertion site of the GOI. The asterisk indicates a potential scar resulting from the cloning strategy.
the efficiency of translation initiation [51]. It is generally assumed that the mRNA sequence of the 5′UTR of an
expression vector is optimized in order to maximize the
translational efficiency of the GOI within a particular
host. However, the mRNA structure of the initial ~40
coding nucleotides can be crucial to obtain high protein
expression levels, given that in general strong secondary
structure (due to e.g. high GC content) inhibits efficient
translation initiation [22,24,52–55,56]. An additional
global observation is the relatively higher abundance of
rare codons in the initial ~30–50 codons [58]. The corre-
lation between the presence of rare codons and weak
mRNA structure in the 5′ end of coding sequences is still
under debate [55].

In typical bacterial vectors, the 5′ end of the coding
sequence (excluding the GOI) often encodes secretion
signals, affinity tags, fusion proteins or linker remnants
resulting from cloning, but the mRNA structure in this
region is rarely scrutinized (Figure 2). This may explain
why the expression of some proteins is strongly affected
by the placement of a tag on either the N-terminus or C-
terminus or by the length and sequence of the linker
region between tag and first codon of the transgene. N-
terminal tags may have evolved towards optimal mRNA
structure, thus increasing expression of some proteins
with a peculiar codon sequence at the start of the gene.
For example, it has been shown that addition of a leader
sequence can overcome the notoriously low expression
of AT-rich genes in E. coli [59]. On the basis of the recent
evidence on the importance of codon usage at the trans-
lation initiation site, the choice of any expression vector
should include a review of this region. If the protein
expression is suboptimal, it could be worthwhile to opti-
mize elements involved in translation initiation upstream
of the GOI and include these in gene synthesis.

In eukaryotes, translation initiation is an intricate stepwise
process orchestrated by multiple components. Extensive
studies in Saccharomyces cerevisiae in particular revealed
that key regulatory elements, i.e. translation initiation motifs
and secondary structures in the 5′-UTR, modulate protein
abundance [58,60–62]. One particularly successful ap-
proach to manipulate protein yield in mammalian cells
has been the incorporation of viral introns and optimized
Kozak sequences into the translation initiation sequence of
the corresponding expression vectors [63]. Translation
termination can also be promoted by the introduction of
viral introns such as the post-transcriptional regulatory
element derived from woodchuck hepatitis virus [64]. It
is expected that further large-scale investigations into the
regulation of eukaryotic protein translation will yield more
elements that can boost protein expression.

Concluding remarks
Economically attractive access to gene synthesis has
enabled optimization of gene sequences in order to
improve recombinant protein expression. Hence, spe-
cies-specific gene optimization using proprietary algo-
rithms can increase protein yield. Unfortunately,
predicting the output of the different optimization algo-
rithms is still a major challenge, especially for membrane
and secreted proteins, due to the many parameters influ-
encing codon usage. It would be a great advance if
commonly used codon optimization algorithms were im-
proved in the future to incorporate the new insights
obtained in recent years with respect to codon usage.
This will probably involve segmentation of the gene,
where each segment has different requirements for codon
optimization, including segments for translation initia-
tion, domain boundaries and sites for post-translational
modifications.

Optimized protein translation initiation could be
addressed by vector design, rather than codon optimiza-
tion of the gene of interest. Indeed, synonymous ex-
change of a small number of codons or introducing
non-coding secondary structure elements at the 5′-end
of a gene can already be sufficient for a significant increase
in expression levels [65,66]. This can be combined with
gene fragment synthesis, consisting of linear DNA frag-
ments that can be used for direct cloning into expression
vectors or as template for PCR amplification. In this
respect, the Biobricks movement (http://biobricks.org/)
is a promising initiative to deliver direct access to tools to
make proteins from genetically optimized building blocks
[67].

The expectation is that broad studies on the relationship
between codon usage and tRNA abundance as well as
mRNA structure and stability will lead to a set of robust
rules for gene design. This will further enable efficient
protein expression that is affordable for the larger molec-
ular biology community.

Conflict of interest statement
Nothing declared.

Acknowledgements
The authors are very grateful to Xia Sheng (Genscript) and Michel Koch for
providing valuable input into the manuscript. We would like to apologize to
those authors whose work could not be cited here due to space restrictions.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


This paper describes how modification of codon usage in a circadian protein revealed the use of specific codons regulates phosphorylation and structure stability.


18. Quax TE, Claassens NJ, Soli D, van der Oost J: Codon bias as a means to fine-tune gene expression. Mol Cell 2015, 59:149-161. Excellent review illustrating the different types of codon bias and its underlying principles. The authors describe the potential of codon bias to modulate gene expression and explore the challenges associated with exploiting codon bias as a tool for synthetic biology.


Large-scale protein expression study combined with biochemical analysis of optimized synthetic genes. This extensive analysis in E. coli strongly suggests that codon content directly modulates two competing processes, mRNA stability and protein elongation.


40. Chartier M, Gaudreault F, Najmanovich R: Large-scale analysis of conserved rare codon clusters suggests an involvement in
Relevant study underlying the importance of slow protein elongation for membrane protein targeting in E. coli.
This paper reports expression data of an extensive library of synthetic reporter constructs consisting of 14,000 combinations of promoters, ribosome binding sites and 11 N-terminal codons fused to GFP. The magnitude of this study allows for detailed dissection of how individual N-terminal codons effect expression yield of GFP fusions. The authors report an up to 14-fold increase in expression when using N-terminal rare codons.


