Biocatalysis, Protein Engineering, and Nanobiotechnology Biotechnology and Bioengineering
DOI 10.1002/bit.26183

Synthetic Gene Design – The rationale for codon optimization and implications for molecular pharming in plants†

Gina Webster[1], Audrey Y-H. Teh[1],[*] and Julian K-C. Ma[1*]

[1]Molecular Immunology Unit, Institute for Infection and Immunity, St. George's University of London, SW17 0RE, London, UK

*Contributed equally.

Correspondence

Audrey Y-H. Teh
Molecular Immunology Unit,
Institute for Infection and Immunity,
St. George's University of London
SW17 0RE, London, United Kingdom
Tel: +442087255667
E-mail: ateh@sgul.ac.uk

†This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/bit.26183]

Received November 17, 2015; Revision Received August 10, 2016; Accepted September 5, 2016

ABSTRACT

Degeneracy in the genetic code allows multiple codon sequences to encode the same protein. Codon usage bias in genes is the term given to the preferred use of particular synonymous codons. Synonymous codon substitutions had been regarded as “silent” as the primary structure of the protein was not affected; however, it is now accepted that synonymous substitutions can have a significant effect on heterologous protein expression. Codon optimization, the process of altering codons within the gene sequence to improve recombinant protein expression, has become widely practised.

Multiple inter-linked factors affecting protein expression need to be taken into consideration when optimizing a gene sequence. Over the years, various computer programmes have been developed to aid in the gene sequence optimization process. However, as the rulebook for altering codon usage to affect protein expression is still not completely understood, it is difficult to predict which strategy, if any, will design the ‘optimal’ gene sequence.

In this review, codon usage bias and factors affecting codon selection will be discussed and the evidence for codon optimization impact will be reviewed for recombinant protein expression using plants as a case study. These developments will be relevant to all recombinant expression systems, however, molecular pharming in plants is an area which has consistently encountered difficulties with low levels of recombinant protein expression, and should benefit from an evidence based rational approach to synthetic gene design.

KEYWORDS: Synonymous codons, heterologous protein expression, genetic code

1. Introduction

The genetic code is made up of 64 different tri-nucleotide codons: 61 encode amino acids and the remaining 3 encode stop codons (Gustafsson et al. 2004). Given that there are only 20 amino acids, most can therefore be encoded by more than one synonymous codon (the exceptions being methionine and tryptophan) (Gustafsson et al. 2004; Hershberg & Petrov 2008). This degeneracy in the genetic code (Crick et al. 1961) may have evolved as a way to preserve a protein's structural information in case of mutation (Zull & Smith 1990).

The use of synonymous codons is made possible through multivalent transfer-RNAs (tRNAs). tRNAs are 75-95 nucleotide structures required for the translation of the genetic code into protein. Each tRNA is linked (charged) with a specific amino acid and contains an anti-codon triplet (tRNA isoacceptor) which is complementary to a specific messenger RNA (mRNA) codon. There are 21 isoacceptor families in total: one for each amino acid and one for seleno-cysteine. Each family can consist of up to five members, for example there is one tRNA^{trp} but five tRNA^{leu} and these members can differ in their anticodon sequence as well as in tRNA backbone residue sequence (Goodenbour & Pan 2006).These multivalent tRNAs allow for binding to synonymous codons through movement (or 'wobble') in the 5' anti-codon. Therefore, even though there are 61 codons that encode the amino acids, fewer than 61 tRNAs are required (Cannarozzi et al. 2010). Thus there are approximately 30 tRNA isoacceptors in prokaryotes and 41-55 tRNA isoacceptors in eukaryotes (Kirchner & Ignatova 2014).

Synonymous codon substitutions had originally been regarded as "silent" in nature as they did not affect the amino acid sequence of the protein, and were therefore regarded as unimportant (Chamary et al. 2006). However, in the last three decades, it has become recognized that synonymous codon substitutions can have a significant effect on expression of heterologous proteins in recombinant expression systems, leading to the concept of codon optimization that may be specific for individual host production systems (Chamary et al. 2006; Gustafsson et al. 2004; Plotkin & Kudla 2011).

The process of “codon optimization” remains something of a dark art, offered by a number of DNA synthesis companies. Commercial algorithms used for codon optimization for different species are never disclosed and the relative benefits are often inconsistent. Indeed, genes that were codon optimized for another host have been reported to express better than the same gene that was codon optimized for the intended host (Lanza et al. 2014; Maclean et al. 2007; Rybicki et al. 2014). For example, human codon optimized human papillomavirus type 16 (HPV-16) L1 capsid protein gene resulted in higher levels of protein expression than the plant codon optimized version in *N. benthamiana* (Maclean et al. 2007).

In this review the basis for codon usage bias will be discussed and the evidence for codon optimization impact will be reviewed for recombinant protein expression using plants as a case study.

2. Codon usage bias can increase or decrease recombinant protein expression

Codon usage bias has been shown to correlate to gene expression levels, in species such as *C. elegans, D. melanogaster* & *A. thaliana*, (Duret & Mouchiroud 1999; Gouy & Gautier 1982) and it has been proposed as an important variable for enhancing recombinant protein expression in heterologous expression systems (Gustafsson et al. 2004; Kane 1995). Initially, research groups tried to optimize the codons within the gene by replacing rare codons with synonymous counterparts more frequently used by the host (Mueller et al. 2006). For example, the expression of an insect control protein gene in plants was increased by up to 100-fold by codon optimizing the sequence to more closely resemble the codon usage in plants (discussed in detail in section 3.2) (Perlak et al. 1991). Similarly, nucleotide sequences of L1 and L2 capsid proteins of papillomavirus (PV) were codon optimized to increase expression in mammalian cells (Zhou et al. 1999). Wild-type and codon-optimized sequences were transcribed in Cos-1 cells but only the codon-optimized sequence was translated (Zhou et al. 1999).

In contrast, expression of virus proteins have been reduced by altering synonymous codons or adjacent codons (“codon pairs”) for their rare counterparts (Coleman et al. 2008; Mueller et al. 2006). The premise was that by decreasing the protein translation of the poliovirus capsid protein, the virus would be attenuated resulting in a novel method for vaccine design that preserves the original amino acid sequence of the vaccine antigen. While similar amounts of virus particles were produced per cell, the virus-particle specific infectivity was decreased by up to 1,000-fold (Mueller et al. 2006). *Renilla* luciferase and Firefly luciferase reporter genes were used to determine the effect of altering codons on the rate of translation. The results determined that the modified capsid coding regions reduced the rate of translation by 80-90% (Mueller et al. 2006). Alteration of codon pairs in the polio capsid protein for their underrepresented counterparts resulted in attenuation of a vaccine candidate in mice, and provided protective immunity after challenge (Coleman et al. 2008). Again, *Renilla* luciferase and Firefly luciferase reporter genes were used and the results strongly suggested that underrepresented codon pairs reduced translation and this was probably sufficient to explain the attenuated phenotype (Coleman et al. 2008). The approach, termed synthetic attenuated virus engineering (SAVE) has also been applied to influenza virus (Mueller et al. 2010) and respiratory syncytial virus (Meng et al. 2014).

2.1. The Origins of Codon Bias

In *Escherichia coli* and *Saccharomyces cerevisiae*, codon usage is biased towards codons that match the most abundant tRNAs in the cell or that bind those tRNAs with optimal binding strength. For example, there are five Leu tRNAs in *E. coli*, but the tRNA with anticodon CAG is most abundant. It correlates with the codon CUG, which is used twenty times more frequently than any of the other Leu codons in highly expressed genes (Sharp et al. 2010). On the other hand, there is a single tRNA for Phe in *E. coli*, which has the anticodon GAA. It can translate both UUC and UUU, yet UUC which is perfectly complementary to the anticodon is used three times as often in highly expressed genes (Sharp et al. 2010).

The assumed mechanism is that the preferred codons tend to be ‘read’ by abundant tRNA molecules whereas rare codons tend to be ‘read’ by scarce tRNAs (Ikemura 1985). Thus, if the cellular tRNA abundances are known for a species it would be possible to predict the codons which are preferred. Although cellular tRNA abundances are unknown for many species (Stoletzki & Eyre-Walker 2007), in the case of *E. coli*, *S. cerevisiae* and *B. subtilis,* where they are known, they correlate to tRNA gene copy numbers (Dong et al. 1996; Ikemura 1985; Kanaya et al. 1999; Percudani et al. 1997). Therefore, in theory gene copy number could be used to predict the most abundant tRNAs (Duret 2000; Kanaya et al. 1999) and this can be used to predict preferred codons. However, it should be noted that this correlation was not found in *D. melanogaster* and humans (Kanaya et al. 2001).

There has been much debate about why some codons are preferentially selected in highly expressed genes. They may enhance translation efficiency, accuracy or a combination of both (Sharp et al. 2010). The traditional view is that efficiency of translation is increased as optimal codons can be translated faster than rare codons (Andersson & Kurland 1990; Ehrenberg & Kurland 1984; Sorensen & Pedersen 1991). This allows the ribosome to move along the mRNA faster and become released faster so they are available to translate other mRNAs. The more efficient use of ribosomes may lead to a quicker growth rate which gives an organism a selective advantage (Kudla et al. 2009; Sharp et al. 2010). An alternative view is that the use of optimal codons increases the accuracy of translation (Akashi 1994; Drummond & Wilke 2008). Translation accuracy is supported by the fact that amino acids that are buried in the secondary protein structure, which are important for protein folding, are normally encoded by preferred codons, (Komar et al. 1999) whereas surface residues, which participate in intermolecular interactions tend not to be (Drummond & Wilke 2008; Stoletzki & Eyre-Walker 2007; Zhou et al. 2009). In *E. coli* and yeast, this association of optimal codons encoding the amino acids important for protein folding was stronger in highly expressed genes (Zhou et al. 2009).

In heterologous protein expression, the gene of interest can be over-expressed and can account for up to 30% of the total protein within the cell (Plotkin & Kudla 2011). This may place strain on the cell and hence, the principles of codon bias for heterologous protein expression may differ substantially from those for endogenous expression. Thus, codons that are preferred for endogenous gene expression may not be the same as those for heterologous gene expression (Fath et al. 2011; Plotkin & Kudla 2011; Welch, Villalobos, et al. 2009). Welch *et al.* synthesized 81 genes encoding two different proteins for expression in *E. coli*, where each set of genes differed only in codon usage (Welch, Govindarajan, et al. 2009). They measured the levels of protein expression and determined that it varied 40-fold between the highest and lowest expressing genes. However, the apparently preferred codons did not correspond to those found in *E. coli's* highly expressed native genes. This contradicts the previously explained gene design principle, which is based on mimicking the codon bias of the highly expressed genes within the host species. The results of Welch *et al.* correlated with amino acid starvation of the tRNAs that recognize the preferred codons. Thus, they hypothesized that as amino acid charging of tRNA becomes limiting, only the tRNAs that are robust to starvation can support high translation levels of heterologous protein expression. Therefore successful codon bias of a gene depends on maintaining high levels of charged tRNAs and minimizing levels of uncharged tRNAs which can inhibit translation (Bonomo & Gill 2005; Dong et al. 1995; Harcum 2002; Welch, Govindarajan, et al. 2009).

This hypothesis is supported by studies modelling how the amino acid charging of tRNA isoacceptors responds when the amino acid becomes rate-limiting for protein synthesis (Elf et al. 2003). They determined the charged levels based on the ratio between the total concentration of isoacceptors and the frequencies at which their synonymous codons appear on the mRNA. Elf *et al.* suggested that isoacceptors with a low ratio will have a low charged fraction and these will read 'starvation-sensitive codons' whereas isoacceptors with a high ratio will have a higher charged fraction under amino acid starvation conditions, and will read ‘starvation-insensitive codons’ (Dittmar et al. 2005; Elf et al. 2003). This theoretical model was experimentally validated by Dittmar *et al.* who used microarray analysis to measure the charged levels of isoacceptors for leucine, threonine and arginine before and during amino acid starvation conditions (Dittmar et al. 2005). They discovered that before amino acid starvation, the majority of tRNAs were fully charged. However, during starvation the isoacceptors for the rate-limiting amino acid were selectively charged, confirming the theoretical predictions of Elf *et al*. (Dittmar et al. 2005). Thus, the theoretical model correctly identifies and orders the amino acid starvation-insensitive isoacceptors and based on Welch *et al.* results (Welch, Govindarajan, et al. 2009), this might be an important design criterion for genes for heterologous protein expression.

One technique used to combat low heterologous protein expression is to over-express infrequent tRNAs in the expression host, for the aforementioned reasons (Grote et al. 2005). For example, *E. coli* that has been genetically modified to over-express tRNAs can be bought commercially from companies such as Novagen & Stratagene. Tegel *et al.* tested the difference in protein expression between *E. coli* BL21 and *E. coli* Rosetta strain and determined that the overall success rate of protein production increased dramatically using the latter (Tegel et al. 2010). The Rosetta strain compensates for rare codons by expressing tRNAs for the AGG, AGA, AUA, CCC, GGA and CUA codons, which are commonly used by eukaryotes but rarely used by *E. coil* (Fu et al. 2007; Tegel et al. 2010). Nevertheless, this method is limited to certain host organisms as they have to be genetically altered.

2.2. The Functional impact of Codon Optimization

Analysis of *E. coli* gene sequences and the proteins they encode have indicated that amino acid sequences encoded by frequently used codons are associated with highly ordered structural elements, such as alpha helices. Similarly, amino acid sequences encoded by rare codons are often associated with domain boundaries (Thanaraj & Argos 1996). This led to the suggestion of two important ideas; the first is that the positioning of abundant or rare codons on the mRNA transcript may not be random, but can influence protein structure; the second is that rare codons at the domain boundaries may cause slow translational progress of the ribosome (ribosomal pausing) allowing for the partial folding of the protein within the ribosomal tunnel prior to the translation of the next structural element, thus minimizing deleterious aggregation (Komar et al. 1999; Shabalina et al. 2013).

Tuller *et al.* predicted that the speed of translation is slower during the first 30-50 codons, named the ‘ramp’, and then increases to a steady speed for the rest of the mRNA (Figure 1) (Tuller et al. 2010). The ramp has been hypothesized to slow translation elongation immediately after initiation, to facilitate uniform spacing between ribosomes and prevent congestion along the mRNA. This is supported by two pieces of evidence. The first is that the ramp is often observed with highly expressed genes which have many ribosomes attached (Fredrick & Ibba 2010). Secondly, a computational prediction of the ribosome density along mRNAs, based on this hypothesis, was subsequently confirmed experimentally by ribosome profiling (Ingolia et al. 2009). The major exception is the codon immediately after the initiation codon, which is predicted to be translated at a high rate. This is suggested to allow rapid release and recycling of the initiator tRNA, which is required as initiation is the rate-limiting step of translation (Tuller et al. 2010).

To summarize, the ribosome binds to the mRNA at the initiation codon, which is the rate limiting step of translation, so the next codon is a preferred codon to allow rapid release of the ribosome and the initiator tRNA. This is followed by rare codons for 30-50 codons which slows the ribosome down. This allows another ribosome to bind to the initiation codon and results in equal spacing of the ribosomes along the mRNA, preventing congestion.

Furthermore, it has been suggested that as the length of the ramp is similar to the length of the polypeptide needed to fill the exit tunnel of the ribosome, it could aid in chaperone-protein interactions (Ban et al. 2000). Thus, as the ribosome migrates from initiation to elongation, the emerging peptide interacts with chaperones, which might help to increase the amount of correctly folded protein. However, this remains to be demonstrated experimentally (Ban et al. 2000; Fredrick & Ibba 2010).

Overall, the importance of rare codons may help to explain why over-expression of rare tRNAs, to improve heterologous protein expression, is not guaranteed to work (Fredrick & Ibba 2010). Over-expression of rare tRNAs can change the translation efficiency of their codons, upsetting the function of the ramp. This may lead to ribosome congestion on the mRNA and premature termination (Fredrick & Ibba 2010; Tuller et al. 2010).

In yeast a codon usage pattern was observed where there was a strong tendency to use the same codon when an amino acid recurs (Cannarozzi et al. 2010). If the same codon was not re-used, the most closely related synonymous wobble codon was preferred. This 'auto-correlation' (Figure 2) was not simply the result of using the preferred codons, rather the use of codons which benefit translation. Therefore, rare codons were just as likely to be re-used. This suggested that re-using codons benefits translation and this was demonstrated experimentally showing that translation of auto-correlated mRNA was approximately 30% faster than anti-correlated mRNA (Cannarozzi et al. 2010). Furthermore, in yeast auto-correlation was strongest in highly expressed genes and pressure for codon correlation was greatest for rare codons, especially rare codons present in highly expressed genes, further indicating a translational advantage for codon correlation. Finally, it was concluded that, as distance between two synonymous codons increased, the pressure for codon auto-correlation declined (Cannarozzi et al. 2010).

These observations led to the suggestion of three different models for tRNA molecules exiting the ribosome (Cannarozzi et al. 2010). The first scenario is that tRNAs diffuse away from the ribosome at a speed relative to translation, thus, they are rapidly interspersed with the other isoacceptors. In this model, there would be no selection pressure for codon correlation, other than that caused by codon bias. The second scenario is that tRNA diffusion is slow compared to translation and acylation. Therefore, the recently used tRNA would still be in the vicinity of the ribosome, and could be recharged *in situ*. This would benefit translation as the ribosome does not need to wait for the arrival of an appropriate tRNA molecule. This model accounts for the observation that distance between identical codons affects codon correlation as the advantage would be strongest for codons that are close together in the gene sequence. The final scenario models tRNA molecules remaining associated with the translation machinery. This model is very similar to model two, except the autocorrelation pressure would decay slower over gene distance (Cannarozzi et al. 2010).

Both of the latter models suggest that the aminoacyl-tRNA-synthetases are associated with the ribosome, as this enzyme is required to re-charge the tRNA with the appropriate amino acid (Fredrick & Ibba 2010). Indeed, aminoacyl-tRNA-synthetases do form ordered complexes with ribosomes in eukaryotes (Deutscher 1984; Kaminska et al. 2009; Mirande et al. 1985), and these complexes have been shown to promote the re-use of charged tRNAs to the ribosome for protein synthesis, increasing translation rates (Kyriacou & Deutscher 2008). This finding combined with the fact that the third scenario explains the results of Cannarozzi *et al.* best, for all genomes tested, as autocorrelation decays more slowly than the diffusion model (second scenario) predicts, has led to the suggestion that tRNAs are recycled through binding to the ribosome, to allow them to be readily available when the next identical codon is ‘read’ (Fredrick & Ibba 2010). It should be noted that the above hypothesis relates to eukaryotes, these models remain to be demonstrated in bacteria (Fredrick & Ibba 2010).

In 1989, a non-random utilization of “codon pairs” in prokaryotic genes was observed, which correlated with gene expression (Gutman & Hatfield 1989). The over-represented codon pairs were translated more slowly than the under-represented codon pairs, thus highly expressed genes tended to avoid over-represented codon pairs (Gutman & Hatfield 1989; Irwin et al. 1995). This codon context bias effect was hypothesized to be due to interactions between adjacent aminoacyl-tRNA molecules in the “A” and “P” sites of the ribosome (Gutman & Hatfield 1989). The relationship between translational efficiency of a codon pair and its degree of bias in *E. coli* was investigated (Irwin et al. 1995). It was observed that the more over-represented the codon pair, the slower the pair was translated and vice versa. This led to the hypothesis that codon context has co-evolved with the abundance and structure of tRNA molecules in order to regulate translational efficiency without the need to alter the amino acid sequence (Irwin et al. 1995). This allows translational pauses to be incorporated into the gene sequence; the same functional significance as mentioned for the use of rare codons.

As discussed earlier, codon pair de-optimization has been used to synthesize novel viral vaccine candidates, validating that codon pair bias influences translational efficiency (Coleman et al. 2008; Mueller et al. 2010). Similarly, Moura *et al.* considered 72 orthologous highly conserved genes from the three domains of life and showed that synonymous mutations were selected to maintain codon context biases (Moura et al. 2011). This result supports the hypothesis that codon context is a factor in determining the evolution of the gene sequence by increasing translational efficiency and accuracy. However, the exact molecular mechanism behind codon context bias is unknown (Moura et al. 2011). Recently, an assay, based on translation of the *his* operon leader peptide, was developed to measure the translation speed of the ribosome *in vivo* (Chevance et al. 2014). The main advantage of this assay over others is that the speed of the ribosome is unaffected by the primary sequence of the *his* operon leader peptide, however, it is still affected by mRNA secondary and tertiary structures. The results indicated that codon context is a major factor affecting translational speed and thus, supported the model where codon context affects the overall stacking energy that is obtained as each charged tRNA enters the ribosome's "A" site which ultimately affects the translational speed of the ribosome. Overall, these experiments suggest that codon context should be a variable considered when redesigning genes for heterologous expression (Chevance et al. 2014).

3. Strategies for Codon Optimization and Implications for Plant Molecular Pharming

In this section, various codon optimization strategies will be discussed as well as computer algorithms that aid in gene design. Examples of codon optimization in regard to plant biotechnology will be described.

Plant molecular pharming is the production of recombinant pharmaceutical proteins using plant biotechnology. It has emerged as a useful technology particularly for products that cannot, for a variety of reasons (e.g. toxicity to host cells or requirement of rapid production times), be manufactured by current fermenter-based systems (Stoger et al. 2014). In 2012 the first plant molecular pharming product, the enzyme taliglucerase α, was approved for use in humans and a number of other pharmaceutical products are currently in the clinical trial development pipeline (Paul & Ma 2011; Paul et al. 2013).

The main limitation of molecular pharming is often that only low levels of recombinant protein expression are achieved (Daniell et al. 2001; Desai et al. 2010; Ma 2000), making this field an obvious candidate for application of codon optimization strategies.

3.1. Development of Algorithms for synthetic genes

Synthetic 'codon optimization' has tended to involve altering rare codons with synonymous codons used at a higher frequency (Angov 2011; Gustafsson et al. 2004), and has been implemented by many research groups using genes from various species in heterologous expression systems (reviewed in Gustafsson et al. 2004).

Although this strategy was successful for some products, in some cases, inclusion bodies of insoluble protein were obtained instead (Angov 2011; Welch, Villalobos, et al. 2009). This was suggested to be due to rapid translation preventing efficient 'self' or chaperone-aided folding. This hypothesis is supported by the mounting evidence that rare codons have a purpose, including causing ribosomal pausing. A slight variation on this codon optimization approach was suggested, called 'codon harmonization', where the codons in the gene sequence are altered to reflect the codon usage bias of the host organism (Angov et al. 2008). This means the gene sequence still retains its preferred and rare codons, these are just altered to the equivalent preferred and rare synonymous codons used by the host organism (Angov et al. 2008; Angov et al. 2011; Angov 2011). The codon harmonization algorithm was applied to genes encoding three *Plasmodium falciparum* proteins to be expressed in *E. coli* and gene expression was subsequently determined to increase from 4 to 1,000 fold that of non-optimized genes (Angov et al. 2008).

Codon optimization alone is not sufficient to optimize synthetic gene design for protein expression. Other gene sequence factors need to be considered. For example, mRNA secondary structure (Carlini et al. 2001; Chen et al. 1999), which has been suggested to physically alter the gene transcript, for example forming hair-pin loops, which can slow down or prevent the ribosome from translating the gene (Carton et al. 2007). Additionally, cryptic splice sites, polyadenylation signals and other regulatory elements may have to be avoided as they can cause undesirable processing of the mRNA (Carton et al. 2007).

Optimizing heterologous protein expression is thus a classic multidimensional optimization problem (Gustafsson et al. 2012; Welch, Villalobos, et al. 2009) and hypothesis-driven sequence design and testing as well as statistical analysis has led to the development of several algorithms for synthetic gene design.

A number of software packages were developed such as Codon Optimizer (Fuglsang 2003), JCat (Grote et al. 2005), SGD (Wu et al. 2006), OPTIMIZER (Puigbò et al. 2007) and ATGme (Daniel et al. 2015). More recently there has been refinement leading to multi-objective computer programs such as EuGene (Gaspar et al. 2012), COOL(Chin et al. 2014), D-Tailor (Guimaraes et al. 2014) and CoStar (Liu et al. 2014) which are discussed in detail in a recent review by Gould *et al*. 2014.

EuGene has two main functions; data gathering and optimization of the gene sequence (Gaspar et al. 2012). The former is achieved by the gene annotations in FASTA or GenBank formats being used to access various databases. Overall, retrieved and calculated data include Codon Adaptation

Index (CAI), GC content, Relative Synonymous Codon Usage (RSCU), protein structures, orthologs and codon pair bias (CPB). Optimization functions of the gene sequence include mRNA free energy optimization, codon usage (based on CAI and RSCU values), codon context (based on CPB), GC content, codon autocorrelation, restriction site elimination, elimination of repetitions and deleterious sites (Table 1) (Gaspar et al. 2012; Gould et al. 2014).

Codon Optimization OnLine (**COOL**) offers various gene optimization parameters based on CAI, individual codon usage, codon context, hidden stop codons, restriction site and repetitious pattern eliminations and GC content (Chin et al. 2014). These parameters are similar to other software programs, however, COOL uses a multi-objective optimization algorithm to generate various Pareto-optimal sequences (see Appendix), which can be viewed on a plot (Chin et al. 2014; Gould et al. 2014). This allows the user to compare best equivalent solutions and the gene sequence that offers the best compromise between design criteria can be chosen (Chin et al. 2014).

DNA-Tailor (**D-Tailor**) includes codon usage optimization based on CAI, GC content optimization, mRNA secondary structure optimization and restriction site and other repetitious pattern eliminations (Table 1) (Guimaraes et al. 2014). First, the template gene sequence is retrieved and analyzed to determine the current sequence fitness in each category to be optimized and the user specifies a predefined set of parameter scores to be obtained (the Design of Experiments (DoE)) (Gould et al. 2014; Guimaraes et al. 2014). The sequence is then synonymously mutated towards improving the fitness of one parameter, while following the user-specified rules (i.e. avoiding user-designated restriction sites, motifs and promoter & terminator sequences). The algorithm then determines if the modified sequence is an improvement on the original sequence by computing the Euclidian distance (see Appendix) of the modified sequence compared to the design criteria. If the modified sequence is an improvement on the original, it is added to the database (Gould et al. 2014; Guimaraes et al. 2014). The main significant difference between D-Tailor and other multi-objective sequence optimization programs is that it allows multiple design targets to be defined as combinations of sequence properties that represent a particular DoE (Guimaraes et al.

2014). This is in contrast to other optimization programs which define design objectives based on the desired response performance, which are linked to sequence properties (Guimaraes et al. 2014).

COStar was developed in 2014 by Liu *et al.* which simultaneously optimizes both local and global properties of a DNA sequence (Liu et al. 2014). It starts by splitting the amino acid sequence into fragments, termed "window peptides" using a sliding window method (see Appendix). Next, all combinations of synonymous codons that can encode the amino acid fragments are determined. These are then filtered to remove any sequences encoding rare codon usage before a weighted directed acyclic graph (see Appendix) is plotted. Finally the D-star Lite algorithm is used to find the best option of assembling all the various nucleotide fragments to achieve the users pre-defined "goals" (Table 1). Liu *et al.* showed that COStar was robust and able to optimize multi-objectives simultaneously and they obtained better results (lower scores of hairpin formation, fewer repetitive sequences & lower variance of GC content) for codon optimizing six specific sequences than using other codon optimization programmes, such as EuGene (Liu et al. 2014).

Many of the codon optimization software lack citations. Of these software packages, citations have been found for JCat and OPTIMIZER for expression systems such as *E. coli* (Fahimi et al. 2016; Guo et al. 2016; Karkhah & Amani 2016; Zhao et al. 2016), *S. cerevisiae* (Ask et al. 2013; Guadalupe-Medina et al. 2013; Li et al. 2015; Milne et al. 2015; Solis-Escalante et al. 2013), *N. benthamiana* (Binder et al. 2014), HEK293 cells (Shah et al. 2015), *B. subtilis* (Reilman et al. 2014), *C. crescentus* (Ko et al. 2013), *P. putida* (Dammeyer et al. 2013; Dammeyer et al. 2011), *S. typhimurium* (Manuel et al. 2011), *S. lividans* (Dubeau et al. 2009), wheat (Mihálik et al. 2015), transplastomic tobacco plants (chloroplast translation system) (Occhialini et al. 2015), *P. berghei* (Singer et al. 2015), *S. frugiperda* (Geisler et al. 2015), *S. pneumoniae* (Overkamp et al. 2013), *A. marginale* (Pierlé et al. 2013), *R. pomeroyi* (Green et al. 2013), *L. acidophilus* (Askelson et al. 2014), *C. reinhardtir* (Erpel et al. 2016), *Y. lipolytica* (Matthaus et al. 2014), *S. elongates* (van der Woude et al. 2016) & baculovirus (Maghodia et al. 2016). While it is encouraging that codon optimization software programs are used for a variety of species and purposes, most of the papers do not compare yields of native and codon optimized sequences, so yield comparisons cannot be made.

EuGene, D-Tailor and COOL are being used by several research groups and gene synthesis companies (personal communications), however as the tools are relatively new additions to the literature there are few citations. One group has utilized EuGene to harmonize the codons of *Pvs*48/45 gene from *P. vivax* for expression in *E. coli* and obtained a yield of protein of ~1mg/L of bacterial culture (Arevalo-Herrera et al. 2015). However, the native sequence was not used, so comparison with the codon optimized sequence could not be determined. Hopefully, in the near future researchers will start to see the potential of the software and their capabilities. Overall, we feel it would be fruitful if all researchers published the codon optimization software utilized and even better if they compared the native and codon optimized sequences so as to highlight if and how much codon optimization makes a difference.

Such processes are used by DNA synthesis companies to design their algorithms for optimizing gene sequences for recombinant protein expression. Few are open about their methodologies; however, an exception is DNA 2.0 (USA) who have published their algorithm design (Gustafsson et al. 2004; Welch, Villalobos, et al. 2009). Their process involved a codon usage table of highly expressed genes for the host organisms to set a threshold level to narrow down the codon choice. Candidate sequences were designed by selecting codons at random, using their probabilities from the codon usage table, and then passing the sequences through filters to ensure other design criteria are met. Firstly, unfavorable codon pairs and extreme GC content were eliminated. Next, repetitive sequences were eliminated before unfavorable mRNA structures were removed. Finally, restriction sites are included or excluded as required by the researcher (Gustafsson et al. 2004). Then expression levels were tested and the algorithm appropriately altered. This experiment-based optimization has been used to improve the gene design algorithms for multiple protein expression hosts, including *E. coli, P. pastoris, S. cerevisiae*, mammalian CHO cells and HEK293 cells (Gustafsson et al. 2012).

3.2. Codon Optimization in plant biotechnology

There are only a handful of reported examples of codon optimization being utilized to increase protein expression in plants (Batard et al. 2000; Jensen et al. 1996; Li et al. 2007; Mason et al. 1998; Perlak et al. 1991; Suo et al. 2006; Thomas & Walmsley 2014).

In 1991, expression of a *Bacillus thuringiensis* insect control protein gene was increased in tobacco and tomato plants by a combination of gene truncation and codon optimization (Perlak et al. 1991). The wild-type gene had localized regions of A+T richness, which resembled plant introns, ATTTA sequences, which destabilize plant mRNA, codons rarely used in plants and possible plant polyadenylation signals. By altering the codon usage across the gene in favor of plant optimal codons and/or gene truncation, protein expression was increased by up to 100-fold compared to the wild-type (Perlak et al. 1991).

In 1996, a hybrid bacterial glucanase gene was codon optimized for expression in barley (Jensen et al. 1996). The (1,3-1,4)-β-glucans from barley are the major polymers in endosperm cell walls and their degradation is required for enzymes to be utilized to provide the growing embryo with nutrients (Fincher 1975). The degradation is also required in malting and monogastric animal feed (Bamforth 1982; Graham et al. 1989; Hesselman & Aman 1986). However improved thermostability of the glucanase is required to survive the higher temperatures used for kiln drying of green malt. The hybrid (1,3-1,4)-β-glucanases from *Bacillus* species have improved thermostability at pH5.0 but were not expressed well in barley (Borriss & Zemek 1981). However, altering the gene for preferred (i.e. more frequently used) codons of barley (141 of 215 codons changed) led to successful expression of the enzyme during germination (Jensen et al. 1996).

In 2000, the 5’-end of the wheat cytochrome P450 gene sequence was engineered to improve the expression of the protein in tobacco (Batard et al. 2000). Cytochrome P450s form one of the largest families of genes in plants (Batard et al. 2000). P450s have been indicated to provide resistance to herbicides for monocotyledonous plants. Thus, many people were interested in the functional and structural characterization of P450s as well as efficiently expressing them in heterologous organisms (Batard et al. 2000). Expression of the wild-type genes in tobacco led to very low protein expression levels; no detectable bands by Western blot and <5 pmol/s/mg cinnamate 4-hydroxylase (C4H) activity. The wheat coding sequence has a high level of GC content and a strong codon usage bias, which is very different to tobacco. Batard *et al.* adjusted the 5'-terminus of the P450 gene sequence to resemble the codon usage of tobacco, which was sufficient to increase the levels of protein expression, from low or undetectable to detectable by Western blot and an average 5-fold increase in C4H activity (Batard et al. 2000).

In 2006, the gene from bone morphogenetic protein (BMP) 2 was optimized, using DNAstar software, by three different methods; the first altered the 6 'rarest' codons with their preferred counterparts in tobacco plants, the second altered the codon usage to make the CAI and A+T content of the gene more similar to the housekeeping genes of tobacco plants and the third altered the rare codons with preferred codons ignoring the A+T content (Suo et al. 2006). BMPs are important in the development of organs and tissues and BMP2 plays a vital role in bone repair. The three codon optimized versions of the human BMP2 gene & the native gene were fused with GUS reporter gene and expressed under the influence of the 35S promoter from Cauliflower Mosaic Virus. All three variants increased GUS activity by over 2-fold compared to the native gene. However, under the more powerful double CaMV35S promoter with AMV enhancer, there was no increase in GUS activity with the codon optimization variants compared to the native construct, indicating that other factors, such as promoters, can also have an effect on protein expression (Suo et al. 2006).

The *cry6A* gene was codon optimized to increase protein expression in tomato roots (Li et al. 2007). Cry6A is a protein produced by *Bacillus thuringiensis* and was used here to increase resistance to plant-parasitic nematodes (Li et al. 2007). The codon optimization strategy used was to decrease or avoid the use of rare codons by altering them to a combination of their synonymous counterparts while maintaining the codon diversity. Additionally, potential polyadenylation sequences, destabilizing sequences or intron sites were removed as well as inserting a plant translation initiation sequence at the 5’ end of the gene and two proline codons at the 3’ end of the gene to protect from proteases. The wild-type gene sequence did not result in detectable protein expression, whereas, the codon optimization strategy resulted in protein levels that were detectable by Western Blot (Li et al. 2007).

E. coli heat labile enterotoxin B subunit (LT-B) was codon optimized for expression in transgenic potato plants, as an edible vaccine against enterotoxigenic *E. coli* (Mason et al. 1998). The codon optimized sequence was designed to contain codons that have a greater than 5% usage rate for potato genes. This optimized sequence retained the amino acid sequence of the native gene except for a substitution of Asn (codon 2) to Val to provide an *Nco*I site by the translation start. Expression of LT-B increased from 4.2-7.1 μg/g for transgenic tubers transformed with bacterial LT-B gene to 7.3-17.2 μg/g for transgenic tubers transformed with plant codon-optimized LT-B gene. Potato tubers expressing codon optimized LT-B generated strong serum and gut mucosal immune responses in mice and provided immunized mice with partial protection from a subsequent challenge(Mason et al. 1998).

More recently, the gene for human epidermal growth factor (hEGF) was codon optimized for expression in *N. benthamiana* (Thomas & Walmsley 2014). hEGF is a 53 amino acid protein with three disulfide bonds that promotes proliferation of epithelial cells. It can enhance the healing of multiple injuries, such as burns or diabetic ulcers. However, despite its potential, hEGF is only used for the treatment of diabetic foot ulcers, due to the high cost of production and the requirement for multiple doses. The sequence was optimized by using the most commonly used codons for *N. benthamiana*. If two or more synonymous codons were used at a similar frequency for the same amino acid, the codon with the highest GC content, especially in the third position, was given preference. This was due to these characteristics being noted in highly expressed native proteins for the host. This codon optimization resulted in 32 nucleotides being changed, spread across 32 of the 53 codons, resulting in a slight increase in overall GC content from 50.4% to 52.7% and an increase in the third position GC from 56.4% to 80%. hEGF expression was targeted to the plant vacuole by addition of a 3’ signal peptide from tobacco chitinase A and to the endoplasmic reticulum (ER) by addition of a KDEL signal peptide. The codon optimized sequence targeted to the vacuole increased the yield of hEGF by 34%, from an average yield of 623.7 ng/ml to 837.1 ng/ml. However, the codon optimized sequence targeted to the ER did not result in an increased yield of hEGF. This was hypothesized to be due to reaching a physiological limit on the amount of hEGF that can be retained in the ER which was not amenable to improvement by codon optimization (Thomas & Walmsley 2014).

4. Conclusions and Future Considerations

Molecular pharming is an emerging technology and there have been many hurdles to address, not least the issue of recombinant protein yields. As the impact of downstream elements of elite plant line selection, horticulture and processing has become better understood, the research emphasis has shifted to other areas including synthetic gene design.

Our understanding of codon bias in most organisms is developing, but still rudimentary. Further mechanistic insight is necessary to drive codon optimization algorithm improvements to the point where researchers can make a reliable and rational assessment on likely benefits.

At present for recombinant protein expression, the impact of codon optimization remains quite unpredictable. Indeed, reports of functionally significant yield increases (10x or greater) are relatively few. This is certainly the case in plant molecular pharming, where benefits of 2-3 fold are commonly reported. Anecdotally, most scientists using plant biotechnology to express heterologous proteins tend to synthesize “codon optimized” genes, using a preferred supplier based on recommendation from past experience. Few are aware of the codon optimization algorithm employed by the preferred supplier, generally accepting a company’s approach on trust. For many academic researchers then, codon optimization is somewhat of a “black box” entrusted to DNA synthesis companies and the effectiveness of codon optimization is more or less accepted on a “trial and error" basis. This is an understandable approach, because few can afford the time and expense of comparing multiple synonymous genes as part of a product development path. A comprehensive and systematic study is warranted however to assess the value of codon optimization for molecular pharming. However, algorithms for codon optimization are varied, and becoming more complex and sophisticated, making these type of investigations increasingly difficult.

Acknowledgements

We would like to thank the generous financial support of the European Research Council, through its Future-Pharma Advanced Grant Award, and the Hotung Foundation. This work was also partly supported by the Wellcome Trust.

5. References

Akashi H. 1994. Synonymous Codon Usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 136: 927–935.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiological reviews 54: 198–210.

Angov E. 2011. Codon usage: nature’s roadmap to expression and folding of proteins. Biotechnology journal 6: 650–9.

Angov E, Hillier CJ, Kincaid RL, Lyon JA. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PloS one 3: e2189.

Angov E, Legler PM, Mease RM. 2011. Adjustment of codon Usage Frequencies by Codon Harmonization Improves Protein Expression and Folding. 705: 1-13.

Arevalo-Herrera M, Vallejo AF, Rubiano K, Solarte Y, Marin C, Castellanos A, Cespedes N, Herrera S. 2015. Recombinant Pvs48/45 antigen expressed in E. Coli generates antibodies that block malaria transmission in anopheles albimanus mosquitoes. PLoS ONE 10: 1–16.

Ask M, Mapelli V, Höck H, Olsson L, Bettiga M. 2013. Engineering glutathione biosynthesis of Saccharomyces cerevisiae increases robustness to inhibitors in pretreated lignocellulosic materials. Microbial cell factories 12: 87–97.

Askelson TE, Campasino A, Lee JT, Duong T. 2014. Evaluation of phytate-degrading Lactobacillus culture administration to broiler chickens. Applied and Environmental Microbiology 80: 943–950.

Bamforth C. 1982. Barley Beta-glucans. Their role in malting and brewing. Brewers Digest 57: 22–27.

Ban N, Nissen P, Hansen J, Moore P, Steitz T. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289: 905–20.

Batard Y, Hehn A, Nedelkina S, Schalk M, Pallett K, Schaller H, Werck-Reichhart D. 2000. Increasing expression of P450 and P450-reductase proteins from monocots in heterologous systems. Arch Biochem Biophys 379: 161–9.

Binder A, Lambert J, Morbitzer R, Popp C, Ott T, Lahaye T, Parniske M. 2014. A modular plasmid assembly kit for multigene expression, gene silencing and silencing rescue in plants. PLoS ONE 9: e88218.

Bonomo J, Gill RT. 2005. Amino acid content of recombinant proteins influences the metabolic burden response. Biotechnology and bioengineering 90: 116–26.

Borriss R, Zemek J. 1981. beta-1,3-1,4-glucanase in spore-forming microorganisms. IV. Properties of some Bacillus-beta-glucan-hydrolases. Zentralbl Bakteriol Naturwiss 136: 63–9.

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. Cell 141: 355–367.

Carlini D, Chen Y, Stephan W. 2001. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics 159: 623–633.

Carton JM, Sauerwald T, Hawley-Nelson P, Morse B, Peffer N, Beck H, Lu J, Cotty A, Amegadzie B, Sweet R. 2007. Codon engineering for improved antibody expression in mammalian cells. Protein expression and purification 55: 279–86.

Chamary J V, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature reviews. Genetics 7: 98–108.

Chen Y, Carlini D, Baines J, Parsch J, Braverman J, Tanda S, Stephan W. 1999. RNA secondary structure and compensatory evolution. Genes Genet Syst 74: 271–86.

Chevance FF V, Le Guyon S, Hughes KT. 2014. The effects of codon context on in vivo translation speed. PLoS genetics 10: e1004392.

Chin JX, Chung BK-S, Lee D-Y. 2014. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics 30: 2210–2.

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320: 1784–7.

Crick F, Barnett L, Brenner S, Watts-Tobin R. 1961. General Nature of the genetic code for proteins. Nature 192: 1227–32.

Dammeyer T, Steinwand M, Krüger S-C, Dübel S, Hust M, Timmis KN. 2011. Efficient production of soluble recombinant single chain Fv fragments by a Pseudomonas putida strain KT2440 cell factory. Microbial cell factories 10: 11–19.

Dammeyer T, Timmis KN, Tinnefeld P. 2013. Broad host range vectors for expression of proteins with (Twin-) Strep-tag, His-tag and engineered, export optimized yellow fluorescent protein. Microb Cell Fact 12: 49–60.

Daniel E, Onwukwe GU, Wierenga RK, Quaggin SE, Vainio SJ, Krause M. 2015. ATGme: Open-source web application for rare codon identification and custom DNA sequence optimization. BMC bioinformatics 16: 303-309.

Daniell H, Streatfield SJ, Wycoff K. 2001. Medical molecular farming: production of antibodies, biopharmaceuticals and edible vaccines in plants. Trends in plant science 6: 219–26.

Desai PN, Shrivastava N, Padh H. 2010. Production of heterologous proteins in plants: strategies for optimal expression. Biotechnology advances 28: 427–35.

Deutscher M. 1984. The eucaryotic aminoacyl-tRNA synthetase complex: suggestions for its structure and function. J Cell Biol 99: 373–7.

Dittmar K, Sørensen M, Elf J, Ehrenberg M, Pan T. 2005. Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO reports 6: 151–7.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. Journal of molecular biology 260: 649–63.

Dong H, Nilsson L, Kurland CG. 1995. Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. Journal of bacteriology 177: 1497–504.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Dubeau MP, Ghinet MG, Jacques P-E, Clermont N, Beaulieu C, Brzezinski R. 2009. Cytosine deaminase as a negative selection marker for gene disruption and replacement in the genus Streptomyces and other actinobacteria. Applied and Environmental Microbiology 75: 1211–1214.

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in genetics 16: 287–9.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 96: 4482–4487.

Ehrenberg M, Kurland C. 1984. Costs of accuracy determined by a maximal growth rate constraint.

Q Rev Biophys 17: 45–82.

Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 300: 1718–22.

Erpel F, Restovic F, Arce-Johnson P. 2016. Development of phytase-expressing chlamydomonas reinhardtii for monogastric animal nutrition. BMC Biotechnology 16: 29–36.

Fahimi H, Sadeghizadeh M, Mohammadipour M. 2016. In silico analysis of an envelope domain III-based multivalent fusion protein as a potential dengue vaccine candidate. Clinical and Experimental Vaccine Research 5: 41–49.

Fath S, Bauer AP, Liss M, Spriestersbach A, Maertens B, Hahn P, Ludwig C, Schäfer F, Graf M, Wagner R. 2011. Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. PloS one 6: e17596.

Fincher G. 1975. Morphology and chemical composition of barley endosperm cell walls. J Inst Brew 81: 116–122.

Fredrick K, Ibba M. 2010. How the sequence of a gene can tune its translation. Cell 141: 227–9.

Fu W, Lin J, Cen P. 2007. 5-Aminolevulinate production with recombinant Escherichia coli using a rare codon optimizer host strain. Appl Microbiol Biotechnol 75: 777–82.

Fuglsang A. 2003. Codon optimizer: a freeware tool for codon optimization. Protein Expression and Purification 31: 247–249.

Gaspar P, Oliveira JL, Frommlet J, Santos M a S, Moura G. 2012. EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics (Oxford, England) 28: 2683–4.

Geisler C, Mabashi-Asazuma H, Kuo C-W, Khoo K-H, Jarvis DL. 2015. Engineering beta 1,4-galactosyltransferase I to reduce secretion and enhance N-glycan elongation in insect cells. J Biotechnol 193: 52–65.

Goodenbour JM, Pan T. 2006. Diversity of tRNA genes in eukaryotes. Nucleic acids research 34: 6137–46.

Gould N, Hendy O, Papamichail D. 2014. Computational tools and algorithms for designing customized synthetic genes. Frontiers in bioengineering and biotechnology 2: 1–14.

Gouy M, Gautier C. 1982. Codon usage in bacteria : correlation with gene expressivity. Nucleic acids research 10: 7055–7074.

Graham H, Fadel J, Newman C, Newman R. 1989. Effect of pelleting and beta-glucanase supplementation on the ileal and fecal digestibility of a barley-based diet in the pig. J Anim Sci 67: 1293–8.

Green RT, Todd JD, Johnston AWB. 2013. Manganese uptake in marine bacteria; the novel MntX transporter is widespread in Roseobacters, Vibrios, Alteromonadales and the SAR11 and SAR116 clades. Isme J 7: 581–91.

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 research 33: W526–31.

Guadalupe-Medina V, Wisselink HW, Luttik MA, de Hulster E, Daran J-M, Pronk JT, van Maris AJ. 2013. Carbon dioxide fixation by Calvin-Cycle enzymes improves ethanol yield in yeast. Biotechnology for biofuels 6: 125–137.

Guimaraes JC, Rocha M, Arkin AP, Cambray G. 2014. D-Tailor: automated analysis and design of DNA sequences. Bioinformatics 30: 1087–1094.

Guo L, Chen X, Li L-N, Tang W, Pan Y-T, Kong J-Q. 2016. Transcriptome-enabled discovery and functional characterization of enzymes related to (2S)-pinocembrin biosynthesis from

Ornithogalum caudatum and their application for metabolic engineering. Microbial Cell Factories 15: 27–45.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends in biotechnology 22: 346–53.

Gustafsson C, Minshull J, Govindarajan S, Ness J, Villalobos A, Welch M. 2012. Engineering genes for predictable protein expression. Protein expression and purification 83: 37–46.

Gutman G, Hatfield GW. 1989. Nonrandom utilization of codon pairs in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 86: 3699–703.

Harcum SW. 2002. Structured model to predict intracellular amino acid shortages during recombinant protein overexpression in E. coli. Journal of biotechnology 93: 189–202.

Hershberg R, Petrov D. 2008. Selection on codon bias. Annual review of genetics 42: 287–99.

Hesselman K, Aman P. 1986. The effect of β-glucanase on the utilization of starch and nitrogen by broiler chickens fed on barley of low- or high-viscosity. Anim. Feed Sci. Technol 15: 83–93.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution 2: 13–34.

Ingolia N, Ghaemmaghami S, Newman J, Weissman J. 2009. Genome-wide analysis in vivo Translation with Nucleotide Resolution Using Ribosome Profiling. Science 324: 218–223.

Irwin B, Heck JD, Hatfield GW. 1995. Codon Pair Utilization Biases Influence Translational Elongation Step Times. J. Biol. Chem. 270: 22801–22806.

Jensen L, Olsen O, Kops O, Wolf N, Thomsen K, von Wettstein D. 1996. Transgenic barley expressing a protein-engineered, thermostable (1,3-1,4)-beta-glucanase during germination. PNAS USA 96: 3487–3491.

Kaminska M, Havrylenko S, Decottignies P, Le Marechal P, Negrutskii B, Mirande M. 2009. Dynamic Organization of Aminoacyl-tRNA Synthetase Complexes in the Cytoplasm of Human Cells. J Biol Chem 284: 13746–54.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. Journal of molecular evolution 53: 290–298.

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: 143–55.

Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coil. Current opinion in biotechnology 6: 494–500.

Karkhah A, Amani J. 2016. A potent multivalent vaccine for modulation of immune system in atherosclerosis : an in silico approach. Clin Exp Vaccine Res 5: 50–59.

Kirchner S, Ignatova Z. 2014. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature Reviews Genetics 16: 98–112.

Ko JH, Llopis PM, Heinritz J, Jacobs-Wagner C, Öll DS. 2013. Suppression of amber codons in Caulobacter crescentus by the orthogonal Escherichia coli histidyl-tRNA synthetase/tRNAHis pair. PLoS ONE 8: 2–8.

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

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

Kyriacou S, Deutscher M. 2008. An important role for the multienzyme aminoacyl-tRNA synthetase complex in mammalian translation and cell growth. Mol Cell 29: 419–27.

Lanza AM, Curran K a, Rey LG, Alper HS. 2014. A condition-specific codon optimization approach for improved heterologous gene expression in Saccharomyces cerevisiae. BMC systems biology 8: 33-43.

Li H, Ban Z, Qin H, Ma L, King AJ, Wang G. 2015. A Heteromeric Membrane-Bound Prenyltransferase Complex from Hop Catalyzes Three Sequential Aromatic Prenylations in the Bitter Acid Pathway. Plant Physiology 167: 650–659.

Li X-Q, Wei J-Z, Tan A, Aroian R V. 2007. Resistance to root-knot nematode in tomato roots expressing a nematicidal Bacillus thuringiensis crystal protein. Plant biotechnology journal 5: 455–64.

Liu X, Deng R, Wang J, Wang X. 2014. COStar: a D-star Lite-based dynamic search algorithm for codon optimization. Journal of theoretical biology 344: 19–30.

Ma JK. 2000. Genes, greens, and vaccines. Nature biotechnology 18: 1141–2.

Maclean J, Koekemoer M, Olivier a J, Stewart D, Hitzeroth II, Rademacher T, Fischer R, Williamson A-L, Rybicki EP. 2007. Optimization of human papillomavirus type 16 (HPV-16) L1 expression in plants: comparison of the suitability of different HPV-16 L1 gene variants and different cell-compartment localization. The Journal of general virology 88: 1460–1469.

Maghodia AB, Geisler C, Jarvis DL. 2016. Characterization of an Sf-rhabdovirus-negative Spodoptera frugiperda cell line as an alternative host for recombinant protein production in the baculovirus-insect cell system. Protein Expression and Purification 122: 45–55.

Manuel ER, Blache CA, Paquette R, Kaltcheva TI, Ishizaki H, Ellenhorn JDI, Hensel M, Metelitsa L, Diamong D. 2011. Enhancement of Cancer Vaccine Therapy by Systemic Delivery of a Tumor Targeting Salmonella-based STAT3 shRNA Suppresses the Growth of Established Melanoma Tumors. Cancer research 71: 4183–4191.

Mason HS, Haq T a, Clements JD, Arntzen CJ. 1998. Edible vaccine protects mice against Escherichia coli heat-labile enterotoxin (LT): potatoes expressing a synthetic LT-B gene. Vaccine 16: 1336–43.

Matthaus F, Ketelhot M, Gatter M, Barth G. 2014. Production of lycopene in the non-carotenoid-producing yeast Yarrowia lipolytica. Applied and Environmental Microbiology 80: 1660–1669.

Meng J, Lee S, Hotard A, Moore M. 2014. Refining the Balance of Attenuation and Immunogenicity of Respiratory Syncytial Virus by Targeted Codon Deoptimization of Virulence Genes. mBio 5: e01704–14.

Mihálik D, Klčová L, Ondreičková K, Hudcovicová M, Gubišová M, Klempová T, Čertík M, Pauk J, Kraic J. 2015. Biosynthesis of Essential Polyunsaturated Fatty Acids in Wheat Triggered by Expression of Artificial Gene. International Journal of Molecular Sciences 16: 30046–30060.

Milne N, van Maris AJA, Pronk JT, Daran JM. 2015. Comparative assessment of native and heterologous 2-oxo acid decarboxylases for application in isobutanol production by Saccharomyces cerevisiae. Biotechnology for Biofuels 8: 204–219.

Mirande M, Le Corre D, Waller J. 1985. A complex from cultered Chinese hamster ovary cells containing nine aminoacyl-tRNA synthetases. Thermolabile leucyl-tRNA synthetase from the tsH1 mutant cell line is an integral component of this complex. Eur J Biochem 147: 281–9.

Moura GR, Pinheiro M, Freitas A, Oliveira JL, Frommlet JC, Carreto L, Soares AR, Bezerra AR, Santos MAS. 2011. Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. PloS one 6: e26817.

Mueller S, Coleman JR, Papamichail D, Ward CB, Nimnual A, Futcher B, Skiena S, Wimmer E. 2010. Live Attenuated Influenza Vaccines by Computer-Aided Rational Design. Nature Biotechnology 28: 723–726.

Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E. 2006. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. Journal of virology 80: 9687–96.

Occhialini A, Lin MT, Andralojc PJ, Hanson MR, Parry MA. 2015. Transgenic tobacco plants with improved cyanobacterial Rubisco expression but no extra assembly factors grow at near wild-type rates if provided with elevated CO. Plant J 85: 148–160.

Overkamp W, Beilharz K, Weme RDO, Solopova A, Karsens H, Kovács ÁT, Kok J, Kuipers OP, Veening JW. 2013. Benchmarking various green fluorescent protein variants in Bacillus subtilis, Streptococcus pneumoniae, and Lactococcus lactis for live cell imaging. Applied and Environmental Microbiology 79: 6481–6490.

Paul M, Ma JK. 2011. Plant-made pharmaceuticals : Leading products and production platforms. Biotechnology and Applied Biochemistry 58: 58–67.

Paul M, Teh A, Twyman R, Ma J. 2013. Target product selection - where can Molecular Pharming make the difference? Curr Pharm Des 19: 5478–85.

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268: 322–30.

Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA. 1991. Modification of the coding sequence enhances plant expression of insect control protein genes. Proceedings of the National Academy of Sciences of the United States of America 88: 3324–8.

Pierlé SA, Hammac GK, Palmer GH, Brayton KA. 2013. Transcriptional pathways associated with the slow growth phenotype of transformed Anaplasma marginale. BMC genomics 14: 272–280.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature reviews. Genetics 12: 32–42.

Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S. 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic acids research 35: W126–31.

Reilman E, Mars RAT, Van Dijl JM, Denham EL. 2014. The multidrug ABC transporter BmrC/BmrD of Bacillus subtilis is regulated via a ribosome-mediated transcriptional attenuation mechanism. Nucleic Acids Research 42: 11393–11407.

Rybicki EP, Williamson A-L, Meyers A, Hitzeroth II. 2014. Vaccine farming in Cape Town. Human Vaccines 7: 339–348.

Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research 41: 2073–94.

Shah K, Cheng Y, Hahn B, Bridges R, Bradbury N, Mueller D. 2015. Synonymous Codon Usage Affects the Expression of Wild Type and F508del CFTR. J Mol Biol 427: 1464–1479.

Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365: 1203–12.

Singer M, Marshall J, Heiss K, Mair G, Grimm D, Mueller A-K, Frischknecht F. 2015. Zinc finger nuclease-based double-strand breaks attenuate malaria parasites and reveal rare microhomology-mediated end joining. Genome Biol 16: 249–267.

Solis-Escalante D, Kuijpers NGA, Bongaerts N, Bolat I, Bosman L, Pronk JT, Daran JM, Daran-

Lapujade P. 2013. amdSYM, A new dominant recyclable marker cassette for Saccharomyces cerevisiae. FEMS Yeast Research 13: 126–139.

Sorensen M, 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: 265–80.

Stoger E, Fischer R, Moloney M, Ma JK-C. 2014. Plant molecular pharming for the treatment of chronic and infectious diseases. Annual review of plant biology 65: 743–68.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Molecular biology and evolution 24: 374–81.

Suo G, Chen B, Zhang J, Duan Z, He Z, Yao W, Yue C, Dai J. 2006. Effects of codon modification on human BMP2 gene expression in tobacco plants. Plant Cell Reports 25: 689–697.

Tegel H, Tourle S, Ottosson J, Persson A. 2010. Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). Protein expression and purification 69: 159–67.

Thanaraj TA, Argos P. 1996. Protein secondary structural types are differentially coded on messenger RNA. Protein science 5: 1973–83.

Thomas DR, Walmsley AM. 2014. Improved expression of recombinant plant-made hEGF. Plant cell reports 33: 1801–14.

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: 344–54.

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009. Design parameters to control synthetic gene expression in Escherichia coli. PloS one 4: e7002.

Welch M, Villalobos A, Gustafsson C, Minshull J. 2009. You’re one in a googol: optimizing genes for protein expression. Journal of the Royal Society, Interface 6: S467–76.

van der Woude AD, Perez Gallego R, Vreugdenhil A, Puthan Veetil V, Chroumpi T, Hellingwerf KJ. 2016. Genetic engineering of Synechocystis PCC6803 for the photoautotrophic production of the sweetener erythritol. Microbial Cell Factories 15: 60–72.

Wu G, Bashir-Bello N, Freeland SJ. 2006. The Synthetic Gene Designer: a flexible web platform to explore sequence manipulation for heterologous expression. Protein expression and purification 47: 441–5.

Zhao H, Zhou J, Zhang K, Chu H, Liu D, Poon VK-M, Chan CC-S, Leung H-C, Fai N, Lin Y-P, Zhang AJ-X, Jin D-Y, Yuen K-Y, Zheng B-J. 2016. A novel peptide with potent and broad-spectrum antiviral activities against multiple respiratory viruses. Scientific Reports 6: 22008.

Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. Journal of virology 73: 4972–82.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Molecular biology and evolution 26: 1571–80.

Zull J, Smith S. 1990. Is genetic code redundancy related to retention of structural information in both DNA strands? Trends Biochem Sci 15: 257–61.

Figure 1: *Codon choice is not random, but it is highly selected across organisms to optimize protein expression. Codons recognized by low-abundance tRNAs are highly represented at the beginning of the gene. This suggests that the ribosome translates more slowly over the first 30-50 codons, named the 'ramp'.*[Modified from (Fredrick & Ibba 2010)]

Figure 2: *The arrangement of synonymous codons along the gene controls translation speed of the ribosome. The two shades of blue represent two different synonymous codons (encoding the same amino acid). When the same codons are arranged along the mRNA, auto-correlated, the translation is faster than when synonymous codons are interspersed, anti-correlated.* [Modified from (Fredrick & Ibba 2010)]

Appendix

1. Pareto-optimal sequences: the algorithm makes optimal decisions which are a compromise between two or more conflicting objectives, for example, minimizing the cost of buying a car while maximizing comfort and fuel consumption. In most cases, there does not exist a single solution, instead there exists a number of Pareto optimal solutions. All Pareto-optimal solutions are equally as good; however, a human decision maker may have subjective preferences and be able to choose a single solution.

2. Euclidean distance: straight-line distance between two points, which can be calculated using Pythagorean Theorem; distance ((x1,y1),(x2,y2)) = $\sqrt{((x1-x2)^2 + (y1-y2)^2)}$.

3. Sliding window method: Forms a sub-list of a collection. If the computer array is [1 2 3 4 5 6 7 8] then a sliding window of size 3 would be [1 2 3] [2 3 4] [3 4 5] [5 6 7] [6 7 8].

4. Weighted directed acyclic graph: A directed graph is a graph where each edge can be followed from one vertex to another. A directed acyclic graph is a directed graph that cannot start and end at the same vertex. A weighted graph is a graph that has a numeric label associated with each edge. Weighted directed acyclic graphs are used to solve the shortest path problem, which is the problem of finding a path between two vertices in a graph such that the sum of the weights of its constituent edges is minimized. Some algorithms are simplified when using weighted directed acyclic graphs instead of general graphs.

Table 1: Overview of gene design features for multi-objective optimization programmes

[Modified from Gould *et al.* 2014]

	Codon Autocorrelation Adjustment	Codon Context	mRNA secondary structure	GC/AT content	Hidden stop codons	Motif Avoidance	Repetitious base removal	Restriction site manipulation
EuGene http://bioinformatics.ua.pt/eugene	✓	✓	✓	✓	✓		✓	✓
COOL http://bioinfo.bti.a-star.edu.sg/COOL/	✓	✓	✓	✓	✓	✓	✓	✓
D-Tailor https://sourceforge.net/projects/dtailor/			✓	✓				✓
COStar http://life.sysu.edu.cn/COStar/COStar.html			✓	✓		✓	✓	✓

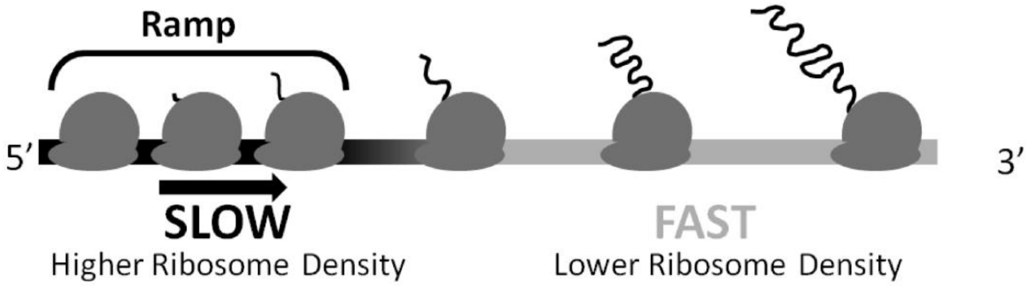
Ramp
5′
3′
SLOW
Higher Ribosome Density
FAST
Lower Ribosome Density

Figure 1

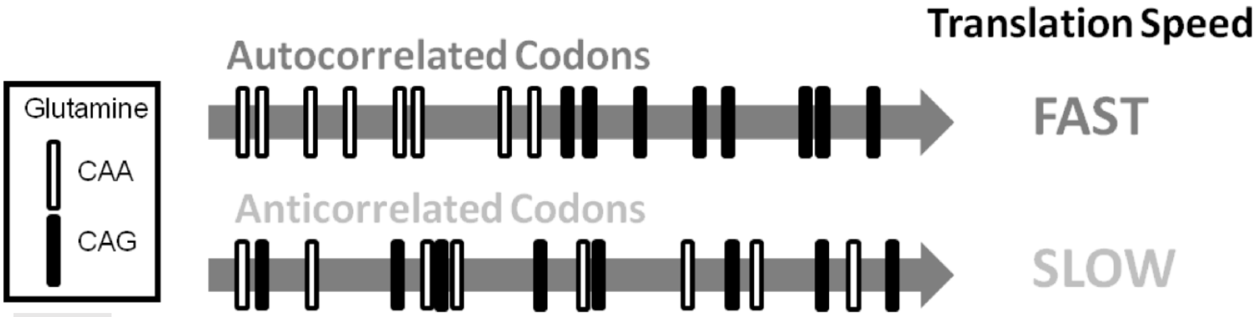
Translation Speed
Autocorrelated Codons
FAST
Anticorrelated Codons
SLOW
Glutamine
CAA
CAG

Figure 2