Codon Bias and Heterologous Protein Expression

Claes Gustafsson*, Sridhar Govindarajan & Jeremy Minshull
DNA 2.0, Inc. 1455 Adams Drive, Menlo Park, CA 94025
* Corresponding author: cgustafsson@dnatwopointo.com

The expression of functional proteins in heterologous hosts is a cornerstone of modern biotechnology. Unfortunately proteins are often difficult to express outside their original context. They may contain codons that are rarely used in the desired host, come from organisms that use non-canonical code, or contain expression-limiting regulatory elements within the coding sequence. Improvements in the speed and cost of gene synthesis facilitate the complete redesign of entire gene sequences to maximize the likelihood of high protein expression. Redesign strategies including modification of translation initiation regions, alteration of mRNA structural elements and use of different codon biases are discussed.

In 1977 when Genentech scientists and their academic collaborators produced the first human protein (somatostatin) in a bacterium [1], expression of proteins in heterologous hosts played a critical role in the launch of the entire biotechnology industry. At the time, only the amino acid sequence of somatostatin was known, so the Genentech group synthesized the 14 codon long somatostatin gene using oligonucleotides instead of cloning it from the human genome. Itakura and co-workers designed these oligonucleotides based on three criteria. First, codons favored by the phage MS2 were used preferentially. Not much of the *Escherichia coli* (*E. coli*) genome DNA sequence was known at the time, but the MS2 phage had just been sequenced and was assumed to provide a good guide to the codons used in highly expressed *E. coli* genes. Second, care was taken to eliminate undesirable inter- and intra-molecular pairing of the overlapping oligonucleotides as this would compromise the gene synthesis process. Third, sequences rich in GC followed by AT rich sequence was avoided, as it was believed it could terminate transcription. The result was the first production of a functional polypeptide from a synthetic gene.

Now a quarter of a century later, most genes are cloned from cDNA libraries or directly by polymerase chain reaction (PCR) from the organism of origin. *De novo* gene synthesis is largely avoided because of perceived high costs in time and effort [2]. Despite its prevalence, PCR-based cloning often requires templates that may not be trivial to access (cDNA templates must generally be used for organisms with introns), gene-specific PCR conditions, re-sequencing of PCR product and site-directed mutagenesis to repair PCR errors. The real fun, though, begins after the amplified gene is cloned into an expression vector: often the protein is not expressed or expressed only at very low levels. Much work has been done to improve the expression of cloned genes, including optimization of host growth conditions and the development of new host strains, organisms and cell free systems [3]. Despite the advances that these approaches have made, they have skirted a significant underlying problem: the DNA sequence used to encode a protein in one organism is often quite different from the sequence that would be used to encode the same protein in another organism.

Why do different organisms prefer different codons?

The genetic code uses 61 nucleotide triplets (codons) to encode 20 amino acids and three to terminate translation. Each amino acid is therefore encoded by between one (Met and Trp) and six (Arg, Leu and Ser) synonymous codons. These codons are "read" in the ribosome by complementary tRNAs which have been charged with the appropriate amino acid. The degeneracy of the genetic code allows many alternative nucleic acid sequences to encode the same protein. The frequencies with which different codons are used vary significantly between different organisms, between proteins expressed at high or low levels within the same organism, and sometimes even within the same operon [4].

There is continuing speculation regarding the evolutionary forces that have produced these differences in codon preferences [5]. Codon distribution respond to genome GC content and the changes in codon usage are at least partly explained by a mutation/selection equilibrium between the different synonymous codons in each organism [6]. Some researchers have hypothesized that codon biases that tend to reduce the diversity of isoacceptor tRNAs reduce the metabolic load and are therefore beneficial to organisms that spend part of their lives under rapid growth conditions [7].

Whatever the reasons for codon bias, it has become increasingly clear that codon biases can have profound impacts on the expression of heterologous proteins [8].

Visualizing codon biases

The correlation that has been observed between the codon bias of a gene and its expression levels has been used to define a codon adaptation index [9]. This measure of codon usage is derived from a reference set of highly expressed genes to score the extent to which an organism prefers specific codons. The index can be used to predict the expression levels of endogenous genes from genome sequence data [10]. However because the index measures only the degree of preference but not the nature of that preference, it cannot be used to assess the likely compatibility between a gene and a candidate host. The gene may have a strong bias resulting in high codon adaptation indices, but these preferences may be for quite different codons.

Principal component analysis can be used to compress the high dimensional information into a two-dimensional map. This provides a more convenient way to visualize differences in codon preferences between different organisms. Figure 1 shows the average codon preferences of the genomes from eight commonly studied organisms represented on such a map. As can be seen in the figure, *Streptomyces coelicolor* (*S. coelicolor*) has the most extreme codon usage profile. In

this organism almost every "wobble" position (the third base in each codon, where much of the degeneracy of the genetic code resides) is a G or C, resulting in *S. coelicolor's* high GC content (71%). The figure also shows that *Saccharomyces cerevisiae* (*S. cerevisiae*), *Caenorhabditis elegans* (*C. elegans*) and *Arabidopsis thaliana* (*A. thaliana*) cluster in this map, indicating that they share similar codon preferences and suggesting that *S. cerevisiae* would be a good candidate for expressing native *A. thaliana* or *C. elegans* genes.

Figure 1 also makes immediately obvious the considerable divergence between *E. coli* and human codon preferences. This confirms what many researchers have learned through extensive experimentation: *E. coli* is not the optimal host for expressing proteins encoded with human codon usage profile.

The codon distribution in the map helps to visualize the codons that are used differentially by each of the 8 organisms. For example mammalian genes commonly use AGG and AGA codons for Arg (each are used for 11.2% of Arg codons in human genes) whereas these are very rarely used in *E. coli* (2.1% and 2.4% respectively). Thus in Figure 1 AGG and AGA both contribute to positive deviations in principal component 2 (PC2) as is seen for the overall human codon preference. In contrast *E. coli* prefers the CGT Arg codon (used 16.4% of the time, compared with 4.5% usage in human genes), so CGT contributes to negative deviations in PC2, as is seen for the overall *E. coli* codon bias. A map of 'codon usage space' is therefore useful as it quickly identifies infrequently used codons in genes derived from each organism that will be potentially problematic when attempting heterologous expression.

How does codon bias affect protein expression?

Codon usage has been identified as the single most important factor in prokaryotic gene expression [11]. The reason for this is almost certainly because preferred codons correlate with the abundance of cognate tRNAs available within the cell. This relationship serves to optimize the translational system and to balance codon concentration with isoacceptor tRNA concentration [12]. In *E. coli*, for example, the $tRNA^{Arg}_4$ that reads the infrequently used AGG and AGA codons for Arg is present only at very low levels. It is likely that codon usage and tRNA isoacceptor concentrations have coevolved, and that the selection pressure for this coevolution is more pronounced for highly expressed genes than genes expressed at low levels [13].

The coevolution of isoacceptor tRNAs with codon frequencies has even led in some cases to departures from the canonical genetic code [14]. While comparative genomics studies are shedding new light on the ongoing evolution of the genetic code [15,16], the existence of slightly different codes in different organisms is a very significant barrier to heterologous expression. Indeed some organisms, notably the ciliates that have played an important role in the elucidation of telomere biology, possess tRNAs that read the canonical stop codons TAA and TAG as Glu, making these genes impossible to express heterologously.

Improving expression by modifying the host

If the negative effect of different codon biases on heterologous gene expression results from different tRNA levels, one solution appears to be to expand the host's intracellular tRNA pool. This can be done by over-expressing genes encoding the rare tRNAs. For *E. coli*, the primary targets to facilitate expression of human genes are the argU gene encoding the minor $tRNA^{Arg}_4$ that reads AGG and AGA codons, $tRNA^{Ile}_2$ that reads AUA, $tRNA^{Leu}_3$ that reads CUA and CUG, and $tRNA^{Pro}_2$ that reads CCC and CCU [8]. *E. coli* strains over-expressing these tRNA genes are commercially available from companies such as Stratagene (www.stratagene.com) and Novagen (www.emdbiosciences.com/html/NVG/home.html). Several laboratories have shown that expression yields of proteins whose genes contain rare codons can be dramatically improved when the cognate tRNA is increased within the host [8].

Even though tRNA over-expression initially appears as an attractive solution, there are caveats. Different tRNAs may need to be over-expressed for genes from different organisms and the strategy is less appealing for hosts more difficult to manipulate than *E coli*. There may also be metabolic effects of changing a cell's tRNA concentrations. Perhaps most important, though, is the question of how increasing the tRNA concentration will affect amino acylation and tRNA modifications and thus whether the composition of the over-expressed protein will be consistent.

Transfer RNA molecules are extensively processed prior to amino-acylation and participation in the translational process. More than 30 modified nucleotides have been found in *E. coli* tRNAs; some are present at the same position for all tRNAs, others are found in one or a few different tRNAs [17]. Many of the tRNA modifications scattered throughout the tRNA molecule and especially those located in the anticodon loop, have been shown to improve reading frame maintenance [18]. One purpose of these modifications is thought to be to reduce translational frameshifts: the lack of some tRNA modifications has been experimentally linked to missense and nonsense errors during translation [17], for example tRNAs lacking methylation of tRNA at the N-1 position of guanosine (m^1G) at position 37 result in translational frameshifts [19].

A problem with the tRNA over-expression strategy, then, is that producing a fully functional tRNA requires other cellular components that may be in limiting supply when the tRNA alone is over-produced. When $tRNA^{Leu}_1$ is over-expressed in *E. coli* the tRNA is significantly under-modified in at least two ways: m^1G at position 37 and pseudouridine (Ψ) at position 32. Only 40% of the $tRNA^{Leu}_1$ molecules are amino acylated, the strain grows very slowly and the ribosomal step time is reduced two - three fold [20]. Similarly over-expressed $tRNA^{Tyr}$ results in a decrease of the 2-methylthio-N-6-isopentenyl adenosine (ms^2i^6A) modification at position 37 and a tRNA that is less efficient *in vitro* [21]. Loss of the ms^2i^6A modification following $tRNA^{Phe}$ over-expression led to decreased fidelity of translation [22].

Translational missense substitution frequencies can increase with more than an order of magintude as a function of under-acetylated tRNA. One particular concern over such loss of fidelity is the possibility that the resulting heterogeneous mixture of proteins might induce an immune response if introduced into vertebrates [23].

In addition to translational fidelity and host metabolic load issues, the tRNA over-expression strategy is not terribly flexible. It is much more difficult to engineer fungal or mammalian host cells than *E coli*. In eukaryotic

cells the tRNA expression is driven by copy number, not promoter strength, further complicating the issue. For some applications such as the emerging field of DNA vaccines, host engineering is quite out of the question. The alternative approach is to modify the gene to be expressed.

Results from codon optimization

In general, the more codons that a gene contains that are rarely used in the expression host, the less likely it is that the heterologous protein will be expressed at reasonable levels [8]. Low expression levels are exacerbated if the rare codons appear in clusters or in the N-terminal part of the protein. A common strategy to improve expression is therefore to alter the rare codons in the target gene so that they more closely reflect the codon usage of the host, without modifying the amino acid sequence of the encoded protein. Techniques for achieving this range from sequential site-directed mutagenesis steps [24] to resynthesis of the entire gene [25].

In Table 1, we have attempted to identify all publications where protein expression levels from natural gene sequences are compared with their codon-optimized counterparts in identical systems. The methods for codon optimization differ in each case, but all have replaced one or more codon that is rarely used in the host with one that is more frequently used.

Many of the published codon optimization reports involve expressing mammalian proteins in *E. coli*. In several instances, increases in expression levels achieved are dramatic. Two papers describe proteins that were effectively undetectable when expressed from the native genes. After codon-optimization, expression levels of between 10% and 20% of total *E. coli* soluble protein were obtained [26,27]. More typical increases in expression for codon-optimized mammalian proteins in *E. coli* are between five and 15-fold and can frequently yield as much as 5% of the *E coli* soluble protein.

Another very successful application of codon optimization, generally by complete resynthesis of the gene, is in enhancing the expression of viral genes in mammalian cell lines. Viruses are a particularly interesting example because their codons are often constrained by a completely different pressure: their very dense information load is frequently accommodated using overlapping reading frames. Many viral genes also encode *cis*-acting negative regulatory sequences within the coding sequence. When expression of only one protein is required, the gene can be resynthesized with a host codon bias that also disrupts the regulatory elements thereby enhancing protein production [28]. Viral codon optimization is often performed for DNA vaccine research to increase the immunogenicity of the target. In many published studies the immune response to the injected DNA is measured but not the protein concentration. Some of these examples have been omitted from Table 1, which only lists publications where the protein concentration is measured directly.

Gene resynthesis is also essential for heterologous expression of genes from organisms that use non-canonical codes. These include pathogens such as *Candida albicans* [29] and ciliate model organisms such as *Tetrahymena* [30]. Elimination of codons that would be read as termination signals or different amino acids is essential not just to improve expression levels, but to achieve any expression at all of the encoded protein.

Beyond codon bias

Although the codon bias in a gene plays a large role in its expression, it would be misleading to suggest that this is the only factor involved. The choice of expression vectors and transcriptional promoters are also important [3]. The nucleotide sequences surrounding the N-terminal region of the protein appear particularly sensitive, both to the presence of rare codons [31,32] and to the identities of the codons immediately adjacent to the initiation AUG [33,34]. There is also some interplay between translation and mRNA stability which has not been completely deconvoluted [2], although reduced translational efficiency may be accompanied by a lower mRNA level because decreased ribosomal protection of the mRNA will increase its exposure to endo-RNAses. The structure of the 5' end of the mRNA also has a significant effect [35], and strategies using short upstream open reading frames for translational coupling of target genes have proved successful in improving the efficiency of expression of some problem genes [36].

It should also be noted that efficient translation is necessary but not sufficient to produce a functional protein. The polypeptide chain must fold correctly, in some cases form appropriate disulphide bonds and even undergo post-translational modifications such as glycosylation. For these processes the absence of the correct redox environment, chaperonins, normal association partners or modifying enzymes will provide additional challenges. These issues are beyond the scope of this article: we will content ourselves for the time being with efficiently producing the polypeptide.

Gene design considerations

Designing a gene *de novo* can be both liberating and daunting. At the least constrained end of the choice spectrum there are an enormous number of DNA sequences that can encode a single amino acid sequence. Each amino acid can be encoded by an average of 3 different codons, so there are around 3^{100} ($\sim 5 \times 10^{47}$) nucleotide sequences that would all produce the same 100 amino acid protein. How many of these possible sequences will result in high levels of heterologous protein expression? At the other end of this spectrum only a single nucleotide sequence is possible. Here only one codon – the one used most frequently by the host - is used for each amino acid.

The 'One amino acid – one codon' approach has several drawbacks. First, a strongly transcribed mRNA from such a gene will generate high codon concentrations for a subset of the tRNA, resulting in imbalanced tRNA pool, skewed codon usage pattern and the potential for translational error [23]: heterologously expressed proteins may be produced at levels as high as 60% of total cell mass, making the use of a single tRNA pool a significant problem. Introducing silent mutations in a 'One amino acid – one codon' optimized gene can increase protein expression four-fold [37]. Second, with no flexibility in codon selection, it is impossible to avoid repetitive elements and secondary structures in the gene and mRNA which may inhibit ribosome processivity through mRNA stem-loops [35]. Repetitive elements may also affect the ease of gene synthesis, making it more troublesome if performed in-house or more expensive and time-consuming if outsourced. Severe repetitive elements may also affect the stability of a gene in its host. Third, it is often desirable to incorporate or exclude sequence elements such as restriction sites from the

sequence to facilitate subsequent manipulations. These are impossible to accommodate if the codon usage is rigidly fixed.

Conclusions

The genetic information encoded in an open reading frame goes far beyond simply stating the order of the amino acids in the protein. It is now estimated that alternative splicing comprises 40-60% of all human multiexon genes, antisense transcription occurs in 10-20% of all genes, mRNA editing is common (at least in neural cells), regulatory elements abundant and mRNA degradation signals through RNAi and otherwise are identified throughout the human genome. As we start to peel through the different layers of complex and integrated information present in the coding regions of DNA, we can start making more informed decisions on how to design genes and genetic networks.

The design and use of synthetic genes offers a mechanism by which researchers can assume much greater control of heterologous protein expression. As well as manipulating codon biases, peptide tags can be added, splice sites removed and restriction sites placed as desired. The cost and fidelity of gene synthesis appears to be following a trajectory similar to that seen for synthetic oligonucleotides over the past two decades, making their use increasingly cost-effective. This trend will allow scientists to focus more on science rather than on obtaining the tools with which to work. The biotechnology industry is thus en-route to closing the circle to its distant past; the genetic engineering tools pioneered by the Genentech group and their academic collaborators in long-ago 1977 will once again become state-of-the-art.

Acknowledgements

We would like to thank Drs. Jon Ness, Tony Cox, Ramasubbu Venkatesh and Tom Vigdal, all from DNA 2.0 Inc., for discussions on codon optimization and helpful comments on the manuscript.

Gene Origin	Protein Name	Host	Improvement	Ref
H. sapiens	IL2	*E. coli*	16 fold	[38]
C. tetani	Fragment C	*E. coli*	Four fold	[39]
B. thuringiensis	CryIA(b), CryIA(c)	*L. esculentum*	100 fold	[40]
B. thuringiensis	CryIA(b), CryIA(c)	*Nicotiana tabacum*	Below detection vs. >0.1% of tot. protein	[40]
M. musculus	IG kappa chain	*S. cerevisiae*	> 50 fold	[41]
Bacillus hybrid	(1,3-1,4)-β-glucanase	*H. vulgare*	Below detection vs. 40ng per 2×10^5 protoplasts	[42]
H. sapiens	TnT	*E. coli*	10 and 40 fold (two different constructs)	[43]
HIV	Gp120	*H. sapiens*	>40 fold	[44]
A. victoria	GFP	*H. sapiens*	Below detection vs. substantial signal	[44]
A. victoria	GFP	*H. sapiens*	22 fold	[45]
A. victoria	Mutated GFP	*C. albicans*	Below detection vs. strong band in western	[29]
M. musculus	c-Fos	*E. coli*	Below detection vs. 20% of soluble protein	[27]
S. oleracea	plastocyanin	*E. coli*	1.2 fold	[46]
H. sapiens	neurofibromin	*E. coli*	three fold	[47]
L. monocytogenes	LLO	*M. musculus*	100 fold	[48]
H. sapiens	M2-2	*E. coli*	140 fold	[49]
R. prowazekii	Tlc	*E. coli*	No effect	[50]
BPV1	L1 and L2	mammalian	$> 1 \times 10^3$ fold	[51]
H. sapiens	PC-TP	*E. coli*	Trace levels vs. 10% of cytosolic protein	[26]
H. sapiens	hCG-β	*Dictyostelium*	Four-five fold	[52]
T. aestivum	CYP73A17	*S. cerevisiae*	Four, seven and 13 fold (three different constructs)	[53]
T. aestivum	CYP73A17	*N. tabacum*	Five fold	[53]
HIV	gag	*H. sapiens*	> 322 fold	[54]
Dermatophagoides	ProDer p1	*P. troglodytes*	Five-10 fold	[55]
HIV	gag	*H. sapiens*	1.5-two fold	[28]
Plasmodium	EBA-175 region II and MSP-1	*M. musculus*	Four fold	[56]
Tn10/Herpes simplex virus	rtTA	*M. musculus*	> 20 fold	[57]
HPV	L1	*H. sapiens*	$1 \times 10^4 - 1 \times 10^5$ fold	[58]
C. diphtheriae – mammal hybrid	DT	*P. pastoris*	0 vs 10mg L^{-1}	[59]
P1 phage	Cre	Mammalian	1.6 fold	[60]
A. equina	Equistatin	*P. pastoris*	Two fold	[61]
H. sapiens	IL-6	*E. coli*	Three fold	[62]
H. sapiens	Glucocerebrosidase	*Pichia pastoris*	Eight and 10 fold (two different constructs)	[63]
Schistosoma mansoni	SmGPCR	*H. sapiens*	Barely detectable vs. strong band in western	[64]
C. elegans	GluClα1, GluClβ	*R. norvegicus*	Six-nine fold	[65]
Herpesvirus	U51	Mammalian	10-100 fold	[66]
HIV	gag , pol, env, nef	*H. sapiens*	>250x, >250x, >45x, >20x respectively	[67]
H. sapiens	IL-18	*E. coli*	Five fold	[68]
HPV	E5	Mammalian	Six-nine fold	[69]
HPV	E7	Mammalian	20-100 fold	[70]
Plasmodium	F2 domain of EBA175	*E. coli*, *Pichia pastoris*	Four fold and nine fold	[71]

Table 1. Compilation of publications where gene expression of codon optimized and wildtype sequences have been compared head-to-head and the produced protein yield has been measured.

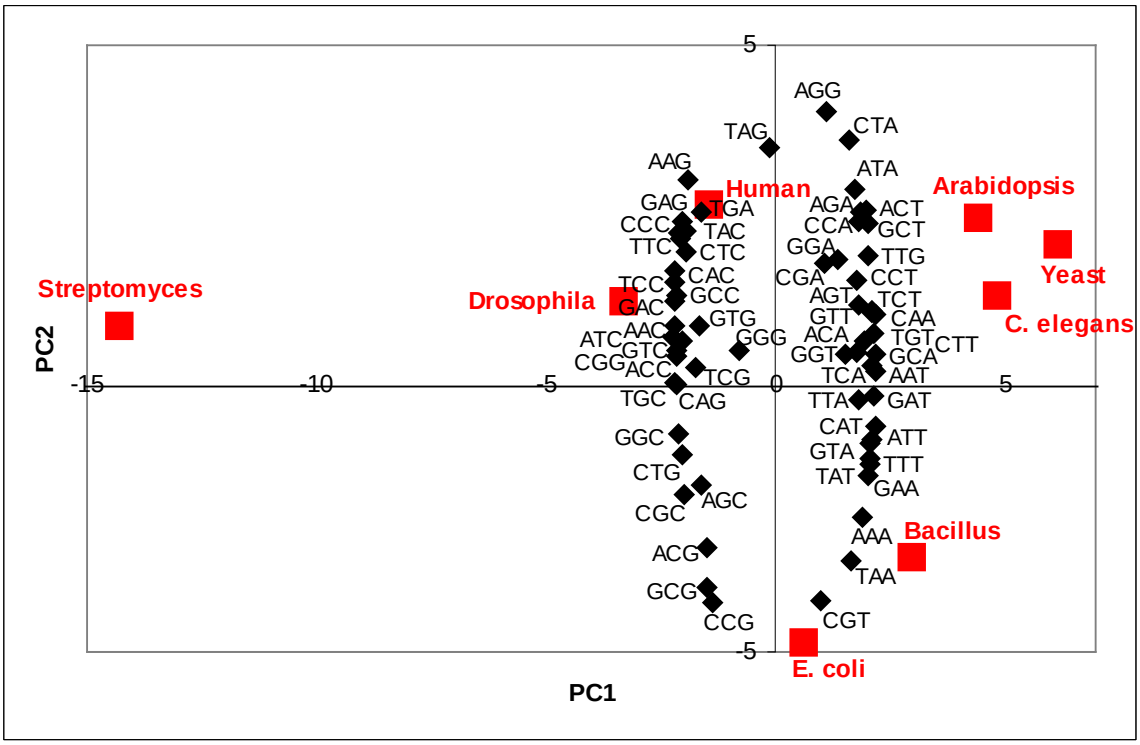
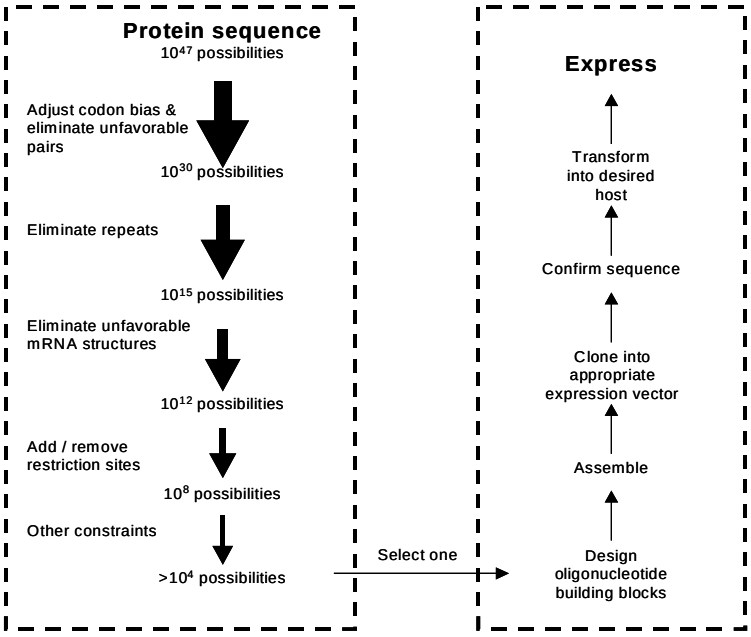


Figure 1. Graphical representation of "codon usage space". Principal component analysis (PCA) involves a mathematical procedure that transforms a number of correlated variables (here codon frequencies) into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The frequencies with which each codon is used in all proteins of eight commonly studied organisms (www.kazusa.or.jp/codon/) were tabulated in a 8 rows/organisms x 62 columns/codons and subjected to principal component analysis to produce a map of "codon usage space". The two codons ATG and TGG that uniquely encode Met and Trp respectively have been omitted. Two dimensions were identified that accounted for 70% (PC1) and 12% (PC2) of the total codon variability information respectively. The black diamonds represent the loads, i.e. the contribution of each codon to the two principal component dimensions (for example codons GAT and CAG contribute nothing to PC2 but have approximately equal negative and positive contributions to PC1). The values of the codon loads have been normalized to that of the organism distribution. The red squares show the preferences of each organism plotted within this space. The plot was made using MatLab from Mathworks (www.mathworks.com)



The gene design process

The procedure developed at DNA 2.0 Inc. (www.dnatwopointo.com) for designing a gene sequence to encode a specific protein is shown in figure. The process involves using an initial codon usage table to propose candidates, then a successive set of filters to eliminate those sequences that do not also comply with additional design constraints.

1. Constructing and using a codon usage table. The large amount of genomic sequence now available has made it possible to derive the codon usage for any organism. An excellent compilation can be found at www.kazusa.or.jp/codon/. For expression in *E. coli*, for example, codon usage from highly expressed (type II) genes are available www.faculty.ucr.edu/~mmaduro/codonusage/codontable.htm [72]. These tables can be adapted for gene design in two steps. First, a threshold level is set. That is, all frequencies below a certain value (typically between 5% and 10%) are set to zero, so that rare codons are completely eliminated. Second, the remaining frequencies are normalized so that the summed frequencies for codons for each amino acid equal 100%.

Hybrid codon usage tables can be constructed for a protein that is to be expressed in more than one host. Codons that are below the threshold in either host are eliminated. The frequencies for the remaining codons can be calculated by simply using the frequencies for the most restrictive organism, or by calculating a mean value for each codon in all of the desired hosts.

Once the codon usage table has been constructed, candidate sequences are enumerated *in silico* by selecting codons at random with probabilities obtained from the codon usage table. Each designed sequence is then passed through subsequent filters to ensure a match with additional design criteria.

2. Eliminating unfavorable codon pairs and extreme GC content. The GC content of genes and the frequency with which adjacent codons occur (codon pair frequency) are both factors that are correlated to codon usage frequency. The codon pair frequency can deviate significantly from what would be expected from just the statistical distribution of each single codon. Codon pairs that are avoided in highly expressed *E. coli* genes can be found on the web (www.bio21.bas.bg/codonpairs) [73] and is used as a criterion to reject candidate designs.

3. Eliminating repetitive sequences. Direct repeats can be detected by standard methods such as a BLAST comparison [74] of the sequence against itself. Candidate designs that contain significant lengths of direct repeats are eliminated.

4. Avoiding unfavorable mRNA secondary structures. Stable mRNA structures, particularly at the 5' end of the transcript, have been implicated in reduced gene expression [2,35]. The potential of a transcribed mRNA to adopt such a structure can be identified using free energies calculations. Software for performing such analyses can be found at www.bioinfo.rpi.edu/applications/mfold [75].

5. Avoiding and including restriction sites. The presence or absence of selected restriction sites is often important to facilitate subsequent gene manipulations such as swapping between vectors, exchanging protein domains and adding or removing peptide tags or fusion partners. Candidate sequences can be tested to ensure the correct placement or elimination of restriction sites.

6. Other constraints. Additional constraints that can be used to sift the gene design solutions through includes avoiding cryptic splice sites and regulatory elements, immuno-stimulatory or immuno-suppressive elements (for DNA vaccines) [76], RNA methylation signals, selenocystein incorporation signals and many more depending on the biological system used and specific concerns. Gene designs can also be used to maximize genetic distances from endogenous gene homologs (to minimize risk of *in vivo* recombination) or patented sequences (to avoid patent infringement).

As shown in Figure, each of these filters reduce the number of possible sequences, but many possible sequences generally remain even with five or six constraints in addition to the codon bias.

Refs

1 Itakura, K. et al. (1977) Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science* 198 (4321), 1056-1063

2 Wu, X. et al. (2004) Codon optimization reveals critical factors for high level expression of two rare codon genes in Escherichia coli: RNA stability and secondary structure but not tRNA abundance. *Biochem Biophys Res Commun* 313 (1), 89-96.

3 Higgins, S.J., Hames, B. D. (1999) *Protein Expression: A Practical Approach*, Oxford University Press

4 Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* 10 (22), 7055-7074.

5 Grosjean, H. and Fiers, W. (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18 (3), 199-209.

6 Knight, R.D. et al. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* 2 (4), RESEARCH0010. Epub 2001 Mar 0022.

7 Andersson, G.E. and Kurland, C.G. (1991) An extreme codon preference strategy: codon reassignment. *Mol Biol Evol* 8 (4), 530-544.

8 Kane, J.F. (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol* 6 (5), 494-500.

9 Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15 (3), 1281-1295.

10 Carbone, A. et al. (2003) Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19 (16), 2005-2015.

11 Lithwick, G. and Margalit, H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res* 13 (12), 2665-2673.

12 Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 151 (3), 389-409.

13 Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* 325 (6106), 728-730.

14 Massey, S.E. et al. (2003) Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. *Genome Res* 13 (4), 544-557.

15 Santos, M.A.S. et al. (2004) Driving change: the evolution of alternative genetic codes. *Trends Genet.* 20, 95-102

16 Knight, R.D. et al. (2001) Rewiring the keyboard: evolvability of the genetic code. *Nat Rev Genet* 2 (1), 49-58.

17 Björk, G.R. (1996) Stable RNA modification. In *Escherichia coli and Salmonella: Cellular and Molecular Biology.* (Neidhardt, F.C. et al., eds.), pp. 861-886, ASM Press,

18 Urbonavicius, J. et al. (2001) Improvement of reading frame maintenance is a common function for several tRNA modifications. *Embo J* 20 (17), 4863-4873.

19 Li, J.N. and Björk, G.R. (1995) 1-methylguanosine deficiency of tRNA influences cognate codon interaction and metabolism in Salmonella typhimurium. *J. Bact* 177, 6593-6600

20 Wahab, S.Z. et al. (1993) Effects of tRNA(1Leu) overproduction in Escherichia coli. *Mol Microbiol* 7 (2), 253-263.

21 Gefter, M.L. and Russell, R.L. (1969) Role modifications in tyrosine transfer RNA: a modified base affecting ribosome binding. *J Mol Biol* 39 (1), 145-157.

22 Wilson, R.K. and Roe, B.A. (1989) Presence of the hypermodified nucleotide N6-(delta 2-isopentenyl)-2-methylthioadenosine prevents codon misreading by Escherichia coli phenylalanyl-transfer RNA. *Proc Natl Acad Sci U S A* 86 (2), 409-413.

23 Kurland, C. and Gallant, J. (1996) Errors of heterologous protein expression. *Curr Opin Biotechnol* 7 (5), 489-493.

24 Kink, J.A. et al. (1991) Efficient expression of the Paramecium calmodulin gene in Escherichia coli after four TAA-to-CAA changes through a series of polymerase chain reactions. *J Protozool* 38 (5), 441-447.

25 Nambiar, K.P. et al. (1984) Total synthesis and cloning of a gene coding for the ribonuclease S protein. *Science* 223 (4642), 1299-1301.

26 Feng, L. et al. (2000) High-level expression and mutagenesis of recombinant human phosphatidylcholine transfer protein using a synthetic gene: evidence for a C-terminal membrane binding domain. *Biochemistry* 39 (50), 15399-15409.

27 Deng, T. (1997) Bacterial expression and purification of biologically active mouse c-Fos proteins by selective codon optimization. *FEBS Lett* 409 (2), 269-272.

28 Graf, M. et al. (2000) Concerted action of multiple cis-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. *J Virol* 74 (22), 10822-10826.

29 Cormack, B.P. et al. (1997) Yeast-enhanced green fluorescent protein (yEGFP)a reporter of gene expression in Candida albicans. *Microbiology* 143 (Pt 2), 303-311.

30 Collins, K. and Gandhi, L. (1998) The reverse transcriptase component of the Tetrahymena telomerase ribonucleoprotein complex. *Proc Natl Acad Sci U S A* 95 (15), 8485-8490.

31 Hoekema, A. et al. (1987) Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. *Mol Cell Biol* 7 (8), 2914-2924.

32 Deana, A. et al. (1998) Silent mutations in the Escherichia coli ompA leader peptide region strongly affect transcription and translation in vivo. *Nucleic Acids Res* 26 (20), 4778-4782.

33 Sato, T. et al. (2001) Codon and base biases after the initiation codon of the open reading frames in the Escherichia coli genome and their influence on the translation efficiency. *J Biochem (Tokyo)* 129 (6), 851-860.

34 Stenstrom, C.M. and Isaksson, L.A. (2002) Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. *Gene* 288 (1-2), 1-8.

35 Griswold, K.E. et al. (2003) Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm. *Protein Expr Purif* 27 (1), 134-142.

36 Ishida, M. et al. (2002) Overexpression in Escherichia coli of the AT-rich trpA and trpB genes from the hyperthermophilic archaeon Pyrococcus furiosus. *FEMS Microbiol Lett* 216 (2), 179-183.

37 Klasen, M. and Wabl, M. (2004) Silent point mutation in DsRed resulting in enhanced relative fluorescence intensity. *BioTechniques* 36 (2), 236-237

38 Williams, D.P. et al. (1988) Design, synthesis and expression of a human interleukin-2 gene incorporating the codon usage bias found in highly expressed Escherichia coli genes. *Nucleic Acids Res* 16 (22), 10453-10467.

39 Makoff, A.J. et al. (1989) Expression of tetanus toxin fragment C in E. coli: high level expression by removing rare codons. *Nucleic Acids Res* 17 (24), 10191-10202.

40 Perlak, F.J. et al. (1991) Modification of the coding sequence enhances plant expression of insect control protein genes. *Proc Natl Acad Sci U S A* 88 (8), 3324-3328.

41 Kotula, L. and Curtis, P.J. (1991) Evaluation of foreign gene codon optimization in yeast: expression of a mouse IG kappa chain. *Biotechnology (N Y)* 9 (12), 1386-1389.

42 Jensen, L.G. et al. (1996) Transgenic barley expressing a protein-engineered, thermostable (1,3-1,4)-beta-glucanase during germination. *Proc Natl Acad Sci U S A* 93 (8), 3487-3491.

43 Hu, X. et al. (1996) Specific replacement of consecutive AGG codons results in high-level expression of human cardiac troponin T in Escherichia coli. *Protein Expr Purif* 7 (3), 289-293.

44 Haas, J. et al. (1996) Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr Biol* 6 (3), 315-324.

45 Zolotukhin, S. et al. (1996) A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J Virol* 70 (7), 4646-4654.

46 Ejdeback, M. et al. (1997) Effects of codon usage and vector-host combinations on the expression of spinach plastocyanin in Escherichia coli. *Protein Expr Purif* 11 (1), 17-25.

47 Hale, R.S. and Thompson, G. (1998) Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in Escherichia coli. *Protein Expr Purif* 12 (2), 185-188.

48 Uchijima, M. et al. (1998) Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium. *J Immunol* 161 (10), 5594-5599.

49 Johansson, A.S. et al. (1999) Use of silent mutations in cDNA encoding human glutathione transferase M2-2 for optimized expression in Escherichia coli. *Protein Expr Purif* 17 (1), 105-112.

50 Alexeyev, M.F. and Winkler, H.H. (1999) Gene synthesis, bacterial expression and purification of the Rickettsia prowazekii ATP/ADP translocase. *Biochim Biophys Acta* 1419 (2), 299-306.

51 Zhou, J. et al. (1999) Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J Virol* 73 (6), 4972-4982.

52 Vervoort, E.B. et al. (2000) Optimizing heterologous expression in dictyostelium: importance of 5' codon adaptation. *Nucleic Acids Res* 28 (10), 2069-2074.

53 Batard, Y. et al. (2000) Increasing expression of P450 and P450-reductase proteins from monocots in heterologous systems. *Arch Biochem Biophys* 379 (1), 161-169.

54 zur Megede, J. et al. (2000) Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 gag gene. *J Virol* 74 (6), 2628-2635.

55 Massaer, M. et al. (2001) High-level expression in mammalian cells of recombinant house dust mite allergen ProDer p 1 with optimized codon usage. *Int Arch Allergy Immunol* 125 (1), 32-43.

56 Narum, D.L. et al. (2001) Codon optimization of gene fragments encoding Plasmodium falciparum merzoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. *Infect Immun* 69 (12), 7250-7253.

57 Valencik, M.L. and McDonald, J.A. (2001) Codon optimization markedly improves doxycycline regulated gene expression in the mouse heart. *Transgenic Res* 10 (3), 269-275.

58 Leder, C. et al. (2001) Enhancement of capsid gene expression: preparing the human papillomavirus type 16 major structural gene L1 for DNA vaccination purposes. *J Virol* 75 (19), 9201-9209.

59 Woo, J.H. et al. (2002) Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in Pichia pastoris. *Protein Expr Purif* 25 (2), 270-282.

60 Shimshek, D.R. et al. (2002) Codon-improved Cre recombinase (iCre) expression in the mouse. *Genesis* 32 (1), 19-26.

61 Outchkourov, N.S. et al. (2002) Optimization of the expression of equistatin in Pichia pastoris. *Protein Expr Purif* 24 (1), 18-24.

62 Li, Y. et al. (2002) Cloning and hemolysin-mediated secretory expression of a codon-optimized synthetic human interleukin-6 gene in Escherichia coli. *Protein Expr Purif* 25 (3), 437-447.

63 Sinclair, G. and Choy, F.Y. (2002) Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, Pichia pastoris. *Protein Expr Purif* 26 (1), 96-105.

64 Hamdan, F.F. et al. (2002) Codon optimization improves heterologous expression of a Schistosoma mansoni cDNA in HEK293 cells. *Parasitol Res* 88 (6), 583-586.

65 Slimko, E.M. and Lester, H.A. (2003) Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels. *J Neurosci Methods* 124 (1), 75-81.

66 Bradel-Tretheway, B.G. et al. (2003) Effects of codon-optimization on protein expression by the human herpesvirus 6 and 7 U51 open reading frame. *J Virol Methods* 111 (2), 145-156.

67 Gao, F. et al. (2003) Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro expression and immune responses in DNA-vaccinated mice. *AIDS Res Hum Retroviruses* 19 (9), 817-823.

68 Li, A. et al. (2003) Optimized gene synthesis and high expression of human interleukin-18. *Protein Expr Purif* 32 (1), 110-118.

69 Disbrow, G.L. et al. (2003) Codon optimization of the HPV-16 E5 gene enhances protein expression. *Virology* 311 (1), 105-114.

70 Cid-Arregui, A. et al. (2003) A synthetic E7 gene of human papillomavirus type 16 that yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization studies. *J Virol* 77 (8), 4928-4937.

71 Yadava, A. and Ockenhouse, C.F. (2003) Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. *Infect Immun* 71 (9), 4961-4969.

72 Henaut, A. and Danchin, A. (1996) Analysis and predictions from *Escherichia coli* sequences. In *Escherichia coli and Salmonella typhimurium cellular and molecular biology* (Vol. 2) (Neidhardt F, C. et al., eds.), pp. 2047-2066, ASM press

73 Boycheva, S. et al. (2003) Codon pairs in the genome of Escherichia coli. *Bioinformatics* 19 (8), 987-998.

74 Altschul, S.F. et al. (1990) Basic local alignment search tool. *J Mol Biol* 215 (3), 403-410.

75 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31 (13), 3406-3415.

76 Satya, R.V. et al. (2003) A pattern matching algorithm for codon optimization and CpG motif engineering in DNA expression vectors. In *The Second International IEEE Computer Society Computational Systems Bioinformatics Conference* (Titsworth, F., ed.), pp. 294:305, The Institute of Electrical and Electronics Engineers, Inc.