Protein Expression Using Synthetic Genes

Tutorial: Guidelines for Effectively Using the Prot-2-DNA Codon-Optimized Algorithm

Claes Gustafsson, Ph.D.

The expression of functional proteins in heterologous hosts is a cornerstone of modern biotechnology. This is the basis for R&D activities such as drug target identification/validation as well as much of today's biotechnology manufacturing processes.

Despite significant progress in the fields of expression vectors, fermentation processes, and refolding protocols, heterologous protein expression is still very much a hit-or-miss endeavor.

The underlying problem is often that the DNA sequence of the gene encodes characteristics such as codon bias, cis-regulatory elements, and restriction sites of the organism from where it is derived, instead of characteristics needed for the expression system and organism where it is intended to be expressed.

Recent improvements in the speed, reliability, and cost of gene synthesis now facilitate the complete de novo synthesis of entire gene sequences to maximize the likelihood of high protein expression.

Immediate access to codon-optimized human genes through the PlanetGene portal, or custom-synthesized genes through providers such as **DNA 2.0** (Menlo Park, CA) leads to large cost savings in the biotech and pharmaceutical industry. This enables researchers to do many projects in parallel and maximize the pharmaceutical industry pipeline value.

In 1977, when **Genentech** scientists produced the first human protein (somatostatin) in a bacterium, protein expression of synthetic genes 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.

The Genentech group synthesized the 14 codon long somatostatin gene using oligonucleotides instead of cloning it from the human genome. They designed the sequence of the oligonucleotides based on the codons favored by the phage MS2.

Figure 1. Screenshot of Prot-2-DNA codon optimization software showing the human p21-activated kinase optimized for baculovirus Sf9 expression.

Not much of the *E. coli* genome sequence was known at the time, but the MS2 phage had just been sequenced and was correctly assumed to provide a good guide to the codon distribution in highly expressed *E. coli* genes. The result was the first production of a functional polypeptide from a synthetic gene, and the results created the biotech industry.

Now, a quarter of a century later, most genes are cloned from cDNA libraries or directly by PCR from the organism of origin. De novo gene synthesis is largely avoided because of perceived high costs and slow turnaround.

Despite its prevalence, PCR-based cloning is not necessarily the quick and easy solution to cloning. It requires templates that may not be trivial to access, optimization of gene-specific PCR conditions, re-sequencing of PCR product, and site-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 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.

Despite the advances that these approaches have made, they cannot avoid the underlying problem that was solved by the Genentech pioneers: the DNA sequence used to encode a protein in one organism is different from the sequence that would encode the same protein in another organism.

Codon Bias Affects Protein Expression

The genetic code is degenerate: 20 amino acids, and termination is encoded by 64 codons. All but two amino acids are coded for by two or more codons, and the codon preference varies widely between different organisms. For example, the amino acid Leucin is coded by UUA, UUG, and CUN (where N is any nucleotide). *E. coli* prefers the CUG codon, whereas yeast prefers UUA and UUG, and mammalian cells prefer CUC or CUG.

Preferred codons correlate well with the level of cognate tRNAs available within the cell. This relationship serves to optimize the translational system and balance codon concentration with correlating isoacceptor tRNA concentration. In *E. coli*, for example, the tRNA Arg4 that reads the infrequently used AGG and AGA codons for Arg is present only at low levels.

Increasing levels of the rare *tRNA* gene can sometimes increase the heterologous protein expression. This solution comes with a serious caveat—protein heterogeneity. Transfer RNA molecules are extensively decorated by more than 30 nucleotide modifications.

Figure 2. Screenshot of PlanetGene, an online collection of synthetic genes. A codon optimized human protein (MEK partner 1) is retrieved and flanking sequences tailored to the customer request.

These modifications are important for translational fidelity, aminoacyl recognition, and frameshift maintenance. Increasing the levels of the rare tRNA molecules results in tRNA undermodification, which, in turns, leads to translational misincorporation.

Figure 3. PlanetGene codon optimized genes in the form of lyophilized plasmid and stab. Each synthetic gene comes with double-stranded DNA sequencing tracefiles and a quality assurance verification.

Human genes codon-optimized for expression in *E. coli* typically increase expression levels 5–15 fold and have, in some cases, increased expression levels from undetectable to between 10–20% of *E. coli* cell mass. Even human genes codon-optimized for expression in mammalian cells often show a substantial increase in expression levels.

Gene Design Considerations

Designing a gene de novo can be both liberating and daunting. There are an enormous number of DNA sequences that all can encode the same single amino acid sequence. Each amino acid is encoded by an average of three different codons, so there are approximately 3^{100} ($\sim 5 \times 10^{47}$) nucleotide sequences that would all produce the same 100 amino acid protein. Which and how many of these possible sequences will result in high levels of heterologous protein expression?

The codon optimization algorithm Prot-2-DNA developed by DNA 2.0 involves using an initial codon usage table to propose candidate sequences, then a successive set of filters to eliminate those sequences that do not also comply with additional design constraints. A user-friendly version of Prot-2-DNA is available free of charge from DNA 2.0 *(Figure 1)*.

The codon usage table is constructed by calculating the codon frequency within an organism or a group of proteins (e.g. highly expressed *E. coli* proteins). These codon usage tables are adapted for gene design in two steps. First, a threshold level is set to completely eliminate all rare codons. Second, the remaining frequencies are normalized so that the summed frequencies for codons for each amino acid equal 100%.

Hybrid codon usage tables, such as the human/*E.coli* table used for the PlanetGene collection, can be constructed for a protein that is to be expressed in more than one host. Once the codon usage table has been constructed, candidate sequences are enumerated in silico by selecting codons using a Monte Carlo-based algorithm 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.

The filters remove any sequence that has unfavorable codon pairs, extreme GC content, repetitive sequences, or unfavorable mRNA secondary structures. The filters can also be customized to eliminate cryptic splice sites, internal ribosome binding sites, selenocystein incorporation signals, and certain restriction sites to facilitate downstream modifications.

The gene design algorithm 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).

Access to Codon-Optimized Genes

The PlanetGene portal (www.planetgene.com) is a collection of more than 25,000 synthetic human genes that have been codon-optimized for protein expression in *E coli* as well as mammalian cells using the Prot-2-DNA algorithm *(Figure 2)*.

This allows the user to easily move the synthetic genes between the two experimental systems, while retaining good expression levels and minimizing the heterogeneity of the protein product.

Each of the synthetic human genes has alternative 5′ and 3′ flanking sequences to accommodate an easy cloning procedure into most common expression systems, including N or C-terminal tags, fusion constructs, or different promoters *(Figure 3)*. The PlanetGene catalog can be searched by BLAST, keywords, or Genbank accession numbers.

Conclusions

Access to PlanetGene synthetic codon-optimized genes offers a mechanism by which researchers can assume much greater control of heterologous protein expression. In addition to adjusting the codon bias, the synthetic genes are also devoid of such elements as repetitive DNA and mRNA secondary structure.

Each gene is flanked by sequences to quickly move it into any expression system of choice to make the entire molecular biology consistent with an efficient R&D process.

The cost and fidelity of gene synthesis is 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 enroute to closing the circle to its distant past; the genetic engineering tools pioneered by the Genentech group and their academic collaborators in 1977 will once again become state-of-the-art. **GEN**

Claes Gustafsson, Ph.D., is co-founder and vp of operations at DNA 2.0. E-mail: cgustafsson@dnatwopointo.com. Website: www.dnatwopointo.com.