Review

Engineering genes for predictable protein expression

Claes Gustafsson, Jeremy Minshull, Sridhar Govindarajan, Jon Ness, Alan Villalobos, Mark Welch *

DNA2.0 Inc., 1140 O'Brien Drive, Suite A, Menlo Park, CA 94025, USA

A R T I C L E    I N F O

Article history:
Received 28 December 2011
and in revised form 27 February 2012
Available online 8 March 2012

Keywords:
Heterologous protein expression
Bioengineering
Synthetic biology
Gene optimization
Codon usage

A B S T R A C T

The DNA sequence used to encode a polypeptide can have dramatic effects on its expression. Lack of readily available tools has until recently inhibited meaningful experimental investigation of this phenomenon. Advances in synthetic biology and the application of modern engineering approaches now provide the tools for systematic analysis of the sequence variables affecting heterologous expression of recombinant proteins. We here discuss how these new tools are being applied and how they circumvent the constraints of previous approaches, highlighting some of the surprising and promising results emerging from the developing field of gene engineering.

Introduction

Low cost production of proteins in heterologous hosts is a fundamental capability on which biotechnology depends. Enzyme-catalyzed industrial processes are increasingly common in applications ranging from food processing to manufacture of small molecule pharmaceuticals. The modern molecular biology toolbox itself consists largely of heterologously expressed DNA modifying enzymes. Even manufacturers of high value protein therapeutics such as insulin and monoclonal antibodies are sensitive to the costs of making protein, particularly as patents expire.

The fermentation and purification steps of the protein production process are reasonably well understood and optimized by chemical process engineering approaches in which influential parameters are varied systematically and the results used to build ever improving processes. The resulting statistical models relate critical input variables such as media components and temperature to system performance and productivity. Robust and accurate models are useful in revealing system limitations, allowing engineers to increase quality, throughput and efficiency while simultaneously mitigating risk and reducing costs. In contrast to the rich
knowledge of variables affecting fermentation and purification, relatively little effort has been made to optimize the genetic components of the expression system, and that effort is usually directed towards a few defined regulatory elements rather than to the coding sequence of the gene itself.

The variables within the open reading frame of a gene that affect its expression are numerous and often interdependent. The relative frequencies of different codons used to represent each amino acid, the propensity of the 5′-end of the mRNA to fold into stable secondary structures, the fraction of the mRNA composed of G and C bases, and the presence of cryptic transcriptional terminators, and others have been proposed as primary determinants of a gene’s expression. Until recently researchers have been limited either to using naturally occurring sequences, or to making a single new version in which an alternative DNA sequence is used to encode the same amino acid sequence, so evidence to support any hypothesis has been sparse, anecdotal and biased in favor of experiments which yielded positive (and therefore publishable) data.

In the last few years, researchers have used hypothesis-driven sequence design and testing, as well as techniques of gene synthesis and statistical analysis to begin exploring the factors within the coding region of a gene that affect its expression. In this review we summarize the current state of understanding of these factors, and describe experimental techniques that are refining this understanding and identifying new factors that control gene expression. We illustrate how modern synthetic biology tools can be used to make a gene’s nucleotide sequence a parameter that can be tuned for expression as easily as dissolved oxygen concentration or pH is tuned for fermentation today.

**Systematic biological engineering**

Maximizing heterologous protein expression is, in principle, a classic multidimensional optimization problem. Multivariate optimization is well established in fields outside biotechnology: examples include such diverse non-biological applications as the design of text on credit card offers [1], player drafting strategy for major league baseball [2], and predicting customers’ movie rental preferences [3]. Examples in biological sciences include prediction of therapy efficacy from genomic data [4], identifying genes that functionally interact from microarray data [5], and QSAR small molecule drug design [6]. As different as these applications are, they have all been successfully approached and optimized using similar mathematical tools and approaches to what we describe here for the application towards gene engineering.

**Fig. 1** illustrates some of the variables that affect protein expression, and the levels within the system at which they operate. To build a statistical model that can predict the behavior of the system, input variables (here, rows in **Fig. 1**) are related to output variables (here, protein expression). Fundamental considerations in model building are described in the text box. Mechanistic understanding of the system can aid in variable selection, but it is not required to build predictive models. Instead the method requires the ability to experimentally make different combinations of variables, measure the performance of the system in each case, and evaluate the relative contribution of each variable to system performance.

For any multidimensional optimization problem, success depends on identifying the relevant variables to alter, and on efficiently sampling the variables such that their individual effects can subsequently be untangled from one another. Domain expertise, validated historical data, and practical constraints are typically used to select variables and their ranges to explore. To combine variables most efficiently it is standard practice in engineering disciplines to employ process optimization tools such as Design of Experiment (DoE) [7], Quality by Design (QbD), Lean Manufacturing and Six Sigma. These tools provide a framework for experimentation where several variables and variable combinations can be simultaneously tested. Multivariate analysis methods can then be used to identify the effect of each on the performance of the system.

Although this kind of systematic multivariate optimization is familiar to fermentation process engineers, it has been notably lacking in the practice of biological systems engineering. Bioengineering has instead been driven primarily by trial and error, usually with relatively little sampling of the potential variables that could impact productivity. One early exception is a systematic DoE study of the impact of promoter sequence substitutions which identified nucleotides in the promoter correlated with promoter strength; this information was then used to engineer new and stronger promoters [8]. The widespread adoption of de novo gene synthesis as a source of genetic constructs, and the exponential growth of known natural sequences are stimulating more systematic bioengineering approaches. The following sections briefly summarize developments in each of the bioengineering domains.

**Strain engineering**

We here define strain engineering as the modification or augmentation of host genotype to confer some improved property (e.g., growth rate, carbon source utilization, internal cellular environment) that improves protein expression productivity. Early strain improvement technologies consisted of whole genome mutagenesis and selection for desired phenotypes. More recently, with the availability of genome sequence data for many expression hosts and increased knowledge of gene functions, strain engineering based on mechanistic understanding has become increasingly prominent.

In parallel to rational strain engineering, successful efforts have been made to alter hosts using libraries of genes, gene fragments or engineered proteins that are screened for their ability to confer an improved phenotype. The Stephanopoulos lab at MIT has developed a process to introduce libraries of mutagenized RNA polymerase sigma or RNA polymerase alpha factors resulting in *Escherichia coli* strains with altered expression profiles [9,10]. Similar work has also been performed in the Harvard lab of George Church using oligonucleotide directed mutations [11] and artificial zinc fingers [12]. Mammalian cell lines have been engineered using RNAi molecules to alter the protein production pathways and secretion efficiency for therapeutic proteins [13,14]. Recent development of TALE nucleases and zinc fingers allows for precise and efficient editing of

---

**Abbreviations used:** DoE, Design of Experiment; QbD, Quality by Design; CHO, Chinese hamster ovary; GST, glutathione transferase; MMF, maltose binding protein; Tnj, thioredoxin; ORF, open reading frame; CAI, Codon Adaptation Index; tAI, tRNA Adaptation Index; eGFP, enhanced green fluorescent protein; scFv, single chain antibody; NN, Neural Nets; SVM, Support Vector Machine.
genomes [15]. Full genome sequence data for Chinese hamster ovary (CHO) cells [16] and Pichia pastoris [17] has recently become available and should accelerate the use of such editing technologies in these platform expression hosts. These genome editing technologies so far have been primarily applied for purposes other than optimization of protein production, but they offer a tremendous opportunity to systematically explore strain modifications for protein expression optimization. An approach of sampling combinations of various genetic modifications for improved expression of representative recombinant proteins should yield novel production strains that facilitate high yield expression of a broad range of target proteins.

**Vector engineering**

Most protein expression vectors employed today are an assemblage of natural genetic parts chosen from a limited pool of available, convenient control elements (promoters, terminators, translation initiation signals) and propagation elements (selection markers, replication origins). These vectors have not been systematically optimized for production – they are simply functional constructions that have been employed for lack of better options or because of historical precedent. There is great potential for systematic engineering not simply of individual vector parts but of part combinations to optimize vector properties.

New promoters, replication origins, cloning sites and selection tags now appear almost daily from academic labs and companies such as New England Biolabs, Promega and Life Technology. Efforts are ongoing by Biofab (www.biofab.org), Addgene (www.addgene.org) and others to catalog and standardize these genetic vector elements and organize distribution processes. Solubility tags such as glutathione transferase (GST), maltose binding protein (MBP) and thioredoxin (Trx) are often used independently or in combination with purification and detection tags such as His-, Strep- and FLAG-tags. The abundance of convenient fusion tags allow for multiple of options when assessing solubility, stability, folding and expression issues [18].

While the number of potentially useful vector elements is expanding, more work is needed to optimize these elements as well as combinations of elements in vectors for heterologous protein expression. With increasingly efficient de novo DNA synthesis [19,20], and computational tools that facilitate the control and design of synthetic gene constructs, vector elements can be stored and distributed as virtual information, to be incorporated independently into designed synthetic constructs based on their performance rather than having to compromise for the best available combination of genetic elements found in the lab freezer with the appropriate cloning sites.

**Open reading frame engineering**

Synonymous substitutions in the open reading frame (ORF) that encodes a protein encompass the least well understood set of variables: they are potentially numerous, and they are intimately intertwined with the protein they encode. Thus, while it is experimentally straightforward to take a natural promoter and test its ability to drive transcription of a wide variety of other genes, or to take a replication origin and test its ability to replicate in a variety of host organisms, until recently it has been much more difficult to recode an open reading frame using synonymous substitutions and test the effect in a meaningful way.

Early heterologous protein expression efforts (such as the pioneering work by Genentech 1977 to express human growth hormone in bacteria) relied on complete gene design and synthesis since only the amino acid sequence of the human hormone was known at the time [21]. With the advent of lambda gt11 cloning and PCR cloning in the late 1980s, gene engineering primarily involved minor alterations such as the inclusion and elimination of restriction sites into natural sequences to facilitate cloning. With increased knowledge of genome sequences, biases in codon usage and other sequence features became evident and provided rationale for gene sequence modification to improve protein expression. Over the past few decades, observed natural sequence biases have been the primary basis for gene engineering approaches. Such approaches have occasionally proven to be useful for boosting protein expression over that of recalcitrant wild-type genes, in some cases by several orders of magnitude. However, the benefits of such approaches have been inconsistent and, in some cases, such “optimization” has been shown to worsen protein expression [22].

It should not be surprising that copying natural biases does not consistently result in good heterologous protein expression. It is important to recognize that natural genes have evolved in the context of natural selection. Accordingly, the evolved features and biases seen in natural gene sequences may not reflect a host’s preferences for high level production of a heterologous protein. Successful overexpression in some cases can reach levels over 30% of the total cellular protein – orders of magnitude higher expression levels than the most abundant natural proteins [23]. Heterologous overexpression in E. coli can overburden cell metabolism, leading to pleiotropic effects that may significantly reduce protein yield [24–27]. Selective pressures on naturally evolved coding sequences are likely complex and include interactions with the entire system in which the genes have evolved: transposon resistance, mRNA processing, RNAi regulatory and assembly into higher order DNA structures are all intertwined with the protein coding information.

**Gene design variables**

There are many ways a gene’s sequence can influence protein expression, including effects on mRNA levels, translation velocity, efficiency of initiation, and rates of charged tRNA consumption. Each of these different influences is exerted through distinct interactions with cellular machinery and each can be associated with different types and combinations of sequence elements. For example, AT-rich sequences within the gene could cause premature transcriptional termination and reduced mRNA levels, mRNA structure near the ribosome binding site could reduce translational initiation, and rare codon clusters within the mRNA could cause translational pausing. All of these sequence elements would have the observable effect of reducing protein expression. Variables in the context of biological engineering therefore can range from single base pair changes to combinatorial variation of bases in an element to replacing large sequence elements in toto. The broad range of testable variables for gene engineering puts a particular emphasis on the methods of variable selection and experimental design when exploring gene sequence-function relationships.

Not only do the possible variables of gene design range in type and influence, but they also can show high degrees of co-variation and complex interdependence. Gene sequence variables range from local sequence motifs to global characteristics such as codon usage frequency. Any particular substitution in a sequence can be interpreted as a change in multiple variables. For example, a synonymous substitution of a single codon will also necessarily influence local and global codon usage frequency, nucleotide frequency, local and global mRNA structure, and local sequence patterns. If an effect is seen upon any particular substitution, then the explanation might be related to any one or a combination of the intertwined variables. Distinguishing individual variables becomes even more complicated when several simultaneous substitutions are made, as is typical in most gene optimization strategies.
An comprehensive (at the time) collection of 41 published examples comparing protein expression levels from wild-type genes versus genes “optimized” by various algorithms was published in 2004 [22]. Since then, the number of published examples has increased by orders of magnitude [28]. In the majority of published examples the studies resulted in increased protein expression level after gene “optimization”, although improvements vary from none to >1000-fold. Essentially all these studies, however, are comparisons of only two genes differing in several gene variables. Because such limited studies cannot distinguish the relative contribution of the multiple variables actually being sampled, it is difficult to assess the robustness and applicability of the design algorithms for the improvement of other genes. Furthermore, it cannot be demonstrated that the algorithm has achieved an optimal combination of gene variables to result in the best possible expression level. Indeed, recent studies of the impact of gene design on heterologous protein expression have shown that some of the primary tenets of common gene design algorithms do not significantly explain variance in heterologous protein expression yield [29–31]. In order to derive truly robust algorithms, systematic assessment of all relevant variables is necessary. In the next sections we describe potential gene engineering variables in more detail and current evidence for their importance for heterologous protein production. We then discuss gene design and synthesis strategies for exploring gene variable space and identifying gene design principles for optimal expression of recombinant proteins.

Variables for ORF engineering

Variables for gene engineering can broadly be classified as local or global variables. Herein, we define local variables to be those variables which are dependent on local sequence patterns within the gene. In general, local variables affect protein expression at the level of the gene DNA or mRNA. Examples are RNA degradation motifs and mRNA secondary structures. In contrast, global variables are measures of aggregate gene features. Examples include codon usage frequency and GC%. Some variable types can be expressed as either local or global. For example, one might consider mRNA structure locally in a window near the ribosome binding site of the gene; alternatively one might consider mRNA structure globally as average structure strength over sequence windows across the gene. Likewise, a global variable such as GC% may be assessed in specific regions of a gene (e.g., a window of codons at the start of the ORF). This distinction between local and global classes is somewhat arbitrary, but serves to illustrate that gene sequence information relevant for protein expression exists in several interdependent layers and that there are many ways to parameterize gene sequence variables.

Local variables for ORF engineering

A complete review of local sequence variables suggested to affect heterologous protein expression is beyond the scope of this article and previous works have covered this subject comprehensively [22,32–35]. There are a number of local variables (sequence motifs) that are considered to be deleterious for heterologous protein expression and many gene design algorithms seek to exclude these from gene sequences. This leads to an inherent conflict in the variable selection process: if too many deleterious variables are present, none of the genes will express, and there will be no data with which to build a statistical model. However, eliminating too many sequence motifs based on preconceived ideas about what is important rather than experimental evidence can severely constrain other correlated gene variables that are actually relevant.

Local variables that are generally considered deleterious include RNase sites, DNA recombination sites, transcriptional terminators, and transcription factor recognition sequences. Some motifs depend on the host organism: cryptic splicing and internal polyA signals could affect genes to be expressed in eukaryotic hosts, while Shine–Dalgarno-like sequences may affect genes to be expressed in prokaryotes. Avoiding motifs with relatively ambiguous sequence composition, such as RNase E sites [36] and polyadenylation signals [37], can be particularly problematic as complete removal of such motifs is difficult and will impose great constraints on other sequence variables.

A variable commonly regarded as local and deleterious is the occurrence in the gene of any codon utilized rarely in the host organism transcriptome. Such codons are generally decoded by rare tRNAs and the presumed logic has been that translational elongation, and thus protein production, may be slowed by overuse of low-concentration tRNAs [38]. Engineering E. coli strains to over-produce rare tRNAs has been shown to improve protein expression from a number of genes with elevated numbers of rare E. coli codons [39–41]. However, genes with several rare codons have been shown to express at a high level in E. coli without increased production of rare tRNAs [28,42]. Further, a number of examples have been described where the increased presence of rare codons in the gene correlates with increased (not decreased, as might be expected) protein expression [43–45]. Overall, studies have observed little correlation between the number of rare codons in a gene and either the protein expression level or the expression improvement upon overproduction of rare tRNAs [46]. These exceptions do not necessarily disprove the idea that rare codons can limit translation rate or that this limit can be overcome with increased levels of cognate tRNAs; however, they do show that the impact of rare codons on protein expression is not fully understood.

One of many possible reasons that the relationship between rare codon occurrence and protein expression is complex is that sequence context may be critical. Studies have indicated that the impact of rare codons is greater when such codons are clustered (i.e., when multiple rare codons are used within a span of a few codons) [38]. Clusters of rare AGA or AGG (arginine) codons may cause ribosomal pausing [47], ribosomal frameshifting [48], decreased translational velocity [49] and amino acid mis-incorporation [50] in E. coli. Also, the impact of the presumed pausing at rare codons may depend on position within the ORF. The deleterious impact of rare codons may be more pronounced when found in the initial few codons of the open reading frame [51,52].

A number of other context-dependent codon usage variables have been suggested by previous work. Cannarozzi, et al., observed biases in local codon usage in genomic E. coli sequences that show a statistical preference for repeated use of the same cognate tRNA within clusters of like amino acids [53]. The authors suggest that tRNA diffusion relative to recharging might limit translation efficiency and tRNA re-use could promote faster elongation. Several groups have characterized significant biases in codon pair frequencies in genomes [54,55]. Measurements of in vivo translational step times suggests that codon pair context can significantly influence translational elongation rates in E. coli [56,57], although the impact of codon pair context on heterologous expression yield is unclear [58,59]. In our studies, we observed no correlation between multiple parameterizations of codon pair usage and expression in E. coli [30]. Another contextual factor that may be significant is position within the ORF with respect to structure of the encoded protein [60,61]. Recent data suggests that strategic placement of certain codons might influence folding efficiency of the encoded protein, presumably by modulating translation rate at critical points in elongation [44,62–65].

An even more complex local variable is RNA secondary structure. Predicted structures in the 5’UTR of genes have been shown to impair protein expression in eukaryotes [66]. Strong RNA structures
predicted to form in the region of the RBS or within the initial coding sequence of the gene ORF have been shown to impair expression in prokaryotes, presumably by interfering with ribosomal binding and translational initiation [29,67]. Study of gene design variables related to mRNA structure is complicated by the fact that measures of structure strength currently rely on computational RNA folding algorithms that are limited in their ability to predict form and strength of actual mRNA structures in vivo. One problem is that commonly used programs cannot reliably predict tertiary structure, which can contribute significantly to overall structure strength and form. Ribosomes possess an intrinsic helicase activity that allows translation through even very strong hairpins [68]. Structures in actively translated messages are repeatedly unwound as ribosomes progress [69]. The structures that do form would be transient, dynamically translated messages are repeatedly unwound as ribosomes and structures can form. Emerging tools such as ribosome profiling might significantly alter which genes in a library are expressed in vivo and its relationship to translational efficiency, helping us to identify more meaningful descriptors of mRNA structure for protein expression. We are clearly still lacking a good understanding of the relationship between predicted mRNA structure and expression level, making this an important class of gene variable to continue to explore.

Global variables for ORF engineering

All of the local sequence variables discussed above must also be considered within the context of global variables. The two global variables most commonly considered in ORF engineering are codon usage frequency and G + C nucleotide content (GC%). The common use of these variables stems from observations of significant codon and/or nucleotide bias in genes, or certain classes of genes in several organisms [72]. As with local variables, global variables have not been systematically explored, and most current “optimization” methods are based on weak experimental evidence or none at all.

Almost all ORF engineering uses design principles that are taken directly from natural genes within the expression host. In the simplest form this means just copying the codon usage frequency or GC% bias of the expression host organism. A variation on this approach uses the codon usage of genes that are highly expressed in the expression host [33]. In 1987, Sharp and Li proposed a quantitative measure of the codon bias in highly expressed genes which they named the Codon Adaptation Index (CAI) [73]. CAI is a measure of how often ‘favored’ codons are used within a gene. ‘Favored’ codons refer to the most frequent codon for a given amino acid either across all genes in the host or only within a subset of naturally highly expressed genes in the host. A gene with a CAI of 1 uses only the most favored codon throughout the entire gene. Gene engineering toward a high CAI therefore dramatically narrows the codon diversity in the gene as the 61 possible codons are reduced to a set of only the 20 most frequent codons (and indirectly narrows all variables correlating with codon bias). A similar parameterization of codon usage called the tRNA Adaptation Index (tAI) has been proposed where this index weights codons not by their relative frequencies in the host, but by their cognate tRNA concentrations, based on tRNA assignment data [74].

The rationale for copying host gene codon biases is straightforward: if the synthetic gene is similar in codon usage to host genes, it presumably will utilize the host tRNA supply in balance with the host transcriptome and therefore be well accommodated and expressed. This is supported by the observation that codon usage frequency bias is generally correlated with cognate tRNA concentration in the cell [72,75,76]. The primary assumption in this logic is that total tRNA concentration is static and correlated to translation elongation rate and that this rate can be expression-limiting. Although in some hosts, including E. coli, CAI and tAI show significant correlation with natural gene expression levels, recent studies do not support these parameterizations of codon usage as strong determinants of heterologous protein expression level in E. coli [30,31,45,29]. This is potentially a case where correlations observed in nature reflect the co-evolution of host systems and gene sequences, but may not predict the outcome of the unnatural conditions and burden of heterologous protein overexpression.

One concern with using variables like the adaptation indices and GC% is that they are ambiguous composites of other variables. For example, two very dissimilar genes, one containing a mixture of high and low-frequency codons and one using exclusively intermediate frequency codons would have the same CAI value. Similarly, the GC% is a composite number that is ambiguous about the individual G%, C%, A%, and T%. GC% also is ambiguous with respect to codon usage frequency, as it does not measure which amino acids are encoded by codons ending in A/T and which are encoded by codons ending in G/C. If individual codon frequencies do have an effect on expression, these ambiguous compressions can lose significant information about their contributions. In systematic gene engineering, care must be taken to define variables unambiguously so that the results are decipherable and true knowledge is gained.

We suggest that the most unambiguous global codon usage variable is the frequency with which each codon is used to encode each amino acid within the protein. If composite variables that are intertwined with codon usage are to be explored, these should be unambiguously defined and validated in studies where the relative impacts of the component variables (i.e., codon usage frequencies and nucleotide frequencies) can be assessed.

Recent experiments with coding sequence variables

The past few years have seen the beginnings of more systematic experimental testing of the effects of sequence coding variables. This hypothesis-testing approach is providing data that will force significant revisions of current design prejudices based merely on observations of natural systems.

In one study, Kudla et al., measured expression in E. coli of a library of 154 genes for enhanced green fluorescent protein (eGFP) that varied in synonymous codon usage [29]. To assess gene variable preferences for heterologous protein expression, the authors used a biased randomization approach in which multiple semi-random oligonucleotides were assembled to form the eGFP coding sequences. The oligonucleotides used in the library synthesis were designed to produce global and regional variation in GC% and CAI. When expressed in E. coli, the eGFP protein levels varied 250-fold across the library as measured by fluorescent intensity. Based on multivariate data analysis, the authors concluded that the majority of the difference in protein expression can be explained by mRNA folding near the translational start of the eGFP ORF. No significant correlation was found between protein expression and GC% or CAI, the gene variables that were primarily targeted for diversification.

In 2010, Allert et al. used a designed variation approach to create 285 genes encoding three different proteins (120 + 39 + 126 gene variants) [31]. These genes were designed to sample global codon usage according to CAI and regional GC% based on observed biases seen in natural E. coli genes. The relative expression levels were assessed using in vitro E. coli extracts and ranked on a scale from 0 (no band on a protein gel) through 3 (strong band). The data were then fit to a function proposed to model expression. The
authors concluded that the majority of the difference in in vitro expression could be attributed to low GC content and low predicted mRNA structure in the 5' end of the ORF.

In a recent study, we used a designed systematic variation approach outlined in Fig. 2. In these experiments we created two sets of /C24 gene variants encoding a single chain antibody (scFv) and a DNA polymerase [30]. The genes were designed using Design of Experiments methods to minimize co-variation in the design variables (i.e., maximized orthogonality). Individual codon choice was randomized in silico to maximize the sequence diversity so that each variant was equally distant from any other variants as measured in Hamming distance (the number of sequence differences at the nucleotide level). Each gene variant was synthesized as an independent gene and the relative protein expression measured. When expressed in E. coli, both the protein expression levels of the scFv and the DNA polymerase varied over two orders of magnitude. In agreement with the study by Kudla et al., and Allert et al., we did not observe any significant correlation between heterologous protein expression and CAI. With our different method of codon usage sampling, however, we could find significant correlation between heterologous protein expression and the frequencies of codons used to encode a subset of amino acids. In contrast to the Kudla and Allert studies, we did not see any correlation between predicted mRNA

C. Gustafsson et al. / Protein Expression and Purification 83 (2012) 37–46
structure near the 5’ end and heterologous protein expression; however, this may be explained by the fact that the scFv and the polymerase gene codings showed low propensity to form structure in this region.

Information derived from the systematic analysis described above has been used to identify coding sequence variables affecting protein expression. For heterologous expression in *E. coli* multivariate analysis as described herein identified and quantified the frequencies of specific codons for about six amino acids to correctly predict the observed differences in expression. It is at this point not clear what the biochemical basis is for the correlation. It may reflect a physiological shock to the host cells as they attempt to synthesize large amount of a single protein, biasing the consumption of the aminoacyl-tRNA population: most of the best codons for high expression correspond to those tRNA that have been shown to remain highly charged under starvation conditions [30,77,78]. In all cases to date, expression is highly correlated with codon usage of a subset of codons. Consistent with Kudla et al., and Allert et al., we do not see a general preference for codons used at highest frequency in the genome or in the highly expressed gene subset of the host.

Much further research is needed to fully understand the experimentally-observed relationships between gene features and expression level; however, the observed correlations can already serve as the basis for reliable design algorithms as well as providing direction for gene improvement strategies. Fig. 3 shows how at DNA2.0 we have applied the results of our studies to improve gene design for *E. coli*. A regression model of the impact of codon usage is used to predict improved codon usage for the design of new genes. A handful of examples are shown where we have observed dramatic improvement in expression by applying codon usage predicted to be improved relative to a high codon adaption index bias. We have now employed experiment-based algorithms to optimize thousands of genes for *E. coli* and have observed a strong improvement in reliability and yield versus our previous algorithms based on genomic data.

In addition to *E. coli*, we have also employed experiment-based optimization to improve gene design algorithms for more than a dozen different protein expression hosts including *P. pastoris*, *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Agaricus bisporus*, mammalian CHO and HEK293 cells, insect cells, *Clostridium* species, and dicot and monocot plants (unpublished; some results

Fig. 4: Sequence diversity in three synthetic gene engineering libraries. Un-rooted tree generated by neighbor-joining, based on the pairwise Hamming distances (number of nucleotide substitutions) among synthetic genes in three different kinds of library. (A) Figure S1 from [29] (Reprinted with permission from AAAS). The library encodes 168 green fluorescent protein gene generated by degenerate oligonucleotide randomization. (B) One of three libraries generated in [31]. This library encodes 120 triose phosphate isomerase gene variants generated by a rational designed variation scheme. (C) A set of 48 individually designed eGFP gene variants created by systematic codon usage variation and codon choice randomization. The scale for each tree indicates the correspondence between branch length and proportion of nucleotide sites that differ between gene sequences. Homology based neighbor-joining trees B and C were generated in MEGA [86].

Fig. 5: Codon usage correlation for two synthetic gene engineering libraries. Colors indicate the degree to which each pair of codons is correlated in global usage frequency either positively (red) or negatively (blue) among genes of the library. Library 5B is same library as 4B and Library 5C is same library as 4C. See legend of Fig. 4 for library details. The DNA sequences for gene variants in 4A [29] were not publicly disclosed, prohibiting inclusion in this comparison. Plots were generated using PLS Toolbox software (Eigenvector, Inc.) in a Matlab environment.
is available at www.dna20.com). In every case we have seen a dramatic dependence of expression on synonymous codon usage that we can capture in gene design algorithms. Based on these results and those emerging throughout the field, it is clear that much valuable knowledge can be gained by experimental interrogation of the impacts of gene design on expression. We thus expect a dramatic improvement of gene design algorithms for any host of interests as methods for host interrogation evolve.

**The future of coding sequence engineering**

A successful engineering approach to coding sequence design requires that gene expression can be predicted with a manageable set of variables. The searchable ORF engineering space is large and highly correlated. There can be more than a googol (10^{100}) different ways to encode a specific protein sequence – clearly far beyond what can be exhaustively searched by any technology [34]. Still, solutions within this space appear plentiful, judged by the number of reported successes of ORF re-coding [22], although one must bear in mind the many failures that go unreported because of the nature of the scientific literature.

The availability of molecular biology tools, particularly of de novo synthetic genes, has opened the door to a new era of ORF engineering. Application of experimental design and multivariate analysis methods commonly utilized in traditional engineering practices and fermentation can now be efficiently used to experimentally optimize gene variables. Fig. 2 illustrates how such a “machine learning” approach is applied to gene engineering. Predefined values of variables are distributed in a test matrix according to Design of Experiments methodology. Each sample is subsequently synthesized and the corresponding protein expression is assessed in the relevant biological system. The resulting data is used to build predictive statistical models and/or the process is repeated with another iteration using knowledge gained to alter the search space (i.e., variables sampled and their value ranges). See box ‘Building predictive models’ for a more detailed description of the process.

The efficiency and outcome of the search is highly dependent on the means by which variation is created in the genes. A Design of Experiments method seeks to maximize search efficiency by maximizing the number of variables assessed per test while minimizing co-variation between variables within the test set. A key advantage of this is that the number of test samples required to assess the variable space is minimized [79]. This can be critical as many real-world applications are difficult to assess with high-throughput assays. Minimally the sample set should be larger than the number of relevant independent variables to explore [80]. In practice, some degree of over sampling is usually warranted to overcome experimental error and “noise” from unaccounted for variables or variable combinations. Thus, to assess the relative contribution of 10 variables, one would want to search using at least 11 but perhaps >30 systematically varied test samples.

In contrast, a random variation approach will be considerably less efficient. Although such approaches have the advantage that diversity can be created with little investment in gene design and synthesis, there are serious limitations which can compromise experiments and their interpretation. One is that many more individuals need to be tested when variation is random versus a properly designed and synthesized test set. For example, for a library to have a >99.5% probability of sampling all combinations of four variables with two values each, one would need to test more than 108 variants randomized in these variables, assuming a ‘perfect’ randomized library. Since true randomization is often difficult to achieve in practice one usually would want to significantly over sample to be sure to sample variation that might be under represented. A synthetic library could sample the 2^4 = 16 possible variants with exactly 16 test genes. Thus, the most efficient way to search the space is to strategically design and synthesize individual test genes (see Figs. 4 and 5).

Another advantage of systematically designed and synthesized libraries is full control of the hierarchy of gene variables. Limited or full randomization of synonymous codon choice at positions in a gene can provide high local sequence diversity, but can only poorly sample many global gene variables which are averages of many codon choices, such as overall codon usage frequency. In silico design and gene synthesis are not constrained by any gene sequence variable allowing exploration of any variable type and range.

Figs. 4 and 5 illustrate the advantages of using gene design and synthesis to create variation. Fig. 4 shows how a designed library can control local sequence variation relative to alternative non-systematic methods. The library in Fig. 4A was obtained by assembling oligonucleotides that contained degeneracies at the third “wobble” position of each codon. The library in Fig. 4B was the product of rational designed variation. The biases of these libraries are clear in the presence of very dense clusters of related sequences. This indicates that some areas of sequence space are being preferentially explored which might limit interpretation of the results. For contrast, compare the even sequence space coverage obtained in a set of individually designed genes, shown in Fig. 4C.

Control over global codon frequencies is illustrated in Fig. 5. Covariation between usage frequencies of different codons is shown for two different library designs. In the plots shown, each of the 61 sense codons is represented on the x and y axes, and covariation in usage between one codon and each other codon is indicated by coloration of the resultant grid: red squares indicate where two codons covary positively, blue squares show where they covary negatively. The sea of red and blue seen for the library in Fig. 5B can be compared with the minimal covariation seen between codons in a set of genes that are designed to diversify codon usage and then individually synthesized, shown in Fig. 5C. In this latter library the areas of intense blue are negative covariation between codons where there are only two possibilities: for example CAT and CAC to encode histidine. Here it is unavoidable that when one codon is used more, the other must be used less. Covariation in the Fig. 5B library is largely due to an intentional global bias in CAI and GC%. While such a library can assess the impact of the particular bias used to vary the genes, it cannot provide information on other measures of global codon usage. The lack of covariation in the Fig. 5C library allows essentially unambiguous assessment of any global codon frequency variable (including CAI).

Experimental design of sequences, direct synthesis and multivariate analysis of the resultant data is the most promising path towards understanding sequence design principles. It is also an efficient combination of tools for obtaining useful solutions to individual expression problems. By interrogating the search space with high precision and great information efficiency, the number of experimental samples can be drastically reduced while producing data that provides the most information regarding the contribution of individual variables to system performance. The same underlying approach can be applied to protein engineering [81–83] and pathway engineering [84], and we expect them to be equally applicable to genome engineering [85]. As the complexity and number of variables in the designed system increases, systematic sampling will become even more critical for understanding and optimization. Understanding the variables and design parameters for re-coding open reading frames will provide an excellent test case in this brave new world of synthetic biology.
References

[1] C. Fishman, This is a marketing revolution, Fast Company 24 (1999) 204.

Acknowledgments

This work was supported by NSF SBR Grant No. 0750206, the NIH Roadmap Center Grant No. P50 GM073210 and DNA2.0 Grant No. 1. We thank Josh Silverman, Calysta Biosystems for input on the manuscript.