ACS SyntheticBiology

Letter

pubs.acs.org/synthbio

Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector−Coding Sequence Junction

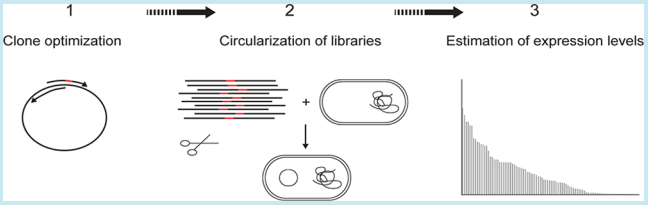
Kiavash Mirzadeh,[†] Virginia Martínez,[‡] Stephen Toddo,[†] Suchithra Guntur,[†] Markus J. Herrgård,[‡] Arne Elofsson,[†,§] Morten H. H. Nørholm,*[,‡] and Daniel O. Daley*[,†]

[†]Center for Biomembrane Research, Department of Biochemistry and Biophysics, [§]Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden

[‡]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark

Supporting Information

ABSTRACT: Protein production in *Escherichia coli* is a fundamental activity for a large fraction of academic, pharmaceutical, and industrial research laboratories. Maximum production is usually sought, as this reduces costs and facilitates downstream purification steps. Frustratingly, many coding sequences are poorly expressed even when they are codon-optimized and expressed from vectors with powerful genetic elements. In this study, we show that poor expression can be caused by certain nucleotide sequences (e.g., cloning scars) at the junction between the vector and the coding sequence. Since these sequences lie between the Shine−Dalgarno sequence and the start codon, they are an integral part of the translation initiation region. To identify the most optimal sequences, we devised a simple and inexpensive PCR-based step that generates sequence variants at the vector−coding sequence junction. These sequence variants modulated expression by up to 1000-fold. FACS-seq analyses indicated that low GC content and relaxed mRNA stability (ΔG) in this region were important, but not the only, determinants for high expression.



KEYWORDS: *protein production, translation initiation, mRNA structure, codon optimization, vector−coding sequence junction, membrane protein*

In a typical bacterial protein production experiment, the coding sequence (CDS) to be expressed is cloned into a vector that contains well-defined genetic elements, such as an origin of replication, an antibiotic resistance marker, and sequences that control transcription and translation.[1,2] High levels of production are usually attained using high-copy vectors and promoters that permit high levels of transcription upon induction. It is also important to use a translation initiation region (TIR) that contains a Shine−Dalgarno (SD) sequence that is purine-rich and optimally spaced from the start codon by a linker region.[3−6] This promotes interaction with the 16S rRNA during the initiation of translation.[7]

Translation initiation is considered to be the rate-controlling step of translation,[3,4,8,9] and the nucleotide sequence of the TIR can have a direct effect on protein production. For example, nucleotide changes to the SD can affect protein production levels by as much as 600-fold.[10] The linker region between the SD sequence and the AUG is not thought to play a specific role;[9] however, there are reports that nucleotide changes in this region can also affect protein production.[11,12] The choice of start codon can also affect translation initiation, with AUG being more effective than alternative start codons.[6,9] Nucleotide sequences in the TIR can also modulate translation initiation by forming mRNA secondary structures that reduce the possibility of interaction with the ribosome.[3,13−18] Prediction algorithms that randomize the TIR *in silico* and select a sequence context that has a relaxed mRNA structure[19] or that combine relaxed mRNA structure with favorable interactions with the 16S RNA are available.[20−22] While these approaches are to some degree successful, there are areas in which they could be improved.[23]

The initial phase of elongation is also thought to be important for protein production, as synonymous codon substitutions in the second and third codons can affect protein production levels.[24−27] However, it is not clear whether this affects expression by modulating (1) ribosomal speed (i.e., through a translational RAMP[28,29]), (2) mRNA folding around the TIR,[13,18,30] or (3) a combination of both.

Despite the fact that all elements in a protein production experiment might have been selected for high-level expression, some CDSs are still difficult to express, and the reasons for this are not clear.[12,20,31−33] This context-dependent problem limits the rational design of expression experiments, causing time delays and increasing costs. To some degree, it can be addressed by utilizing ribosome binding site (RBS) calculators that select sequence contexts using thermodynamic principles (reviewed in ref 23), approaches such as transcriptional and translational coupling,[10,34,35] or translational fusions.[15,18] In this study, we devised a PCR-based optimization step that modifies the junction between the vector and the coding sequence. The approach allowed us to express CDSs that were previously

Received: February 16, 2015
Published: May 7, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

thought to be difficult to express, even though optimized genetic elements had been used. Thus, we were able to provide an experimental tool for addressing context-dependent expression.

The most commonly used vectors for protein production are the T7-based *pET* range, which we have used in this study. They allow expression of the CDS in strains of *Escherichia coli* that contain a lysogenized *DE3* phage fragment encoding the T7 RNA polymerase. Examples of such strains include BL21(*DE3*) and derivatives that have been selected or engineered for high-level production.[1,2] Previously, we cloned 502 CDSs for *E. coli* membrane proteins into a modified version of the *pET28a* vector (herein called $pET28a^{XhoI}$).[32] The CDSs were genetically fused to a region encoding a TEV-GFP-His_8 tag (Figure 1a) so that whole cell fluorescence measurements could be used to estimate expression levels (Supporting Information Figure S1). While working with two difficult to express CDSs, $araH^{WT}$ and $narK^{WT}$, we noted that the choice of restriction site used for cloning could affect expression levels significantly. For example, when an *Xho*I site was present, $araH^{WT}$ and $narK^{WT}$ both expressed poorly (Figure 1b, lanes 1 and 4). In contrast, when the *Xho*I site was replaced with either *Eco*RI or *Dra*I sites, the level of expression dramatically increased (Figure 1b, lanes 2, 3, 5, and 6). A possible explanation for this observation is that the combination of *Xho*I recognition sequence and the 5′ end of the $araH^{WT}/narK^{WT}$ resulted in unfavorable secondary structure when transcribed into mRNA. Since the 5′ restriction enzyme site sits between the SD sequence and the start codon, it encodes part of the RBS, and strong mRNA structure in this region could affect translation initiation and therefore protein production.[13−17,29] Analysis of mRNA structure in this region supported this hypothesis, as it indicated that changing the restriction site relaxed the mRNA structures within the RBS (Figure 1c). On the basis of these data, we reasoned that cloning scars (defined here as sequences outside of the CDS that were used for cloning) could have a significant effect on expression.

To find optimal sequences at the vector−CDS junction, we designed a simple PCR step. In the experiment, the six nucleotides upstream of the AUG start codon, which contain the restriction enzyme recognition sequence, were changed in all possible combinations by PCR amplification of the original plasmid using degenerate primers (Figure 2a,b). We also changed the six nucleotides downstream of the AUG start codon so that synonymous codons could be sampled. The latter design concept was included because synonymous codon choice in the +2 and +3 positions can affect the expression of $araH^{WT}$ and $narK^{WT}$ (ref 24 and Supporting Information Figure S3) as well as other CDSs.[25−27] Note that we deliberately did not change the SD sequence, as it was already considered to be strong. The libraries that we generated were called $pET28a^{OPT}$-$araH^{WT\text{-}OPT}$ and $pET28a^{OPT}$-$narK^{WT\text{-}OPT}$, and it was mathematically possible that they contained 24 576 and 49 152 different vector−CDS junctions, respectively. Analysis of expression levels from 96 randomly selected colonies in these libraries indicated a 350-fold difference between the lowest and the highest expressing clones (Figure 2c,d). This difference was not caused by cell-to-cell variation, as little variation was observed when we assayed 96 colonies of the original unoptimized clones (see inset boxes). Significantly, the highest expressing clones from both of these small-scale screens were in excess of 60 mg/L, which is extremely high for a membrane protein.

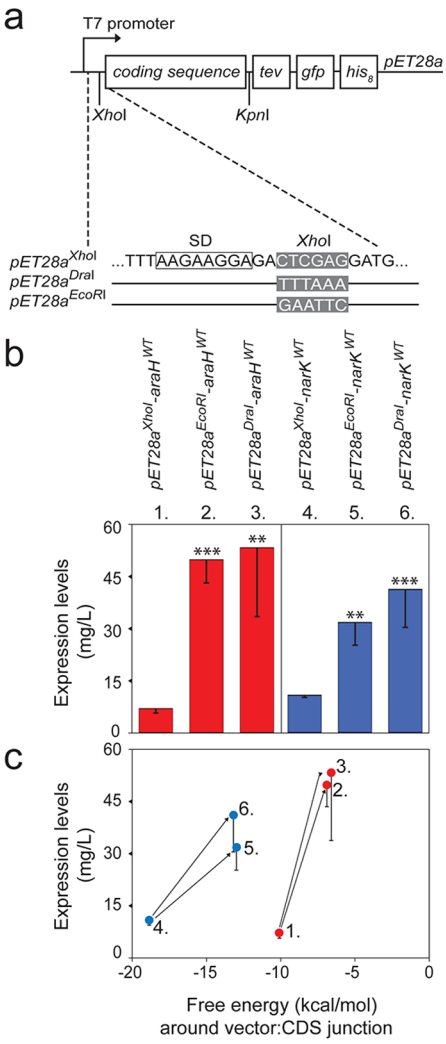


Figure 1. Nucleotide sequences used for cloning affect expression levels. (a) Overview of the expression cassette used in this study. CDSs were cloned into a derivative of the *pET28a* vector (called $pET28a^{XhoI}$ here and *pGFPe* elsewhere[32]) using *Xho*I and *Kpn*I restriction endonucleases. They were genetically fused to a region encoding the tobacco etch virus protease recognition sequence (TEV), green fluorescent protein (GFP), and an octa-histidine purification tag (His_8). In some experiments, the original *Xho*I site, located between the Shine−Dalgarno sequence (SD) and the AUG start codon, was changed to *Dra*I or *Eco*RI. (b) Comparison of expression levels for $araH^{WT}$ (red) and $narK^{WT}$ (blue) when expressed from the original vector ($pET28a^{XhoI}$) or vectors where the 5′ *Xho*I site was changed ($pET28a^{DraI}$ or $pET28a^{EcoRI}$). The constructs were transformed into the BL21(*DE3*)*pLysS* strain, and expression was induced with 1.0 mM IPTG for 5 h at 25 °C. To estimate the amount of protein produced (in mg/L), whole cell fluorescence was compared to a standard curve obtained with purified GFP. These estimates were not influenced by free GFP, as we only detected full-length fusion proteins when cellular extracts were analyzed by western blotting and in-gel fluorescence (Supporting Information Figure S2). Error bars represent the standard deviation from three biological replicates. Statistical significance was determined by an unpaired two-tailed Student's *t* test assuming unequal variance. Three stars indicate a probability of $P < 0.001$. Two stars indicate a probability of $P < 0.01$. (c) Prediction of mRNA stability around the AUG start codon (i.e., −20 to +37). The free energy (ΔG) associated with mRNA folding was calculated in kcal/mol using mFold[40] and plotted against the expression level. Numbers correspond to clones described in (b).

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

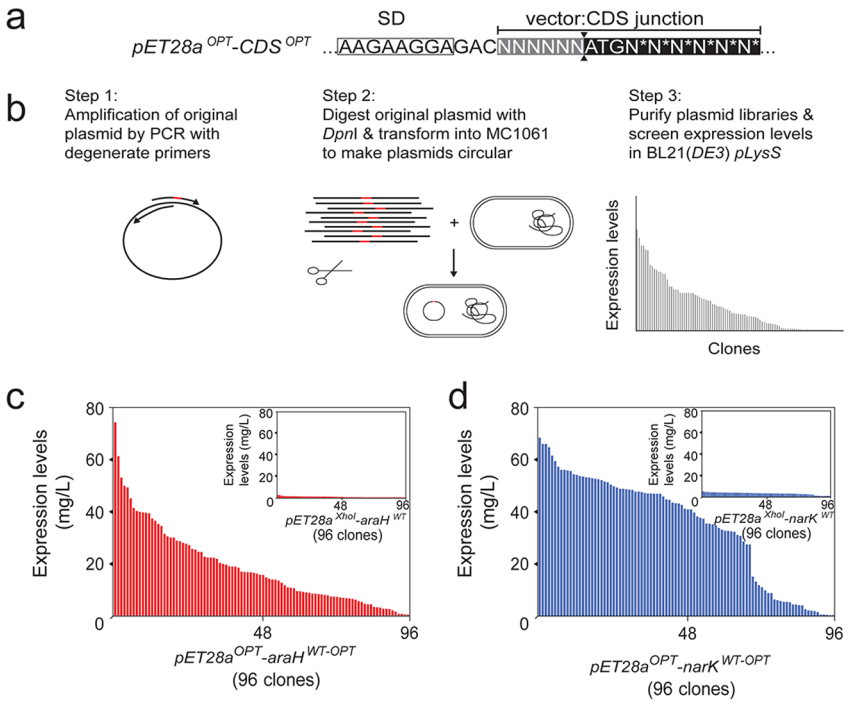


Figure 2. Optimization at the vector−coding sequence junction enables expression of CDSs that were previously thought to be difficult to express. (a) Overview of the random mutagenesis approach that generated libraries of vector−CDS junctions. Randomization of the six nucleotides upstream of the ATG allowed all possible nucleotides (denoted N), whereas the six nucleotides downstream were restricted to nucleotides that did not change the sequence of the encoded protein (denoted N*). (b) Overview of the workflow used to generate vector−coding sequence junctions. (c, d) Comparison of expression levels from 96 randomly selected clones in the $pET28a^{OPT}$-$araH^{WT\text{-}OPT}$ and $pET28a^{OPT}$-$narK^{WT\text{-}OPT}$ libraries (red and blue, respectively). The clones were assayed as described in Figure 1. The inset boxes show 96 randomly selected colonies of the original unoptimized clone (i.e., the mother plasmid).

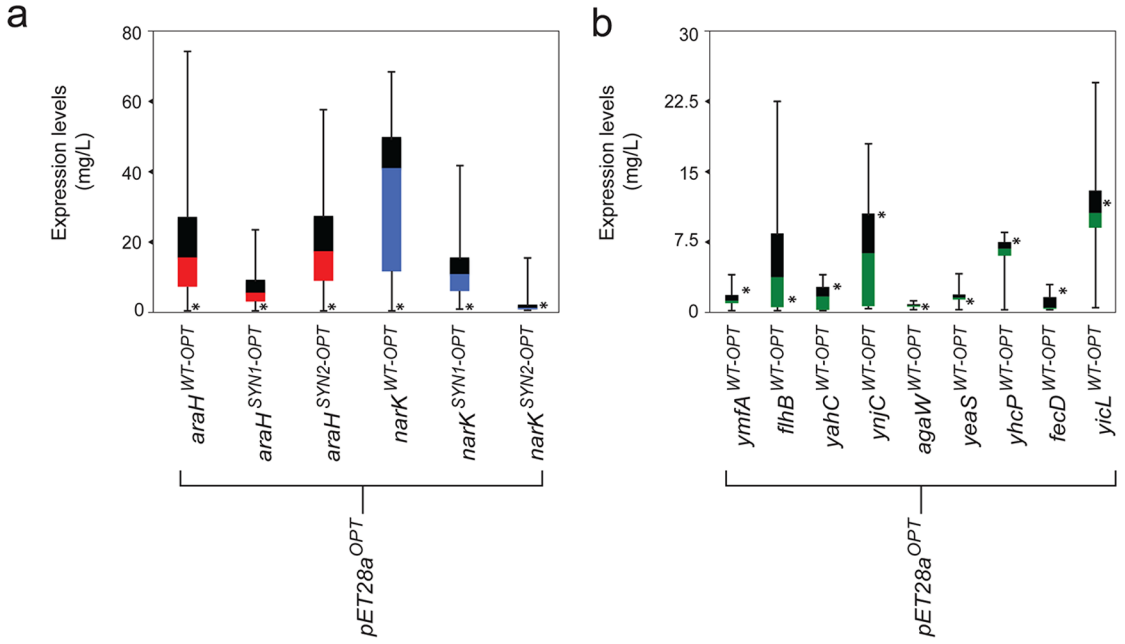


Figure 3. Optimization of the vector−CDS junction increases expression of a range of CDSs. (a) Box and whisker plot showing the differences in expression from 96 clones in the $pET28a^{OPT}$-$araH^{WT\text{-}OPT}$ and $pET28a^{OPT}$-$narK^{WT\text{-}OPT}$ libraries, as well as libraries generated with synthetic versions of each CDS that had been optimized by commercial vendors.[24] The top and bottom ends of the line represent the highest and lowest expression levels observed. Expression levels of the original unoptimized clones are marked with an asterisk (*). (b) Same as (a) except that the libraries were generated from nine additional CDSs from our previously synthesized library.[32] These CDSs all encode *E. coli* membrane proteins.

The generality of our optimization step was demonstrated using four synthetic CDSs (Figure 3a). These CDSs had been codon-optimized by two different commercial vendors but still did not express to high levels. We also carried out the optimization step on nine additional CDSs, chosen from our membrane protein-TEV-GFP-His library (Figure 3b). In all of these

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

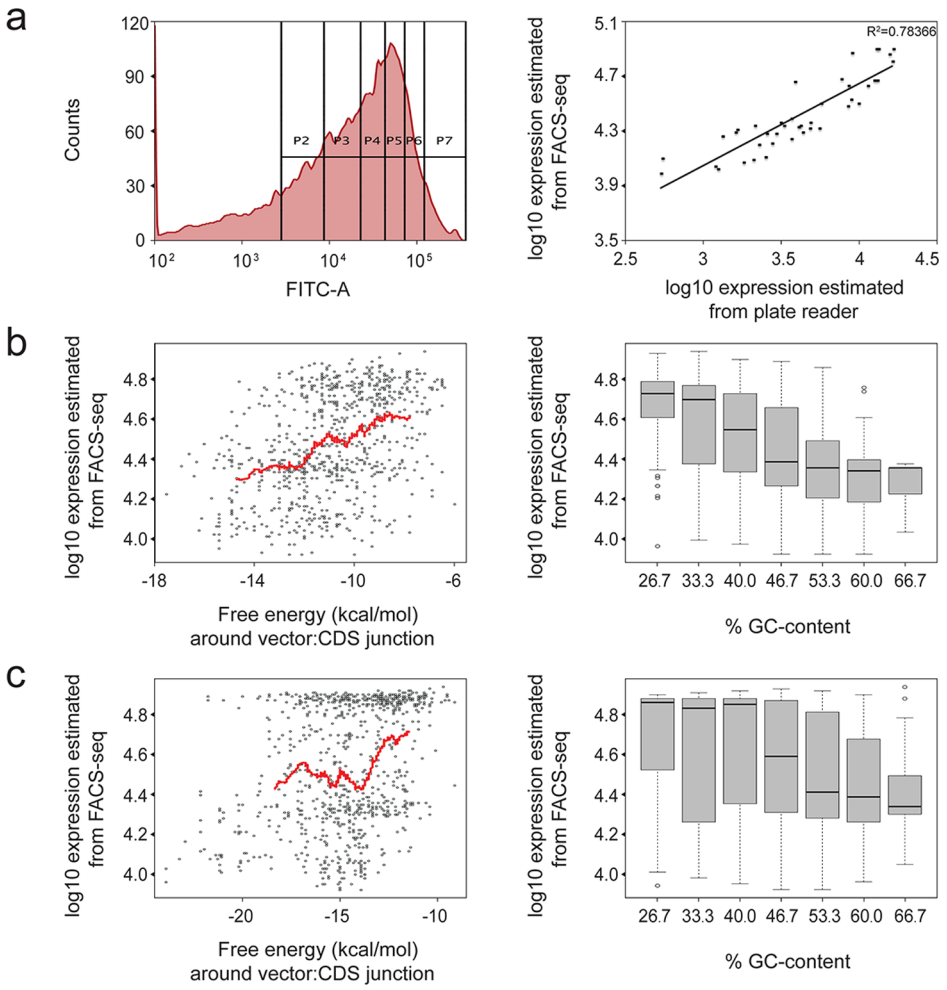


Figure 4. FACS-seq of vector−CDS junctions indicate that relaxed mRNA structure and low GC content are important, but not the only, determinants of expression. (a) Left panel, *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ and *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ cell libraries were mixed and sorted into six bins (P2−P7) by FACS. PCR amplicons covering the vector−CDS junction from each bin were then sequenced. Right panel, 41 colonies from the bins were randomly picked, and fluorescence was measured in a plate reader (Supporting Information Table S3). These values correlated with the expression values assigned from FACS-seq (R^2 = 0.78) and thus validated our approach. (b) Left panel, analysis of mRNA stability (Δ*G*) around the AUG start codon (i.e., −20 to +37) in the *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ library was calculated using mFold[40] and plotted against the expression level obtained by FACS-Seq. The red line is a running average over 100 samples. Right panel, GC-content in the *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ library was plotted against the expression level obtained by FACS-seq. (c) Same as (b) except that the *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ library was analyzed.

experiments, we observed a range of expression levels when we created libraries of vector−CDS junctions and assayed 96 clones. These differences ranged from 339-fold (*pET28*^{OPT}-*araH*$^{SYN2\text{-}OPT}$) to 7-fold (*pET28*^{OPT}-*agaW*^{OPT}). Significantly, the optimization step invariably enabled us to identify clones that expressed to a higher level than that of the original unoptimized clone (see *).

To identify vector−CDS junctions that gave rise to high expression, we analyzed nucleotide sequences in the *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ and *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ libraries. In this experiment, we used fluorescence activated cell sorting (FACS) to sort clones from the libraries into six bins according to the GFP expression level (Figure 4a, left panel). This analysis indicated that the difference between the lowest and the highest expressing clones in the libraries was on the order of 1000-fold. We then used deep sequencing to determine the presence and frequency of different vector−CDS junctions in the different bins. The sequences from each bin were PCR-amplified using mixtures of barcoded oligonucleotides. Barcoded libraries were then pooled and sequenced on the Illumina sequencing platform. This combined FACS and high-throughput sequencing approach is referred to as FACS-seq, and it allowed us to identify 13 914 (56% of possible total) and 25 861 (53% of possible total) unique vector−CDS junctions from each of the *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ and *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ libraries, respectively (Supporting Information Tables S1 and S2). For each identified sequence with more than 100 reads distributed across the bins in the FACS-seq data set (789 and 881 sequences for *araH* and *narK* respectively), we estimated the GFP expression level by computing the weighted average of bin mean GFP levels using normalized frequencies of the sequence across the bins as weights (see Methods). These estimates were validated by randomly assaying 41 clones across the six bins using a plate reader (Figure 4a, right panel).

Next, we sought to identify sequence characteristics in the vector−CDS junctions that contributed to low or high expression in the *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ and *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ libraries. We noted a trend between expression level in the FACS-seq experiment and mRNA stability (Δ*G*) around the

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

TIR when we analyzed the sequences with mFold (Figure 4b,c, left panels). We also observed a nucleotide bias in the sequences that gave high expression. Specifically, those vector−CDS junctions with a GC content below 40% gave higher levels of expression (Figure 4b,c, right panels). These observations are consistent with a large body of work suggesting that mRNA structure and GC content around the AUG start codon could affect translation initiation and therefore protein production.[13−17,29,30] However, it is important to note that many outliers were present, indicating that other variables contribute to high expression. One of these variables could be synonymous codon choice; however, our libraries sampled only a few amino acids and thus lack the diversity required to assess codon choice comprehensively.

In this study, we designed a one-step optimization approach that selectively modified the junction that is formed between a vector and the 5′ end of a coding sequence during cloning. In the optimization, we modified the six nucleotides upstream of the start codon, which is where cloning scars would normally be. We also modified synonymous codons in the second and third positions of the CDS as they have been shown to affect expression.[24−27] Thus, the modified region encompassed a large proportion of the TIR but not the SD sequence. When the libraries were screened, we noted a wide range of expression levels. In most experiments, we could identify clones with significantly higher expression than that of the original unoptimized clone.

The approach is simple and inexpensive, requiring only a single PCR that could be implemented during cloning or in a postcloning optimization step (as we have done here). We therefore believe that it could become an important part of the molecular biology toolbox, which can be used in conjunction with other tools for protein production, such as optimized genetic elements, synthetically designed genes, and strains selected for high expression. The downside to the approach is the need for screening methods to identify clones that lead to high expression. In this study, we used a translation fusion to a fluorescent protein; however, the reporter protein could also be translationally coupled (see ref 34).

On-going work aims to better understand the molecular code for designing vector−CDS junctions that result in high expression, as it would eliminate the need for screening. Data presented here from *circa* 1700 vector−CDS junctions (representing two CDSs) indicated that that low GC content and relaxed mRNA stability (ΔG) in this region were important, but not the only, variables to consider for high expression. Additional experiments that consider different synonymous and nonsynonymous codons in the 5′ end might provide this insight. There may also be as yet unknown parameters that need to be considered.

METHODS

Molecular Cloning. All CDSs were harbored in a vector derived from *pET28a* (called $pET28a^{XhoI}$ here and *pGFPe* elsewhere), as previously described.[24,32] They were genetically fused at the 3′ end to a sequence that encoded the tobacco-etch virus (TEV) protease recognition site, green fluorescent protein (GFP), and an octa-histidine purification tag (His_8). Site-directed mutagenesis was performed using the QuikChange site-directed mutagenesis kit (Stratagene), and constructs were verified by DNA sequencing (Eurofins MWG Operon, Germany).

Libraries of different vector−CDS junctions were amplified from the original clone using a reverse primer located in the $pET28a^{XhoI}$ vector (5′-TCTCCTTCTTAAAGTTAAACAAA-ATTATTCTAGAGGGGAATTGTTATCCG-3′) and a degenerate forward primer. The forward primer contained the following design principles: (1) The six nucleotides upstream of the AUG start codon were changed in all possible combinations. (2) The six nucleotides downstream of the AUG start codon were changed in combinations that allowed for all synonymous codon substitutions. (3) The flanking regions were 16−23 nucleotides. (4) The 5′ end of the forward primer matched the 5′ end of the reverse primer (15 nucleotides) so that the PCR products could circularize by homologous recombination when transformed into *E. coli*. Note that in some cases it was not possible to design a single forward primer, so two primers were used and PCR products were mixed. Amplification by PCR was carried out using Q5 polymerase (New England Biolabs) in a program that consisted of 95 °C for 2 min and then 30 cycles of 95 °C for 45 s, 48−68 °C for 45 s (using a gradient block), and 72 °C for 6.5 min. A final elongation step of 68 °C was then carried out for 12 min. Q5 polymerase is a high-fidelity polymerase whose error rate is so low that it is difficult to measure in a statistically significant manner (see manufacturer specifications); thus, we do not believe that PCR errors in the coding sequence or vector contribute to differences in expression levels. Following the PCR, 40 units of *Dpn*I were added to 20 μL of the PCR product, and it was then transformed into the *E. coli* strain MC1061 (to facilitate circularization). The transformation was transferred to 100 mL of Luria−Bertani (LB) media containing 50 μg/mL kanamycin and 34 μg/mL chloramphenicol and incubated at 37 °C with shaking for 16 h. Purification of the plasmid library was carried out using the E.Z.N.A. DNA midi kit (Omega Biotek). To ensure that there was significant diversity in the library, we plated a small aliquot and sequenced the DNA isolated from five random colonies. In every case, we received five different sequences.

Expression. Plasmids and plasmid libraries were transformed into BL21(*DE3*)*pLysS* using standard protocols. Overnight cultures were prepared by inoculating a single colony in 800 μL of LB liquid media containing 50 μg/mL kanamycin and 34 μg/mL chloramphenicol. Cultures were incubated in a 2.2 mL 96-well plate at 37 °C with shaking for 16 h. Cultures were then back-diluted (1:50) into 5 mL of LB plus antibiotics in a 24-well growth plate and incubated as before until an OD_{600} of approximately 0.3 was reached. Expression was induced by addition of 1.0 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and incubation for 5 h at 25 °C with shaking. The OD_{600} was measured, and cells were harvested by centrifugation at 3220*g* for 10 min, resuspended in buffer [50 mM Tris-HCl, pH 8.0, 200 mM NaCl, 15 mM EDTA], and transferred to a 96-well optical bottom plate. Fluorescence was read in a Spectramax Gemini (Molecular Devices) at an excitation wavelength of 485 nm and an emission wavelength of 513 nm. The amount of GFP produced (in mg/L) was calculated using a standard curve obtained from purified GFP mixed with whole cells (to account for quenching). When the $pET28a^{OPT}$-$araH^{WT\text{-}OPT}$ and $pET28a^{OPT}$-$narK^{WT\text{-}OPT}$ libraries were assayed, the five most highly expressed and five most poorly expressed clones were sequenced so that we could ensure that there was diversity in the experiment. In every case, we received 10 different sequences.

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

In-Gel Fluorescence and Western Blotting. A volume of cells corresponding to an OD_{600} of 0.2 units was collected, resuspended in Laemlli loading buffer, and analyzed by 12% SDS-PAGE. Fluorescence emitting from the SDS-PAGE was detected using a LAS-1000 (Fuji Film). The gel was then immediately transferred to a nitrocellulose membrane using a semidry transfer-blot apparatus (Bio-Rad) and probed with an antibody against purified GFP. Antibody binding was detected using an anti-rabbit IgG horseradish peroxidase-linked whole antibody (from donkey) and a SuperSignal West femto luminol/enhancer solution (Thermo Scientific).

Cell Sorting of Libraries with FACS. *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ and *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ plasmid libraries were transformed into BL21(*DE3*)*pLysS*, and protein expression was induced as described above. Subsequently, cell libraries were mixed and sorted by FACS into different bins according to their level of fluorescence. Six gates were defined to sort the libraries, and a total of 10^4 cells were sorted in each gate with precision single-cell mode at a rate of approximately 1500 events per second. The collected cells were grown overnight in 2 mL of LB liquid medium containing 50 μg/mL kanamycin. For each of the sorted pools, 1 mL of cell culture was stored at −80 °C and 1 mL was harvested for plasmid DNA preparation.

High-Throughput Sequencing and Analysis. The plasmids in the collected cells were isolated and used for PCR amplification, as previously described.[36] Specifically, barcoded primers containing overhang sequences compatible with Illumina Nextera XT indexing (Supporting Information Table S4) were used to amplify 185 base pair DNA fragments containing the vector−CDS junction regions for each of the sorted pools. Amplification by PCR was carried out using Phusion hot start II HF Pfu polymerase (Thermo Fisher Scientific) in a program that consisted of 98 °C for 30 s and then 25 cycles of 98 °C for 10 s, 50−60 °C for 30 s (using a gradient block), and 72 °C for 30 s, with a final elongation step at 72 °C for 10 min. Following PCR amplification, amplicons were purified using AMPure XP beads (Beckman Coulter, CA) and pooled in equal quantities to a total amount of 1 μg. Sequencing adapters were integrated into the amplicons by PCR (98 °C for 3 min and then 25 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, with a final elongation step at 72 °C for 5 min) using Illumina Nextera XT Index (Illumina no. FC-131-1001). Deep sequencing was performed on a MiSeq benchtop sequencer (Illumina, San Diego, CA) using 151 bp paired-end reads.

Illumina sequencing data was analyzed by first merging the paired end reads using Flash[37] (minimum overlap 40 bp; maximum overlap 150 bp) and then extracting the relevant variable sequence fragment using qiime[38] (removing sequences <120 bp and allowing for 2 mismatches in the left-hand flanking region and one mismatch in the right-hand flanking region). Finally, sequences with indels (i.e., length larger or smaller than the desired 15 bp) were filtered out using seqtk (https://github.com/lh3/seqtk). The raw counts of distinct 15 bp sequences in each FACS bin were determined using a custom Python script available upon request from the authors. In order to estimate the GFP expression level for each 15 bp sequence, we used the method first introduced by Sharon et al.[39] This method computes the weighted mean expression level for each sequence s (f_s) using the formula $f_s = (\sum_b n_{b,s}/n_b \times e_b)/(\sum_b n_{b,s}/n_b)$, where e_b is the mean expression level of FACS bin b, n_b is the total count of all sequences in bin b, and $n_{b,s}$ is the count of specific sequence s in bin b. For further analysis, only sequences with more than 100 merged reads distributed across the FACS bins were used.

From each of the sorted bins, 5−9 individual colonies were isolated, and the region of interest was Sanger-sequenced. Finally, a total of 41 different plasmids were identified and transformed into BL21(*DE3*)*pLysS* to analyze protein expression. The sequences of the vector−CDS junctions in individual isolates are shown in Supporting Information Table S3.

Free Energy Measurements. The free energy (ΔG) associated with mRNA folding was calculated in kcal/mol using mFold[40] with default settings in a window from −20 to +37 relative to the A of the AUG start codon.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1: Comparison of expression levels for 397 coding sequences (CDSs) of *E. coli* inner membrane proteins. Figure S2: Estimates of protein production levels are not influenced by free-GFP. Figure S3: Synonymous codon substitutions in the 5′ coding sequence (CDS) of *araH*^{WT} and *narK*^{WT} affect expression. Table S1: FACS-Seq data obtained from *pET28a*^{OPT}-*araH*$^{WT\text{-}OPT}$ library. Table S2: FACS-Seq data obtained from *pET28a*^{OPT}-*narK*$^{WT\text{-}OPT}$ library. Table S3: Vector−CDS junctions of individual isolates from the FACS bins. Table S4: Oligonucleotides used in this study. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.5b00033.

■ AUTHOR INFORMATION

Corresponding Authors

*(M.H.H.N.) Tel: +45 45 25 80 26. E-mail: morno@biosustain.dtu.dk.

*(D.O.D.) Tel: +46 8 162 910. E-mail: ddaley@dbb.su.se.

Author Contributions

K.M., S.T., S.G., V.M., M.H.H.N., and D.O.D. designed the research; K.M., S.T., S.G., V.M., M.J.H., and A.E. performed the research; M.H.H.N. and D.O.D. supervised the research; and K.M. and D.O.D. wrote the paper.

Notes

The authors declare the following competing financial interest(s): Work presented in this publication is considered under a patent submitted by CloneOpt AB (SE1451553-0). K.M., S.T., and D.O.D. are founders and shareholders in CloneOpt AB.

■ ACKNOWLEDGMENTS

This work was supported by a grant from the Swedish Research Council to D.O.D. and A.E. and by the Novo Nordisk Foundation to V.M., M.J.H., and M.H.H.N. Anna Koza is thanked for technical assistance.

■ REFERENCES

(1) Sorensen, H. P., and Mortensen, K. K. (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. Biotechnol. 115*, 113−128.

(2) Rosano, G. L., and Ceccarelli, E. A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol. 5*, 172.

(3) McCarthy, J. E., and Gualerzi, C. (1990) Translational control of prokaryotic gene expression. *Trends Genet. 6*, 78−85.

(4) Laursen, B. S., Sorensen, H. P., Mortensen, K. K., and Sperling-Petersen, H. U. (2005) Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev. 69*, 101−123.

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965

(5) Egbert, R. G., and Klavins, E. (2012) Fine-tuning gene networks using simple sequence repeats. *Proc. Natl. Acad. Sci. U.S.A. 109*, 16817−16822.
(6) Gold, L. (1988) Posttranscriptional regulatory mechanisms in *Escherichia coli*. *Annu. Rev. Biochem. 57*, 199−233.
(7) Shine, J., and Dalgarno, L. (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature 254*, 34−38.
(8) Gualerzi, C. O., and Pon, C. L. (1990) Initiation of mRNA translation in prokaryotes. *Biochemistry 29*, 5881−5889.
(9) Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene 361*, 13−37.
(10) Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods 10*, 354−360.
(11) Hui, A., Hayflick, J., Dinkelspiel, K., and de Boer, H. A. (1984) Mutagenesis of the three bases preceding the start codon of the beta-galactosidase mRNA and its effect on translation in *Escherichia coli*. *EMBO J. 3*, 623−629.
(12) Liebeton, K., Lengefeld, J., and Eck, J. (2014) The nucleotide composition of the spacer sequence influences the expression yield of heterologously expressed genes in *Bacillus subtilis*. *J. Biotechnol. 191*, 214−220.
(13) Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Bluthgen, N. (2013) Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol. 9*, 675.
(14) Hall, M. N., Gabay, J., Debarbouille, M., and Schwartz, M. (1982) A role for mRNA secondary structure in the control of translation initiation. *Nature 295*, 616−618.
(15) Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science 324*, 255−258.
(16) Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet. 15*, 469−479.
(17) Plotkin, J. B., and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet. 12*, 32−42.
(18) Goodman, D. B., Church, G. M., and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science 342*, 475−479.
(19) Care, S., Bignon, C., Pelissier, M. C., Blanc, E., Canard, B., and Coutard, B. (2008) The translation of recombinant proteins in *E. coli* can be improved by in silico generating and screening random libraries of a −70/+96 mRNA region with respect to the translation initiation codon. *Nucleic Acids Res. 36*, e6.
(20) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946−950.
(21) Na, D., Lee, S., and Lee, D. (2010) Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst. Biol. 4*, 71.
(22) Seo, S. W., Yang, J. S., Kim, I., Yang, J., Min, B. E., Kim, S., and Jung, G. Y. (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng. 15*, 67−74.
(23) Reeve, B., Hargest, T., Gilbert, C., and Ellis, T. (2014) Predicting translation initiation rates for designing synthetic biology. *Front. Bioeng. Biotechnol 2*, 1.
(24) Norholm, M. H., Toddo, S., Virkki, M. T., Light, S., von Heijne, G., and Daley, D. O. (2013) Improved production of membrane proteins in *Escherichia coli* by selective codon substitutions. *FEBS Lett. 587*, 2352−2358.
(25) Bivona, L., Zou, Z., Stutzman, N., and Sun, P. D. (2010) Influence of the second amino acid on recombinant protein expression. *Protein Expression Purif. 74*, 248−256.
(26) Looman, A. C., Bodlaender, J., Comstock, L. J., Eaton, D., Jhurani, P., de Boer, H. A., and van Knippenberg, P. H. (1987) Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *Escherichia coli*. *EMBO J. 6*, 2489−2492.
(27) Stenstrom, C. M., Jin, H., Major, L. L., Tate, W. P., and Isaksson, L. A. (2001) Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene 263*, 273−284.
(28) Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell 141*, 344−354.
(29) Tuller, T., and Zur, H. (2015) Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res. 43*, 13−28.
(30) Allert, M., Cox, J. C., and Hellinga, H. W. (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames. *J. Mol. Biol. 402*, 905−918.
(31) Cheong, D. E., Ko, K. C., Han, Y., Jeon, H. G., Sung, B. H., Kim, G. J., Choi, J. H., and Song, J. J. (2014) Enhancing functional expression of heterologous proteins through random substitution of genetic codes in the 5′ coding region. *Biotechnol. Bioeng. 112*, 822−826.
(32) Daley, D. O., Rapp, M., Granseth, E., Melen, K., Drew, D., and von Heijne, G. (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science 308*, 1321−1323.
(33) Mutalik, V. K., Guimaraes, J. C., Cambray, G., Mai, Q. A., Christoffersen, M. J., Martin, L., Yu, A., Lam, C., Rodriguez, C., Bennett, G., Keasling, J. D., Endy, D., and Arkin, A. P. (2013) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods 10*, 347−353.
(34) Mendez-Perez, D., Gunasekaran, S., Orler, V. J., and Pfleger, B. F. (2012) A translation-coupling DNA cassette for monitoring protein translation in *Escherichia coli*. *Metab. Eng. 14*, 298−305.
(35) Marino, J., Hohl, M., Seeger, M. A., Zerbe, O., and Geertsma, E. R. (2015) Bicistronic mRNAs to enhance membrane protein overexpression. *J. Mol. Biol. 427*, 943−954.
(36) Lee, J. S., Kallehauge, T. B., Pedersen, L. E., and Kildegaard, H. F. (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep. 5*, 8572.
(37) Magoc, T., and Salzberg, S. L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics 27*, 2957−2963.
(38) Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods 7*, 335−336.
(39) Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol. 30*, 521−530.
(40) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406−3415.

DOI: 10.1021/acssynbio.5b00033
ACS Synth. Biol. 2015, 4, 959−965