Published online 27 June 2015

Nucleic Acids Research, 2015, Vol. 43, No. 14 ***7137–7151***
doi: 10.1093/nar/gkv635

A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons

Tian Tian[1] and Howard M. Salis[1,2,*]

[1]Department of Biological Engineering, Pennsylvania State University, University Park, PA 16802, USA and
[2]Department of Chemical Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received January 10, 2015; Revised June 02, 2015; Accepted June 08, 2015

ABSTRACT

Natural and engineered genetic systems require the coordinated expression of proteins. In bacteria, translational coupling provides a genetically encoded mechanism to control expression level ratios within multi-cistronic operons. We have developed a sequence-to-function biophysical model of translational coupling to predict expression level ratios in natural operons and to design synthetic operons with desired expression level ratios. To quantitatively measure ribosome re-initiation rates, we designed and characterized 22 bi-cistronic operon variants with systematically modified intergenic distances and upstream translation rates. We then derived a thermodynamic free energy model to calculate *de novo* initiation rates as a result of ribosome-assisted unfolding of intergenic RNA structures. The complete biophysical model has only five free parameters, but was able to accurately predict downstream translation rates for 120 synthetic bi-cistronic and tri-cistronic operons with rationally designed intergenic regions and systematically increased upstream translation rates. The biophysical model also accurately predicted the translation rates of the nine protein *atp* operon, compared to ribosome profiling measurements. Altogether, the biophysical model quantitatively predicts how translational coupling controls protein expression levels in synthetic and natural bacterial operons, providing a deeper understanding of an important post-transcriptional regulatory mechanism and offering the ability to rationally engineer operons with desired behaviors.

INTRODUCTION

Engineering a genetic system often requires the coordinated expression of its proteins; for example, to express multi-subunit recombinant biologics, engineer genetic circuits to control cellular processes, catalyze bioconversions with multi-enzyme pathways, or assemble multi-protein organelles (1–5). When expression level ratios undergo transient or permanent changes, a genetic system's behavior can deviate from its best possible performance; active biologic titers will decrease, genetic circuits will incorrectly process a signal, and a metabolic pathway's activity will be reduced.

Through genetic modifications, several approaches have been developed to control and coordinate protein expression levels. In eukaryotes, equimolar expression of proteins can be achieved by fusing subunit coding sequences together with a self-cleaving 2A peptide linker (6). In prokaryotes, by incorporating different combinations of promoters, the expression levels of two operons were varied to increase taxadiene production titers (7). Libraries of structured ribosome binding sites and RNAse binding sites were combinatorially inserted into the intergenic regions of 2- and 3-protein operons to improve amorphadiene biosynthesis (8). Combinations of plasmid origins of replication and ribosome binding sites were used to vary the expression level ratios of 15 genes grouped into three modules to increase fatty acid production (9). Several *cis*-acting, self-cleaving ribozymes were characterized and subsequently introduced into the 5′ UTRs of engineered genetic circuits to insulate their input-output transfer functions from promoter changes (10). The endoRNAse Csy4 from the *Pseudomonas aeruginosa* CRISPR system was employed to cleave mRNAs at a 28 nucleotide recognition sequence, separating 2-protein operons into mono-cistronic mRNAs that exhibited improved consistency in expression (11). Using the RBS Library Calculator, the multi-dimensional translation rate space of a 3-enzyme terpenoid biosynthesis pathway was characterized, mapped, and modeled to identify optimally balanced metabolic pathways (3). Using the same approach, a 5-enzyme synthetic Entner-Doudoroff pathway was rationally designed and systematically optimized, resulting in a 25-fold higher NADPH regeneration rate in *Escherichia coli* (5).

*To whom correspondence should be addressed. Tel: +1 814 865 1931; Fax: +1 814 863 1031; Email: salis@psu.edu

Here, we employ the mechanism of translational coupling to coordinate the expression of multiple proteins in bacterial operons. We develop a biophysical model to predict expression level ratios in natural operons and to design synthetic operons that utilize translational coupling to achieve desired expression level ratios.

Translational coupling evolved in prokaryotic operons to satisfy several objectives that remain relevant when engineering genetic systems (12). First, using a single promoter, the expression of multiple genes in an operon can be regulated in response to changing metabolite levels or environmental conditions (13). By combining such co-regulation with metabolite-responsive promoters, recent studies have shown that autonomous control of multi-enzyme pathways can substantially increase their activity by eliminating transient imbalances in flux and the accumulation of toxic intermediates (14,15). Second, most operons have very short intergenic regions between their protein coding sequences to reduce the number of binding sites for endoribonucleases and therefore increase their mRNA transcript's stability (16). 50% of operons in *E. coli* MG1655 have intergenic distances of 10 nucleotides or less and 34% of operons contain negative intergenic distances where upstream and downstream protein coding sequences overlap (17). Third, many operons have conserved gene orders, due to the need to sequentially express functionally related proteins, and to direct their co-translational folding and assembly into multi-subunit protein complexes (18,19). It is within these constraints that translational coupling has become a ubiquitous mechanism enabling prokaryotes to control the expression level ratios of individual proteins within operons affecting wide range of biological processes, including housekeeping metabolism (20–25), natural product biosynthesis (26), nitrogen fixation (27), chemotactic signaling (28), post-translational protein modification (29) and both Sec-mediated and Type III protein secretion (4,30).

Translational coupling occurs when the translation rate of an upstream protein coding sequence (CDS) controls the translation rate of a downstream protein coding sequence through ribosome-mRNA interactions at intergenic regions, particularly when operons have adjacent or overlapping CDSs, where the ribosome binding site for the downstream CDS is located within upstream CDS. The magnitude of translational coupling is controlled by two separate mechanisms. First, ribosomes terminating translation of the upstream CDS can dissociate and re-initiate translation of the downstream CDS at a certain rate, called ribosome re-initiation (31). Second, ribosomes elongating along the upstream CDS will unfold mRNA structures, and the absence of these mRNA structures can increase the translation of the downstream CDS, called *de novo* initiation (24). Notably, elongating ribosomes hydrolyze GTP to ratchet forward along the mRNA, and therefore have additional energy to unfold mRNA structures (32). In contrast, prior to translation initiation, an assembling 30S ribosome does not have a source of external energy to unfold mRNA structures.

The mechanism of translational coupling has been utilized to control protein expression levels in operons. Selected lactococcal genes were translationally coupled to *E. coli lacZ* to control their expression in *Lactococcus lactis* (33). A long, multi-domain olefin megasynthase was translationally coupled to a reporter protein or antibiotic resistance marker as a way to monitor and improve its translation elongation via synonymous codon mutations (34). A collection of promoters and translational coupling cassettes were also shown to improve the rank-ordering of expression levels for a variety of protein coding sequences (35). Finally, translational coupling between reporter proteins was quantitatively perturbed and characterized by varying ribosome binding sites and intergenic distances between 31 and 850 nucleotides (36). Currently, while sequence features that play an important role in translational coupling have been noted, a mechanistic model that yields testable and quantitative predictions has not been proposed or experimentally validated.

In this work, we develop a physics-based, quantitative model of translational coupling that predicts translation rates according to an operon's mRNA sequence. Our biophysical model accounts for both ribosome re-initiation and *de novo* initiation, including the effects of changing upstream translation rates, intergenic distances, and overlapping or intergenic sequences. We systematically measure ribosome re-initiation rates by constructing and characterizing 22 bi-cistronic operon variants. We then constructed 120 bi-cistronic and tri-cistronic operon variants to critically test model predictions, comparing measured and predicted downstream CDS translation rates while systematically increasing upstream CDS translation rates across a $>10\,000$-fold range. We show that the model can predict the extent of translational coupling in natural bacterial operons and we illustrate how to use the model to rationally design intergenic regions to genetically hard-code desired expression level ratios.

MATERIALS AND METHODS

Strains, media and cloning

Escherichia coli strains were cultured using Luria-Bertani (LB) media (10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl), M9 minimal media (6 g/l Na_2HPO_4, 3 g/l KH_2PO_4, 0.5 g/l NaCl, 1 g/l NH_4Cl, 0.24 g/l $MgSO_4$, 0.011 g/l $CaCl_2$, 0.05 g/l leucine, pH 7.4) with 0.4% glucose, or SOC media (20 g/l tryptone, 5 g/l yeast extract, 2.4 g/l $MgSO_4$, 0.58 g/l NaCl, 0.19 g/l KCl, 3.6 g/l glucose) as described. 0.25 μg/ml biotin was added to the M9 minimal media when culturing *E. coli* EcNR1 strain (37). As described, media was supplemented with 50 μg/ml ampicillin or chloramphenicol for selections.

To construct the initial bi-cistronic and tri-cistronic operons, genetic systems were assembled from PCR-amplified parts using extension primers, followed by DNA assembly using the chew back, anneal and repair method (46). The constructed bi-cistronic operons contained a σ^{70} constitutive promoter (BioBrick J23100), a synthetic RBS sequence, a codon-optimized mRFP1 fluorescent protein coding sequence, an intergenic sequence, a codon-optimized GFPmut3b fluorescent protein coding sequence, and an efficient transcriptional terminator (BioBrick B1006). Upstream RBS sequences were replaced by digestion with XbaI and SacI, followed by ligation using annealed

oligonucleotides with complementary overhangs. mRFP1-GFPmut3b intergenic regions were replaced by digestion with XhoI and EcoRI, followed by ligation using annealed oligonucleotides with complementary overhangs. We constructed tri-cistronic operons by inserting a codon-optimized Cerulean CFP coding sequence between the J23100 promoter and mRFP1 RBS of selected bi-cistronic operons by digestion with AatII and XbaI, followed by digestion and ligation of amplified PCR product with complementary ends. All bi-cistronic and tri-cistronic operons were carried on an ColE1 vector with a chloramphenicol resistance marker. All cloned operons were verified by sequencing.

Growth and fluorescence measurements

Growth and fluorescence measurements were performed in 96-well high-throughput format. A deep 96-well plate containing 750 μl LB and 50 μg/ml chloramphenicol was inoculated from single colonies and grown overnight at 37°C with 200 rpm. orbital shaking. A fresh 96-well microtiter containing 200 μl supplemented M9 minimal media was incubated at 37°C in a spectrophotometer (Tecan M1000) with high orbital shaking using a 1:80 dilution. OD_{600} measurements were recorded every 10 min. Once a culture reached an OD_{600} of 0.15–0.20 (3–4 h), a sample of each culture was transferred to a new microtiter plate containing 200 μl PBS and 2 mg/ml kanamycin (CalBioChem). This media replacement strategy was repeated twice more using fresh, pre-warmed plates containing supplementary minimal media (a 1:40 dilution requiring 8–10 h of growth). At least three samples were taken for each culture. The fluorescence distribution of each sample was measured by a flow cytometry using blue (488 nm) and green (532 nm) lasers (BD Fortessa). Fluorescence detectors were chosen to obtain <1% spectral overlap between CFP, mRFP1, and GFPmut3b fluorescence spectrums. The fluorescence distribution of each sample was measured with flow cytometer. The arithmetic average of each distribution was taken and the background autofluorescence of wild-type DH10B cells was subtracted from each sample. This procedure was repeated at least twice for each construct.

RNA extraction and RT-qPCR measurements

Selected transformed strains were inoculated from single colony streaks into 5 ml LB culture containing 50 μg/ml chloramphenicol and incubated overnight at 37°C. Cultures were then 1:10 diluted into 20 ml supplemented minimal media and incubated at 37°C until their OD_{600} reached 0.5, followed by removing 2 ml of culture to perform RNA extraction. RNA was isolated from cells using the Total RNA Purification Kit (NORGEN #17200) as per the manufacturer's protocol. To remove DNA contamination, the extracted RNA was DNase-treated using the TURBO DNA-free Kit (Ambion). 0.5 μg RNA per sample was used as template for complementary DNA (cDNA) synthesis using the ABI High Capacity RT Kit (PN #4368813). Separate, custom TaqMan probes were designed and ordered (Life Technologies) to bind to the middles of the GFPmut3b and mRFP1 coding sequences. RT-qPCR was then performed in triplicate in a 25 μl total reaction volume, containing 12.5 μl of TaqMan 2X Universial Mix (PN# 4324018) and 5 μl of the cDNA template, using an ABI 7300 RT-qPCR thermocycler. Simultaneously, RT-qPCR was used to measure the cultures' 16S rRNA levels, which were used an endogenous internal control to normalize the mRNA levels of GFPmut3b and mRFP1.

Statistical analysis and data sources

Squared Pearson correlation coefficients were calculated (R^2) to determine the linear correlation between model predictions and experimental measurements. Two-tailed *T*-tests were calculated (*P* values) to determine the statistical significance of these comparisons. All comparisons were statistically significant with a less than 5% probability of a random correlation ($P < 0.05$). Ribosome profiling measurement data was obtained from the Supplementary Information of (38), including the number of mapped RNA-seq reads per gene (RPKM units) and the number of mapped ribosome-bound reads per gene (arbitrary units). The number of bound ribosomes per unit transcript is their ratio (arbitrary units).

RESULTS

A mechanistic model of translational coupling in bacterial operons

Consider a multi-cistronic bacterial operon containing a promoter, a 5′ untranslated region (UTR), and at least two protein coding sequences (CDSs), separated by intergenic regions, followed by a transcriptional terminator (Figure 1A). The complete biophysical model calculates the translation rates for all CDSs within an operon, according to the sequences of the 5′ UTR, the CDSs and the intergenic regions. The ribosome's interactions with a mRNA of arbitrary sequence are quantified according to a previously developed, multi-term free energy model (39–41). Here, we assume that each CDS has been sufficiently codon-optimized so that their translation elongation rates are high, and therefore their translation initiation rates are the rate-limiting step in the overall translation process. The model has five unknown coefficients that we determine by measuring the expression levels of rationally designed operon variants. The model also does not account for changes in transcription rate or sequence elements that could affect mRNA stability, and therefore there is a proportionality constant that relates translation initiation rates to protein expression levels.

The translation initiation rate of the first CDS is calculated according to $r_1 \propto \exp(-\beta \Delta G_{\text{total}})$, where ΔG_{total} is the 30S ribosomal subunit's total binding free energy to the 5′ UTR. Using a superscript to denote the first CDS, this total binding free energy is determined according to the model

$$\Delta G^{(1)}_{\text{total}} = \Delta G_{\text{mRNA:rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}} \quad (1)$$

which quantifies the energy needed to unfold mRNA structures that overlap with the ribosome's footprint ($\Delta G_{\text{mRNA}} < 0$); the energy released when the ribosome's 16S rRNA binds to the mRNA and when alternative non-inhibitory mRNA

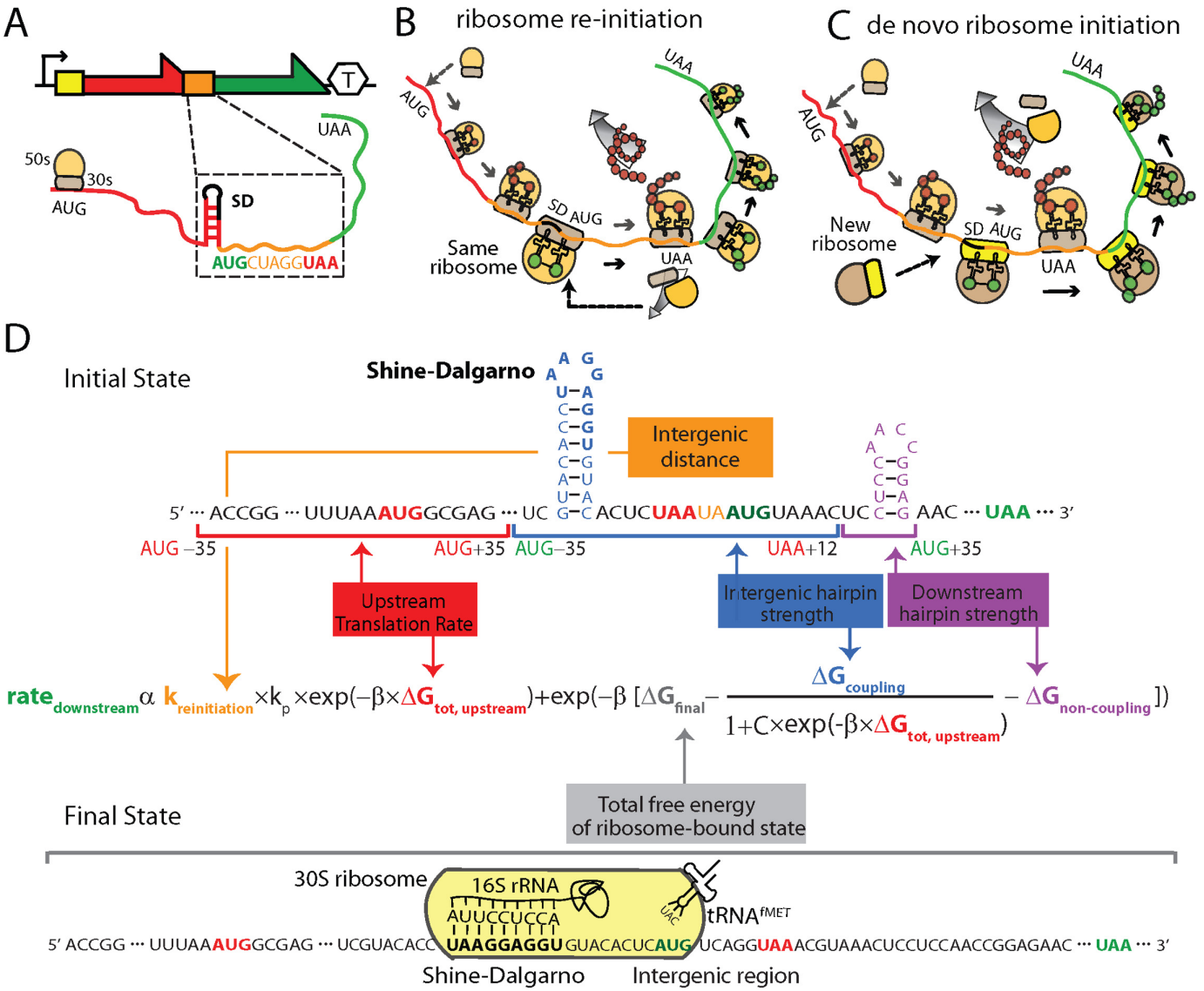

Figure 1. A Mechanistic Model of Translational Coupling. (**A**) Translation of a bi-cistronic operon. The schematic shows the (black arrow) promoter, the (yellow box) 5′ UTR, the (red) upstream coding sequence, the (orange box) intergenic region, the (green) downstream coding sequence, and the (hexagon) transcriptional terminator. The intergenic region contains an inhibitory RNA structure, a Shine-Dalgarno (SD) sequence, and overlapping coding sequences. (**B**) Ribosome re-initiation occurs when an upstream elongating ribosome dissociates and reassembles at an intergenic region, followed by translation initiation of the downstream coding sequence. (**C**) Ribosome de novo initiation occurs when cytosolic ribosomes assemble onto the mRNA at intergenic regions and initiate translation of the downstream coding sequence. (**D**) The proposed biophysical model of translational coupling quantifies the molecular interactions that control both ribosome re-initiation and de novo initiation rates according to an inputted mRNA sequence. Model components and their corresponding sequence region are color coded, including the free energies of inhibitory RNA structures (coupled and non-coupled), the intergenic distance, the free energy of the ribosome-bound state, and the upstream coding sequence's translation rate.

structures refold after ribosome binding ($\Delta G_{\mathrm{mRNA-rRNA}} < 0$); the energy released when the initiator $\mathrm{tRNA}^{\mathrm{fMet}}$ base pairs to the start codon ($\Delta G_{\mathrm{start}} < 0$); an energetic penalty for ribosomal stretching or compression caused by a long or short spacer region between the 16S rRNA binding site and start codon ($\Delta G_{\mathrm{spacing}} > 0$); and an energy penalty determined by the mRNA standby site's interactions with the ribosome's platform domain (41) ($\Delta G_{\mathrm{standby}} > 0$). The free energy of the final state, $\Delta G_{\mathrm{final}}$, is the sum of the first four energy terms. Using only the mRNA's sequence as an input, these free energies are calculated using a semi-empirical model of RNA interactions (42), RNA folding algorithms (43) and previous empirical measurements of the ribosome's flexibility and standby site interactions (39). The apparent Boltzmann factor β has been measured in four studies to have a consistent value of 0.45 ± 0.05 mol/kcal (3,39,41,44).

The translation initiation rate of the second CDS is then calculated by summing together two sources of translational coupling: ribosome re-initiation (Figure 1B) and ribosome de novo initiation (Figure 1C) according to

$$r_2 \propto r_{\mathrm{reinitiation}}^{(2)} + \exp\left(-\beta \Delta G_{\mathrm{total}}^{(2)}\right) \qquad (2)$$

In our model, we propose that the ribosome re-initiation rate depends proportionally on the upstream CDS' translation rate as well as the relative probability of ribosome re-assembly, which depends on the intergenic distance. This relationship can be expressed as

$$r^{(2)}_{\text{reinitiation}} = k_P \; k_{\text{reinitiation}}\,(d_{1,2}) \exp\left(-\beta \Delta G^{(1)}_{\text{total}}\right) \qquad (3)$$

where the coefficient $k_{\text{reinitiation}}$ quantifies the intergenic distance dependence and k_{P} the proportionality constant between the ribosome assembly rate and the translation initiation rate. The intergenic distances are calculated according to $d_{ij} = x_{\text{start}}$ - $x_{\text{stop}} - 3$, where x_{start} and x_{stop} are the positions of the first nucleotides in the jth CDS's start codon and the ith CDS's stop codon, respectively. Intergenic distances are negative when out-of-frame protein coding sequences overlap; for example, the intergenic sequence 5′-AUGAUCGAUAA-3′ has a distance $d = -11$. The most common overlapping intergenic sequence is 5′-AUGA-3′ with a distance $d = -4$. In the following sections, we measured how the intergenic distance controls $k_{\text{reinitiation}}$, and test if intergenic sequence has a substantial effect on $k_{\text{reinitiation}}$. We then treat $k_{\text{reinitiation}}$ as constant, and characterize several bi-cistronic operons to measure k_{P}, which we then treat as a constant.

The rate of ribosome *de novo* initiation for the second CDS is then determined using a modified free energy model that automatically identifies and specially treats intergenic RNA structures that both inhibit translation initiation and overlap with the first CDS sequence. These RNA structures will be actively unfolded by ribosomes during translation elongation of the first CDS, enabling a second ribosome to bind to the intergenic region without having to unfold these specific RNA structures. As a result, the translation initiation rate of the second CDS will increase according to the amount of free energy needed to unfold these RNA structures. The first CDS's translation rate controls how likely these RNA structures will remain unfolded, and therefore the free energy bonus for the second ribosome. Accordingly, we calculate the total binding free energy for new ribosomes initiating translation at the second CDS using

$$\begin{aligned}\Delta G^{(2)}_{\text{total}} &= \Delta G_{\text{mRNA:rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} \\ &+ \Delta G_{\text{standby}} - \Delta G_{\text{noncoupling}} - \Delta G_{\text{coupling}} F_{\text{coupling}}\end{aligned} \qquad (4)$$

where the free energy needed to unfold the mRNA in its initial state (ΔG_{mRNA}) has been decomposed into two terms: a $\Delta G_{\text{coupling}}$ that quantifies the unfolding free energies of all inhibitory RNA structures that overlap with the first CDS sequence, including the last nucleotide of the stop codon; and a $\Delta G_{\text{noncoupling}}$ that quantifies the unfolding free energy of all other RNA structures. RNA structures are considered inhibitory if they block the standby site or overlap with the Shine–Dalgarno sequence, the spacer region, or the downstream footprint region of the ribosome. RNA structures will be partially inhibitory if only a portion of them must be unfolded to unblock the standby site or otherwise enable ribosome binding. Notably, like the free energy model for the first CDS, RNA structures that are non-inhibitory will have free energy contributions to both $\Delta G_{\text{noncoupling}}$ and $\Delta G_{\text{mRNA–rRNA}}$, effectively canceling out their contribution to ΔG_{total}.

We quantify an elongating 70S ribosome's ability to unfold inhibitory RNA structures using a dimensionless fraction F_{coupling} that can be interpreted as the fraction of time that a ribosome is actively translocating across an RNA structure, forcing it to remain unfolded. Conceptually, F_{coupling} can be derived by assuming that the RNA structure has two states (fully folded or unfolded) and that the rate of upstream translation proportionally increases the amount of unfolded state. With these assumptions and a conservation balance, the following simple expression relates the first CDS's translation rate to F_{coupling} according to:

$$F_{\text{coupling}} = \frac{1}{1 + C \exp\left(\beta \Delta G^{(1)}_{\text{total}}\right)} \qquad (5)$$

Equation (5) introduces a parameter C, which we call the ribosome-assisted unfolding coefficient. We initially treat this parameter empirically and experimentally measure its value, but we can also derive an expression for C to provide an initial estimate. First, the average number of elongating ribosomes translating a coding sequence is determined by its translation initiation rate (r), its elongation rate (r_{elong} is about 60 nt/s), and the coding sequence's length L according to the expression (rL/r_{elong}), which assumes that ribosomes initiate and terminate translation under steady-state conditions. If we assume that elongating ribosomes are uniformly distributed across a coding sequence with a footprint of 30 nucleotides (fp), then each one will occupy a fraction of the coding sequence's length (fp/L). Therefore, the total fraction of coding sequence that is occupied by an elongating ribosome will be the product of these expressions ($r\,fp/r_{\text{elong}}$), which has dimensionless units. We then separate this expression into a predicted translation initiation rate, which is $\exp(-\beta \Delta G_{\text{total}})$, and the coefficient C, which has an estimated value of 0.50 ($fp\ /r_{\text{elong}}$). Additional factors could affect the value of C, such as a slower elongation rate or the RNA structure's geometry or refolding rate.

For operons with three or more CDSs, translation initiation rates are calculated in a nested, iterative manner. Using Equation (3), the ribosome re-initiation rate for the third CDS is calculated, which depends on the calculated translation rate for the second CDS. Using Equations (4) and (5), the third CDSs' ribosome de novo initiation rate is calculated, which depends on the intergenic sequence between the second and third CDSs as well as the second CDS's calculated translation rate. Together, Equation (2) determines the third CDSs' translation initiation rate. Iterations continue for the remaining CDSs in the operon.

Quantitative measurements of the ribosome re-initiation rate

We first quantitatively measured the extent of the ribosome's re-initiation, and its dependence on the intergenic sequence and distance, by rationally designing synthetic intergenic regions, controlling upstream translation rates, and measuring downstream translation in a bi-cistronic mRFP1 and GFPmut3b reporter system (Figure 2A). Eleven intergenic regions were constructed and inserted into the bi-cistronic operon with intergenic distances that varied from $d_{12} = -25$ (5′-

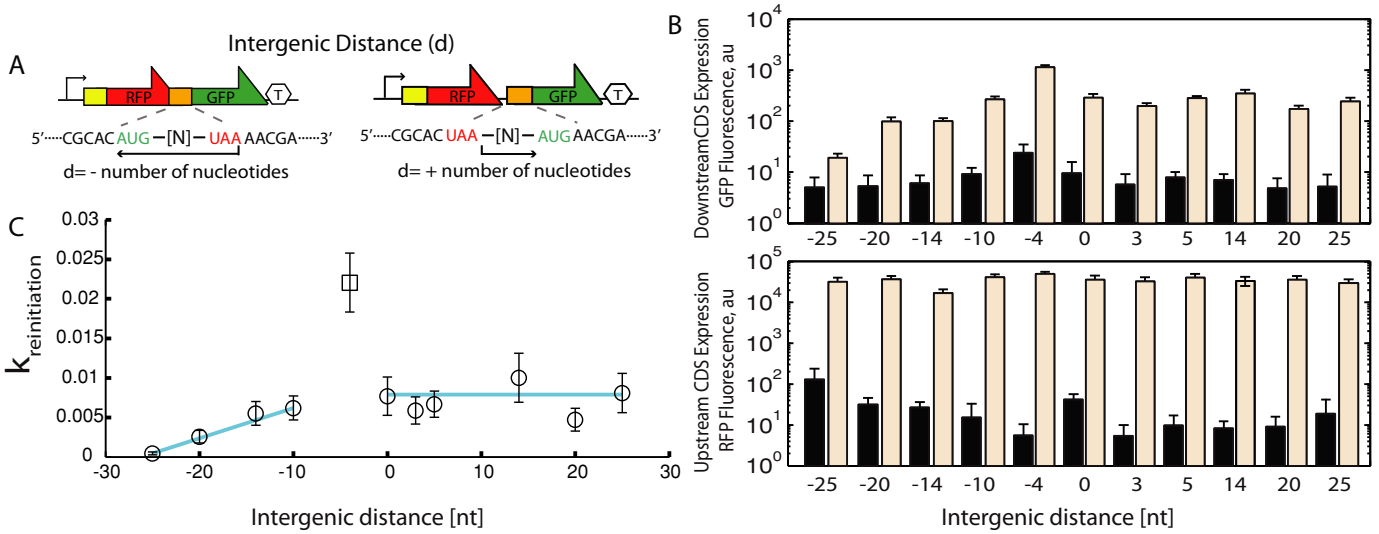

Figure 2. Measuring and modeling ribosome re-initiation rates. (**A**) The intergenic distance is calculated according to the positions of the stop and start codons in the upstream and downstream coding sequences, respectively, within rationally designed bi-cistronic operons. (**B**) Fluorescence measurements show the expression levels of the (bottom) mRFP1 and (top) GFPmut3b reporters after the mRFP1 translation rates were modified to be either (black bars) very low or (beige) very high. (**C**) Analysis of these measurements reveals how the re-initiation coefficient ($k_{\text{reinitiation}}$) depends on intergenic distance according to a three parameter model ($R^2 = 0.91$, $P = 5.94 \times 10^{-12}$). Values and error bars are the mean and SD of at least two replicates.

AUGCCGCAGUAUACCGGUCAAG**UAA**-3′) to $d_{12} = 25$ (5′- **UAA**CAUACGUAUCCCACAUCCCACACAU**AUG**-3′). These intergenic regions were designed to minimize the other source of translational coupling, the ribosome's de novo initiation rate, by preventing the formation of mRNA structures and ensuring that the 30S ribosomal subunit bound poorly to the downstream CDS's ribosome binding site, as quantified by a high ΔG_{total} free energy (lower ribosome affinity). We then designed two synthetic ribosome binding site sequences controlling mRFP1 expression with translation rates of 10 and 94 000 au on the RBS Calculator proportional scale.

The eleven intergenic sequences and two upstream RBS sequences were combined to construct 22 bi-cistronic operons. mRFP1 and GFPmut3b fluorescence levels were monitored during long-time *E. coli* DH10B cultures maintained in the exponential growth phase, and recorded using flow cytometry ('Materials and Methods' section). All single-cell fluorescence distributions were unimodal. All sequences and measurements are listed in Supplementary Data. As expected, the expression of the upstream CDS (mRFP1) increased substantially as a result of increasing its translation rate, on average 3160-fold. Concomittantly, the expression of the downstream CDS (GFPmut3b) also increased substantially, from between 4- and 49-fold, and the magnitude of the increase depended on the intergenic distance (Figure 2B).

A intergenic distance-dependent model of the ribosome re-initiation rate

We quantified the relationship between the ribosome re-initiation rate and intergenic distance by comparing the increase in upstream CDS expression to the increase in downstream CDS expression. For each intergenic distance, we calculate the apparent value of $k_{\text{reinitiation}}$ by dividing the difference in downstream CDS expression ($\text{GFP}_{\text{high}} - \text{GFP}_{\text{low}}$ fluorescence) by the difference in upstream CDS expression ($\text{mRFP1}_{\text{high}} - \text{mRFP1}_{\text{low}}$ fluorescence), which yields a ratiometric coefficient that is independent of our model calculations (Figure 2C). We found that $k_{\text{reinitiation}}$ has a linear relationship with intergenic distance between −25 and 25, except for a distance of −4. The resulting model has three fitted parameters and accurately calculates the downstream CDS translation rates for the 22 bi-cistronic operons ($R^2 = 0.91$, $P = 5.94 \times 10^{-12}$) (Supplementary Figure S1). Ribosome re-initiation was largely constant for d between 0 and 25 nucleotides with $k_{\text{reinitiation}} = 0.0072 \pm 0.0018$, and it decreased as the intergenic distance became more negative than $d = -10$ with a constant slope of 0.0004 per nucleotide ($0.00040 \pm 3 \times 10^{-5}$). At an intergenic distance of −4, however, the re-initiation rate was substantially higher with $k_{\text{reinitiation}} = 0.0220$, due to enhanced tRNA–mRNA binding. In the Discussion section, we propose a mechanistic model that explains how ribosome re-initiation rates depend on intergenic distances in terms of the ribosome's scanning and detachment rates.

The fluorescence measurements and therefore the absolute values of $k_{\text{reinitiation}}$ depended on a single unknown proportionality constant K that reflects the selected reporter proteins and the flow cytometry parameters used to measure their fluorescence levels. To determine the value of this proportionality constant, we recorded mRFP1 and GFPmut3b fluorescences in mono-cistronic operons on the same vector and found a 10.3-fold lower amount of GFPmut3b fluorescence, compared to mRFP1 fluorescence, when adjusting for different rates of protein expression. This apparent K is directly related to the constant k_{P} that is multiplied by $k_{\text{reinitiation}}$ to arrive at the ribosome's re-initiation rate; K multiplied by k_{P} must equal one. As a result, we set k_{P} to 10 and treat it as a constant for the remainder of this study. Below, we utilized all operon measurements and a sensitivity

analysis to find the value of k_P, which resulted in a similar answer.

Predicting translational coupling rates in bi-cistronic operons

Next, we investigated the biophysical model's ability to predict the rate of translational coupling by designing and characterizing 76 bi-cistronic operon variants with six different intergenic regions. The intergenic regions were designed to have inhibitory mRNA structures that overlap with the upstream CDS and become unfolded by upstream elongating ribosomes. Three of the regions have the same intergenic distance ($d = -4$) with translationally coupled mRNA structures that have increasingly higher free energies of unfolding ($\Delta G_{coupling} = -5.3$, -12.6 and -16.9 kcal/mol) (Figure 3A). In particular, the latter two mRNA structures have identical shapes, but with different nucleotide compositions and duplex stabilities. The other three regions have varying intergenic distances ($d = +3$, -1, -11) with similarly stable overlapping, inhibitory mRNA structures ($\Delta G_{coupling} = -8.9$ or -11.1 kcal/mol) (Figure 3B). For all intergenic regions, we then designed 12–13 synthetic RBS sequences to systematically increase the translation rate of the upstream CDS from 1 to 391 000 on the RBS Calculator proportional scale. The translation rates of the downstream CDS also depend on the basal (non-coupled) translation rates, determined the ΔG_{final} and $\Delta G_{non\text{-}coupling}$ free energies. All sequences, measurements, and calculations are listed in Supplementary Data.

We monitored the upstream (mRFP1 fluorescence) and downstream (GFPmut3b fluorescence) CDS expression levels of the 76 bi-cistronic operons during long-time cultures maintained in the exponential growth phase, and recorded single-cell fluorescence distributions using flow cytometry (**Methods**). The upstream RBSs controlling upstream CDS translation caused mRFP1 fluorescence to increase by 21 000- to 44 000-fold with well-predicted translation initiation rates ($R^2 = 0.57$) (Supplementary Figure S2), averaged over the six intergenic regions. The exceptions are the 391 000 au RBS sequences, which yielded similar mRFP1 fluorescences as the 94 000 au RBS sequences, potentially due to the introduction of another rate-limiting step in gene expression, such as translation elongation. From the lowest to highest upstream CDS expression levels, the downstream CDS expression increased by 57-, 126- and 95-fold, respectively, for the first three intergenic regions (Figure 3A) and 54-, 125- and 55-fold, respectively, for the second three intergenic regions (Figure 3B). Consistent with the proposed biophysical model, there was a sigmoidal relationship between upstream and downstream CDS expression levels.

It was expected that changes in mRNA level may play a role in the observed changes in downstream CDS expression as poorly translated coding sequences become accessible to RNAse binding and cleavage activity. Particularly at low upstream CDS translation rates, destabilization of the entire mRNA will affect both upstream and downstream CDS expression. To investigate this possibility, we measured the mRNA levels of a bi-cistronic operon as the upstream CDS's translation was increased from a very low to very high rate. We selected an operon variant (pTTBd-4-17, Figure 3A, right) and employed qRT-PCR to measure *mRFP1* and *gfpmut3b* mRNA levels in comparison to endogenous 16S rRNA levels ('Materials and Methods' section). We found that the mRNA levels for the upstream CDS dropped by about 2-fold as its translation rate was substantially reduced (Figure 3D). Due to the effects of mRNA processing, the downstream CDS mRNA level also decreased by about 1.7-fold when the upstream CDS's translation rate was low. Therefore, while changes in mRNA stability are a factor, they cannot explain the >50-fold change in downstream CDS expression as upstream CDS translation rates were systematically varied.

We then performed quantitative comparisons between these measurements and biophysical calculations to determine if a single model could explain how changing the upstream CDS translation rate controls downstream CDS translation rates. First, we utilized these measurements to determine the value of the ribosome-assisted unfolding coefficient, C, which relates the rate of upstream elongating ribosomes to the fraction of time that an RNA structure has been unfolded by a ribosome. The parameter C controls the steepness of the sigmoidal curve that relates the upstream and downstream CDSs' translation rates via the de novo initiation mechanism. We determined that C is 0.81 $\pm$ 0.17 by normalizing the 76 GFPmut3b fluorescence measurements, subtracting the re-initiation rates as a source of translational coupling, and finding the best-fit value of C that minimized the difference between normalized measurements and predicted downstream CDS translation initiation rates. Fitting of this single parameter resulted in accurate translation initiation rate predictions for the downstream CDS across the 76 measurements with an average Pearson R^2 of 0.78 (Figure 3C). In contrast, if the mechanism of translational coupling was not included within the biophysical model, then the model predictions were not comparable to the measured translation rates (average $R^2 = 0.01$) (Supplementary Figure S3).

We observed that changing the intergenic region could have a proportional effect on the absolute amount of downstream CDS expression beyond the observed mechanism of translational coupling. Using the biophysical model to distinguish the known from unknown interactions, we found that two of the six intergenic regions had over-predicted Gfpmut3b expression levels with a maximum proportional change of 7.2-fold across all upstream CDS translation rates.

Predicting translational coupling rates in tri-cistronic operons

We further evaluated the biophysical model's predictions by designing, constructing, and characterizing 44 tri-cistronic operon variants employing the CFP, mRFP1 and GFPmut3b reporters. Four different sets of intergenic regions were constructed where the first intergenic region between CFP and mRFP1 contained two different inhibitory RNA structures with $\Delta G_{coupling}$ energies of -14.5 and -9.1 kcal/mol, while the second intergenic region between mRFP1 and GFPmut3b contained inhibitory RNA structures with $\Delta G_{coupling}$ energies of -12.6 and -5.3 kcal/mol. To systematically increase the expression of the most upstream CDS, we designed and inserted 11 synthetic ribosome binding sites controlling CFP expression with trans-

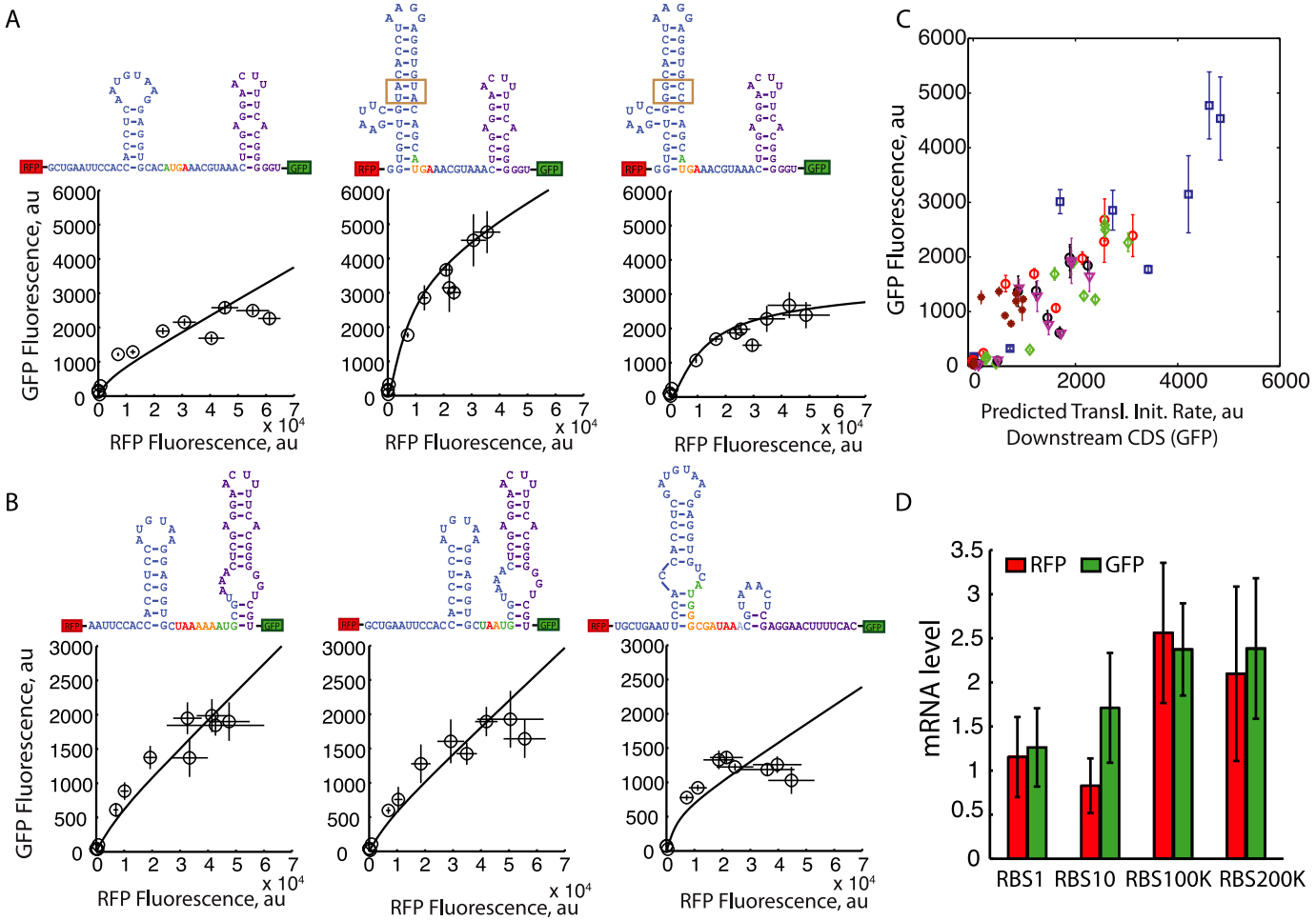

Figure 3. Predicting translational coupling in bi-cistronic operons. The intergenic regions of bi-cistronic reporter operons were designed and characterized to quantitatively determine how inhibitory RNA structures control translational coupling. (black lines) Biophysical model predictions using $C = 0.81$ and $k_P = 10$ are shown alongside (black circles) upstream mRFP1 and downstream GFPmut3b expression level measurements as *mRFP1* translation rates are systematically increased. Nucleotide colors are the same as in Figure 1. (**A**) Calculated $\Delta G_{coupling}$ ($\Delta G_{non\text{-}coupling}$) free energies are −5.3 (−20.1), −12.6 (−16.3) and −16.9 (−16.3) kcal/mol, respectively. Boxes highlight a four nucleotide mutation between two operon variants. (**B**) Three additional bi-cistronic operons were designed with varying intergenic distances ($d = +3, -1$ and -11, respectively) and inhibitory RNA structures. Calculated $\Delta G_{coupling}$ ($\Delta G_{non\text{-}coupling}$) free energies are −8.9 (−12.6), −8.9 (−9.8) and −11.1 (−17.7) kcal/mol, respectively. (**C**) Predicted translation initiation rates are compared to measured downstream CDS (GFPmut3b) expression levels. The apparent proportionality factors are (blue squares, pTTBd-4-13) 9.8, (red circles, pTTBd-4-17) 5.5, (green diamonds, pTTBd-4-5) 17.3, (black circles, pTTBd3-9) 31.0, (pink triangles, pTTBd-1-9) 30.8 and (brown stars, pTTBd-11-11) 4.3 with Pearson coefficients R^2 of 0.85 ($P = 5 \times 10^{-5}$), 0.86 ($P = 8 \times 10^{-6}$), 0.84 ($P = 3 \times 10^{-5}$), 0.77 ($P = 2 \times 10^{-4}$), 0.74 ($P = 4 \times 10^{-4}$) and 0.60 ($P = 0.0069$), respectively. Values and error bars are the mean and SD of three replicates. (**D**) The mRNA levels of the *mRFP1* and *gfpmut3b* coding sequences within four variants of the bi-cistronic operon pTTBd-4-17 were measured as the *mRFP1* translation rates were increased from 1 to 200 000 au on the RBS Calculator proportional scale, showing the effects of upstream translation on mRNA stability in an operon. Values and error bars are the mean and SD of at least two replicates.

lation initiation rates from 10 to 268 000 on the RBS Calculator proportional scale. We then monitored the CFP, mRFP1 and GFPmut3b expression levels of the 44 operon variants during long-time cultures maintained in the exponential growth phase, and recorded single-cell fluorescence distributions using flow cytometry ('Materials and Methods' section).

As the upstream CDS translation rates were increased, translational coupling between CFP and mRFP1 increased mRFP1 expression by between 10.5- to 31.3-fold (Figure 4). Simultaneously, translational coupling between mRFP1 and GFPmut3b increased GFPmut3b expression by between 41.8- and 83.3-fold. The relationships between upstream and downstream CDS translation rates were well predicted by the biophysical model's calculations (Figure 4).

Specifically, CFP expression levels increased proportionally to the translation initiation rate of its designed RBS sequences (average $R^2 = 0.67$) (Figure 5AD), except at a translation rate of 268 000 au where expression reached a plateau. To predict the *mRFP1* translation rates, the predicted *cfp* translation rates and the *cfp-mRFP1* intergenic sequence were then fed into the biophysical model of translational coupling (Equations (1-5)) using the previously determined parameter values ($C = 0.81$, $k_P = 10$). Compared to the measured expression levels, the biophysical model accurately predicted *mRFP1* translation rates on a propor-

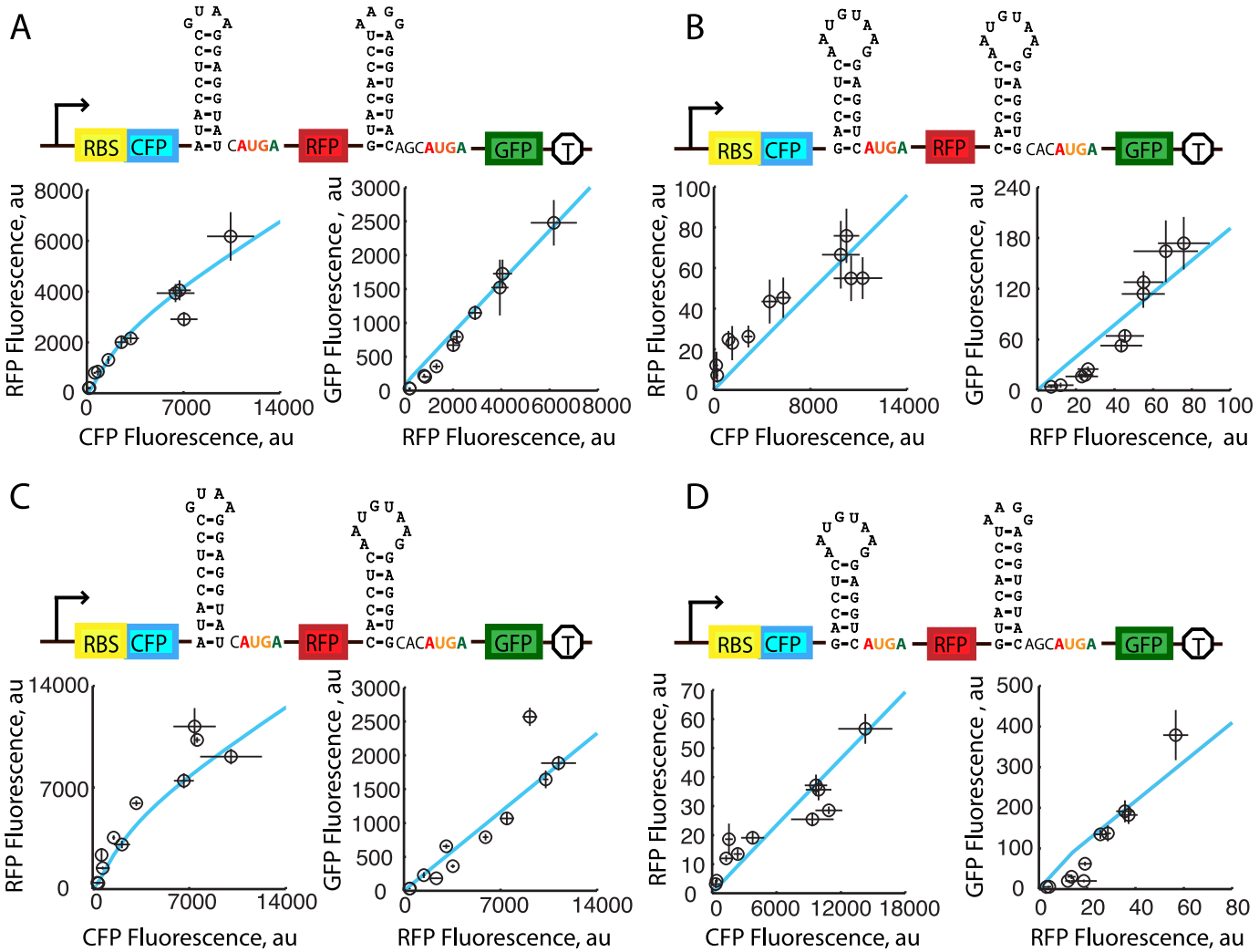

Figure 4. Expression measurements of tri-cistronic operons. The intergenic regions of four tri-cistronic operons were rationally designed and characterized to evaluate the biophysical model's predictions. (blue lines) Biophysical model predictions using $C = 0.81$ and $k_P = 10$ are shown alongside (black circles) expression level measurements of CFP, mRFP1, GFPmut3b reporters as the *cfp* translation rates are systematically increased. Calculated $\Delta G_{coupling}$ ($\Delta G_{non\text{-}coupling}$) free energies for the two intergenic regions are (**A**) −14.5 (−15.4) and −12.6 (−16.3); (**B**) −9.1 (−16.5) and −5.3 (−20.1); (**C**) −14.5 (−15.4) and −5.3 (−20.1); (**D**) −9.1 (−16.5) and −12.6 (−16.3) kcal/mol. All have intergenic distances $d = -4$. Values and error bars are the mean and SD of at least two replicates.

tional scale (average $R^2 = 0.82$) (Figure 5BE), particularly for the intergenic regions that yielded higher expression levels (Figure 4ACD), where the Pearson R^2 comparisons were 0.94 ($P = 3 \times 10^{-6}$), 0.91 ($P = 2 \times 10^{-5}$) and 0.90 ($P = 3 \times 10^{-5}$). The *gfpmut3b* translation initiation rates were predicted using the same approach; the predicted *mRFP1* translation rates and the *mRFP1-gfpmut3b* intergenic sequence were fed into the biophysical model and the translation initiation rates of *gfpmut3b* were calculated. Compared to the measured expression levels, the biophysical model was also able to accurately predict *gfpmut3b* translation rates on a proportional scale (average $R^2 = 0.73$) (Figure 5CF). As before, excluding the mechanism of translational coupling from the biophysical model resulted in a sharp reduction in accuracy; *mRFP1* and *gfpmut3b* translation rates were not well-predicted (R^2 values < 0.01) (Supplementary Figure S4). Overall, incorporating translational coupling into the biophysical model was essential to accurately predicting translation initiation rates within multicistronic operons.

Importantly, though the bi-cistronic and tri-cistronic operons have distinctly different 5′ UTR and intergenic sequences as well as gene orders, the model was able to predict their translation initiation rates with similar accuracy. For example, the addition of the *cfp* protein coding sequence as the first gene within the tri-cistronic operon altered the translation initiation of the 5′ UTR by enabling the formation of mRNA structures that inhibited ribosome binding. These changes in upstream CDS translation rate then propagated forward via translational coupling to affect the translation rates of both the *mRFP1* and *gfpmut3b* CDSs. By quantifying the strengths of the molecular interactions controlling translation initiation and translational coupling, the model was able to account for these significant changes in operon sequence and architecture.

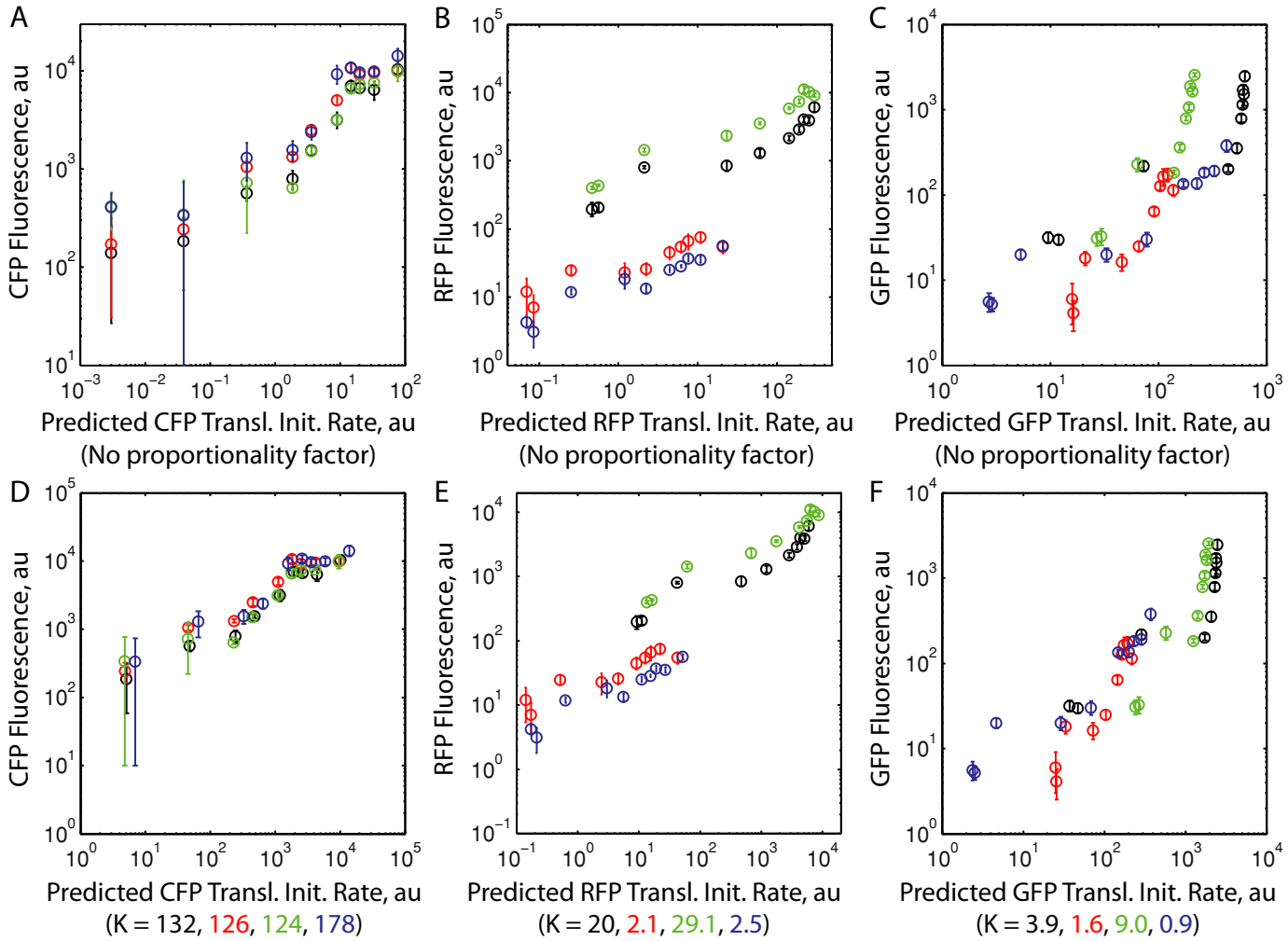

Figure 5. Biophysical model predictions for tri-cistronic operons. (**A**) A comparison between expression measurements of tricistronic operons and the biophysical model's predictions of translational coupling without using a proportionality factor. (**B**) The same comparison is performed after multiplying the predicted translation rates by proportionality factors as shown. Average Pearson R^2 values for comparison are 0.67, 0.82 and 0.73, left-to-right respectively, and are unchanged by the value of the proportionality factor. Colors represent the operon variants (black) pTTTd-4-15-13, (red) pTTTd-4-9-5, (green) pTTTd-4-15–5 and (blue) pTTTd-4-9-13. Values and error bars are the mean and SD of at least two replicates.

We then performed a sensitivity analysis on the conversion factor between ribosome assembly and re-initiation rate (k_P) and the ribosome-assisted unfolding constant (C) to determine how changing their values altered the model's translation initiation rate predictions. We quantified the range of parameter values that yielded translation rate predictions within the 95% confidence interval, using the 120 bi-cistronic and tri-cistronic operon variants, totaling 360 translationally coupled expression levels, as the experimental measurements. We found that the best-fit values for k_P and C were 14 and 0.81, respectively, and that the 95% confidence ranges for these parameters were [8,33] and [0.67,1] (Supplementary Figure S5). Changing the value of C, based on the bi-cistronic operon measurements, did not improve the model's ability to predict the translation initiation rates in the tri-cistronic operons. Similarly, the best-fit value of k_P largely agreed with the previous measurement of k_P that relied on separate measurements of mono-cistronic mRFP1 and GFPmut3b expression.

However, we found that modifications to the intergenic regions had an effect on the absolute expression levels of downstream CDSs within the tri-cistronic operons (Figure 5). For example, after accounting for model-predicted differences in translation rates, there was an 8-fold decrease in mRFP1 expression *at all upstream cfp translation rates* (Figure 5B) when the inhibitory RNA hairpin in the *cfp-mRFP1* intergenic region had a shortened stem (Figure 4A versus Figure 4D). This modification to the cfp-mRFP1 intergenic region also lowered GFPmut3b expression at all upstream *cfp* translation rates by 4.5-fold, though the mRFP1-gfpmut3b intergenic region was unchanged. The reductions in mRFP1 and GFPmut3b expression, independent of the *cfp* translation rate, suggested the presence of other interactions that have a multiplicative effect on protein expression (e.g. changes to mRNA stability). The magnitude of these interactions are quantified by an apparent proportionality factor K.

In contrast, if the *mRFP1-gfpmut3b* intergenic region was modified while leaving the *cfp-mRFP1* intergenic region unchanged, GFPmut3b expression levels decreased by only a factor of 2.3 at all upstream *cfp* translation rates, while mRFP1 expression levels were largely similar (1.45-fold change) (Figure 4A versus Figure 4C). Therefore, the additional factors controlling proportional changes in expression level, beyond translational coupling, were sequence-specific and only found here within the first intergenic region. Overall, across the four intergenic regions, the absolute levels of *mRFP1* and *gfpmut3b* expression levels differed from model predictions by at most a proportional factor of 13.85- and 10.2-fold, respectively (Figure 5DEF). These differences in the proportionality factor quantify the magnitudes of the interactions, beyond translational coupling, that can affect downstream CDS expression levels, but are not included within the model. Below, we discuss how incorporating additional gene regulatory mechanisms in operons can explain the changes in the apparent proportionality factor.

Predicting translational coupling in a natural operon

Advances in next-generation sequencing-based assays have enabled quantitative measurements of mRNA levels and ribosome occupancies across an organism's transcriptome (37,45). In particular, the RNA-Seq and ribosome profiling techniques were recently applied to *E. coli* MG1655 to determine its genome-wide mRNA levels and ribosome occupancies in complete, minimal, and methionine-free MOPS medias (38). Here, we apply our biophysical model of translation initiation and translational coupling to predict the translation initiation rates of coding sequences within a natural operon and compare to these genome-wide measurements.

Prior to these comparisons, it is important to examine the types of data generated by these sequencing-based techniques and to analyze the assumptions needed to compare these measurements to model predictions. First, RNA-Seq measurements are typically reported using units of reads per kilobase of mRNA per million reads (RPKM). The RPKM unit accounts for differences in CDS length and sequencing coverage, and provides a proportional measurement of mRNA concentration, though the average RPKM across a transcriptome will vary from one organism to another (46). Therefore, mRNA level measurements in units of RPKM exist on a proportional scale.

Similarly, the ribosome profiling technique records the number of bound ribosomes to mRNA transcripts, which is averaged across a CDS's length to provide gene-level ribosome numbers. The average number of ribosomes per mRNA transcript can be quantified by dividing these gene-level ribosome numbers by mRNA level measurements, yielding a quantitative metric for gene-level ribosome density. The conversion from ribosome densities to mRNA translation rates, however, depends on both the mRNA's translation initiation and elongation rates, which confounds a direct comparison between measured ribosome densities and model-predicted translation initiation rates. Instead, in the initial analysis of ribosome profiling measurements (38), it was assumed that all *E. coli* genes have the same translation elongation rates, though this assumption was never experimentally validated and it is not likely to be true. In particular, the measured ribosome densities for 90% of *E. coli* genes varied by only 11.8-fold. If translation elongation rates were the same for all genes, that would imply that translation initiation rates only vary by 11.8-fold across the entire transcriptome, which is also not likely to be true. Therefore, we should only compare measured ribosome densities to predicted translation initiation rates when there is reason to suspect that the coding sequences within an operon have roughly equal translation elongation rates.

We selected the *atpIBEFHAGDC* operon as it was previously used as an illustrative example (38) and because it encodes nine coding sequences without any known internal promoters or transcriptional terminators. The Atp proteins bind together to form a large multimeric complex, composed of the ATP synthase F0 complex ($AtpBE_{10}F_2$) and the ATP synthase F1 complex ($AtpHA_3GD_3C$), and therefore are more likely to have synchronized and similar translation elongation rates. Accordingly, the measured numbers of ribosomes bound to each CDS were roughly proportional to the expected protein stoichiometries within the ATP synthase complex. For example, the number of transcript-bound ribosomes for *atpE* was 10.7-fold higher than *atpB* as a result of a 2.4-fold higher mRNA level and a 4.4-fold higher ribosome occupancy per transcript. Notably, because the *atp* operon features nine genes and four different protein stoichiometries, it provides a larger number of unique data-points for comparison than other operons.

When using the experimentally measured transcriptional start site upstream of *atpI* and the more highly translated GUG start codon for *atpI*, the biophysical model was able to accurately predict the translation initiation and translational coupling rates for all *atp* genes when compared to measured ribosome densities ($R^2 = 0.72$, $P = 0.004$) (Table 1). When translational coupling was not included within the model, the translation initiation rates for *atpF* and *atpA* were less accurately predicted ($R^2 = 0.65$, $P = 0.008$); ribosome re-initiation is responsible for 98% of *atpF* translation initiation, and the intergenic region for *atpA* contains an overlapping inhibitory RNA structure ($\Delta G_{coupling} = -5.9$ kcal/mol) that increases its translation initiation rate by 6.5-fold. Overall, the biophysical model of translation initiation and translational coupling could explain *why* the protein synthesis rates from the *atpIBEFHAGDC* operon are proportional to the expected protein stoichiometries.

Design criteria for intergenic regions to coordinate expression in synthetic operons

Using the biophysical model, we now analyze how intergenic regions can be designed to coordinate protein expression within synthetic bacterial operons. Our applications of interest include the co-expression of multi-subunit proteins and multi-enzyme pathways. First, an operon's upstream and downstream CDSs' translation rates will be linearly related when an intergenic region does not have any inhibitory RNA structures that overlap with the footprint of upstream elongating ribosomes ($\Delta G_{coupling} = 0$). Due to ribosome re-initiation, the slope of this relationship is controlled by the

Table 1. Translational coupling in the *atp* operon

Gene	Ribosome read counts (au)	mRNA read counts (RPKM)	Ribosomes per transcript (au)	Transl. init. rate (with coupling)	Transl. init. rate (without coupling)	Coupling factor
atpI	151	764	0.20	0.014	0.014	1.00
atpB	10508	1695	6.20	2.47	2.47	1.00
atpE	112959	4114	27.46	27.73	27.56	1.01
atpF	17866	3109	5.75	1.98	0.034	57.63
atpH	9335	2938	3.18	2.75	2.15	1.28
atpA	30696	2724	11.27	4.61	0.71	6.54
atpG	9832	1914	5.14	4.88	4.55	1.07
atpD	30603	2177	14.06	4.80	4.46	1.08
atpC	12695	2193	5.79	12.60	12.26	1.03

The biophysical model's predicted translation initiation rates with and without coupling are compared to gene-level ribosome density measurements, previously obtained by Li *et al.* (38).

intergenic distance between these CDSs and varies between zero and 0.022. The relatively small slope can have a large effect on protein expression level ratios when upstream and downstream CDSs have widely different basal translation rates as quantified by the ribosome's interactions with the mRNA ($\Delta G_{\text{final}} - \Delta G_{\text{non-coupling}}$). For example, when an upstream CDS has a 100-fold higher basal translation rate than a downstream CDS (intergenic distance $d = -4$), there will only be a 31-fold difference in translation rate when ribosome re-initiation is taken into account. To prevent such effects, intergenic regions should be designed so that coding sequences overlap by 25 or more nucleotides.

Second, once an intergenic region contains overlapping, inhibitory RNA structures, the operon's protein expression level ratios are controlled by the structures' unfolding energies (quantified by $\Delta G_{\text{coupling}}$), the basal translation rate (determined by $\Delta G_{\text{final}} - \Delta G_{\text{non-coupling}}$), and the rate of ribosome re-initiation. An intergenic region with a high basal translation rate ($\Delta G_{\text{final}} - \Delta G_{\text{non-coupling}} = -20$ kcal/mol) and a weak RNA structure ($\Delta G_{\text{coupling}} = -5$ kcal/mol) will vary the ribosome's *de novo* initiation rate by ∼10-fold as upstream CDS translation rates is increased (Figure 6A). As the RNA structure becomes more stable, the downstream CDS's basal translation rate drops substantially and will only increase if the upstream CDS is highly translated. In principle, translational coupling can increase a downstream CDS's translation by 1000-fold if the inhibitory RNA structure is both highly stable and completely unfolded by upstream elongating ribosomes.

Third, intergenic regions can be designed to obtain desired protein expression level ratios by introducing inhibitory RNA structures with targeted folding energies. By varying $\Delta G_{\text{coupling}}$, a wide range of protein expression level ratios are achievable (Figure 6B). In particular, at a high basal translation rate ($\Delta G_{\text{final}} - \Delta G_{\text{non-coupling}} = -20$ kcal/mol), protein expression level ratios are tunable from 2-fold to 70-fold when modulating $\Delta G_{\text{coupling}}$ from −5 to −30 kcal/mol. When instead designing intergenic regions with a low basal translation rate ($\Delta G_{\text{final}} - \Delta G_{\text{non-coupling}} = -10$ kcal/mol), similar tunability is achieved when varying $\Delta G_{\text{coupling}}$ from −1 to −15 kcal/mol.

DISCUSSION

Translational coupling is a ubiquitous mechanism for controlling protein expression levels in bacterial operons. While its effects have been observed in several cases (4,20–30,34–36), a physics-based model had not been formulated or tested. In this study, we have developed a biophysical model that predicts how translational coupling controls protein expression level ratios within bacterial operons, quantifying the rates of both ribosome re-initiation and de novo initiation for each coding sequence. Our model has five free parameters whose values were determined through the characterization of 142 rationally designed operons with systematically varied characteristics. As a result, the model was able to accurately predict how changing the translation rates of upstream coding sequences controlled the translation of downstream coding sequences in both bi-cistronic and tri-cistronic operons. The model also quantifies how altering intergenic distance and the folding stabilities of intergenic RNA structures controls translational coupling. Importantly, the biophysical model performs these predictions using mRNA sequence as its only input, which allows its calculations to be combined with other sequence-to-function models to predict the function of complex genetic systems.

A key novelty of our study is the rational design of synthetic genetic systems with targeted and systematically varied properties to perturb specific interactions controlling gene expression and system function. Through characterization of these systems, we were able to distinguish separate sets of interactions controlling ribosome re-initiation and de novo initiation and therefore develop a quantitative model for each of these components of translational coupling. To carry out this rational design, it was essential to use a previously developed sequence-to-function model, the RBS Calculator, to design 5′ UTR sequences and systematically control the translation rates of upstream coding sequences (40). As this example illustrates, to develop a more complete understanding of how sequence controls genetic system function, it will become increasingly necessary to bootstrap the predictions of one model to formulate and test the predictions of more complex models.

Analysis of our results has also provided a more complete picture of the ribosome-ribosome interactions that are responsible for ribosome re-initiation. First, the re-

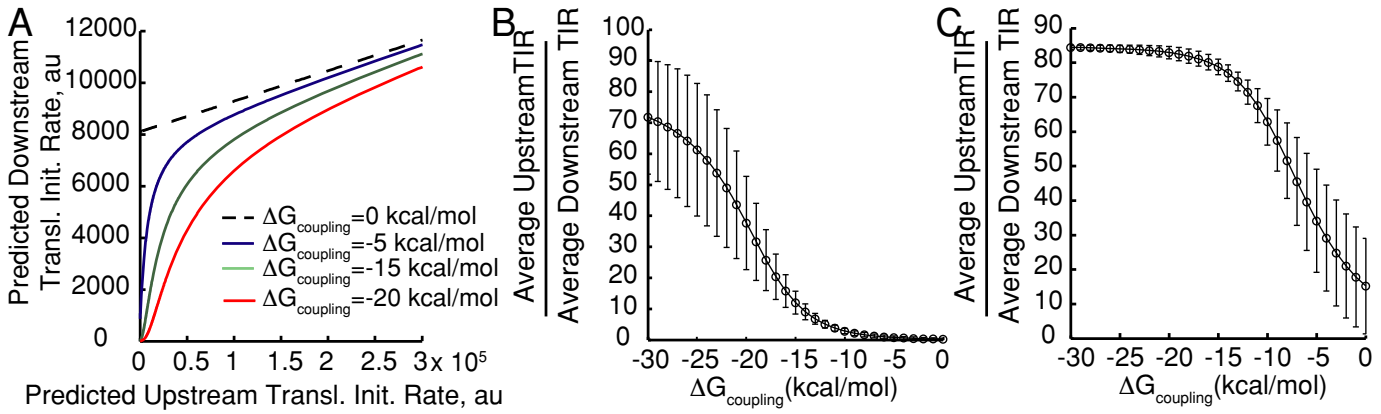

Figure 6. Translational coupling parameters that control protein ratios. (**A**) Predicted rates of upstream and downstream CDS translation rates as an intergenic region's $\Delta G_{coupling}$ energy is varied. (**B**) The predicted relationship between translation rate ratios and $\Delta G_{coupling}$ when ($\Delta G_{final} - \Delta G_{non\text{-}coupling}$) is −20 kcal/mol. (**C**) The predicted relationship between translation rate ratios and $\Delta G_{coupling}$ when ($\Delta G_{final} - \Delta G_{non\text{-}coupling}$) is −10 kcal/mol. Circles and bars represent the mean and standard deviations of predictions when evaluated across a 10 000-fold range of upstream CDS translation rates.

initiation rate is substantially higher ($k_{reinitiation} = 0.022$) at $d = -4$ because the 30S ribosomal complex's $tRNA^{fMet}$ anti-codon loop (3′-UACU-5′) forms four Watson–Crick base pairings with the intergenic sequence (5′-AUGA-3′) (47) and scanning is not required to re-initiate translation. Second, when scanning is necessary for ribosome re-initiation, the rate of translation appears to be governed by a one-dimensional random walk along the mRNA with spontaneous detachment. After the 70S ribosome reaches a stop codon and dissociates, the 30S ribosomal subunit uses its positively charged platform domain to remain loosely bound to the mRNA. Through non-specific interactions, the 30S ribosome scans along the mRNA in both directions and has a small chance of detaching completely from the mRNA, returning to the cytostolic ribosome pool. Forward movement is unimpeded and appears to have a sufficiently high velocity as the probability of ribosome re-assembly, as quantified by $k_{reinitiation}$, is similar for short distances up to 25 nucleotides ($k_{reinitiation} = 0.007$). In another study, for longer intergenic distances up to 350 nucleotides, the ribosome re-initiation rate decreased by 1.4- to 2.0-fold for every 100 nucleotides (36). However, we found that a 30S ribosome's reverse movement is impeded as it scans across an upstream coding sequence. It if collides with an elongating 70S ribosome, the 30S ribosome will detach from the mRNA and return to the cytostolic pool. The probability of avoiding a collision and re-initiating translation decreased by 11.6-fold when the upstream and downstream coding sequence overlapped by 25 nucleotides. Interestingly, no additional sequence specificity, such as a Shine-Dalgarno sequence, appears to be needed for a ribosome to re-initiate translation at a start codon after it successfully scans to its location, though the presence of a Shine-Dalgarno sequence will increase the basal (non-coupled) translation rate.

Using Markov modeling, we can relate these measured re-initiation rates to the 30S ribosome's per-nucleotide scanning and detachment rates, as quantified by the transition rates a_{scan} and a_{off}, respectively. The probability of a ribosome scanning forward by one nucleotide without detachment is $a_{scan}/\left(a_{scan} + a_{off}\right)$. The probability of N such forward moves to find a start codon is $\left[a_{scan}/\left(a_{scan} + a_{off}\right)\right]^N$. If we assume that scanning is forward-only, then a 2-fold decrease in re-initiation per 100 nucleotides would indicate that the ribosome's scanning rate is 144-fold faster than its dissociation rate. If we assume that scanning is bidirectional, but without detachment collisions between different 30S scanning ribosomes, then the scanning rate is 72-fold faster than its dissociation rate. Conversely, the probability of a ribosome scanning backwards N nucleotides, while not encountering an elongating ribosome, had much higher detachment rates. For reverse movement, the scanning rate was only 4.9-fold faster than the detachment rate.

Next, we further illustrate how elongating ribosomes unfold mRNA structures within intergenic regions and control de novo initiaton of downstream CDSs. First, the presence of mRNA structures within the upstream CDS could have influenced the ribosome's translation elongation rate. However, because elongating ribosomes actively hydrolyze GTP and have an external energy source during translocation, we expect that most RNA structures are not stable enough to pause translocation. Accordingly, we did not observe any appreciable changes in 1st CDS expression level as we altered the stabilities of RNA structures within the first intergenic region, particularly in the bottom three nucleotides that would most affect translocation times.

Second, the ribosome's translation initiation rate, elongation rate, and footprint length will all determine how likely an RNA structure will become unfolded. To model this relationship, we derived the ribosome-assisted unfolding coefficient C, which depends on the ribosome's footprint length and elongation rate, and converts the upstream CDS's translation initiation rate into the fraction of time that an intergenic RNA structure remains unfolded. Interestingly, our theoretical estimate for C, based on a footprint of 30 nucleotides and an elongation rate of 60 nucleotides per second, largely agrees with our empirical measurements (0.50 versus 0.81). We would expect to see a difference if our reporter coding sequences had slower translation elon-

gation rates or if partial refolding of intergenic RNA structures took place.

Third, according to our model, we expected to see a sigmoidal relationship between upstream and downstream translation rates, where very high upstream CDS translation rates lead to a plateau in downstream CDS expression. However, it was also possible that over-crowding of elongating ribosomes at intergenic regions could block the binding of free cytostolic ribosomes, leading to lower downstream CDS expression levels. In several cases, we observed the expected expression plateaus, but did not measure any appreciable decrease in downstream CDS expression that would indicate that ribosomal over-crowding was a significant factor. As a result, it appears that elongation rates of these coding sequences were sufficiently fast, compared to the highest initiation rates, for ribosomes to clear the intergenic region.

Lastly, by comparing model predictions to measurements, we can distinguish between the known and unknown interactions that control protein expression levels. Here, we observed that changing the intergenic sequence unexpectedly resulted in proportional changes in downstream CDS expression levels across all upstream CDS translation rates as quantified by a proportionality factor that varied by at most 7.2- and 14.2-fold for the bi-cistronic and tri-cistronic operons. These differences in CDS expression levels at the second and third gene positions could be caused by a combination of several factors: lower downstream mRNA levels from RNA polymerase fall off, lower mRNA stabilities from RNAse activity, increased translation through coupling between transcription and translation (48), and mis-calculated ribosome binding free energies. Assuming that mis-calculated translation initiation and coupling rates are the only source of error, the maximum free energy error in ΔG_{final}, $\Delta G_{\text{coupling}}$ or $\Delta G_{\text{non-coupling}}$ would be 4.4 and 5.9 kcal/mol, respectively. If the stability of the transcript decreased by 2-fold as a result of introducing an RNA structure that was cleaved by an RNAse, then the maximum free energy error would be 2.8 and 4.4 kcal/mol, respectively. Transcription-translation coupling would result in a 40% increase in downstream CDS expression level per 1 kb of distance from the CDS's start codon to the mRNA's transcriptional start site (48). If we account for this effect, then we would expect to observe an increase in mRFP1 and Gfpmut3b expression level by a factor of 1.32 and 1.6 (48). By combining the lower mRNA stability and the effect of transcription-translation coupling, the maximum free energy error estimate becomes 2.2 and 3.3 kcal/mol, respectively. Changes in DNA copy number, operon transcriptional direction, first-gene mRNA stability, and RNA folding kinetics will also influence overall protein expression levels.

With the incorporation of translational coupling, we have developed a more comprehensive and accurate sequence-to-function biophysical model of translation initiation for multi-cistronic operons. By accounting for all known interactions, biophysical models are more capable of identifying knowledge gaps, designing experiments, and making testable predictions than observational correlations. Improving the accuracy of these models is an essential step towards engineering genetic systems without trial-and-error. A software implementation of the biophysical model of translational coupling (version 1.0) is available at http://salislab.net/software as part of a new design method, called the Operon Calculator.

ACKNOWLEDGEMENTS

The authors would like to thank the Genomics Core Facility and Flow Cytometry Core Facility at Penn State for technical assistance.
Author contributions: TT performed the experiments. T.T. and H.M.S. developed the model, analyzed the results and wrote the manuscript.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Air Force Office of Scientific Research [FA9550-14-1-0089]; Office of Naval Research [N00014-13-1-0074]; NSF Career Award [CBET-1253641]; DARPA CLIO program [N66001-12-C-4017]; DARPA Young Faculty Award [N66001-10-1-4019 to H.M.S.]. Funding for open access charge: Extramural funding from NSF.
Conflict of interest statement. None declared.

REFERENCES

1. Temme,K., Zhao,D. and Voigt,C.A. (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 7085–7090.
2. Schlatter,S., Stansfield,S.H., Dinnis,D.M., Racher,A.J., Birch,J.R. and James,D.C. (2005) On the optimal ratio of heavy to light chain genes for efficient recombinant antibody production by CHO cells. *Biotechnol. Progr.*, **21**, 122–133.
3. Farasat,I., Kushwaha,M., Collens,J., Easterbrook,M., Guido,M. and Salis,H.M. (2014) Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol. Syst. Biol.*, **10**, 731.
4. Button,J.E. and Galán,J.E. (2011) Regulation of chaperone/effector complex synthesis in a bacterial type III secretion system. *Mol. Microbiol.*, **81**, 1474–1483.
5. Ng,C.Y., Farasat,I., Maranas,C.D. and Salis,H.M. (2015) Rational design of a synthetic Entner Doudoroff pathway for improved and controllable NADPH regeneration. *Metab. Eng.*, **29**, 86–96.
6. Fang,J., Qian,J.-J., Yi,S., Harding,T.C., Tu,G.H., VanRoey,M. and Jooss,K. (2005) Stable antibody expression at therapeutic levels using the 2A peptide. *Nature Biotechnol.*, **23**, 584–590.
7. Ajikumar,P.K., Xiao,W.-H., Tyo,K.E., Wang,Y., Simeon,F., Leonard,E., Mucha,O., Phon,T.H., Pfeifer,B. and Stephanopoulos,G. (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science*, **330**, 70–74.
8. Pfleger,B.F., Pitera,D.J., Smolke,C.D. and Keasling,J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, **24**, 1027–1032.
9. Xu,P., Gu,Q., Wang,W., Wong,L., Bower,A.G., Collins,C.H. and Koffas,M.A. (2013) Modular optimization of multi-gene pathways for fatty acids production in E. coli. *Nat. Commun.*, **4**, 1409.
10. Lou,C., Stanton,B., Chen,Y.-J., Munsky,B. and Voigt,C.A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.*, **30**, 1137–1142.
11. Qi,L., Haurwitz,R.E., Shao,W., Doudna,J.A. and Arkin,A.P. (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.*, **30**, 1002–1006.
12. Fang,G., Rocha,E.P.C. and Danchin,A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, **9**, 4.

13. Price,M.N., Huang,K.H., Arkin,A.P. and Alm,E.J. (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, **15**, 809–819.
14. Zhang,F., Carothers,J.M. and Keasling,J.D. (2012) Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat. Biotechnol.*, **30**, 354–359.
15. Xu,P., Li,L., Zhang,F., Stephanopoulos,G. and Koffas,M. (2014) Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 11299–11304.
16. Ehretsmann,C.P., Carpousis,A.J. and Krisch,H.M. (1992) Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes Dev.*, **6**, 149–159.
17. Keseler,I.M., Mackie,A., Peralta-Gil,M., Santos-Zavaleta,A., Gama-Castro,S., Bonavides-Martínez,C., Fulcher,C., Huerta,A.M., Kothari,A. and Krummenacker,M. (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.*, **41**, D605–D612.
18. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
19. Chartier,M., Gaudreault,F. and Najmanovich,R. (2012) Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics*, **28**, 1438–1445.
20. Rex,G., Surin,B., Besse,G., Schneppe,B. and McCarthy,J. (1994) The mechanism of translational coupling in Escherichia coli. Higher order structure in the atpHA mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes. *J. Biol. Chem.*, **269**, 18118–18127.
21. Lesage,P., Chiaruttini,C., Graffe,M., Dondon,J., Milet,M. and Springer,M. (1992) Messenger RNA secondary structure and translational coupling in the Escherichia coli operon encoding translation initiation factor IF3 and the ribosomal proteins, L35 and L20. *J. Mol. Biol.*, **228**, 366–386.
22. Hellmuth,K., Rex,G., Surin,B., Zinck,R. and McCarthy,J. (1991) Translational coupling varying in efficiency between different pairs of genes in the central region of the atp operon of Escherichia coli. *Mol. Microbiol.*, **5**, 813–824.
23. Schümperli,D., McKenney,K., Sobieski,D.a. and Rosenberg,M. (1982) Translational coupling at an intercistronic boundary of the Escherichia coli galactose operon. *Cell*, **30**, 865–871.
24. Oppenheim,D.S. and Yanofsky,C. (1980) Translational coupling during expression of the tryptophan operon of Escherichia coli. *Genetics*, **95**, 785–795.
25. Aksoy,S., Squires,C.L. and Squires,C. (1984) Translational coupling of the trpB and trpA genes in the Escherichia coli tryptophan operon. *J. Bacteriol.*, **157**, 363–367.
26. Mathiesen,G., Huehne,K., Kroeckel,L., Axelsson,L. and Eijsink,V.G.H. (2005) Characterization of a new bacteriocin operon in sakacin P-producing Lactobacillus sakei, showing strong translational coupling between the bacteriocin and immunity. Genes. **71**, 3565–3574.
27. Govantes,F., Andujar,E. and Santero,E. (1998) Mechanism of translational coupling in the nifLA operon of Klebsiella pneumoniae. *EMBO J.*, **17**, 2368–2377.
28. Løvdok,L., Bentele,K., Vladimirov,N., Müller,A., Pop,F.S., Lebiedz,D., Kollmann,M. and Sourjik,V. (2009) Role of translational coupling in robustness of bacterial chemotaxis pathway. *PLoS Biol.*, **7**, e1000171.
29. Grabowska,A.D., Wandel,M.P., Lasica,A.M., Nesteruk,M., Roszczenko,P., Wyszyńska,A., Godlewska,R. and Jagusztyn-Krynicka,E.K. (2011) Campylobacter jejuni dsb gene expression is regulated by iron in a Fur-dependent manner and by a translational coupling mechanism. *BMC Microbiol.*, **11**, 166.
30. Nakatogawa,H., Murakami,A. and Ito,K. (2004) Control of SecA and SecM translation by protein secretion. *Curr. Opin. Microbiol.*, **7**, 145–150.
31. Spanjaard,R.A. and Jan,v.D. (1989) Translational reinitiation in the presence and absence of a Shine and Dalgarno sequence. *Nucleic Res.* **17**, 5501–5507.
32. Qu,X., Wen,J.-D., Lancaster,L., Noller,H.F., Bustamante,C. and Tinoco,I. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature*, **475**, 118–121.
33. van de Guchte,M., Kok,J. and Venema,G. (1991) Distance-dependent translational coupling and interference in Lactococcus lactis. *Mol. Gen. Geneti.: MGG*, **227**, 65–71.
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. 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. *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10**, 354–360.
36. Levin-Karp,A., Barenholz,U., Bareia,T., Dayagi,M., Zelcbuch,L., Antonovsky,N., Noor,E. and Milo,R. (2013) Quantifying translational coupling in E. coli synthetic operons using RBS modulation and fluorescent reporters. *ACS Synth. Biol.*, **2**, 327–336.
37. Wang,H.H., Isaacs,F.J., Carr,P.A., Sun,Z.Z., Xu,G., Forest,C.R. and Church,G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
38. Li,G.-W., Burkhardt,D., Gross,C. and Weissman,J.S. (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, **157**, 624–635.
39. 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.
40. Salis,H.M. (2011) The ribosome binding site calculator. *Methods Enzymol*, **498**, 19–42.
41. Borujeni,A.E., Channarasappa,A.S. and Salis,H.M. (2013) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.*, **42**, 2646–2659.
42. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
43. Lorenz,R., Bernhart,S.H., Zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorith. Mol. Biol.*, **6**, 26.
44. Levine,E., Zhang,Z., Kuhlman,T. and Hwa,T. (2007) Quantitative characteristics of gene regulation by small RNA. *PLoS Biol.*, **5**, e229.
45. Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, **15**, 205–213.
46. Wagner,G.P., Kin,K. and Lynch,V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.
47. Schmitt,M., Manderschied,U., Kyriatsoulis,a., Brinckmann,U. and Gassen,H.G. (1980) Tetranucleotides as effectors for the binding of initiator tRNA to Escherichia coli ribosomes. *Eur. J. Biochem./ FEBS*, **109**, 291–299.
48. Lim,H.N., Lee,Y. and Hussein,R. (2011) Fundamental relationship between operon organization and gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10626–10631.