HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2010 April 04.

Published in final edited form as:

Nat Biotechnol. 2009 October ; 27(10): 946–950. doi:10.1038/nbt.1568.

Automated Design of Synthetic Ribosome Binding Sites to Precisely Control Protein Expression

Howard M. Salis[1], **Ethan A. Mirsky**[2], and **Christopher A. Voigt**[1,*]

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, 94158

[2]Graduate Group in Biophysics, University of California San Francisco, San Francisco, CA, 94158

Abstract

Microbial engineering often requires fine control over protein expression; for example, to connect genetic circuits 1-7 or control flux through a metabolic pathway 8-13. We have developed a predictive design method for synthetic ribosome binding sites that enables the rational control of a protein's production rate on a proportional scale. Experimental validation of over 100 predictions in *Escherichia coli* shows that the method is accurate to within a factor of 2.3 over a range of 100,000-fold. The design method also correctly predicts that reusing a ribosome binding site sequence in different genetic contexts can result in different protein expression levels. We demonstrate the method's utility by rationally optimizing a protein's expression level to connect a genetic sensor to a synthetic circuit. The proposed forward engineering approach will accelerate the construction and systematic optimization of large genetic systems.

Keywords

synthetic biology; translation; optimization; metabolic engineering; genetic circuit; RNA secondary structure

Introduction

Microbial engineering is a time-consuming procedure that often requires multiple rounds of trial-and-error genetic mutation. As it becomes possible to construct larger pieces of synthetic DNA 14, including whole genomes 15, automated methods for genetic circuit assembly and metabolic pathway optimization will be critically important. As genetic systems grow in size and complexity, the application of a trial-and-error approach to optimizing these systems is more difficult.

A genetic system's function is optimized by varying the sequences of its regulatory elements to control the expression levels of its protein coding sequences. Each rate-limiting step in

* Corresponding author: cavoigt@picasso.ucsf.edu, Phone: 415-502-7050.

Author Contributions: HMS and CAV designed the study and wrote the manuscript. HMS developed the method. HMS and EAM performed the experiments.

gene expression offers the opportunity for rationally modulating the protein expression level. In bacteria, ribosome binding sites (RBSs) and other regulatory RNA sequences are effective control elements for translation initiation 16-19. As a consequence, they are commonly mutated to optimize genetic circuits, metabolic pathways, and the expression of recombinant proteins.

Previous studies have generated libraries of RBS sequences with the goal of optimizing the function of a genetic system 1, 7, 18. Generation and selection of a sequence library can become impractical as the number of participating proteins increases, especially if measuring the function requires a low-throughput assay or screen 6. For example, randomly mutating 4 nucleotides of an RBS generates a library of 256 sequences. The library size increases combinatorially with the number of proteins in the engineered system (16.7 million sequences for 3 proteins, 2.8×10^{14} sequences for 6 proteins).

A biophysical model of translation initiation would aid the optimization process by enabling the design of an RBS sequence to obtain a desired translation initiation rate. Using thermodynamics, the free energies of key molecular interactions involved in translation initiation have been characterized 20, 21. Thermodynamic models are made possible by measuring the sequence-dependent energetic changes during RNA folding and hybridization 22-26. These methods have enumerated and characterized the attributes of a RBS sequence that affect its translation initiation rate, but a predictive model that combines all of the interactions together has not been created and tested.

Bacterial translation consists of four phases: initiation, elongation, termination, and ribosome turnover (Figure 1A) 27. In most cases, translation initiation is the rate-limiting step. The translation initiation rate is determined by the summary effect of multiple molecular interactions, including the hybridization of the 16S rRNA to the RBS sequence, the binding of $tRNA^{fMET}$ to the start codon, the distance between the 16S rRNA binding site and the start codon, and the presence of RNA secondary structures that occlude either the 16S rRNA binding site or the standby site 20, 21, 28-31.

We have developed an equilibrium statistical thermodynamic model to quantify the strengths of the molecular interactions between the 30S complex and an mRNA transcript and to predict the resulting translation initiation rate. The thermodynamic model describes the system as having two states separated by a reversible transition (Figure 1B). The initial state is the folded mRNA transcript and the free 30S complex. The final state is the assembled 30S pre-initiation complex on an mRNA transcript. The difference in Gibbs free energy between these two states is quantified by the Gibbs free energy change ΔG_{tot}. The ΔG_{tot} depends on the mRNA sequence surrounding a specified start codon and will become more negative when attractive interactions are present and more positive when mutually exclusive secondary structures are present.

The translation initiation rate r is related to the ΔG_{tot} according to

$$r \propto \exp\left(-\beta \Delta G_{\mathrm{tot}}\right) \quad (1)$$

where β is the Boltzmann factor for the system. The derivation of Equation 1 is presented in the Supplementary Methods. Importantly, Equation 1 describes the differences in translation initiation rate that result from differences in mRNA sequence. The amount of expressed protein E is proportional to the translation initiation rate where the proportionality factor K accounts for any ribosome-mRNA molecular interactions that are independent of mRNA sequence and any translation-independent parameters, such as the DNA copy number, the promoter's transcription rate, the mRNA stability, and the protein dilution rate (Supplementary Figure 1).

Given a specific mRNA sequence surrounding a start codon, called the subsequence, the ΔG_{tot} is predicted according to the energy model:

$$\Delta G_{\mathrm{tot}}=\Delta G_{\mathrm{mRNA:rRNA}}+\Delta G_{\mathrm{start}}+\Delta G_{\mathrm{spacing}}-\Delta G_{\mathrm{standby}}-\Delta G_{\mathrm{mRNA}} \quad (2)$$

where the reference state is a fully unfolded subsequence with $\Delta G_{ref} = 0$.

The $\Delta G_{mRNA:rRNA}$ term is the energy released when the last 9 nucleotides (nt) of the *E. coli* 16S rRNA – 3′-AUUCCUCCA-5′ – hybridizes and co-folds to the mRNA subsequence ($\Delta G_{mRNA:rRNA} < 0$). Intra-molecular folding within the mRNA is allowed. All possible hybridizations between the mRNA and 16S rRNA are considered to find the highest affinity 16S rRNA binding site. The binding site minimizes the sum of the hybridization free energy $\Delta G_{mRNA:rRNA}$ and the penalty for non-optimal spacing $\Delta G_{spacing}$. Thus, the algorithm can identify the 16S rRNA binding site regardless of its similarity to the consensus Shine-Dalgarno sequence.

The ΔG_{start} term is the energy released when the start codon and the initiating tRNA anti-codon loop – 3′-UAC-5′ – hybridize together. The $\Delta G_{spacing}$ is the free energy penalty caused by a non-optimal physical distance between the 16S rRNA binding site and the start codon ($\Delta G_{spacing} > 0$). When this distance is increased or decreased from an optimum of 5 nt (or ∼17 Å) 29, the 30S complex becomes distorted, resulting in a decreased translation initiation rate.

The ΔG_{mRNA} is the work required to unfold the mRNA subsequence when it folds to its most stable secondary structure, called the minimum free energy structure ($\Delta G_{mRNA} < 0$). The $\Delta G_{standby}$ is the work required to unfold any secondary structures sequestering the standby site ($\Delta G_{standby} < 0$) after 30S complex assembly. We define the standby site as the 4 nucleotides upstream of the 16S rRNA-binding site, which is its location in a previously studied mRNA 28.

To calculate the $\Delta G_{mRNA:rRNA}$, ΔG_{start}, ΔG_{mRNA}, and $\Delta G_{standby}$ free energies, we use the NUPACK suite of algorithms, developed by Pierce and coworkers 32, with the Mfold 3.0 RNA energy parameters 22, 23. These free energy calculations do not have any additional fitting or training parameters and explicitly depend on the mRNA sequence. In addition, the free energy terms are not orthogonal; changing a single nucleotide can potentially affect multiple energy terms.

We designed a series of experiments to quantify the relationship between the aligned spacing *s* and the free energy penalty $\Delta G_{spacing}$. Thirteen synthetic RBSs are created where the aligned spacing is varied from 0 to 15 nucleotides while verifying that the $\Delta G_{mRNA:rRNA}$, ΔG_{mRNA}, ΔG_{start}, and $\Delta G_{standby}$ free energies remain constant (Supplementary Table I). The translation initiation rates of RBS sequences are measured using a fluorescent protein measurement system (**Methods**). Steady-state fluorescence measurements are performed on *E. coli* cultures over a 24 hour period. Under these conditions, the average fluorescence measurement is expected to be proportional to the translation initiation rate *r*.

The quantitative relationship between the aligned spacing and $\Delta G_{spacing}$ is obtained from the fluorescence measurements (**Methods**). According to the data, it is conceptually useful to treat the 30S complex as a model barbell connected by a rigid spring, where either stretching or compressive forces cause a reduction in entropy and an increase in the $\Delta G_{spacing}$ penalty. We empirically fit these measured $\Delta G_{spacing}$ values to either a quadratic ($s > 5$ nt) or a sigmoidal function ($s < 5$ nt). Following this parameterization, we tested the accuracy of these equations on an additional set of synthetic RBS sequences (Supplementary Figure 2).

For an arbitrary mRNA transcript, the thermodynamic model (Equation 2) is evaluated for each AUG or GUG start codon. The algorithm considers only a subsequence of the mRNA transcript, consisting of 35 nucleotides before and after the start codon. This subsequence includes the RBS and part of the protein coding sequence. The model predictions do not improve when longer subsequences are considered (Supplementary Figure 3).

The development of the thermodynamic model makes certain assumptions. Contributions related to the ribosomal S1 protein's potential preference for pyrimidine-rich sequences are omitted from the free energy model33. The model also assumes that the reversible transition between the initial and final state of 30S complex assembly reaches chemical equilibrium on a physiologically relevant timescale and without any long-lived intermediate states. The presence of overlapping or neighboring start codons, overlapping RBS and protein coding sequences, regulatory RNA binding sites, or RNAse binding sites also pose a challenge to the predictive accuracy of the thermodynamic model. The presence of multiple in-frame start codons, each with significant translation initiation, may distort its predictive accuracy. A genetic system can be designed to avoid many of these complications.

The thermodynamic model can be used in two ways. First, it can predict the relative translation initiation rate of an existing RBS sequence for a particular protein coding sequence on an mRNA transcript. We refer to this as “reverse engineering” because the RBS sequence already exists. Second, it can be used in conjunction with an optimization algorithm to identify a synthetic RBS sequence that is predicted to translate a given protein coding sequence at a user-selected rate. We refer to this mode as “forward engineering” because it generates a *de novo* sequence according to a user's specifications.

We use the thermodynamic model to predict the translation initiation rates of 28 existing RBS sequences (Figure 2A) that were obtained from a natural genome or taken from a list of commonly used sequences (Supplementary Table I). The lengths of these sequences, as

measured by the distance from the transcriptional start site to the fluorescent protein's start codon, vary from 24 to 42 nucleotides. The steady-state protein fluorescences from the sequences are then assayed in the measurement system (**Methods**). The growth rates of the cell cultures did not correlate with protein fluorescence (Supplementary Figure 4). According to the theory (Equation 1), we expect a linear relationship between the predicted ΔG_{tot} and the log protein fluorescence. Using linear regression, the squared correlation coefficient R^2 is 0.54 with Boltzmann factor $\beta = 0.45 \pm 0.05$ mol/kcal (Figure 2B). The average error is $\langle|\Delta\Delta G|\rangle = 2.1$ kcal/mol (Figure 2C).

While these commonly used RBS sequences vary the protein expression by 1500 fold, the thermodynamic model predicts that both stronger and weaker RBSs are possible. For example, one of these RBS sequences contains a strong 16S rRNA binding site ($\Delta G_{mRNA:rRNA} = -15.2$ kcal/mol), but did not yield a high protein expression level due to a strong mRNA secondary structure and non-optimal spacing ($\Delta G_{mRNA} = -11.4$, $\Delta G_{spacing} = 1.73$ kcal/mol). By optimizing the RBS sequence towards a selected ΔG_{tot}, we gain the ability to rationally control the translation initiation rate over a wide range with a proportional effect on the protein expression level.

Using the thermodynamic model, we developed an optimization algorithm that automatically designs an RBS sequence to obtain a desired relative protein expression level. The user inputs a specific protein coding sequence and a desired translation initiation rate. The rate can be varied over five orders of magnitude on a proportional scale. Equation 1 and the experimentally measured $\beta = 0.45$ mol/kcal is used to convert the user-selected translation initiation rate into the target ΔG_{tot}. The method then generates a synthetic RBS sequence according to the desired specifications.

The design method combines the thermodynamic model of translation initiation with a simulated annealing optimization algorithm to design an RBS sequence that is predicted to have a target ΔG_{tot} (Figure 2D). The RBS sequence is initialized as a random mRNA sequence upstream of the protein coding sequence. The method then creates new mRNA sequences by inserting, deleting, or replacing random nucleotides. For each new sequence, the ΔG_{tot} is calculated and compared to the target ΔG_{tot}. The sequences are then accepted or rejected according to the Metropolis criteria and three additional sequence constraints that are based on the model's assumptions (**Methods**). The procedure continues until the synthetic sequence has a predicted ΔG_{tot} to within 0.25 kcal/mol of the target. For a given target ΔG_{tot}, multiple solutions are possible, creating an ensemble of degenerate RBS sequences. The characterization of these ensembles is described in the Supplementary Discussion.

The forward design method is tested by generating 29 synthetic RBS sequences (Supplementary Table I) and comparing their predicted ΔG_{tot} values to the measured protein fluorescences. The coding sequence for a red fluorescent protein is specified and the ΔG_{tot} target is varied from −7.1 to 16.0 kcal/mol. The design method then generates a synthetic RBS sequence for each target ΔG_{tot}. These RBS sequences vary in length from 16 to 35 nucleotides and were highly dissimilar. The steady-state protein fluorescence for each sequence is measured (**Methods**). The growth rates of the cell cultures did not significantly

vary across sequences (Supplementary Figure 4). As expected from the theory (Equation 1), we obtain a linear relationship between the log protein fluorescence and the predicted ΔG_{tot} with $\beta = 0.45 \pm 0.01$ ($R^2 = 0.84$) (Figure 2E). The average error is $\langle|\Delta\Delta G|\rangle = 1.82$ kcal/mol, corresponding to a 2.3-fold error in the protein expression level. The probability distribution of the $\Delta\Delta G$ for a synthetic RBS is well fit by a Gaussian distribution (Figure 2F).

We next tested the ability of the design method to control the translation initiation rates of different proteins. Two chimeric proteins are constructed that fused the first 27 nucleotides from commonly used transcription factors to a red fluorescent protein ($TetR_{27}$-RFP and $AraC_{27}$-RFP). The design method is then used to generate 23 synthetic RBSs with ΔG_{tot} targets ranging from −8.5 to 10.5 kcal/mol (Supplementary Table I). The thermodynamic model correctly predicts the translation initiation rates of the $TetR_{27}$-RFP ($R^2 = 0.54$) and $AraC_{27}$-RFP ($R^2 = 0.95$) chimeric protein coding sequences (Figure 3A). Notably, the linear relationship between the predicted ΔG_{tot} and the log protein fluorescence yields a similar slope $\beta = 0.45 \pm 0.05$ mol/kcal.

A common practice is to reuse the same well-characterized RBS sequence for the expression of different proteins. Interestingly, the thermodynamic model predicts that this can yield dramatically different translation initiation rates. This absence of modularity will occur when the RNA sequence, containing the RBS, forms strong secondary structures with one protein coding sequence, but not another 30.

We designed experiments to test the model's ability to predict the impact of changing the protein coding sequence on the translation initiation rate. We use the design method to generate 14 synthetic RBS sequences; these sequences are then placed upstream of two different protein coding sequences: the fluorescent protein (RFP) and a chimeric fluorescent protein (TF-RFP: $LacI_{27}$-RFP, $TetR_{27}$-RFP, or $AraC_{27}$-RFP). The optimization procedure for these synthetic RBSs was modified to maximize the objective function $|\Delta G_{RFP} - \Delta G_{TF\text{-}RFP}|$, where ΔG_{RFP} and $\Delta G_{TF\text{-}RFP}$ are the predicted ΔG_{tot}'s when the RBS sequence is placed upstream of either the RFP or TF-RFP protein coding sequences, respectively. As predicted by the model, the translation initiation rates of these synthetic RBS sequences greatly change when they are reused with different protein coding sequences (Figure 3B); for example, replacing the fluorescent protein with the $TetR_{27}$-RFP chimera resulted in a 530-fold increase in expression level.

The thermodynamic model can accurately predict these differences in translation initiation rate when the correct protein coding sequence is specified ($R^2 = 0.62$ and 0.51, Figure 3C). When the incorrect protein coding sequence is used, the translation initiation rate is not accurately predicted ($R^2 = 0.04$, 0.02). Consequently, when designing a RBS sequence, the beginning of the protein coding sequence must be included in the thermodynamic calculations.

Altogether, 119 predictions of the design method were tested, revealing that the translation initiation rate can be controlled over at least a 100,000-fold range. The thermodynamic model is most accurate when all free energy terms are included in the ΔG_{tot} calculation (Supplementary Figure 5). By themselves, each free energy term is a poor predictor of the

translation initiation rate (Supplementary Figure 6) and excluding one free energy term from the ΔG_{tot} calculation results in a poorer prediction (Supplementary Figure 7). According to the distribution of the method's error (Figure 2F), an optimized RBS sequence has a 47% probability of expressing a protein to within 2-fold of the target. The probability increases to 72%, 85%, or 92% by generating two, three, or four optimized RBS sequences with identical target translation initiation rates (Supplementary Discussion).

We now demonstrate how combining the design method with a quantitative model of a genetic system enables the efficient optimization of its RBS sequences towards a targeted system behavior. Here, our objective is to optimize the connection between the arabinose-sensing P_{BAD} promoter and an AND gate genetic circuit7. The AND gate genetic circuit is regulated by the expression levels of two input promoters (P_{BAD} and P_{sal}) and controls the expression level of an output gene, which is selected to be a *gfp* reporter (Figure 4A). The desired AND logic requires that the output gene is only expressed when both input promoters are active. The digital accuracy of the AND logic is highest when the maximum expression level from the P_{BAD} promoter is an optimal value between underexpression and overexpression. When the promoter is underexpressed, the *gfp* expression is never turned on; when overexpressed, transcriptional leakiness causes *gfp* expression to turn on even in the input's absence.

The quantitative model relates the RBS sequence downstream of the P_{BAD} promoter to the accuracy of the AND gate genetic circuit's function (Figure 4B). We use previously characterized transfer functions7 to relate the arabinose and salicylate concentrations to the expression levels of the P_{BAD} and P_{sal} promoters (I_1 and I_2) (Supplementary Figure 8). The P_{BAD} promoter has a maximum protein expression level of g_{ref} = 590 au at full induction (x = 1.3 mM arabinose) and when using an RBS sequence with a predicted ΔG_{tot} value of ΔG_{ref} = −1.05 kcal/mol. We then substitute I_1 and I_2 into the AND gate genetic circuit's transfer function to determine the output gene's expression level, which is in turn substituted into the fitness function F that quantifies the ability of the genetic system to carry out AND logic (Supplementary Methods).

We can interconvert between the maximum protein expression level of the P_{BAD} promoter and the predicted ΔG_{tot} of its RBS sequence according to the equation,

$$g = g_{\mathrm{ref}} \exp\left(-\beta\left(\Delta G_{\mathrm{tot}} - \Delta G_{\mathrm{ref}}\right)\right) \quad (3)$$

where g is called the gain. The experimentally measured β = 0.45 mol/kcal is utilized. Consequently, we create a quantitative curve $F(\Delta G_{tot})$ that relates the predicted ΔG_{tot} of the P_{BAD} promoter's RBS sequence to the fitness of the genetic system. The fitness curve identifies an optimal region at ΔG_{tot} = −1.17 ± 2 kcal/mol where the genetic system will exhibit the best AND logic with respect to the P_{BAD} promoter's RBS sequence (Figure 4B).

Using the forward engineering mode of the design method, we then generate 2 synthetic RBS sequences targeted to the optimum region of the genetic system's function (predicted ΔG_{tot} = −1.48 and −1.15 kcal/mol). We also design 7 additional synthetic RBSs to test the accuracy of the $F(\Delta G_{tot})$ fitness curve, where the ΔG_{tot} ranged from 0.60 to 17.2 kcal/mol.

Each RBS sequence (32 to 35 nt) is inserted downstream of the P_{BAD} promoter and the resulting genetic circuit's response to varying inducer concentrations is assayed (Figure 4C and **Methods**). The fitness values of these rationally mutagenized genetic systems are then compared to the predictions of the model and design method (Figure 4B). The insertion of two stronger RBS sequences (ΔG_{tot} = −2.5 and −3.0 kcal/mol) cause the genetic system to exhibit a fatal growth defect.

Both optimally designed synthetic RBS sequences result in a successful connection between the arabinose-sensing P_{BAD} promoter and the AND gate genetic circuit (mean fitness > 0.85, Figure 4B). The experimentally determined optimum in the $F(\Delta G_{tot})$ curve is nearby ΔG_{tot} = 0.60 kcal/mol, which is only a 1.8 kcal/mol deviation from the model's prediction. The quantitative model and design method also correctly predict how the fitness of the genetic system deteriorates with an increasing ΔG_{tot}. Thus, our approach enabled us to rationally connect two synthetic genetic circuits together to obtain a target behavior while performing only a few mutations and assays (additional design calculations are located in the Supplementary Discussion).

A central goal of synthetic biology is to program cells to carry out valuable functions. As we construct larger and more complicated genetic systems, models and optimization techniques will be required to efficiently combine genetic parts to achieve a target behavior. To accomplish this, biophysical models that link the DNA sequence of a part to its function will be necessary. As engineered genetic systems scale to the size of genomes, the integration of multiple design methods will enable the design of synthetic genomes on a computer to control cellular behavior.

Materials and Methods

Software Implementation

A software implementation of the design method has been named the RBS Calculator and is available at http://voigtlab.ucsf.edu/software. Visitors may use the RBS Calculator in two ways: first, to predict the translation initiation rate of each start codon on an mRNA sequence (reverse engineering); second, to optimize the sequence of a ribosome binding site to rationally control the translation initiation rate with a proportional effect on the protein expression level (forward engineering). The translation initiation rate is gauged on a proportional scale with a suggested range of 0.1 to 100000, although a larger range is potentially feasible. In reverse engineering mode, the software will warn visitors when ribosome binding sites fail to satisfy the sequence constraints or contain additional sequence complications.

A thermodynamic model of translation initiation

The mRNA subsequence S_1 consists of the max(1, n_{start} − 35) to n_{start} nucleotides and the subsequence S_2 consists of the max(1, n_{start} − 35) to n_{start} + 35 nucleotides, where n_{start} is the position of a start codon. The ΔG_{start} is −1.19 and −0.075 kcal/mol for AUG and GUG start codons, respectively 22.

Using the NuPACK 'subopt' algorithm 32 with Mfold 3.0 parameters at 37°C 22, 23, base pair configurations of the folded 16S rRNA and sequence S_1 are enumerated, starting with the minimum free energy (mfe) configuration and continuing with suboptimal configurations, each with a corresponding $\Delta G_{mRNA:rRNA}$. For each configuration, the aligned spacing between the 16S rRNA binding site and start codon is calculated according to $s = n_{start} - n_1 - n_2$, where n_1 and n_2 are the rRNA and mRNA nucleotide positions in the farthest 3′ base pair in the 16S rRNA binding site. When the 30S complex is stretched (s > 5 nt), the $\Delta G_{spacing}$ is calculated according to the quadratic equation,

$$\Delta G_{\mathrm{spacing}} = c_1 (s - s_{\mathrm{opt}})^2 + c_2 (s - s_{opt}), \quad (4)$$

where s_{opt} = 5 nt, c_1 = 0.048 kcal/mol/nt^2, and c_2 = 0.24 kcal/mol/nt. When the 30S complex is compressed (s < 5 nt), the $\Delta G_{spacing}$ is calculated according to the sigmoidal function,

$$\Delta G_{\mathrm{spacing}} = \frac{c_1}{[1+\exp(c_2 (s - s_{\mathrm{opt}} + 2))]^3}, \quad (5)$$

where c_1 = 12.2 kcal/mol and c_2 = 2.5 nt $^{-1}$. The above parameter values are determined by minimizing the difference between the $\Delta G_{spacing}$ values calculated from the experimental measurements (Supplementary Figure 2) and the evaluation of Equation 4 or 5. For each configuration, the $\Delta G_{spacing}$ is added to the $\Delta G_{mRNA:rRNA}$. The configuration in the list with the lowest free energy is then identified as containing the predicted 16S rRNA binding site with a corresponding $\Delta G_{mRNA:rRNA}$. The protein coding sequence is excluded from S_1 because ribosome binding excludes the formation of downstream secondary structures.

Using the NuPACK 'mfe' algorithm and Mfold parameters, the mfe configuration of sequence S_2 is calculated and its free energy is designated ΔG_{mRNA}. The standby site is the 4 nt region upstream of the 16S rRNA binding site. The energy required to unfold the standby site is determining by calculating the mfe of sequence S_2 with and without preventing the standby site from forming base pairs. The difference between these mfe's is designated $\Delta G_{standby}$. To calculate the mfe of sequence S_2 with a standby site that is constrained to be single-stranded, the sequence is first split into two subsequences, their mfes are each calculated, and then summed together. The two subsequences are the nucleotides $n_{start} - 35$ to $n_3 - 4$ and n_3 to $n_{start} + 35$, where n_3 is the most 5′ base pair in the 16S rRNA binding site and 4 is the standby site length.

The five energy terms are summed together to calculate the ΔG_{tot}. Notably, selecting an alternate reference energy state simply adds a sequence-independent constant to the predicted ΔG_{tot}, which becomes indistinguishable from the proportionality factor K.

The simulated annealing optimization algorithm

An initial RBS sequence is randomly generated and inserted in between a pre-sequence and protein coding sequence to create a sequence S. The ΔG_{tot} of the sequence S is calculated and the objective function $O_{old} = |\Delta G_{tot} - \Delta G_{target}|$ is evaluated. In an iterative procedure, the simulated annealing optimization algorithm randomly deletes, inserts, or replaces a

nucleotide in the RBS sequence. The ΔG_{tot} and objective function O_{new} are then recalculated. If the ΔG_{tot} calculation of S invalidates the sequence constraints, then the mutation is immediately rejected. Otherwise, the mutation is accepted with probability $\max(1, \exp([O_{old} - O_{new}]/T_{SA}))$, where T_{SA} is the simulated annealing temperature. The T_{SA} is continually adjusted to maintain a 5% to 20% acceptance rate.

There are three sequence constraints that prevent the optimization algorithm from generating a synthetic RBS sequence that may invalidate one of the thermodynamic model's assumptions. The first constraint calculates the energy required to unfold the 16S rRNA binding site on the mRNA sequence and rejects the ones that require more than 6 kcal/mol to unfold. The second constraint quantifies the presence of long-range nucleotide interactions. According to a growth model for random RNA sequences 34, the equilibrium probability of nucleotides i and j forming a base pair in solution is proportional to $p = |i - j|^{-1.44}$. For each base pair in sequence S, we calculate p. If the minimum p is less than 6×10^{-3} then the sequence is rejected. Finally, the creation of new AUG or GUG start codons within the RBS sequence is disallowed.

Strains, media, and plasmid construction

The Luria-Bertani (LB) media (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) is obtained from Fisher Scientific (Pittsburgh, PA). The supplemented minimal media contains M9 minimal salts (6.8 g/L Na_2PO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl) from Sigma, 2 mM MgSO4 (Fischer Scientific), 100 μM $CaCl_2$ (Fischer Scientific), 0.4% glucose (Sigma), 0.05 g/L leucine (Acros Organics, Belgium), 5 μg/mL chloramphenicol (Acros Organics), and an adjusted pH of 7.4. The expression system is a ColE1 vector with chloramphenicol resistance (derived from pProTet, Clontech). The expression cassette contains a σ^{70} constitutive promoter (BioBrick J23100), the RBS sequence, followed by the mRFP1 fluorescent protein reporter. XbaI and SacI restriction sites are located before the RBS and after the start codon. An RBS with a desired sequence is inserted into the expression vector using standard cloning techniques. Pairs of complementary oligonucleotides are designed with XbaI and SacI overhangs and the vector is digested with XbaI and SacI restriction enzymes (NEB, Ipswich, MA). Ligation of the annealed oligonucleotides with cut vector results in a nicked plasmid, which is transformed into *E. coli* DH10B cells. Sequencing is used to verify a correct clone.

The AND gate genetic circuit is composed of three plasmids: pBACr-AraT7940, pBR939b, and pAC-SalSer914 with kanamycin, ampicillin, and chloramphenicol resistance markers, respectively. The P_{BAD} promoter maximum expression level was modified by inserting designed synthetic RBSs on plasmid pBACr-AraT7940. Plasmid pBACr-AraT7940 was digested with BamHI and ApaLI enzymes and pairs of oligonucleotides were designed to contain the desired RBS sequence and corresponding overhangs. Ligation, transformation, selection, and sequencing proceeded as described above.

Growth and fluorescence measurements

The fluorescent protein measurement system is composed of a constitutive promoter, a sequence containing a RBS, and the mRFP1 fluorescent protein reporter (Supplementary

Figure 9). An annotated DNA sequence of the system (Genbank format) is available in the Supplementary Data.

Growth and fluorescence measurements are performed in 96-well high throughput format. A 96-well plate containing 200 µl LB and 50 µg/ml chloramphenicol is inoculated, from single colonies, with up to 30 different DH10B *E. coli* cultures in an alternating, staggered pattern that excludes the outer wells. Cultures are incubated overnight at 37°C with 250 RPM orbital shaking. A fresh 96-well plate containing 200µl supplemented minimal media is inoculated by overnight cultures using a 1:100 dilution. This plate is then incubated at 37°C in a Safire2 plate spectrophotometer (Tecan) with high orbital shaking. OD_{600} measurements are recorded every 3 minutes. Once a culture reaches an OD_{600} of 0.15 to 0.20 (4 to 6 hours), a sample of each culture is transferred to a new plate containing 200 µl PBS and 2 mg/ml kanamycin (Acros Organics) for flow cytometry measurements. This media replacement strategy is repeated twice more using fresh, pre-warmed plates containing supplementary minimal media (the first with a 1:10 dilution requiring 8 to 10 hours of growth and the second with a 1:7 dilution requiring 13 to 15 hours of growth). At least three samples are taken for each culture. The fluorescence distribution of each sample is measured with a LSRII flow cytometer (BD Biosciences). We use an ellipse in forward and side scatter space to gate at least 30 000 flow cytometer events. All distributions are unimodal. The autofluorescence distribution of DH10B cells is also measured. The arithmetic mean of each distribution is taken and the mean autofluorescence is substracted.

From single colonies, RBS variants of each AND gate genetic circuit are grown overnight in LB and antibiotics (50 µg/ml ampicillin, 25 µg/ml chloramphenicol, and 25 µg/ml kanamycin). A 96-well plate containing 200 µl LB, antibiotics, and sixteen different inducer concentrations (combinations of 0.0, 1.3×10^{-3}, 8.3×10^{-2}, and 1.3 mM arabinose with 0.0, 6.1×10^{-4}, 3.9×10^{-2}, and 0.62 mM sodium salicylate) are inoculated by overnight cultures using a 1:100 dilution. Plates are grown in a Safire2 plate spectrophotometer (Tecan) with high orbital shaking. OD_{600} and *gfp* fluorescence measurements are recorded every 10 minutes for 14 hours. Background autofluorescence is subtracted from each fluorescence measurement. This procedure is repeated twice for each variant. For each variant, the average and standard deviation of the fluorescence per OD_{600} for each inducer concentration at the final time point are then calculated.

Data analysis

The $\Delta G_{spacing}$ is inferred from the fluorescent protein expression data E in the following way. The RNA sequences used to parameterize the model of $\Delta G_{spacing}$ are predicted to have identical ΔG_{mRNA}, $\Delta G_{mRNA:rRNA}$, $\Delta G_{standby}$, and ΔG_{start} free energies. According to Equation 1, dividing the expression of a sequence with spacing s_1 over another with spacing s_2 and rearranging then yields the relation: $\Delta G_{spacing}(s_1) - \Delta G_{spacing}(s_2) = -\beta^{-1}\log(E_1/E_2)$. The fluorescent protein expression at $s = 5$ nt was considered maximal and $\Delta G_{spacing}(s = 5)$ is accordingly set to zero. Using an experimentally measured value of $\beta = 0.45$ mol/kcal, the model of $\Delta G_{spacing}$ for each s is then determined.

Linear regression is used to determine the accuracy of the theory, which hypothesizes a linear relationship between the log average protein fluorescence E and the predicted ΔG_{tot}

data. The squared correlation coefficient R^2 and slope $-\beta$ are calculated according to $-\beta = (N\Sigma(x_i y_i) - \Sigma x_i \Sigma y_i) / (N\Sigma(x_i^2) - (\Sigma x_i)^2)$ and $R^2 = (N\Sigma(x_i y_i) - \Sigma x_i \Sigma y_i)^2 / [(N\Sigma(x_i^2) - (\Sigma x_i)^2)(N\Sigma(y_i^2) - (\Sigma y_i)^2)]$, where N is the number of average expression levels recorded, y is log *E*, and x is ΔG_{tot}. The standard deviation of β is calculated by substituting the log *E* data with the $\log(E+\delta E)$ and $\log(E-\delta E)$ data (δE : standard deviation of *E*) and calculating the average difference.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to all members of the Voigt lab for technical advice and continued support. This work is supported by the Pew and Packard Foundations, Office of Naval Research, NIH EY016546, NIH AI067699, NSF BES-0547637, NSF TeraGrid TG-MCB080126T, and a Sandler Family Opportunity Award. C.A.V, H.S, and E.M. are part of the NSF SynBERC ERC (www.synberc.org). E.M is supported by an NSF Graduate Research Fellowship and an ASEE National Defense Science and Engineering Graduate Fellowship.

References

1. Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R. A synthetic multicellular system for programmed pattern formation. Nature. 2005; 434:1130–1134. [PubMed: 15858574]
2. Stricker J, et al. A fast, robust and tunable synthetic gene oscillator. Nature. 2008; 456:516–519. [PubMed: 18971928]
3. Friedland AE, et al. Synthetic gene networks that count. Science. 2009; 324:1199–1202. [PubMed: 19478183]
4. Ellis T, Wang X, Collins JJ. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat Biotechnol. 2009; 27:465–471. [PubMed: 19377462]
5. Yokobayashi Y, Weiss R, Arnold FH. Directed evolution of a genetic circuit. Proc Natl Acad Sci U S A. 2002; 99:16587–16591. [PubMed: 12451174]
6. Tabor JJ, et al. A synthetic genetic edge detection program. Cell. 2009; 137:1272–1281. [PubMed: 19563759]
7. Anderson JC, Voigt CA, Arkin AP. Environmental signal integration by a modular AND gate. Mol Syst Biol. 2007; 3:133. [PubMed: 17700541]
8. Dueber JE, et al. Synthetic protein scaffolds provide modular control over metabolic flux. Nat Biotechnol. 2009; 27:753–759. [PubMed: 19648908]
9. Anthony JR, et al. Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Escherichia coli for production of the anti-malarial drug precursor amorpha-4,11-diene. Metab Eng. 2008
10. Atsumi S, Hanai T, Liao JC. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature. 2008; 451:86–89. [PubMed: 18172501]
11. Hawkins KM, Smolke CD. Production of benzylisoquinoline alkaloids in Saccharomyces cerevisiae. Nat Chem Biol. 2008; 4:564–573. [PubMed: 18690217]
12. Lee KH, Park JH, Kim TY, Kim HU, Lee SY. Systems metabolic engineering of Escherichia coli for L-threonine production. Mol Syst Biol. 2007; 3:149. [PubMed: 18059444]
13. Lutke-Eversloh T, Stephanopoulos G. Combinatorial pathway analysis for improved L-tyrosine production in Escherichia coli: identification of enzymatic bottlenecks by systematic gene overexpression. Metab Eng. 2008; 10:69–77. [PubMed: 18243023]
14. Czar MJ, Anderson JC, Bader JS, Peccoud J. Gene synthesis demystified. Trends Biotechnol. 2009; 27:63–72. [PubMed: 19111926]
15. Gibson DG, et al. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science. 2008; 319:1215–1220. [PubMed: 18218864]

16. Isaacs FJ, et al. Engineered riboregulators enable post-transcriptional control of gene expression. Nat Biotechnol. 2004; 22:841–847. [PubMed: 15208640]
17. Carrier TA, Keasling JD. Library of synthetic 5′ secondary structures to manipulate mRNA stability in Escherichia coli. Biotechnol Prog. 1999; 15:58–64. [PubMed: 9933514]
18. Pfleger BF, Pitera DJ, Smolke CD, Keasling JD. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. Nat Biotechnol. 2006; 24:1027–1032. [PubMed: 16845378]
19. Chubiz LM, Rao CV. Computational design of orthogonal ribosomes. Nucleic Acids Res. 2008; 36:4038–4046. [PubMed: 18522973]
20. de Smit MH, van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci U S A. 1990; 87:7668–7672. [PubMed: 2217199]
21. Vellanoweth RL, Rabinowitz JC. The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. Mol Microbiol. 1992; 6:1105–1114. [PubMed: 1375309]
22. Xia T, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998; 37:14719–14735. [PubMed: 9778347]
23. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999; 288:911–940. [PubMed: 10329189]
24. Kierzek R, Burkard ME, Turner DH. Thermodynamics of single mismatches in RNA duplexes. Biochemistry. 1999; 38:14214–14223. [PubMed: 10571995]
25. Miller S, Jones LE, Giovannitti K, Piper D, Serra MJ. Thermodynamic analysis of 5′ and 3′ single- and 3′ double-nucleotide overhangs neighboring wobble terminal base pairs. Nucleic Acids Res. 2008; 36:5652–5659. [PubMed: 18765476]
26. Christiansen ME, Znosko BM. Thermodynamic characterization of the complete set of sequence symmetric tandem mismatches in RNA and an improved model for predicting the free energy contribution of sequence asymmetric tandem mismatches. Biochemistry. 2008; 47:4329–4336. [PubMed: 18330995]
27. Laursen BS, Sorensen HP, Mortensen KK, Sperling-Petersen HU. Initiation of protein synthesis in bacteria. Microbiol Mol Biol Rev. 2005; 69:101–123. [PubMed: 15755955]
28. Studer SM, Joseph S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex. Mol Cell. 2006; 22:105–115. [PubMed: 16600874]
29. Chen H, Bjerknes M, Kumar R, Jay E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nucleic Acids Res. 1994; 22:4953–4957. [PubMed: 7528374]
30. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. Science. 2009; 324:255–258. [PubMed: 19359587]
31. de Smit MH, van Duin J. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. J Mol Biol. 2003; 331:737–743. [PubMed: 12909006]
32. Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA. Thermodynamic Analysis of Interacting Nucleic Acid Strands. SIAM Review. 2007; 49:65–88.
33. Sengupta J, Agrawal RK, Frank J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:11991–11996. [PubMed: 11593008]
34. David F, Hagendorf C, Wiese KJ. A growth model for RNA secondary structures. Journal of Statistical Mechanics: Theory and Experiment. 2008:P04008.

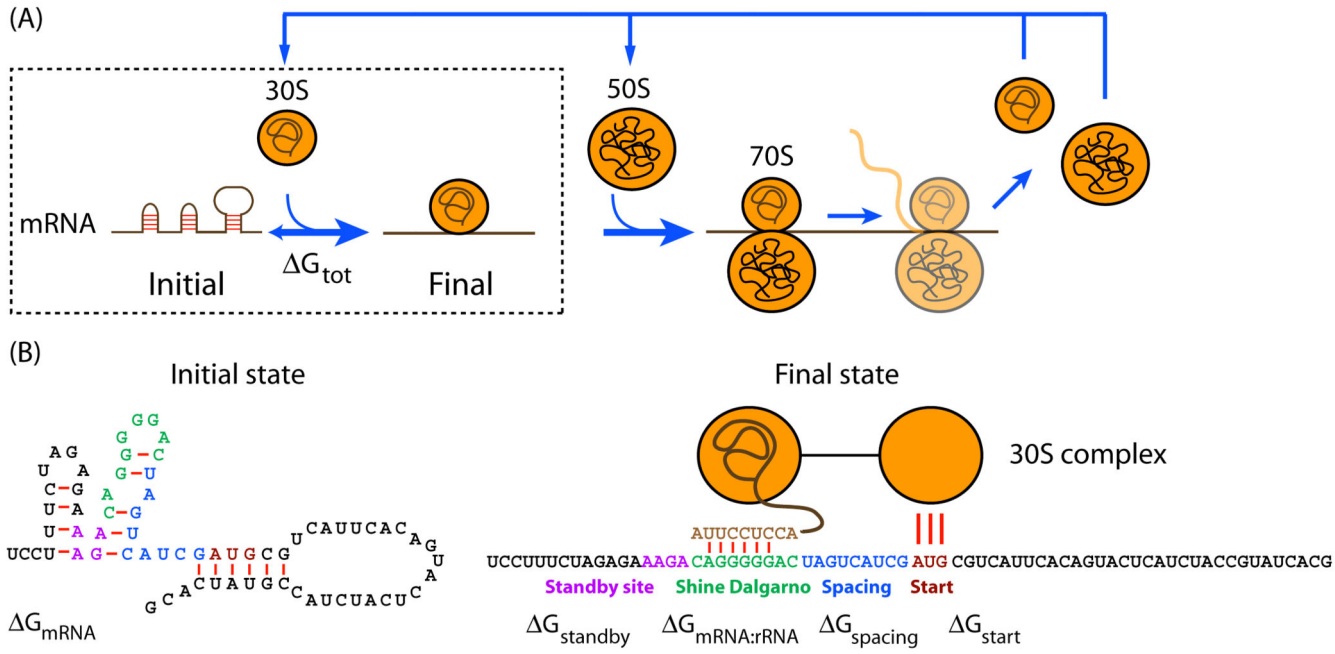

Figure 1.

A thermodynamic model of bacterial translation initiation. (A) The ribosome translates an mRNA transcript and produces a protein in a four step process: the rate-limiting assembly of the 30S pre-initiation complex, translation initiation, translation elongation, translation termination, and the turnover of ribosomal subunits and other factors. (B) The thermodynamic free energy change during the translation initiation step is determined by five molecular interactions that participate in the initial and final states of the system. See text for a description of each free energy term. The Watson-Crick base pairs and G:U wobbles (red lines) are shown.

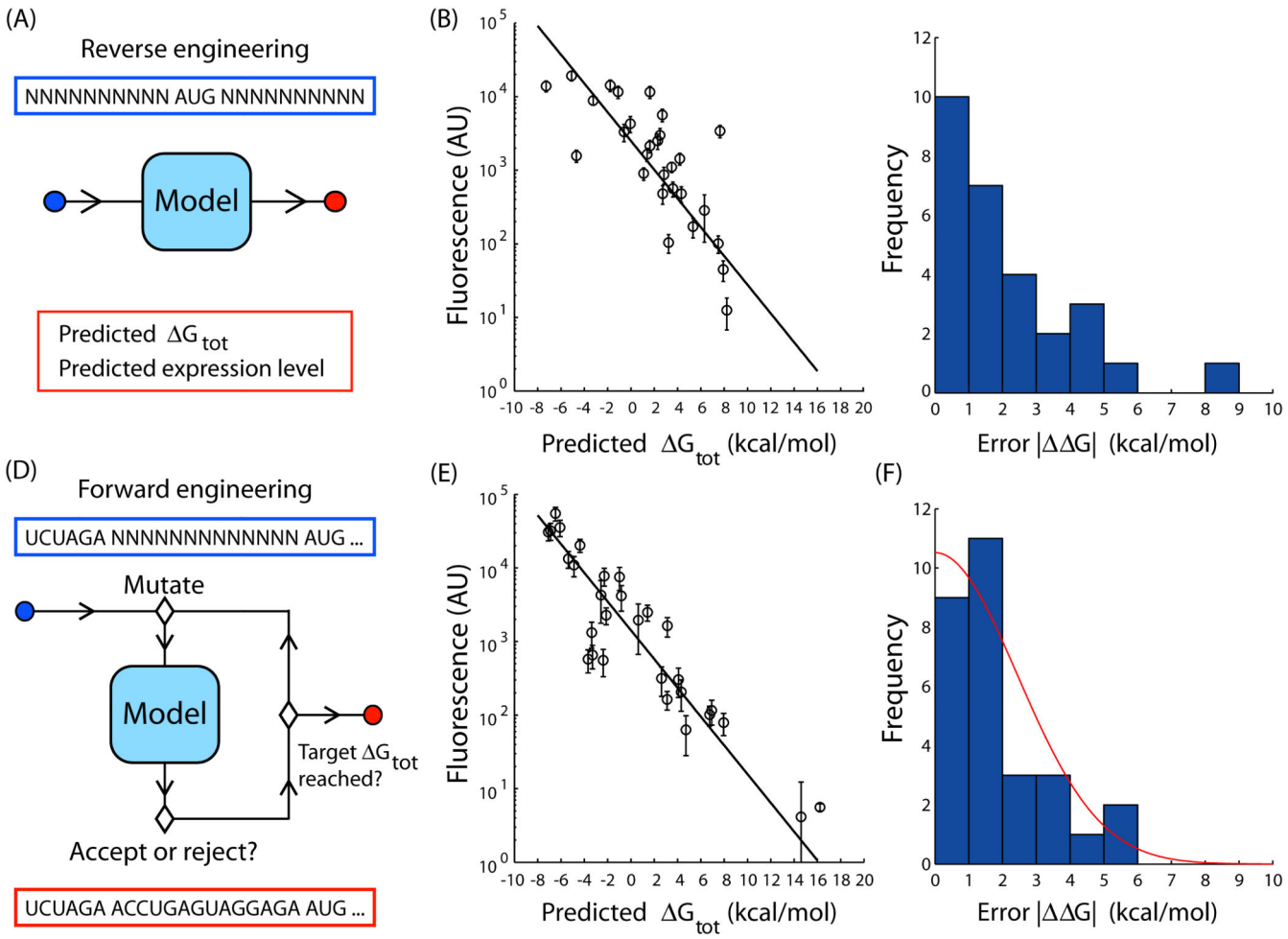

Figure 2.

The design method has two modes of operation: (A) The method can predict the relative translation initiation rate of an existing RBS when placed in front of a protein coding sequence. The method calculates the ΔG_{tot} from the input sequence. According to Equation 1, a linear relationship between the log protein fluorescence and the predicted ΔG_{tot} is expected. (B) The fluorescence levels from 28 natural or existing RBSs in front of the RFP fluorescent protein are measured (circles) and compared to the predicted ΔG_{tot} calculations. The error bars are calculated as the standard deviation of 6 measurements performed on two different days. The expected relationship is obtained (line, $R^2 = 0.54$) with a slope $\beta = 0.45 \pm 0.05$. (C) A histogram shows the distribution of error in the predicted ΔG_{tot}, denoted by $|\Delta\Delta G|$, of the sequences in B. The average of this distribution is 2.11 kcal/mol. (D) An optimization algorithm with Metropolis criteria, the sequence constraints, and simulated annealing uses iterations of mutation and selection to identify an RNA sequence that is predicted to have the target ΔG_{tot}. (E) The fluorescence levels from 29 synthetic RBSs in front of RFP are measured (circles) and compared to the predicted ΔG_{tot} calculations. The error bars are calculated as the standard deviation of at least 5 measurements performed on 2 different days. The expected linear relationship between log protein expression level and predicted ΔG_{tot} is shown (line, $R^2 = 0.84$) with slope $\beta = 0.45 \pm 0.01$. (F) A histogram shows

the distribution of the error, $|\Delta\Delta G|$. The average of the distribution is 1.82 kcal/mol and fits well to a one-sided Gaussian distribution (red line) with standard deviation $\sigma = 2.44$ kcal/mol.

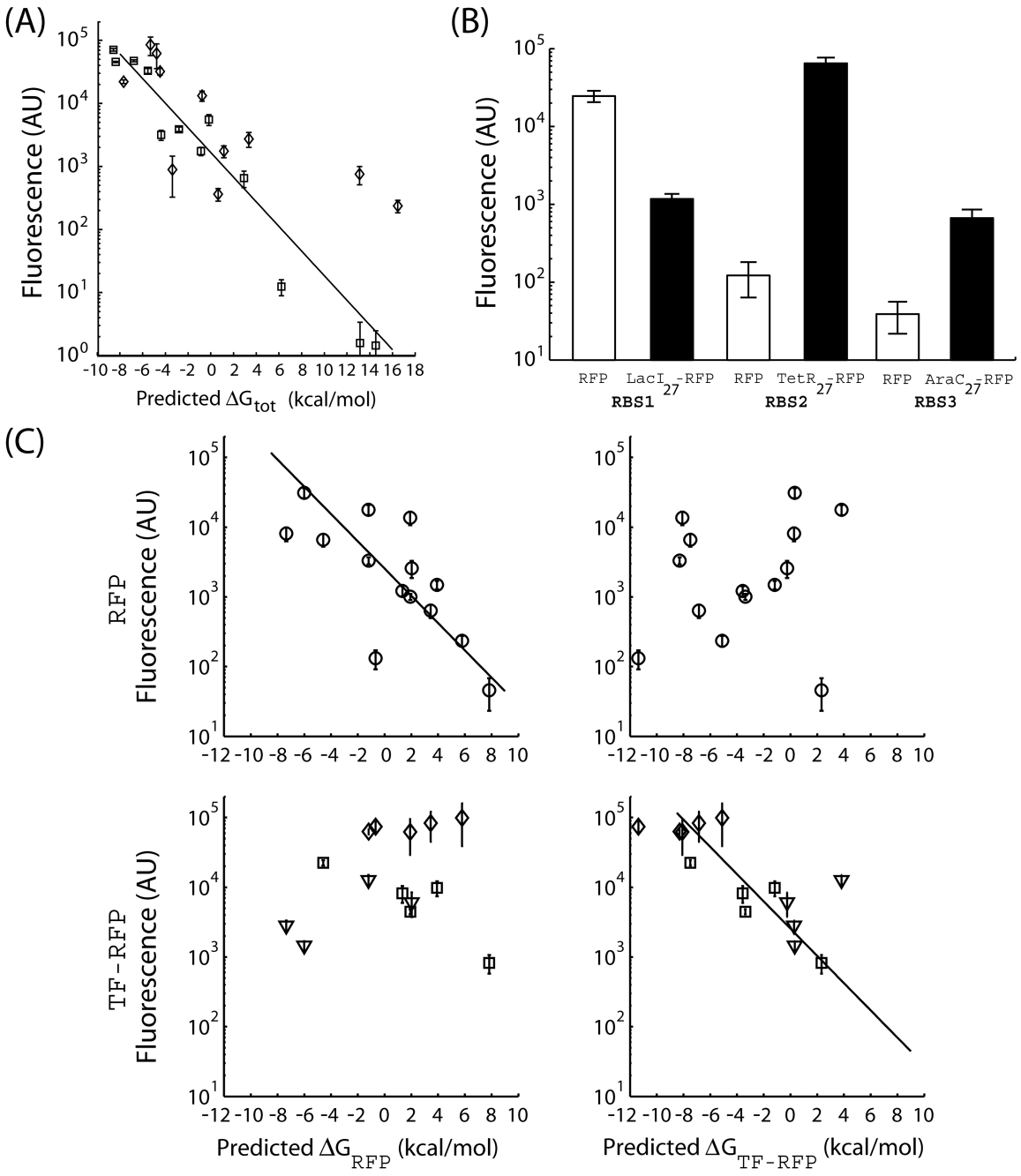

Figure 3.

The design method can control the expression level of different proteins by predicting the impact of changing the protein coding sequence. (A) The fluorescence levels from 23 synthetic RBSs in front of two different protein coding sequences are measured and compared to the predicted ΔG_{tot} calculations. The two proteins are $TetR_{27}$-RFP (diamonds) and $AraC_{27}$-RFP (squares). The expected relationship between the log protein fluorescence and the predicted ΔG_{tot} is obtained for each protein coding sequence ($TetR_{27}$-RFP, R^2=0.54; $AraC_{27}$-RFP, R^2 = 0.95). (B) Reusing the same RBS sequence with two different protein

coding sequences can alter the translation initiation. Fluorescence levels from identical RBS sequences in front of either RFP (white bars) or a chimeric fluorescent protein (either $LacI_{27}$-RFP, $TetR_{27}$-RFP, or $AraC_{27}$-RFP; black bars) are shown. (C) The design method must use the correct protein coding sequence to accurately predict the ΔG_{tot}. The fluorescence levels from 14 pairs of RBS sequences in front of either RFP (black circles) or a chimeric fluorescent protein ($LacI_{27}$-RFP, triangles; $TetR_{27}$-RFP, diamonds; $AraC_{27}$-RFP, squares) are measured. When the correct protein coding sequence is used to calculate the ΔG_{tot}, the expected relationship between log protein fluorescence and ΔG_{tot} is obtained (lines, $R_2 = 0.62$ and $R_2 = 0.51$). Otherwise, the thermodynamic model does not correctly predict the expression level ($R^2 = 0.04$ and 0.02). The error bars calculated as the standard deviation of at least 6 measurements performed on 2 different days.

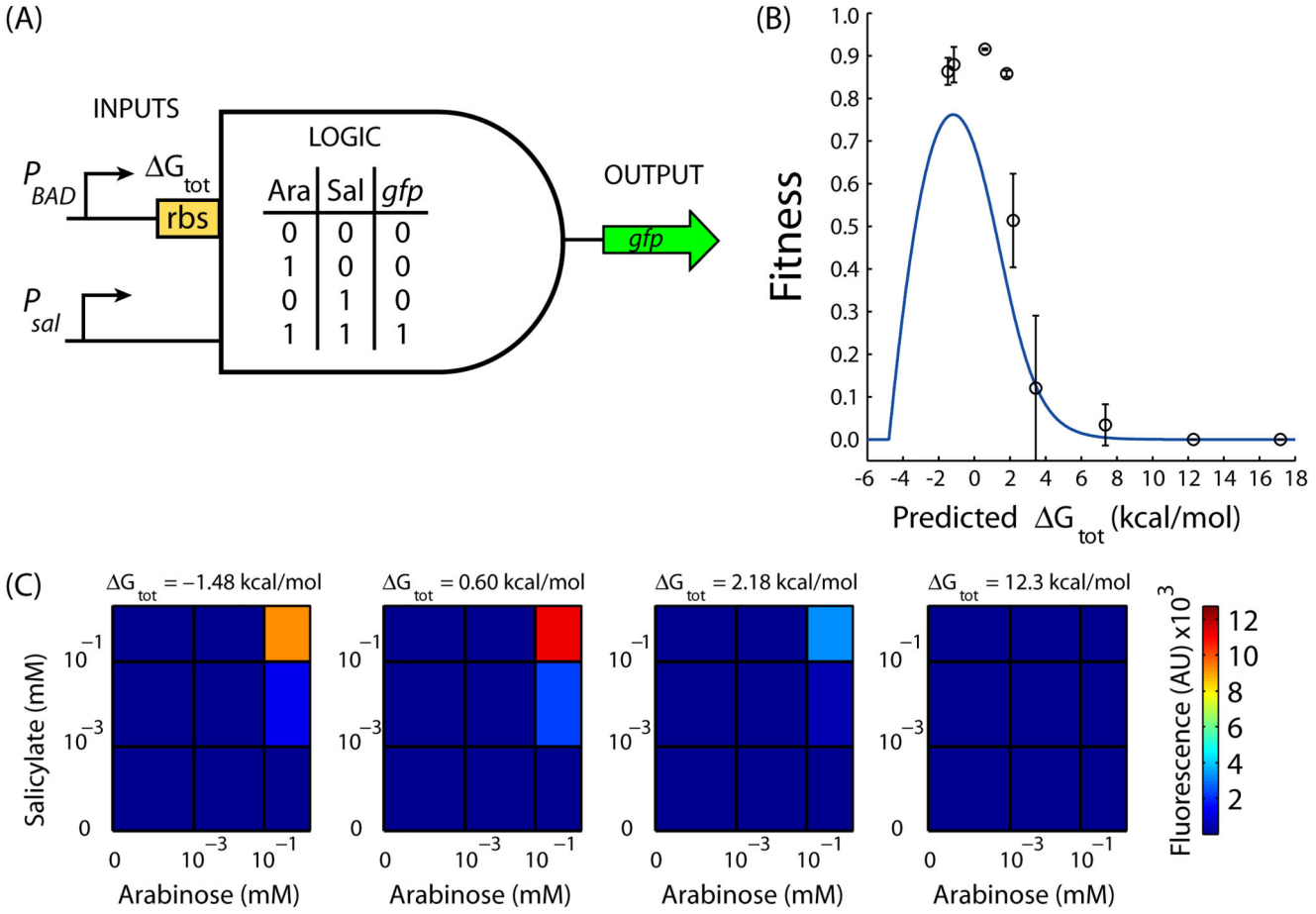

Figure 4.

Optimal connection of a sensor input to an AND gate genetic circuit. (A) A functional AND gate genetic circuit will only turn on the *gfp* reporter output when both the P_{BAD} and P_{sal} promoter inputs are sufficiently induced by arabinose and salicylate, respectively. (B) The quantitative model and design method predict a fitness curve $F(\Delta G_{tot})$ (blue line), relating the predicted ΔG_{tot} of the P_{BAD} promoter's RBS sequence to the quality of the genetic circuit's AND logic. The accuracy of this curve is tested by assaying the fitness of nine genetic circuit variants, each containing a synthetic RBS that was designed to possess a selected ΔG_{tot} (black circles). (C) The amount of *gfp* fluorescence is shown in response to combinations of arabinose (0.0, 1.3×10^{-3}, 8.3×10^{-2}, and 1.3 mM) and salicylate (0.0, 6.1×10^{-4}, 3.9×10^{-2}, and 0.62 mM) for selected AND gate genetic circuits. These genetic circuits contain RBS sequences with predicted ΔG_{tot}'s of 12.3, 2.18, 0.60, and −1.48 kcal/mol. The error bars calculated as the standard deviation of 2 measurements of fitness performed on 2 different days.