nature biotechnology

Automated design of synthetic ribosome binding sites to control protein expression

Howard M Salis[1], Ethan A Mirsky[2] & Christopher A Voigt[1]

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]. To circumvent the need for trial and error optimization, we developed a predictive method for designing synthetic ribosome binding sites, enabling a rational control over the protein expression level. Experimental validation of >100 predictions in *Escherichia coli* showed that the method is accurate to within a factor of 2.3 over a range of 100,000-fold. The design method also correctly predicted that reusing identical ribosome binding site sequences in different genetic contexts can result in different protein expression levels. We demonstrate the method's utility by rationally optimizing protein expression to connect a genetic sensor to a synthetic circuit. The proposed forward engineering approach should accelerate the construction and systematic optimization of large genetic systems.

Trial-and-error mutation to optimize an engineered genetic circuit or metabolic pathway becomes prohibitively inefficient as the system's size and complexity grows. To address this problem, we developed a predictive design method that interconverts between the DNA sequence of a key genetic element—ribosome binding sites—and their function inside a genetic system (controlling the translation initiation rate and the protein expression level). The design method's capabilities enable the systematic optimization of genetic systems, which will be increasingly valuable as it becomes possible to synthesize larger pieces of DNA[14], including whole genomes[15].

In bacteria, ribosome binding sites (RBSs) and other regulatory RNA sequences are effective control elements for translation initiation[16–19], and thereby protein expression. Previous studies have generated libraries of RBS sequences with the goal of optimizing the function of a genetic system[1,7,18]. However, library size increases combinatorially with the number of proteins in the engineered system—for example, randomly mutating four nucleotides of an RBS generates a library of 256 sequences, thus requiring 256^3, or 16.7 million, sequences for three proteins and 256^6, or 2.8×10^{14}, sequences for six proteins.

In contrast to a library-based approach, we combined a biophysical model of translation initiation with an optimization algorithm to predict the sequence of a synthetic RBS sequence that provides a target translation initiation rate on a proportional scale. The model builds on previous work that characterized the free energies of key molecular interactions involved in translation initiation[20,21] and on measurements of the sequence-dependent energetic changes that occur during RNA folding and hybridization[22–26].

Bacterial translation consists of four phases: initiation, elongation, termination and ribosome turnover (**Fig. 1a**)[27]. In most cases, translation initiation is the rate-limiting step. Its rate is determined by 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 (called spacing) and the presence of RNA secondary structures that occlude either the 16S rRNA binding site or the standby site[20,21,28–31].

Our equilibrium statistical thermodynamic model quantifies the strengths of the molecular interactions between an mRNA transcript and the 30S ribosome complex—which includes the 16S rRNA and the $tRNA^{fMET}$—to predict the resulting translation initiation rate, r (equation (1), derived in **Supplementary Methods**).

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

The model describes the system as having two states separated by a reversible transition (**Fig. 1b**). The initial state is the folded mRNA transcript and the free 30S complex. The final state is the assembled 30S pre-initiation complex bound on an mRNA transcript. The difference in Gibbs free energy between these two states (ΔG_{tot}) depends on the mRNA sequence surrounding a specified start codon. ΔG_{tot} is more negative when attractive interactions between ribosome and mRNA are present, and ΔG_{tot} is more positive when mutually exclusive secondary structures are present. β is the apparent Boltzmann constant for the system, which converts thermodynamic free energies to temperature differences. Importantly, equation (1) describes the differences in translation initiation rate that result from differences in mRNA sequence. The amount of expressed protein is proportional to the translation initiation rate where the proportionality factor 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 Fig. 1**).

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, USA. [2]Graduate Group in Biophysics, University of California San Francisco, San Francisco, California, USA. Correspondence should be addressed to C.A.V. (cavoigt@picasso.ucsf.edu).

Received 30 July; accepted 8 September; published online 4 October 2009; doi:10.1038/nbt.1568

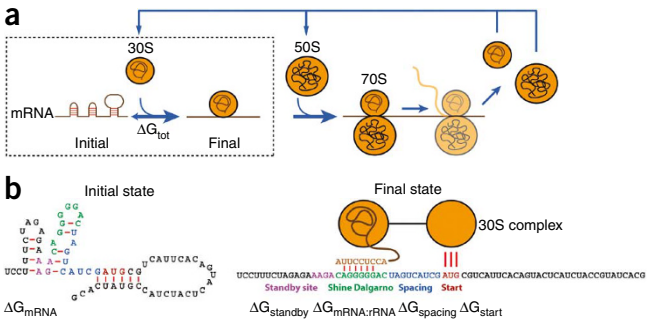

Figure 1 A thermodynamic model of bacterial translation initiation. (**a**) The ribosome translates an mRNA transcript and produces a protein in a multistep process: the assembly of the 30S complex (box), initiation, elongation, termination, and the turnover of ribosomal subunits and other factors. (**b**) The thermodynamic free energy change during 30S complex assembly is determined by five molecular interactions that participate in the initial and final states of the system. The Watson-Crick base pairs and G:U wobbles (red lines) are shown.

Given a specific mRNA sequence—called the sub-sequence—surrounding a start codon, ΔG_{tot} is predicted according to an energy model (equation (2)), where the reference state is a fully unfolded sub-sequence with G = 0.

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

$\Delta G_{mRNA:rRNA}$ is the energy released when the last nine nucleotides (nt) of the *E. coli* 16S rRNA (3′-AUUCCUCCA-5′) hybridizes and co-folds to the mRNA sub-sequence ($\Delta G_{mRNA:rRNA} < 0$). Intramolecular 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 nonoptimal spacing, $\Delta G_{spacing}$. Thus, the algorithm can identify the 16S rRNA binding site regardless of its similarity to the consensus Shine-Dalgarno sequence.

ΔG_{start} is the energy released when the start codon hybridizes to the initiating tRNA anticodon loop (3′-UAC-5′).

$\Delta G_{spacing}$ is the free energy penalty caused by a nonoptimal 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.

ΔG_{mRNA} is the work required to unfold the mRNA sub-sequence when it folds to its most stable secondary structure, called the minimum free energy structure ($\Delta G_{mRNA} < 0$).

$\Delta G_{standby}$ is the work required to unfold any secondary structures sequestering the standby site ($\Delta G_{standby} < 0$) after the 30S complex assembly. We define the standby site as the four nucleotides upstream of the 16S rRNA binding site, which is its location in a previously studied mRNA[28].

To calculate $\Delta G_{mRNA:rRNA}$, ΔG_{start}, ΔG_{mRNA} and $\Delta G_{standby}$, we use the NUPACK suite of algorithms[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. The relationship between the spacing and the $\Delta G_{spacing}$ was empirically determined by measuring the protein expression level driven by synthetic RBSs of varying spacing and fitting a quantitative model to this data (Online Methods, **Supplementary Table 1** and **Supplementary Fig. 2**).

For an arbitrary mRNA transcript, the thermodynamic model (equation (2)) is evaluated for each AUG or GUG start codon. The algorithm considers only the sub-sequence of the mRNA transcript consisting of 35 nucleotides before and after the start codon. This

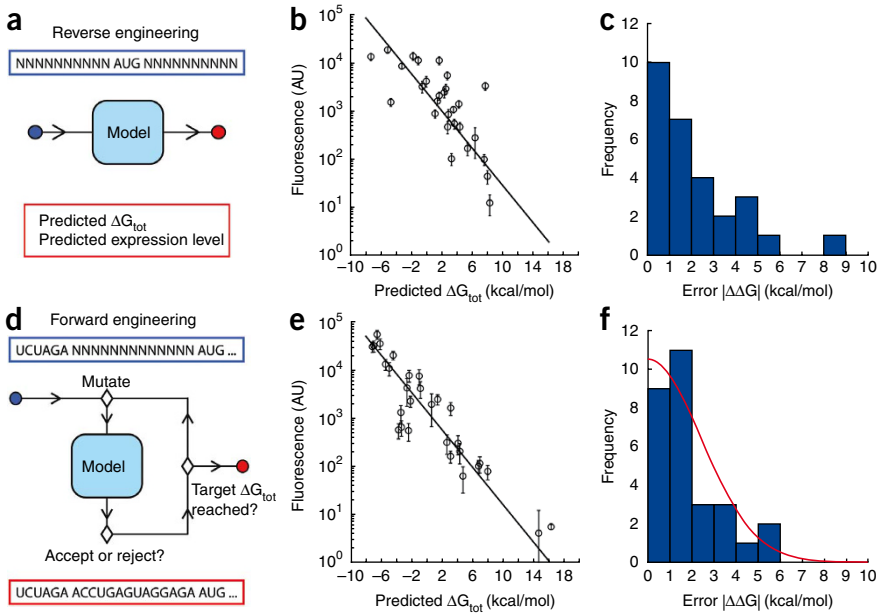

Figure 2 A ribosome binding site design method. (**a**) Reverse engineering. The method predicts the relative translation initiation rate (red) of an RBS upstream of a given protein coding sequence (blue). The ΔG_{tot} is the free energy change before and after the 30S ribosomal complex assembles on the mRNA. Equation (1) predicts a linear relationship between the log protein fluorescence and the predicted ΔG_{tot}. (**b**) Red fluorescence protein reporter expression driven by 28 natural or existing RBSs compared to predicted ΔG_{tot} calculations. Error bars are s.d. of six measurements performed on two different days. Linear regression $R^2 = 0.54$ with slope $\beta = 0.45 \pm 0.05$. (**c**) Histogram of 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**) Forward engineering. A simulated annealing optimization algorithm iteratively mutates an RNA sequence until a target ΔG_{tot} is found. (**e**) RFP expression driven by 29 synthetic RBSs compared to the predicted ΔG_{tot} calculations. Error bars are s.d. of at least five measurements performed on two different days. Linear regression $R^2 = 0.84$ with slope $\beta = 0.45 \pm 0.01$. (**f**) Histogram of the distribution of the error, $|\Delta\Delta G|$ from **e**. The average of the distribution is 1.82 kcal/mol and fits well to a one-sided Gaussian distribution (red line) with s.d. $\sigma = 2.44$ kcal/mol.

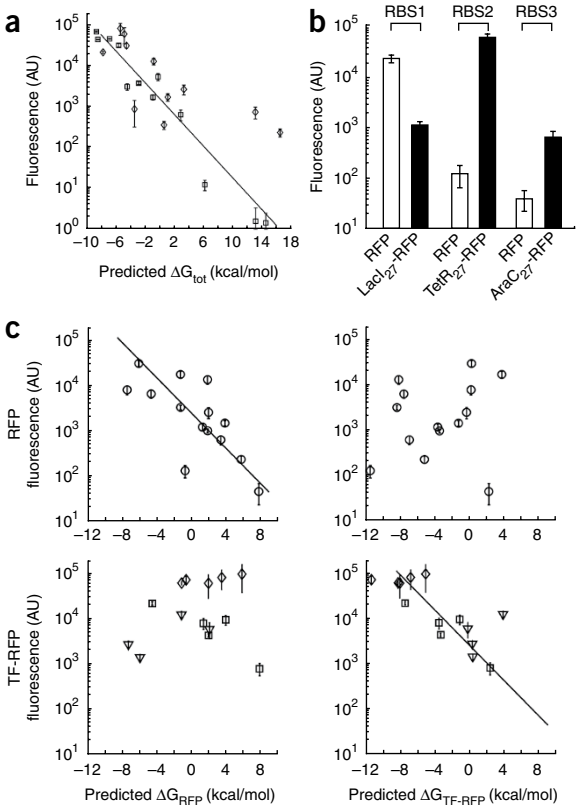

Figure 3 The RBS design method can control the expression level of different proteins by accounting for the influence of the protein coding sequence. (**a**) Expression of TetR$_{27}$-RFP (diamonds) and AraC$_{27}$-RFP (squares) fluorescence reporters driven by 23 synthetic RBSs (TetR$_{27}$-RFP, $R^2 = 0.54$; AraC$_{27}$-RFP, $R^2 = 0.95$). (**b**) Reusing the same RBS sequence with two different protein coding sequences alters translation initiation. Fluorescence levels from identical RBS sequences in front of either RFP (white bars) or a chimeric fluorescent protein (black bars). (**c**) Protein coding sequence is required to accurately predict ΔG$_{tot}$. Fluorescence levels from 14 pairs of RBS sequences in front of either RFP (black circles) or a chimeric fluorescent protein (triangles, LacI$_{27}$-RFP; diamonds, TetR$_{27}$-RFP; squares, AraC$_{27}$-RFP). When the correct protein coding sequence is used to calculate the ΔG$_{tot}$, the relationship between log protein fluorescence and ΔG$_{tot}$ is linear, as expected (upper-left, $R^2 = 0.62$; lower-right, $R^2 = 0.51$). Otherwise, the thermodynamic model does not correctly predict the expression level (lower-left, $R^2 = 0.04$; upper-right, $R^2 = 0.02$). Error bars are s.d. of at least six measurements performed on two different days.

sub-sequence includes the RBS and part of the protein coding sequence. The model predictions do not improve when longer sub-sequences are considered (**Supplementary Fig. 3**).

The thermodynamic model is based on the following assumptions. (i) Contributions related to the ribosomal S1 protein's potential preference for pyrimidine-rich sequences are omitted[33]. (ii) Chemical equilibrium of the reversible transition between the initial and final state of the 30S complex assembly is reached on a physiologically relevant timescale and without any long-lived intermediate states. (iii) The RBS does not overlap with start codons, protein coding sequences, regulatory RNA binding sites or RNAse binding sites. (iv) The protein coding sequence does not contain multiple in-frame start codons with significant amounts of translation initiation.

The thermodynamic model enables both 'reverse engineering' and 'forward engineering'. Reverse engineering involves predicting the relative translation initiation rate of an existing RBS sequence upstream of a particular protein coding sequence. Forward engineering incorporates the model into an optimization algorithm that outputs a *de novo* nucleotide sequence of a synthetic RBS that is predicted to drive protein translation at a specified rate.

To demonstrate reverse engineering, we used the model to predict the translation initiation rates of 28 existing RBS sequences (**Fig. 2a**) obtained from a natural genome or taken from a list of commonly used sequences (**Supplementary Table 1**). The sequences are 24–42 nt long, as measured by the distance from the transcriptional start site to the start codon of a fluorescent reporter protein. Flow cytometry was used to measure steady-state protein expression driven by each RBS. The growth rates of the cell cultures did not correlate with protein fluorescence (**Supplementary Fig. 4**), supporting the assumption that the translation initiation rate and protein expression level are proportional. As expected (equation (1)), the relationship between the predicted ΔG$_{tot}$ and the log protein fluorescence was linear (**Fig. 2b**). Using linear regression, the squared correlation coefficient R^2 was 0.54 with Boltzmann factor $\beta = 0.45 \pm 0.05$ mol/kcal. The average error was $\langle|\Delta\Delta G|\rangle = 2.1$ kcal/mol (**Fig. 2c**).

Although these existing RBS sequences cause 1,500-fold variations in protein expression, 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 (ΔG$_{mRNA:rRNA}$ = −15.2 kcal/mol), but did not yield a high protein expression level owing to a strong mRNA secondary structure and nonoptimal spacing (ΔG$_{mRNA}$ = −11.4, ΔG$_{spacing}$ = 1.73 kcal/mol). Therefore, we explored the possibility that optimizing the RBS sequence toward a selected ΔG$_{tot}$ would enable rational control of the translation initiation rate, and thereby protein expression, over a wider dynamic range.

To demonstrate such forward engineering, we developed an optimization approach 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 approach then generates a synthetic RBS sequence with a target ΔG$_{tot}$ by combining the thermodynamic model of translation initiation with a simulated annealing optimization algorithm (**Fig. 2d**). The RBS sequence is initialized as a random mRNA sequence upstream of the protein coding sequence. Then, new mRNA sequences are created 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 (Online 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 (characterized in the **Supplementary Discussion** and **Supplementary Fig 5**).

We tested the forward design method by generating 29 synthetic RBS sequences to drive expression of a red fluorescent protein (RFP) with target ΔG$_{tot}$ between −7.1 and 16.0 kcal/mol (**Supplementary Table 1**). These RBS sequences varied in length from 16 to 35 nucleotides and were highly dissimilar. The growth rates of the cell cultures did not significantly vary across sequences (**Supplementary Fig. 4**). As expected (equation (1)), we obtained a linear relationship between the log protein fluorescence and the predicted ΔG$_{tot}$ with $\beta = 0.45 \pm 0.01$ ($R^2 = 0.84$) (**Fig. 2e**). The average error was $\langle|\Delta\Delta G|\rangle = 1.82$ kcal/mol, corresponding to a 2.3-fold error in the protein expression level. The

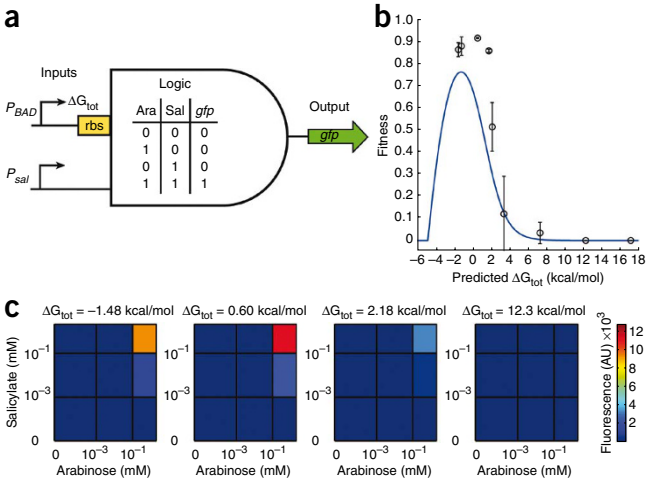

Figure 4 Optimal connection of a sensor input to an AND gate genetic circuit. (**a**) A functional AND-gate genetic circuit will only activate the *gfp* reporter output when both the P_{BAD} and P_{sal} promoter inputs are sufficiently induced by arabinose and salicylate, respectively. (**b**) The quantitative 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 accuracy of the genetic circuit's AND logic. Circles indicate the fitness of nine genetic circuit variants, each containing a synthetic RBS that was designed to possess a selected ΔG_{tot}. Error bars are s.d. of two measurements on 2 different days. (**c**) The fluorescence of *gfp* reporter 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.

probability distribution of the ΔΔG for a synthetic RBS is well fit by a Gaussian distribution (**Fig. 2f**), enabling a statistical analysis of the method's error (**Supplementary Discussion**).

We next tested the ability of the design method to control the translation initiation rates of different proteins. Two chimeric proteins were constructed that fused the first 27 nucleotides from commonly used transcription factors to an RFP ($TetR_{27}$-RFP and $AraC_{27}$-RFP). We then designed 23 synthetic RBSs with ΔG_{tot} targets ranging from −8.5 to 10.5 kcal/mol (**Supplementary Table 1**). The thermodynamic model correctly predicted 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 (**Fig. 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 key aspect of the thermodynamic model is that it explicitly depends on the sequence of the mRNA before and after the start codon. To validate the importance of this dependency, we designed 14 synthetic RBS sequences placed upstream of either RFP or one of three chimeric fluorescent proteins ($LacI_{27}$-RFP, $TetR_{27}$-RFP or $AraC_{27}$-RFP), collectively denoted TF-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 expected, the translation initiation rates of these synthetic RBS sequences change greatly when they are reused with different protein coding sequences (**Fig. 3b**); for example, replacing RFP 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, **Fig. 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 an RBS sequence, the beginning of the protein coding sequence must be included in the thermodynamic calculations. It is likely that this absence of modularity is caused by the formation of strong secondary structures between the RBS-containing RNA sequence and one protein coding sequence but not another[30]. Taken together, these results suggest that reusing the same well-characterized RBS sequence for different proteins—a common practice—is not likely to work reliably.

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 Fig. 6**). By themselves, each free energy term is a poor predictor of the translation initiation rate (**Supplementary Fig. 7**) and excluding one free energy term from the ΔG_{tot} calculation results in a poorer prediction (**Supplementary Fig. 8**). According to the distribution of the method's error (**Fig. 2f**), an optimized RBS sequence has a 47% probability of expressing a protein to within twofold of the target. The probability increases to 72%, 85% or 92% by generating two, three or four optimized RBS sequences, respectively, with identical target translation initiation rates (**Supplementary Discussion**).

Next, we efficiently optimized a complex genetic circuit by combining the RBS design method with a quantitative model of the system. Our objective was to connect the P_{BAD} promoter to the AND-gate genetic circuit[7] and maximize its ability to turn on green fluorescent protein *gfp* expression only when both input promoters (P_{BAD} and P_{sal}) are actively expressed (**Fig. 4a**). The accuracy of the circuit's logic (referred to as its fitness) is highest when the maximum expression level from the P_{BAD} promoter is optimized to a value between underexpression and overexpression. Otherwise, when the promoter is underexpressed, *gfp* expression is never turned on; and when the promoter is overexpressed, transcriptional leakiness causes *gfp* expression even when an input is absent. Our rational approach enables this optimization while minimizing the number of mutations and assays.

The quantitative model relates the RBS sequence downstream of the P_{BAD} promoter to the accuracy of the function of the AND-gate genetic circuit (**Fig. 4b**). We used previously characterized transfer functions[7] 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 Fig. 9**). We then substituted I_1 and I_2 into the transfer function of the AND-gate genetic circuit to determine the output gene's expression, which was in turn substituted into the fitness function F that quantifies the ability of the genetic system to carry out AND logic (**Supplementary Methods**).

Equation (3) defines the relationship between maximum protein expression level of the P_{BAD} promoter (called the gain, g) and the predicted ΔG_{tot} of its RBS sequence, compared to a reference P_{BAD} promoter and RBS sequence.

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

The gain from a reference P_{BAD} promoter was measured to be $g_{ref} = 590$ AU at full induction ($x = 1.3$ mM arabinose) while using a reference RBS sequence whose predicted ΔG_{tot} is $\Delta G_{ref} = -1.05$ kcal/mol. The experimentally measured value of β (0.45 mol/kcal) was used. Using the transfer functions, fitness function and equation (3), we created

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 $\Delta G_{tot} = -1.17 \pm 2$ kcal/mol where the genetic system will exhibit the best AND logic with respect to the P_{BAD} promoter's RBS sequence (**Fig. 4b**).

Using the forward engineering mode of the design method, we generated two synthetic RBS sequences targeted to the optimum region of the genetic system's function (predicted $\Delta G_{tot} = -1.48$ and -1.15 kcal/mol). We also designed seven 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) was inserted downstream of the P_{BAD} promoter and the resulting genetic circuit's response to varying inducer concentrations was assayed (**Fig. 4c**).

The two synthetic RBS sequences designed for the optimal ΔG_{tot} successfully connected the arabinose-sensing P_{BAD} promoter and the AND-gate genetic circuit (mean fitness > 0.85, **Fig. 4b**). The experimentally determined optimum in the $F(\Delta G_{tot})$ curve is approximately $\Delta G_{tot} = 0.60$ kcal/mol, which is only a 1.8 kcal/mol deviation from the model's prediction (**Fig. 4b**). The quantitative model and design method also correctly predicted 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 to obtain a target behavior while performing only a few mutations and assays (for additional design calculations, see **Supplementary Fig.10** and **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, we will have to construct biophysical models that link the DNA sequence of a part to its function. 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.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Software implementation. A software implementation of the design method, the RBS Calculator, is available at http://voigtlab.ucsf.edu/software.

Note: Supplementary information is available on the Nature Biotechnology website.

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, National Institutes of Health (NIH) EY016546, NIH AI067699, NSF BES-0547637, National Science Foundation (NSF) TeraGrid TG-MCB080126T and a Sandler Family Opportunity Award. C.A.V., H.M.S., and E.A.M. are part of the NSF SynBERC Engineering Research Center (http://www.synberc.org/). E.A.M. is supported by an NSF Graduate Research Fellowship and an American Society for Engineering Education National Defense Science and Engineering Graduate Fellowship.

AUTHOR CONTRIBUTIONS

H.M.S and C.A.V designed the study and wrote the manuscript. H.M.S. developed the method. H.M.S. and E.A.M. performed the experiments.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H. & Weiss, R. A synthetic multicellular system for programmed pattern formation. *Nature* **434**, 1130–1134 (2005).
2. Stricker, J. *et al.* A fast, robust and tunable synthetic gene oscillator. *Nature* **456**, 516–519 (2008).
3. Friedland, A.E. *et al.* Synthetic gene networks that count. *Science* **324**, 1199–1202 (2009).
4. Ellis, T., Wang, X. & Collins, J.J. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.* **27**, 465–471 (2009).
5. Yokobayashi, Y., Weiss, R. & Arnold, F.H. Directed evolution of a genetic circuit. *Proc. Natl. Acad. Sci. USA* **99**, 16587–16591 (2002).
6. Tabor, J.J. *et al.* A synthetic genetic edge detection program. *Cell* **137**, 1272–1281 (2009).
7. Anderson, J.C., Voigt, C.A. & Arkin, A.P. Environmental signal integration by a modular AND gate. *Mol. Syst. Biol.* **3**, 133 (2007).
8. Dueber, J.E. *et al.* Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27**, 753–759 (2009).
9. Anthony, J.R. *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.* **11**, 13–19 (2008).
10. Atsumi, S., Hanai, T. & Liao, J.C. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, 86–89 (2008).
11. Hawkins, K.M. & Smolke, C.D. Production of benzylisoquinoline alkaloids in *Saccharomyces cerevisiae*. *Nat. Chem. Biol.* **4**, 564–573 (2008).
12. Lee, K.H., Park, J.H., Kim, T.Y., Kim, H.U. & Lee, S.Y. Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol. Syst. Biol.* **3**, 149 (2007).
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.* **10**, 69–77 (2008).
14. Czar, M.J., Anderson, J.C., Bader, J.S. & Peccoud, J. Gene synthesis demystified. *Trends Biotechnol.* **27**, 63–72 (2009).
15. Gibson, D.G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319**, 1215–1220 (2008).
16. Isaacs, F.J. *et al.* Engineered riboregulators enable post-transcriptional control of gene expression. *Nat. Biotechnol.* **22**, 841–847 (2004).
17. Carrier, T.A. & Keasling, J.D. Library of synthetic 5′ secondary structures to manipulate mRNA stability in *Escherichia coli*. *Biotechnol. Prog.* **15**, 58–64 (1999).
18. Pfleger, B.F., Pitera, D.J., Smolke, C.D. & Keasling, J.D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.* **24**, 1027–1032 (2006).
19. Chubiz, L.M. & Rao, C.V. Computational design of orthogonal ribosomes. *Nucleic Acids Res.* **36**, 4038–4046 (2008).
20. de Smit, M.H. & van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci. USA* **87**, 7668–7672 (1990).
21. Vellanoweth, R.L. & Rabinowitz, J.C. The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli* in vivo. *Mol. Microbiol.* **6**, 1105–1114 (1992).
22. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
23. Mathews, D.H., Sabina, J., Zuker, M. & Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).
24. Kierzek, R., Burkard, M.E. & Turner, D.H. Thermodynamics of single mismatches in RNA duplexes. *Biochemistry* **38**, 14214–14223 (1999).
25. Miller, S., Jones, L.E., Giovannitti, K., Piper, D. & Serra, M.J. Thermodynamic analysis of 5′ and 3′ single- and 3′ double-nucleotide overhangs neighboring wobble terminal base pairs. *Nucleic Acids Res.* **36**, 5652–5659 (2008).
26. Christiansen, M.E. & Znosko, B.M. 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* **47**, 4329–4336 (2008).
27. Laursen, B.S., Sorensen, H.P., Mortensen, K.K. & Sperling-Petersen, H.U. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69**, 101–123 (2005).
28. Studer, S.M. & Joseph, S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol. Cell* **22**, 105–115 (2006).
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.* **22**, 4953–4957 (1994).
30. Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
31. de Smit, M.H. & van Duin, J. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J. Mol. Biol.* **331**, 737–743 (2003).
32. Dirks, R.M., Bois, J.S., Schaeffer, J.M., Winfree, E. & Pierce, N.A. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Rev.* **49**, 65–88 (2007).
33. Sengupta, J., Agrawal, R.K. & Frank, J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl. Acad. Sci. USA* **98**, 11991–11996 (2001).

ONLINE 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 100,000, 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 sub-sequence S_1 consists of the max(1, n_{start} − 35) to n_{start} nucleotides and the sub-sequence 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,

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

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,

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

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 Fig. 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 determined by calculating the mfe of sequence S_2 with and without preventing the standby site from forming base pairs. The difference between these mfes 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 sub-sequences, their mfes are each calculated and then summed together. The two sub-sequences 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 presequence 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 5–20% 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 >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 $< 6 \times 10^{-3}$ 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) was obtained from Fisher Scientific. 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 $MgSO_4$ (Fischer Scientific), 100 µM $CaCl_2$ (Fischer Scientific), 0.4% glucose (Sigma), 0.05 g/l leucine (Acros Organics), 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 (New England Biolabs). 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 an RBS and the mRFP1 fluorescent protein reporter (**Supplementary Fig. 11**). An annotated DNA sequence of the system (GenBank format) is available in the **Supplementary Data**.

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

flow cytometer events. All distributions were unimodal. The autofluorescence distribution of DH10B cells was also measured. The arithmetic mean of each distribution was taken and the mean autofluorescence was subtracted.

From single colonies, RBS variants of each AND gate genetic circuit were 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 16 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) were inoculated by overnight cultures using a 1:100 dilution. Plates were grown in a Safire^2 plate spectrophotometer with high orbital shaking. OD_{600} and *gfp* fluorescence measurements were recorded every 10 min for 14 h. Background autofluorescence was subtracted from each fluorescence measurement. This procedure was repeated twice for each variant. For each variant, the average and s.d. of the fluorescence per OD_{600} for each inducer concentration at the final time point were then calculated.

Empirical determination of $\Delta G_{spacing}$. To quantify the relationship between the free energy penalty $\Delta G_{spacing}$ and the distance between the 16S rRNA binding site and the start codon (called the aligned spacing *s*), we created thirteen synthetic RBSs where the aligned spacing was varied from 0 to 15 nucleotides whereas the $\Delta G_{mRNA:rRNA}$, ΔG_{mRNA}, ΔG_{start} and $\Delta G_{standby}$ free energies remained constant (**Supplementary Table 1**). The translation initiation rates of RBS sequences were measured using a fluorescent protein measurement system. Steady-state fluorescence measurements were performed on *E. coli* cultures over a 24 h 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 $\Delta G_{spacing}$ was obtained from the fluorescence measurements. 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) (**Supplementary Discussion**). After this parameterization, we tested the accuracy of these equations on an additional set of synthetic RBS sequences (**Supplementary Fig. 2**).

The $\Delta G_{spacing}$ was inferred from the fluorescent protein expression data *E* in the following way. The RNA sequences used to parameterize the model of $\Delta G_{spacing}$ were 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 $\Delta G_{spacing}(s = 5)$ was accordingly set to zero. Using an experimentally measured value of $\beta = 0.45$ mol/kcal, the model of $\Delta G_{spacing}$ for each *s* was then determined.

Error analysis. 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 s.d. of β is calculated by substituting the log *E* data with the $\log(E+\delta E)$ and $\log(E-\delta E)$ data (δE: s.d. of *E*) and calculating the average difference.

34. David, F., Hagendorf, C. & Wiese, K.J. A growth model for RNA secondary structures. *J. Stat. Mech. Theor. Exp.* P04008 (2008).

doi:10.1038/nbt.1568