Quantitative Design of Regulatory Elements Based on High-Precision Strength Prediction Using Artificial Neural Network

Hailin Meng[1]◑, Jianfeng Wang[1,2]◑, Zhiqiang Xiong[1], Feng Xu[1], Guoping Zhao[1], Yong Wang[1]*

1 Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **2** State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China

Abstract

Accurate and controllable regulatory elements such as promoters and ribosome binding sites (RBSs) are indispensable tools to quantitatively regulate gene expression for rational pathway engineering. Therefore, *de novo* designing regulatory elements is brought back to the forefront of synthetic biology research. Here we developed a quantitative design method for regulatory elements based on strength prediction using artificial neural network (ANN). One hundred mutated Trc promoter & RBS sequences, which were finely characterized with a strength distribution from 0 to 3.559 (relative to the strength of the original sequence which was defined as 1), were used for model training and test. A precise strength prediction model, NET90_19_576, was finally constructed with high regression correlation coefficients of 0.98 for both model training and test. Sixteen artificial elements were *in silico* designed using this model. All of them were proved to have good consistency between the measured strength and our desired strength. The functional reliability of the designed elements was validated in two different genetic contexts. The designed parts were successfully utilized to improve the expression of BmK1 peptide toxin and fine-tune deoxy-xylulose phosphate pathway in *Escherichia coli*. Our results demonstrate that the methodology based on ANN model can *de novo* and quantitatively design regulatory elements with desired strengths, which are of great importance for synthetic biology applications.

Citation: Meng H, Wang J, Xiong Z, Xu F, Zhao G, et al. (2013) Quantitative Design of Regulatory Elements Based on High-Precision Strength Prediction Using Artificial Neural Network. PLoS ONE 8(4): e60288. doi:10.1371/journal.pone.0060288

Editor: Szabolcs Semsey, Niels Bohr Institute, Denmark

Received January 16, 2013; **Accepted** February 25, 2013; **Published** April 1, 2013

Funding: This work was supported by the National Basic Research Program of China (''973'' Program, grant no. 2012CB721104), the National High Technology Research and Development Program (''863'' Program, grant no. 2012AA02A701), the National Natural Science Foundation of China (grant no. 31170101 and 31100073), ''Technology Innovation Action Plan'' Key Project of Shanghai Science and Technology Commission (grant no. 10dz1910100), and the major Projects of Knowledge Innovation Program of Chinese Academy of Sciences (grant no. KSCX2-EW-J-12). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yongwang@sibs.ac.cn

◑ These authors contributed equally to this work.

Introduction

The coming era of synthetic biology aims at design and construction of complex biological networks to achieve our special goals (e.g., high-level production of clinically valuable natural products), which requires fine-tuning gene expression in the cellular networks to achieve an expected metabolic behaviour [1,2]. Genetic elements with desired strengths/activities, e.g., promoters/RBSs for transcriptional/translational controls, are the most important tools to accurately control the expression of rate-limiting genes in an engineered system. In the last decade, an array of randomly mutated or synthetic promoter libraries with a wide range of strength have been constructed and applied to control protein expression or pathway engineering in *E. coli* and yeast [3,4,5]. However, acquisition of a controllable regulatory element from a random library needs laborious screening and multifaceted characterizations to ensure homogeneity at the single-cell level [5], especially in the situation of multiple genes regulated at different expression levels in one system. More recently, great advances in synthetic biology and its applications in engineering pathways and producing valuable chemicals in microbes have brought back the sequence-activity modelling approaches to the forefront. Researchers have put immense interests to decipher the 'regulatory code' (e.g., $-10/-35$ region, RBS region) that translates DNA sequences into expression strength [6,7,8,9], and built quantitative models for strength prediction and rational design of regulatory elements [10,11,12,13]. For instance, based on position weight matrix (PWM) models, Rhodius VA *et al* [12] scored various motifs of *E. coli* σ^E binding promoters and correlated promoter scores with *in vitro* and *in vivo* measured strength, the correlation coefficient values (R) ranged from 0.57 to 0.77 for *in vitro* and *in vivo* strength fit was achieved. Besides promoter, Salis HM *et al* [10] targeted translation initiation process and developed a equilibrium statistical thermodynamics model for designing synthetic RBSs (linear regression R^2 ranged from 0.54 to 0.91), which correlates the Gibbs free energy variation of translation initiation with the translation rate. The above methods mainly targeting feature motifs or key processes have reached a certain point of success. But as Jensen *et al* [14] and De Mey M *et al* [11] indicated, promoter strength could not be simply linked to anomalies in the feature motifs such as -10 box and/or -35 box, and to the length of spacer. Therefore, De Mey M *et al* [11] established a correlation between the entire sequence and strength by applying partial least squares (PLS) regression method. This model exhibits promising applications for quantitative strength prediction and rational design of promoters, but still has great potential to improve its accuracy. Hence, building precise computational models that can predict the activity of regulatory elements and quantitatively design elements with desired strength is still a real challenge in gene expression area over decades.

Aforementioned quantitative prediction models commonly use linear regression analysis or its derivative methods (e.g., linear correlation of data after logarithm processing) to simplify the complex process for model construction. Thus, it is hard to well reflect the complex non-linear relationship between the sequences and their strengths, which results in a low prediction accuracy and poor generality. In addition, these models are supposed to have the potential, but have not been further developed into *in silico* methods for *de novo* design of elements with desired strength. In contrast to the above methods, we introduced a non-linear modelling methodology, artificial neural network (ANN), to address these issues. ANN is essentially a mathematical model constructed by simulation of the structure and function of human brain neural networks [15,16]. It can be adapted to continuously change the network structure based on input/output information during learning phase, which could reflect the non-linear relationships between quantitative characteristics and related qualitative performance in complex phenomena. Thus, ANNs have been widely used to various biological research fields such as protein structure and stability prediction [17,18,19], RNA secondary structure prediction [20], as well as promoter recognition and structure analysis [21,22,23,24,25,26,27,28]. In this work, we constructed a high-performance ANN model to directly predict the strength of regulatory element from its sequence. Based on this model, we further developed an effective computational platform for quantitative design of novel regulatory elements with desired properties for synthetic biology applications.

Materials and Methods

Strains, plasmids, reagents and general manipulation

All strains and plasmids involved in this study are listed in Table 1. *E. coli* DH10B was used for library construction and strength quantification. *E. coli* BL21(DE3) was served for BmK1 expression and amorphadiene biosynthesis. Enzymes & reagents for DNA manipulation and bacteria culture were purchased from New England Biolabs, Takara, or Oxoid. All primers, designed sequences, codon optimized [29] peptide BmK1 (*bmk*1, Genbank: AAD39510), and amorphadiene synthase (*ads*, Genbank: AAF98444) genes [30], were synthesized by Generay Biotech Ltd. (Shanghai, China). Antibiotics were added according to the resistance marker of plasmids in each culture. The working concentration of ampicillin and kanamycin were 100 mg/L and 50 mg/L, respectively.

Reporter plasmid pJF07 was created by inserting a *gfp* gene [5] into the *Bam*HI & *Eco*RI sites of pTrcHis2B. To create plasmids s14/s05/s21-*bmk*1, the *bmk*1 gene was inserted into plasmids s14/s05/s21-*gfp* to replace the *gfp* gene. The *dxs* gene was obtained by PCR using primers *dxs*F (5′-CATGCCATGGGCATGAGTTTTGATATTGCCAAATACCCG-3′) and *dxs*R (5′-CCGGAATTCACTAGTTTATGCCAGCCACCTT-3′) and genomic DNA of *E. coli* K12 MG1655 as template. The isolated *dxs* gene fragment was cloned into the *Nco*I & *Eco*RI sites of pTrcHis2B to create plasmid pTrcHis2B-*dxs*, or inserted into s14/s05/s21-*gfp* to replace the *gfp* for creation of plasmids s14/s05/s21-*dxs* respectively. The *ads* gene was inserted into the *Nde*I & *Eco*RI sites of pET21c to create plasmid pET21c-*ads*. The *ispA* gene was isolated by PCR using primers (5′-CATGCCATGGGCATGGACTTTCCGCAGCAACTCGAAG-3′) and (5′-CCGGAATTCACTAGTTTATTTATTACGCTGGATGATGTAG-3′) and the genomic DNA of *E. coli* K12 MG1655 as template. The product of *ispA* was inserted into the *Nco*I & *Eco*RI sites of pET28a to create pET28a-*ispA*. The *Xba*I & *Eco*RI excised fragment of pET21c-*ads* was inserted into the *Spe*I & *Eco*RI sites of pET28a-*ispA* to create pET28a-*ispA-ads*. All standard DNA manipulations were performed as described by Sambrook *et al.* [31].

Library construction and characterization

Random mutagenesis of the wild-type Trc promoter & RBS sequence was performed by error-prone PCR using primers TrcF (5′-ATAAGAATGCGGCCGCAACGGTTCTGGCAAATATTCTGAAAT-3′, the restriction site is underlined. The same below) and TrcR (5′-TCCTTTACGCATTGGATCCATGG-3′) and plasmid pJF07 as template according to the Kit's instruction (JBS Error-Prone Kit PP101, Jena Bioscience). The reporter plasmid skeleton was PCR amplified by PrimeSTAR DNA polymerase using primers pJF07F (5′-CGGGATCCAATGCGTAAAGGAGAAGAAC-3′) and pJF07R (5′-ATAAGAATGCGGCCGCATGATGTCGGCGCAAAAAACATTATC-3′) and plasmid pJF07 as template. PCR products of the reporter plasmid skeleton and Trc promoter & RBS excised by *Not*I & *Bam*HI were ligated and transformed into DH10B competent cells. The transformed culture were spread onto LB agar plate and cultivated overnight at 37°C for 16 hours.

For primary screening, transformants were picked out into the 48-deep-well plate and screened through gene fluorescent protein assay (excitation/emission wavelength = 485 nm/535 nm). The conditions for 48-deep-well plate cultivation were as follows: 0.5 ml LB medium with 0.1 mM IPTG in each 5 ml-well at 37°C and 250 rpm for 8 h during exponential phase. The $OD_{600\ nm}$ and green fluorescent signal of 100 µl culture was quantified in a 96-well plate reader (Multiskan FC Microplate Photometer, Thermo Scientific). For the convenience of comparison, we used the relative strength [30] to represent the strength of a mutated sequence, which was defined and calculated as

$$S = \frac{\left(\frac{F}{OD_{600}}\right)_{\text{clone}} - \left(\frac{F}{OD_{600}}\right)_{\text{pTrcHis2B}}}{\left(\frac{F}{OD_{600}}\right)_{\text{pJF07/m000}} - \left(\frac{F}{OD_{600}}\right)_{\text{pTrcHis2B}}},$$

where S is the relative strength of the sequence, F the fluorescent value; pTrcHis2B the blank control, and pJF07/m000 the wild-type Trc promoter & RBS.

One hundred clones with distributed strength were selected and cultivated overnight in LB broth and preserved in 20% glycerol at −80°C for seed culture. Fine quantification of the selected elements was performed in tube (15 mm×150 mm) and assayed by flow cytometry (FACSCalibur flow cytometer, Bection Dickinson). Seventy-five microliters of seed culture was innoculated into 1.5 ml LB with 0.1 mM IPTG and incubated at 37°C and 250 rpm for 3 h at exponential phase. The culture was cooled with ice bath and assayed using clone containing pTrcHis2B as blank control. Each clone was sampled with 20,000 events and the geometric mean (Gmean) of fluorescent signal was calculated using statistics. The relative strength value [32] compared with wild-type Trc promoter & RBS was calculated as

Table 1. Strains and plasmids in this study.

Strains & plasmids	Relevant characteristics	Source
DH10B	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara*, *leu*)7697 *gal*U *gal*K λ- *rps*L *nup*G	Invitrogen
BL21(DE3)	F- *omp*T *hsd*S (*r*BB-*m*B-) *gal dcm* (DE3)	EMD4 Biosciences
pTrcHis2B	Ampicillin resistance marker, Trc promoter	Invitrogen
pJF07	Plasmid pTrcHis2B carrying a *gfp* gene at BamHI/EcoRI sites	This study
pET28a-*isp*A	pET28a derived plasmid carrying a *isp*A gene at NcoI/EcoRI sites	This study
pET21c-*ads*	pET21c derived plasmid carrying a *ads* gene at NdeI/EcoRI sites	This study
pET28a-*isp*A-*ads*	pET28a derived plasmid carrying *isp*A, *ads* for amorphadiene production	This study
s14/s05/s21-*gfp*	Three pTrcHis2B derived plasmids carrying synthetic promoters s14 (0.56), s05 (1.00) and s21 (2.50) followed by a *gfp* gene at BamHI/EcoRI sites, respectively	This study
s14/s05/s21-*bmk*1	Three pTrcHis2B derived plasmids carrying synthetic promoters s14 (0.56), s05 (1.00) and s21 (2.50) followed by a *bmk*1 gene at NcoI/HindIII sites, respectively	This study
s14/s05/s21-*dxs*	Three pTrcHis2B derived plasmids carrying synthetic promoters s14 (0.56), s05 (1.00) and s21 (2.50) followed by a *dxs* gene at NcoI/EcoRI sites, respectively	This study
pTrcHis2B-*dxs*	pTrcHis2B derived plasmid carrying a *dxs* gene at NcoI/EcoRI sites	This study

doi:10.1371/journal.pone.0060288.t001

$$S=\frac{Gmean_{\mathrm{clone}}}{Gmean_{\mathrm{pJF07/m000}}}.$$

Computational platform construction

Matlab 2012a (Mathworks Inc., http://www.mathworks.com/) ran on a personal computer with Microsoft Windows 7 64-bit (Microsoft Inc., http://www.microsoft.com/) operation system. Neural Network Toolbox within Matlab served as the basic tool for artificial neural network (ANN) model construction, data fitting and prediction. All programs used in this work were designed and run upon Neural Network Toolbox and Matlab environment.

Cultivation of recombinant strains and products analysis

Peptide expression was performed in *E. coli* BL21(DE3) at 37°C and 250 rpm, induced with 0.1 mM IPTG at 0.6 of OD_{600} for 3 h, and analyszed by SDS-PAGE. The stained PAGE was imaged by Tanon 2500R gel-imaging system (Tanon Science & Technology Ltd., Shanghai, China). The relative content of BmK1 to total cellular protein was calculated by the GIS 1-D software (Tanon) according to the ratio of the intensity of target peptide band to that of all protein bands.

Recombinant strains *E. coli* BL21(DE3) harbouring plasmid pET28a-*ispA-ads* and s14-*dxs*, or s05-*dxs*, or s21-*dxs* were used for amorphadiene production. Shake-flask fermentation was performed using the following conditions: 2% of inoculation and 10 ml TB medium (12 g/L tryptone, 24 g/L yeast extract, 2.31 g/L KH_2PO_4, 12.54 g/L K_2HPO_4) in 100 ml shake-flask with 2% glycerol, 20% dodecane, and 0.1 mM IPTG at 28°C and 250 rpm for 3 days. After cultivation, dodecane phase was diluted using ethyl acetate to an appropriate concentration and analysed by GC-MS using caryophyllene (Sigma-Aldrich) as internal standard [30].

Results

Construction of Trc promoter & RBS strength library

Trc promoter is commonly used for protein expression in *E. coli* or other prokaryotic systems. To build and train the ANN models, we initially constructed and characterized a mutated Trc promoter library. Considering that protein expression is influenced by both transcription and translation processes, herein the DNA region of Trc promoter plus its RBS (224 bp in total) from a commercial plasmid pTrcHis2B was subjected to random mutagenesis. The mutagenesis rate of the library reached up to about 20%. After initial screening 4,000 clones, 100 mutants with uniformly distributed strengths were chosen to construct a strength-gradient library (Figure 1 and Text S1). All of the mutants were finely quantified and sequenced (Text S2 and Text S3). The library contains 100 sequences (including the wild-type sequence) with the strength ranging from 0 to 3.559 (relative to the strength of the original sequence), of which 20 sequences are positive mutants (relative strength >1.0) and 79 other sequences are negative mutants (relative strength <1.0).

Construction and training of ANN predicting models

The initial ANN model was built as a backpropagation model (BP-ANN model) by using Matlab functions provided by Neural Network Toolbox. The model contains three layers, including an input layer, an output layer and a hidden layer. Neuron numbers of the input layer and the output layer were 896 and 1 (determined by the data conversion rule), respectively. For the hidden layer, the number was variable for optimization. The initial weights for all neuron connections were randomly assigned by Matlab functions.

We evaluated the predicting performance by using the sum squared error (*SSE*) between the prediction value a_i and target strength value t_i as

$$SSE=\sum_{t=1}^{n}(a_i-t_i)^2$$

(where n is the sequence number of training data set or test data set), and defined prediction error as

$$E_i=|a_i-t_i|.$$

The activation functions of the hidden layer and the output layer were set to be a non-linear sigmoid function 'logsig', which

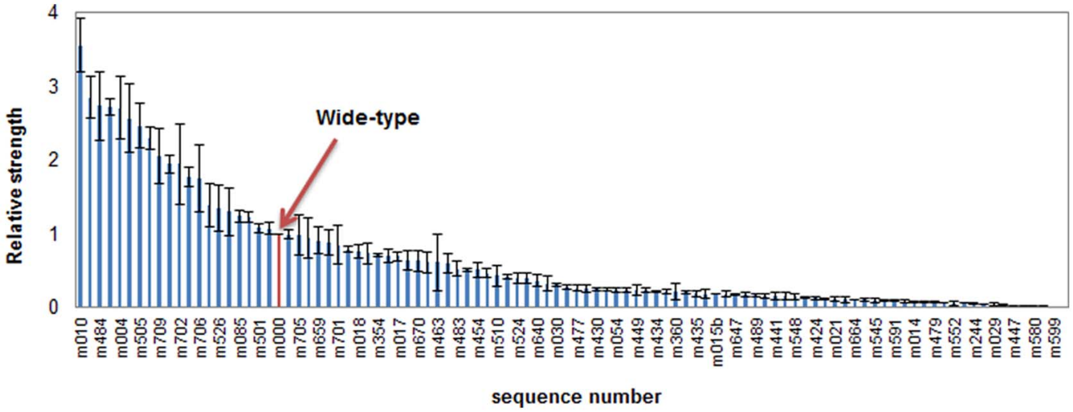


Figure 1. Relative strengths of the constructed Trc promoter & RBS library. The region of Trc promoter & RBS in pTrcHis2B is selected for random mutagenesis by error prone PCR, and mutants with various strength are obtained by detecting the fluorescent intensity of GFP after screening by 48-deep-well plates and flow cytometry assay.
doi:10.1371/journal.pone.0060288.g001

was defined as

$$f(x) = \frac{1}{1+e^{-x}}.$$

For training of BP-ANNs, a set of example pairs was given as (x,y), $x \in X$ and $y \in Y$, and the aim was to find a function $F: X \rightarrow Y$ that can match the examples. Here, X refers to the promoter & RBS sequences and Y refers to their relative strength. In other words, we wished to infer the mapping relationship between sequence and strength by the samples of training data set. Here, the mapping relationship was a 'black box' which can be served as a predicting model for the prediction of test set data. This 'black box' may be constructed after training by a set of sample data.

The original sequence data were translated to digital data and served as the input matrix according to the following rules: A = {1, 0, 0, 0}, G = {0, 1, 0, 0}, C = {0, 0, 1, 0}, and T = {0, 0, 0, 1}. For instance, a given sequence 'ATTGCC' can be translated to a '0-1' digital series of {1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0}.

It must be noted that, since the output range of logsig function lies in (0,1), while the target strength can be greater than 1, so it was necessary to normalize the target strength data through dividing by the maximum strength value and then multiplying by this value after simulation.

The goal of *SSE* value for fitting training data was set to be 0.2. The initial weights for all neuron connections were set randomly and automatically by Matlab functions. In addition, 'traingdx' was adopted as the learning function; the training epochs and momentum factor were set to be 5,000 and 0.95, respectively.

All 100 sequences in the library were randomly split into two data sets (the training set and the test set) to train and test the ANN prediction models. Considering the effect of the size of training set on the prediction performance, training set was sampled from 40 to 90 sequences (51 situations in total) and each corresponding test set contained the rest sequences. The neuron number of the hidden layer was optimized in a range from 5 to 30, and each trained to generate 1,000 models. Consequently, we obtained $51 \times 26 \times 1{,}000 = 1{,}326{,}000$ models. Owing to the random initialization of weights, the trained models have different prediction performance which can be evaluated by their *SSE* and *E* values for prediction of the corresponding test set. Figure 2 describes the overall trends of *SSE* and *E* values as a function of the size of the training data set. Both the maximum and the minimum *SSEs* decline with the increasing size of training data set (Figure 2A), indicating that a certain size of training data set is a requisite for obtaining a model with high predicting performance. For prediction errors *E*, however, both the maximum and the minimum errors do not significantly decline until the size of training set reaches 88 (Figure 2B). Among all generated models, NET90_19_576 (containing 19 hidden layer neurons with a training set's size of 90, see Dataset S1 and Model S1) shows the best performance with the lowest *SSE* of 0.19 and the highest correlation coefficient values of 0.98 for test set prediction. Meanwhile, its correlation coefficient values for fitting the training data set reaches up to 0.98 as well (Figure 3A and 3B), suggesting that this model does not overfit in the process of model training. Besides the high correlation coefficient for the total test set, NET90_19_576 also accurately predicts each element in the test set (Figure 3C). Both the correlation coefficient and the predicting accuracy in our ANN model are significantly improved compared with PLS- [11], PWM- [12] and thermodynamics-based [10] methods. As a comparison, the best result of PWM-based fitting using our data only has an *R* of 0.63 (Figure 3D).

Quantitative design of promoter & RBS sequences with desired strength

Owing to the high correlation and accurate prediction performance, the model NET90_19_576 can be effectively developed into a computational platform for quantitative design of novel regulatory parts. Our quantitative design strategy was achieved by consequential *in silico* mutagenesis on native Trc promoter & RBS sequences coupled with rapid strength prediction using NET90_19_576 model. There are two approaches to introduce mutations: i) introduction of random mutagenesis and ii) only introduction mutation of key points (nucleotides significantly affecting the strength). For the first approach, some 'non-key points' may be introduced as mutations and increase the calculation time. Besides directly obtaining one (or more) sequence(s), it can randomly generate an *in silico* library in an arbitrary scale (e.g., 10,000 sequences). Hence, we can obtain any desired sequence(s) from this pre-constructed computational library. For the second approach, the effect of single point mutation of wild-type sequence on its strength should be evaluated to determine which points are the 'key-points' at first. The strength

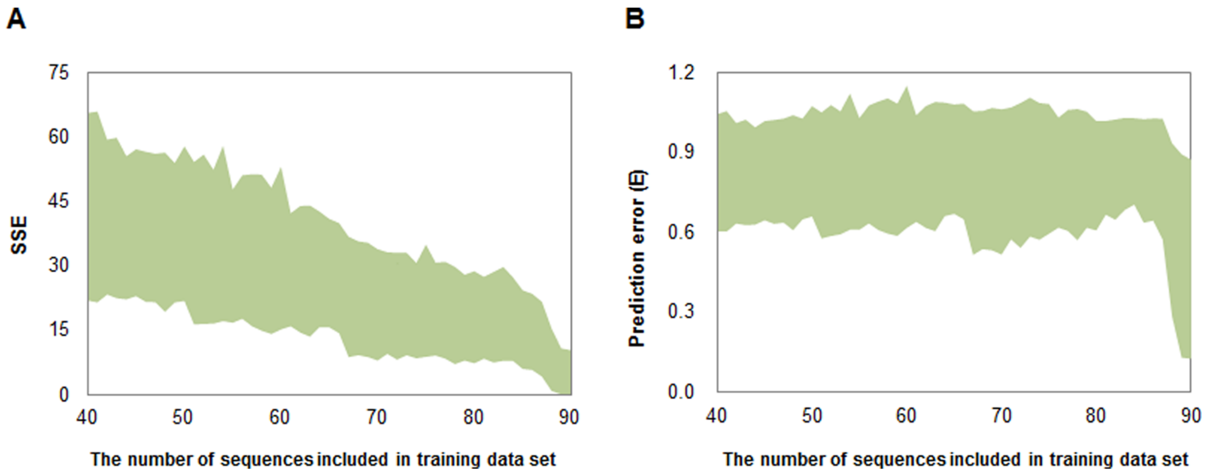


Figure 2. Functional relationship between the prediction performance of ANN models and the scale of training data set. Training data set scale ranges from 40 to 90 sequences. (A) Maximum and minimum *SSE* values of prediction as a function of training data set scale. (B) Maximum and minimum prediction errors as a function of training data set scale.
doi:10.1371/journal.pone.0060288.g002

of mutated sequences may have some correlation with their mutation points. Those sequences with extremely low activity (i.e. m006, m007, m029, etc.) have large amount of mutation points, and most of their functional domains, such as −35 region, −10 region, and RBS region, are destroyed (except for m413, m447, m590, m599). On the contrary, most mutated sequences with high strength have relatively conservative domains. Our results reconfirm that key points significantly affect the sequence strength. We can find out these key points through changing nucleotide one-by-one and use them for designing new element sequences using our computational platform. Figure 4A presents the prediction results of all single point mutations and each point has three

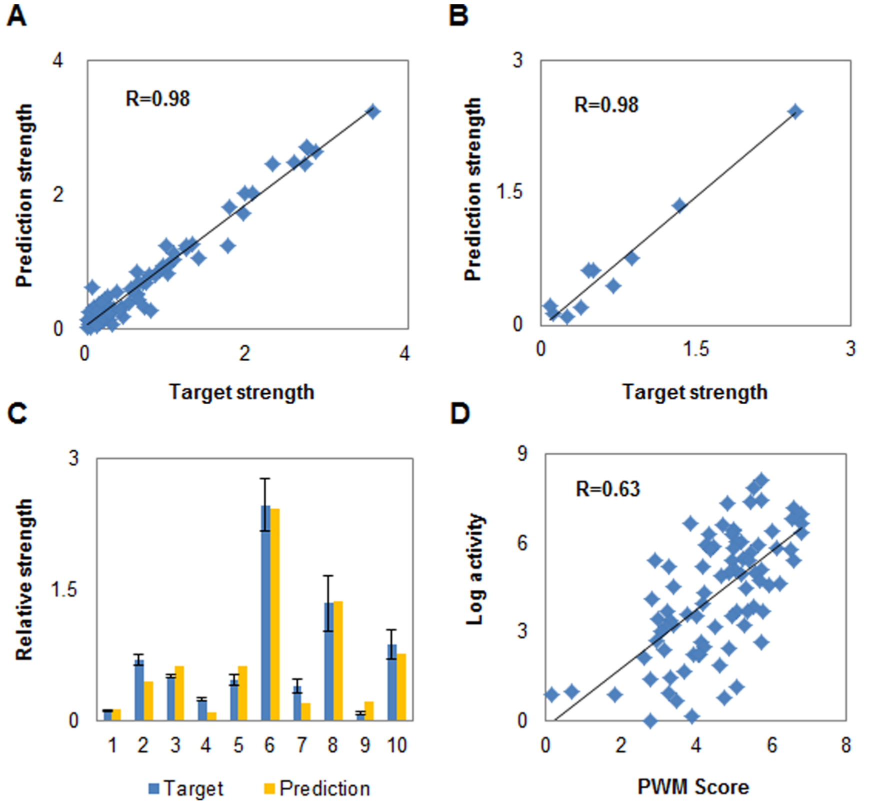


Figure 3. The well trained BP-ANN model NET90_19_576 can finely predict the measured Trc promoter & RBS strengths. (A) The predicted relative strengths of promoter & RBS fit with the measured values using the data of training set. (B) The predicted values fit with the measured values using the data of test set. (C) The comparison results between prediction values and target values (experiment values). (D) The best fitting results of log Trc promoter & RBS activities with their PWM scores.
doi:10.1371/journal.pone.0060288.g003

mutation types. As a result, 135 points are found significantly impacting the sequence strength, in which 15 points (designated as positive points) can significantly enhance $\geq$20% of the strength and 120 other points (designated as negative points) can significantly reduce $\geq$20% of the strength. Sequence logo analysis [33] was also performed to show the most conserved bases among mutations. As a result, most positive impacting points are non-conservative in sequences with high activity (strength $>$1) (11 of 15, Figure 4B) and conservative in sequences with extremely low activity (strength $<$0.1) (8 of 15, Figure 4C); in contrast, most negative impacting points are non-conservative in sequences with extremely low activity (90 of 120, Figure 4C) and conservative in sequences with high activity (91 of 120, Figure 4B) as well. That is to say, it is easier to obtain a positive mutation sequence by changing the positive impacting points, or to obtain a negative mutation sequence by changing the negative impacting points. Moreover, most negative points are found to change nucleotide from AT to GC (Figure 4D). It indicates that an increase in GC content may decrease the sequence activity; a higher energy barrier should be overcome for DNA dissociation with a higher GC when transcription and translation initiate.

To further verify the effectiveness of our design, sixteen novel Trc promoters & RBS sequences (s01–s08 designed from pre-generated library by approach i, s11–s15 generated from random mutagenesis by approach i, and s21–s23 designed by approach ii) were synthesized *in vitro* and quantified their strengths in strain DH10B (Figure 5, Text S1 and S2). The measured values of all sixteen designed elements show good consistency with our desired strength values, suggesting that both of the above two approaches can achieve quantitative design of novel elements under *in silico* environment.

Application of designed elements for protein expression and pathway engineering

The aforementioned work proves that predicting strength of one randomized part and designing a new part are feasible. To further validate the methodology, we need to change the reporter GFP with other metabolic enzymes to test if the designed parts are functionally reliable. Herein we attempted to apply these quantitatively designed regulatory elements in different genetic contexts in strain *E. coli* BL21(DE3), which is protease deficient and suitable for peptide/protein expression. The first case is to optimize heterologous expression of a small peptide BmK1, which is a scorpion toxin secreted by Chinese scorpion *Buthus martensii* Karsch (BmK) and a traditional Chinese medicine for treating ion channelopathies [34,35]. Like most small peptide toxins, BmK1 is extremely difficult to express in prokaryotic host such as *E. coli* [36,37]. The strength of promoters has large effects on the production of target protein in surrogate hosts [38,39], we thus selected three designed elements s14 (strength = 0.56), s05 (strength = 1.0) and s21 (strength = 2.50) to improve BmK1 expression. As shown in Figure 6B, these three elements make great difference for the peptide expression. The expression level of BmK1 was improved from 1.6% (s14) to 9.1% (s21) of total cellular protein with the increase of element strength. This result shows that the strength of our designed elements still agrees with the expression level of the novel genetic context BmKI in BL21(DE3), thus the functional reliability of designed regulatory elements is further verified. In contrast to the expression of GFP without obvious strain growth variations, the growth here was significantly decreased with the accumulation of BmKI, which may probably be due to the cellular toxicity of this small peptide.

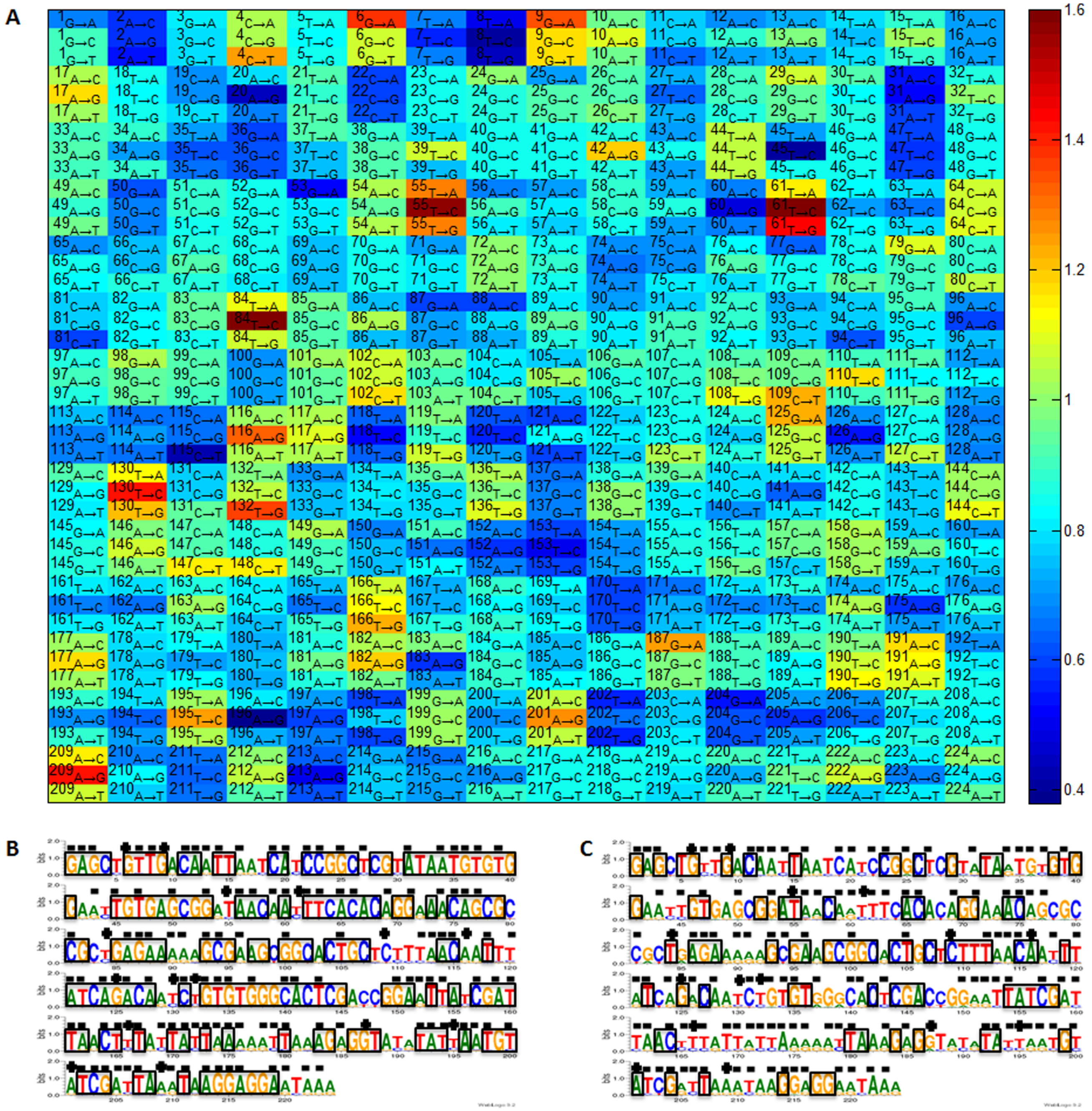
The second case is to fine-tune the expression of 1-deoxy-D-xylulose-5-phosphate synthase gene (*dxs*) in *E. coli* BL21(DE3), which has been known to regulate the metabolic flux of deoxy-xylulose phosphate (DXP) pathway [5]. Here we used three designed promoters (s14, s05 and s21) with different strengths to control the expression of *dxs* for improving the supply of isoprenoid precursors. As shown in Figure 6B, the production of sesquiterpene amorphadiene was enhanced with the decrease of element strength. The weakest element s14 achieves the highest yield of amorphadiene (4.82 mg/L/OD_{600}). This result agrees with the previous report that fine tuning the expression of rate-limiting enzymes can effectively improve pathway's metabolic flux [5] and further verified the functional reliability of our designed regulatory elements. The strain growth here was also decreased with the enhancement of *dxs* expression and the accumulation of amorphadiene. Both cases demonstrate that our methodology is an effective tool for designing and selecting regulatory element with proper strength for fine-tuning the target gene in metabolic engineering processe. The designed elements with proper strength can achieve both optimized specific product yield and strain growth, which can eventually maximize the process productivity.

Discussion

Constructing computational models that can precisely predict the strength of a regulatory element and further quantitatively build regulatory elements with desired strength have been a real challenge in gene expression area over decades. Many non-linear or unknown relationships between the sequences of regulatory elements and their strengths are still waiting to be uncovered [7]. We have introduced a methodology for constructing high-precision predicting model based on artificial neural network, which can finely predict the strength of regulatory element by its sequence. Both the high correlation coefficient and the predicting accuracy confirmed that the model is competent for *de novo* quantitative design of desired regulatory elements. The designed elements can be successfully applied in different genetic contexts.

In contrast to the existing prediction models [10,11,12], the presented method does not depend on comprehensive understanding of the transcription/translation processes. As most current studies focusing on correlating quantitative characteristics with qualitative information of biological behaviours [40,41], our method can quantitatively link the strength of regulatory elements with their sequences based on a finely characterized sequence library. The influence of each nucleotide mutation on sequence strength was evaluated and 135 key points were identified in this work, which are very useful for further study of 'regulatory code' and *de novo* design of elements. Besides, the methodology can be generalized and applied to construct models for predicting and designing more other promoters and RBSs, and even other regulatory elements like terminators.

Previous studies have confirmed that certain promoters can be identified or predicted based on ANN method [21,22,23,24,25,26,27,28], but no further effort was reported for quantitative description of their strength. Here, we constructed a finely characterzied Trc promoter & RBS library for sufficient model training and greatly improved the prediction accuracy compared with previous reported methods (PLS-, PWM- and thermodynamics-based) [10,11,12]. In addition, we built BP-ANNs using feed-forward as the network structure and Back-propagation (BP) as the training algorithm, mainly due to their rigorous mathematical derivation and proof, well generalization, strong non-linear mapping property, and a wide range of adaptability and effectiveness [15,16]. The effectiveness and high accuracy of BP-ANNs in constructing strength prediction models has been proven. Here, the initial model only contains one hidden



D

Mutation type	A→C	A→G	A→T	C→A	C→G	C→T	G→A	G→C	G→T	T→A	T→C	T→G
Negative mutation count	38	32	36	12	12	9	17	13	13	31	31	27
Positive mutation count	0	3	0	0	1	1	4	0	0	1	5	5

Figure 4. Effect of each single point mutation on sequence strength and sequence conservative analysis. (A) Sequence strength influenced by mutation of each single site. Red indicates positive mutation while blue indicates the negative. Deeper color means more significant change of strength. Each box represents one base in the sequence. Figure in the boxes is the location number of this base, while the subscript indicates that this base is mutated to another one (e.g., A→C means A mutated to C, and T→G means T mutated to G, etc.). (B) Conservative analysis of high activity sequences (strength >1). Bases in the boxes are conservative points. '+/−' indicates this point is predicted to be a positive/negative 'key-point'. Same as below. (C) Conservative analysis of extremely low activity sequences (strength <0.1). The analysis was performed using online WebLogo Tool (http://weblogo.threeplusone.com/create.cgi). (D) Count of mutation types of the 'key-points'. Figure in the boxes is the count of negative or positive mutation number.

doi:10.1371/journal.pone.0060288.g004

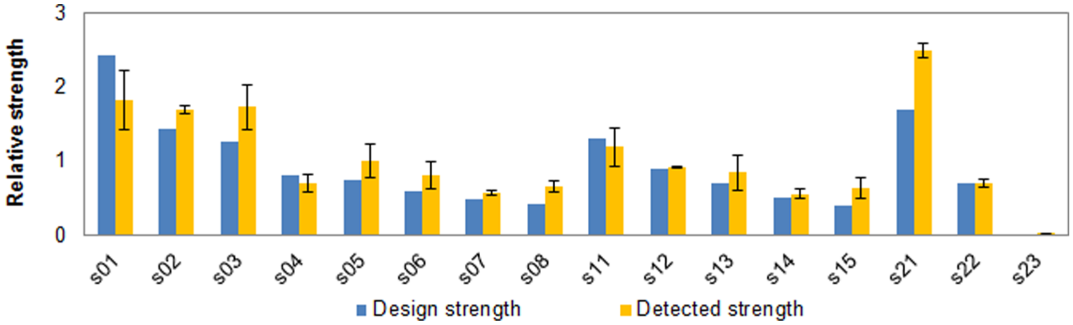


Figure 5. Promoter & RBS sequence design based on ANN prediction model. (A) Sequence with desired strength can be designed by the following strategies: i) 8 out of 10,000 sequences (s01–s08) are randomly selected from an *in silico* Trc promoter & RBS library generated based on ANN predicting model NET90_19_576; ii) sequences (s11–s15) with desired strength can be generated by repeated introduction of random mutations into the wild-type sequence under a certain mutation rate; iii) sequences (s21–s23) with desired strength can be generated by using different combinations of 'key site' mutations based on the prediction of NET90_19_576. All designed sequences were synthesized and their strengths were tested and compared with the design strength.
doi:10.1371/journal.pone.0060288.g005

layer, since in theory it can be approximate to a specific function in an arbitrary precision [16]. Adding more hidden layers may enhance the prediction performance but greatly increase the training time. Instead, optimization of the number of hidden layer neurons can also improve the predicting performance [15]. To address the overfitting problem, we tried *SSEs* ranging from 0.001 to 0.5 for model training and found the optimal value of 0.2. The results demonstrate that higher *SSE* makes lower correlation coefficients of fitting for both training data set and test data set. A lower *SSE* generates a higher *R* value of fitting for training data set, but a much lower *R* value of fitting for test data set. In other words, setting a lower *SSE* value of fitting for training data set can easily result in the overfitting problem and bad prediction performance for the test data set. Therefore, setting a suitable *SSE* is important to avoid overfitting problem and achieve the optimal prediction performance.

During the library construction process, we found that large fraction of clones was negative mutants and the probability of picking a positive mutant was less than 0.5%. In contrast, five designed elements with desired strength >1.0 were experimentally verified. These results demonstrate that the present methodology makes great sense for obtaining large amount of elements with different strength without laborious experimental screenings, especially for those stronger elements. But we cannot design a

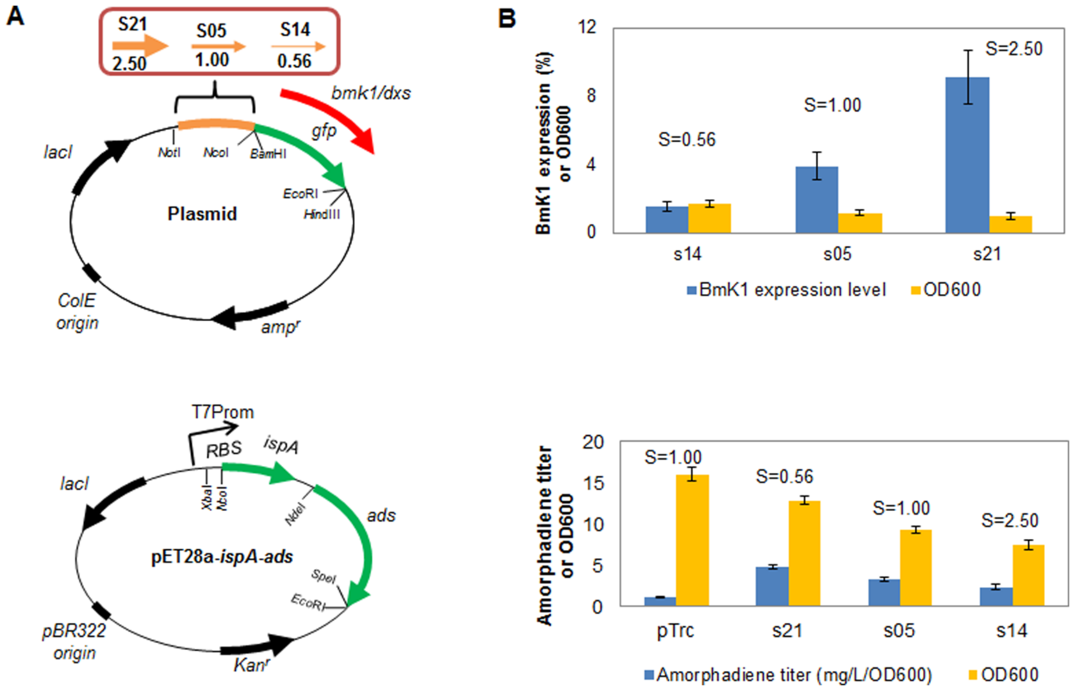


Figure 6. Application of designed elements for peptide BmK1 expression and DXP pathway engineering in *E. coli*. (A) Sketch maps of plasmids for designed elements applications. Plasmids s21-*gfp*, s05-*gfp* and s14-*gfp* contain gene *gfp* between BamHI/EcoRI sites, plasmids s21-*bmk*1, s05-*bmk*1 and s14-*bmk*1 contain gene *bmk*1 between NcoI/HindIII sites, plasmids s21-*dxs*, s05-*dxs* and s14-*dxs* contain gene *dxs* between NcoI/EcoRI sites. (B) Effect of applying designed elements for peptide BmK1 expression and DXP pathway engineering in *E. coli*. The wild-type Trc promoter and RBS (without inserting *dxs* gene) served as the blank control.
doi:10.1371/journal.pone.0060288.g006

high strong element with a relative strength larger than the maximum value of training data set (3.559), which is limited by the strength range of data samples for model training.

With the rapid development of synthetic biology, quantitative characterization and standardization of regulatory elements will be in general valuable in predicting parts in ever increasing genome sequence data [42]. The presented methodology would help us to easily build high performance prediction and design models using these standardized data in literatures/databases (e.g., the Registry of Standard Biological Parts founded by MIT, http://partsregistry.org) without reconstructing libraries by repeated and laborious experiments. In this framework, our methodology for constructing high prediction performance models and quantitative design of regulatory elements has bright prospects for synthetic biology application.

Supporting Information

Text S1 Relative strength value of Trc promoter & RBS elements.
(DOCX)

Text S2 Sequences of Trc promoter & RBS elements.
(DOCX)

Text S3 Sequence alignment of Trc promoter & RBS elements.
(PDF)

Dataset S1 Training data set and test data set for model NET90_19_576.
(DOCX)

Model S1 The ANN model NET90_19_576 provided as Matlab format.
(RAR)

Author Contributions

Conceived and designed the experiments: YW HM JW. Performed the experiments: HM JW. Analyzed the data: HM JW. Wrote the paper: HM JW ZX FX GZ YW.

References

1. Dehli T, Solem C, Jensen PR (2012) Tunable promoters in synthetic and systems biology. Subcell Biochem 64: 181–201.
2. Boyle PM, Silver PA (2012) Parts plus pipes: synthetic biology approaches to metabolic engineering. Metab Eng 14: 223–232.
3. Blount BA, Weenink T, Vasylechko S, Ellis T (2012) Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology. PloS One 7: e33279.
4. Qin X, Qian J, Yao G, Zhuang Y, Zhang S, et al. (2011) GAP Promoter Library for Fine-Tuning of Gene Expression in *Pichia pastoris*. Appl Environ Microb 77: 3600–3608.
5. Alper H, Fischer C, Nevoigt E, Stephanopoulos G (2005) Tuning genetic control through promoter engineering. P Natl Acad Sci USA 102: 12678–12683.
6. Straney R, Krah R, Menzel R (1994) Mutations in the −10 TATAAT sequence of the *gyr*A promoter affect both promoter strength and sensitivity to DNA supercoiling. J Bacteriol 176: 5999–6006.
7. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, et al. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol 30: 521–530.
8. Kiryu H, Oshima T, Asai K (2005) Extracting relations between promoter sequences and their strengths from microarray data. Bioinformatics 21: 1062–1068.
9. Harley CB, Reynolds RP (1987) Analysis of *E.coli* pormoter sequences. Nucleic Acids Res 15: 2343–2361.
10. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol 27: 946–950.
11. De Mey M, Maertens J, Lequeux G, Soetaert W, Vandamme E (2007) Construction and model-based analysis of a promoter library for *E.coli*: an indispensable tool for metabolic engineering. BMC Biotechnol 7: 34.
12. Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σ^E. P Natl Acad Sci USA 107: 2854–2859.
13. Na D, Lee D (2010) RBSDesigner: software for designing synthetic ribosome binding sites that yield a desired level of protein expression. Bioinformatics 26: 2633–2634.
14. Jensen PR, Hammer K (1998) The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters. Appl Environ Microb 64: 82–87.
15. Erb R (1993) Introduction to Backpropagation Neural Network Computation. Pharm Res-Dordr 10: 165–170.
16. Agatonovic-Kustrin S, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharmaceut Biomed 22: 717–727.
17. Kakumani R, Devabhaktuni V, Ahmad M (2008) A two-stage neural network based technique for protein secondary structure prediction. Conf Proc IEEE Eng Med Biol Soc 2008: 1355–1358.
18. Qu W, Sui H, Yang B, Qian W (2011) Improving protein secondary structure prediction using a multi-modal BP method. Comput Biol Med 41: 946–959.
19. Capriotti E, Fariselli P, Casadio R (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20: i63–i68.
20. Koessler DR, Knisley DJ, Knisley J, Haynes T (2010) A predictive model for secondary RNA structure using graph theory and a neural network. BMC Bioinformatics 11 Suppl 6: S21.
21. Wang J, Ungar LH, Tseng H, Hannenhalli S (2007) MetaProm: a neural network based meta-predictor for alternative human promoter prediction. BMC Genomics 8: 374.
22. Askary A, Masoudi-Nejad A, Sharafi R, Mizbani A, Parizi SN, et al. (2009) N4: a precise and highly sensitive promoter predictor using neural network fed by nearest neighbors. Genes Genet Syst 84: 425–430.
23. de Avila ESS, Gerhardt GJ, Echeverrigaray S (2011) Rules extraction from neural networks applied to the prediction and recognition of prokaryotic promoters. Genet Mol Biol 34: 353–360.
24. Demeler B, Zhou GW (1991) Neural network optimization for *E.coli* promoter prediction. Nucleic Acids Res 19: 1593–1599.
25. Horton PB, Kanehisa M (1992) An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. Nucleic Acids Res 20: 4331–4338.
26. Mahadevan I, Ghosh I (1994) Analysis of *E.coli* promoter structures using neural networks. Nucleic Acids Res 22: 2158–2165.
27. O'Neill MC (1992) *Escherichia coli* promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. Nucleic Acids Res 20: 3471–3477.
28. Zhu HM, Wang JX (2006) Predicting eukaryotic promoter using both interpolated Markov chains and time-delay neural networks. Proceedings of 2006 International Conference on Machine Learning and Cybernetics, Vols 1–7: 4262–4267.
29. Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 35: W126–W131.
30. Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat Biotech 21: 796–802.
31. Sambrook J, Russell DW (2001) Molecular cloning. A laboratory manual 3rd. edition. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.
32. Kelly J, Rubin A, Davis J, Ajo-Franklin C, Cumbers J, et al. (2009) Measuring the activity of BioBrick promoters using an *in vivo* reference standard. J Biol Eng 3: 4.
33. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.
34. Fan S, Sun Z, Jiang D, Dai C, Ma Y, et al. (2010) BmKCT toxin inhibits glioma proliferation and tumor metastasis. Cancer Lett 291: 158–166.
35. Fu YJ, An N, Chan KG, Wu YB, Zheng SH, et al. (2011) A model of BmK CT in inhibiting glioma cell migration via matrix metalloproteinase-2 from experimental and molecular dynamics simulation study. Biotechnol Lett 33: 1309–1317.
36. Ingham AB, Moore RJ (2007) Recombinant production of antimicrobial peptides in heterologous microbial systems. Biotechnol Appl Biochem 47: 1–9.
37. Shao JH, Wang YQ, Wu XY, Jiang R, Zhang R, et al. (2008) Cloning, expression, and pharmacological activity of BmK AS, an active peptide from scorpion *Buthus martensii Karsch.* Biotechnol Lett 30: 23–29.
38. Tegel H, Ottosson J, Hober S (2011) Enhancing the protein production levels in *Escherichia coli* with a strong promoter. FEBS J 278: 729–739.
39. Santos K, Duke CMP, Rodriguez-Colon SM, Dakwar A, Fan S, et al. (2007) Effect of promoter strength on protein expression and immunogenicity of an HSV-1 amplicon vector encoding HIV-1 Gag. Vaccine 25: 1634–1646.
40. Levine E, Zhang Z, Kuhlman T, Hwa T (2007) Quantitative Characteristics of Gene Regulation by Small RNA. PloS Biol 5: e229.
41. Bourdon J, Eveillard D, Siegel A (2011) Integrating Quantitative Knowledge into a Qualitative Gene Regulatory Network. PloS Comput Biol 7: e1002157.
42. Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. Nat Biotechnol 26: 787–793.