Systems Biology

A comprehensive web tool for toehold switch design

Andrew Ching-Yuet To[1,†], David Ho-Ting Chu[1,†], Angela Ruoning Wang[1,†], Frances Cheuk-Yau Li[2,†], Alan Wai-On Chiu[3], Daisy Yuwei Gao[2,3], Chung Hang Jonathan Choi[3], Siu-Kai Kong[1], Ting-Fung Chan[1], King-Ming Chan[1] and Kevin Y. Yip[2,3,*]

[1]School of Life Sciences, [2]Department of Computer Science and Engineering, [3]Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Background: Toehold switches are a class of RNAs with a hairpin loop that can be unfolded upon binding a trigger RNA, thereby exposing a ribosome binding site (RBS) and permitting translation of the reporter protein. They have been shown very useful in detecting a variety of targets including RNAs from Zika and Ebola viruses. The base complementation between the toehold switch and the trigger RNA also makes it sensitive to sequence variations. Design of toehold switches involves a series of considerations related to their sequence properties, structures and specificities.

Results: Here we present the first comprehensive web tool for designing toehold switches. We also propose a score for predicting the efficacy of designed toehold switches based on properties learned from ~180 experimentally tested switches.

Availability: The toehold switch web tool is available at https://yiplab.cse.cuhk.edu.hk/toehold/.

1 INTRODUCTION

Toehold switches are a class of riboregulators that can be activated when interacting with the trigger RNA (Green *et al*., 2014). Each toehold switch contains a stable hairpin loop that includes an RBS and a start codon among other stacking pairs (Fig. 1a). Downstream of the hairpin loop is the coding sequence of a gene. The stable loop structure prevents the binding of ribosome and as a consequence translation of the downstream gene is repressed. The upstream region of the hairpin loop contains a single-stranded toehold domain that is designed to be complementary to part of the trigger RNA to be detected (which we describe here as "Part 1"), with the remaining of the trigger RNA ("Part 2") complementary to the domain right next to the toehold. When the trigger RNA is present and interacts with the toehold domain, the hairpin loop is unfolded, exposing the RBS and the start codon, which permits translation of the downstream gene.

Toehold switches can be used to detect trigger RNAs of interest by having a reporter gene downstream of the hairpin. For example, they have been successfully used in cell-free, paper-based sensors for detecting the Zika virus RNA genome (Pardee *et al*., 2016), which provides a convenient way for efficient testing in the field. On the other hand, toehold switches can also be used to construct gene regulatory circuits including the regulation of endogenous genes, which have been shown to have a wide dynamic range and low crosstalk (Green *et al*., 2014).

For a given trigger RNA sequence, the design of a suitable toehold switch involves the consideration of a series of factors:

1. The toehold switch itself should form a stable hairpin loop structure to avoid expression of the downstream gene in the absence of the trigger RNA.
2. The toehold switch-trigger RNA duplex should have a more favorable energy state as compared to the unbound toehold switch and trigger RNA, to ensure the unfolding of the hairpin loop and the activation of translation by the interaction.
3. The toehold switch sequence should be free of unwanted stop codons that forbid the translation of the downstream gene.
4. The trigger RNA recognized by the toehold switch should be unique to this RNA, to avoid off-target effects.

Based on these requirements, we have developed a web tool that guides the user step-by-step in designing toehold switches.

2 WORKFLOW

Our web tool takes as input an RNA sequence of interest and various design parameters, such as the promoter and RBS sequences to be inserted into the toehold switch, length of the trigger RNA (i.e., Part 1 and Part 2, which interact with the toehold switch), and experiment temperature. Users can also choose to produce some optional outputs, such as rare codon count, minimum free energy terms of the monomers and dimmers, and counts of paired and unpaired bases, which could be useful for ranking the resulting list of candidate toehold switches.

Upon receiving the inputs, our web tool carries out a number of steps to design the toehold switches (Fig. 1b). After checking the format and validity of the inputs, it uses a sliding window to consider every x consecutive bases of the input RNA sequence as the potential trigger RNA, where x is the length of this region as specified by the user. A candidate toehold switch is constructed based on the sub-sequence in this window and the other input parameters as follows. The single-stranded toehold domain and the domain right next to it that forms half of the internal loop (the orange region in Figure 1a) are directly complementary to Part 1 and Part 2 of the trigger RNA, respectively. A fixed connecting sequence (the black region) is then added next to it, followed by the ribosome

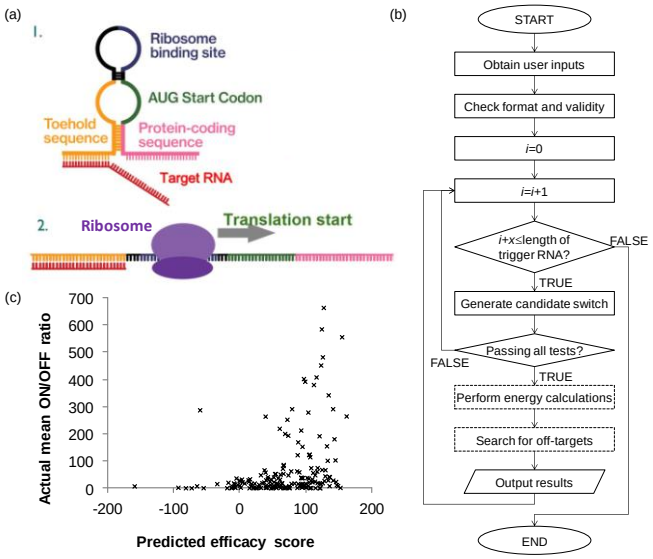

Fig. 1. (a) Sequence components and secondary structure of a toehold switch. Upon binding the trigger RNA (1), the toehold switch will be opened up to allow the translation of the coding sequence (2). (b) Workflow of our web tool for designing toehold switches. (c) A plot of the actual experimentally measured ON/OFF ratios against our efficacy scores for the 181 switches from Green *et al*. (2014).

binding site (the blue region) and the start codon (the green region). After that, Part 2 of the trigger RNA is added (the double-stranded part of the pink region) such that it forms stacking pairs with the domain right next to the toehold domain. Finally, the protein-coding sequence of the reporter gene is added (the single-stranded part of the pink region).

The constructed candidate toehold switch sequence is then subject to a number of tests. First, the trigger RNA is used to search a sequence database to check for highly similar off-target sequences. Currently, we perform this search against the human expressed sequence tags (ESTs). More databases can be added later upon users' requests. The candidate is rejected if such an off-target sequence is found, or if it contains an additional stop codon between the start codon and the downstream gene, or consecutive bases of the same type that would lead to structural instability. If the switch can pass all these tests, additional user-specified constructs such as promoters and restriction sites will be added. Finally, free energy calculations are performed for the switch and the trigger monomers, and their interacting dimer, if the user chooses to output such information.

3 EFFICACY SCORE

Since the outputs contain a variety of information about the designed switches, it is not trivial to rank them based on their expected efficacy. To help users identify the most effective switches, we developed an efficacy score using machine learning. Specifically, we took 181 switches from Supplementary Tables S1 and S3 of Green *et al*. (2014), and computed the free energy terms using our tool. Each of these switches was experimentally tested and the efficacy was represented by an ON/OFF ratio. We used our computed energy terms to construct a linear regression model with ridge regularization to predict these ON/OFF ratios by the LinearRegression class in Weka (Hall *et al*., 2009). It uses the Akaike information criterion (AIC) to select the best model. Based on a 10-fold cross-validation setting, the Spearman correlation between our predicted efficacy scores and the actual measured ON/OFF ratios was 0.4, suggesting that our efficacy score is a reasonable predictor of switch effectiveness (Fig. 1c).

Trained with all 181 switches, the resulting model was:

$$\begin{aligned} 29.421 &\times \text{rbsLinkerMFE} + \\ -10.8327 &\times \text{switchMFE} + \\ 7.6235 &\times \text{switchSwitchDimerMFE} + \\ 6.899 &\times \text{mfeDifference} + \\ -19.7348 &\times \text{toeholdDomainPairedBases} + \\ 210.3128, & \end{aligned}$$

where the selected features are respectively the MFE of the RBS linker, the MFE of the toehold switch, the MFE of switch-switch dimer, the MSE difference (switch monomer + trigger monomer – switch-trigger dimer) and the number of paired bases right next to the toehold domain. We have added this score to our tool as part of the standard output.

We have also used this formula to test the efficacy of 10 toehold switches of a new design reported in Green *et al*. (2017). The Spearman correlation between our computed efficacy scores and the experimentally measured ON/OFF ratios were 0.22, which is lower than the results for the 181 switches in Green *et al*. (2014) but is still substantially higher than the expected value of 0 for a random ordering.

4 CONCLUSION

Our web tool provides a one-stop solution for designing toehold switches, taking into account various important considerations. We plan on adding more features in the future, including molecular dynamics simulation of the interaction between toehold switches and trigger RNAs, and evolutionary conservation of different parts of the input sequence that can help identify strain-specific regions.

Currently, due to the small number of training examples, the performance of the efficacy score cannot be easily improved by replacing the linear regression by some non-linear methods such as artificial neural networks and support vector regression with a non-liner kernel. We will attempt to improve the efficacy score when experimental data of more toehold switches become available.

ACKNOWLEDGEMENT

This work originated from the project of the 2017 iGEM team of The Chinese University of Hong Kong.

REFERENCES

Green A.A., Silver P.A., Collins J.J. and Yin P. (2014) Toehold switches: de-novo-designed regulators of gene expression. *Cell* 159(4), 925-939.

Green A.A., Kim J., Ma D., Silver P.A., Collins J.J. and Yin P. (2017) Complex cellular logic computation using ribocomputing devices. *Nature* 548(7665), 117-121.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10-18.

Pardee K., Green A.A., Takahashi M.K., Braff D., Lambert G., Lee J.W., Ferrante T., Ma D., Donghia N., Fan M., Daringer N.M., Bosch I., Dudley D.M., O'Connor D.H., Gehrke L. and Collins J.J. (2016) Rapid, low-cost detection of Zika virus using programmable biomolecular components. *Cell* 165(5), 1255-1266.