End-to-end computational approach to the design of RNA biosensors for miRNA biomarkers of cervical cancer

Priyannth Ramasami S. Baabu^{1*¶}, Shivaramakrishna Srinivasan^{2*§}, Swetha Nagarajan¹, Sangeetha Muthamilselvan², Raghavv R. Suresh¹, Thamarai Selvi², and Ashok Palaniappan^{2**}

¹*Department of Bioengineering, School of Chemical and Biotechnology, SASTRA Deemed University, Thanjavur, Tamil Nadu - 613 401. India.*

²*Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA Deemed University, Thanjavur, Tamil Nadu - 613 401. India.*

*These authors contributed equally.

**Corresponding Author: apalania@scbt.sastra.edu

Abstract

Cervical cancer is a global public health subject as it affects women in the reproductive ages, and accounts for the second largest burden among cancer patients worldwide with an unforgiving 50% mortality rate. Poor awareness and access to effective diagnosis have led to this enormous disease burden, calling for point-of-care, minimally invasive diagnosis methods. Here, an end-to-end quantitative approach for a new kind of diagnosis has been developed, comprising identification of optimal biomarkers, design of the sensor, and simulation of the diagnostic circuit. Using miRNA expression data in the public domain, we identified circulating miRNA biomarkers specific to cervical cancer using multi-tier screening. Synthetic riboregulators called toehold switches specific for the biomarker panel were then designed. To predict the dynamic range of toehold switches for use in genetic circuits as biosensors, we developed a generic grammar of these switches, and built a multivariate linear regression model using thermodynamic features derived from RNA secondary structure and interaction. The model yielded predictions of toehold efficacy with an adjusted $R^2 = 0.59$. Reaction kinetics modelling was performed to predict the sensitivity of the second-generation toehold switches to the miRNA biomarkers. Simulations showed a linear response between 10nM and 100nM before saturation. Our study demonstrates an end-to-end workflow for the efficient design of genetic circuits geared towards the effective detection of unique genomic signatures that would

[¶]Present address: School of Advanced Materials Science and Engineering, Sungkyunkwan University (SKKU), Suwon, 16419, Republic of Korea

[§]Present address: Life Sciences Division, TCS Research, Tata Consultancy Services, Chennai, Tamil Nadu 600042. India

be increasingly important in today's world. The approach has the potential to direct experimental efforts and minimise costs. All resources including the machine learning toolkit, reaction kinetics simulation, designed toehold sequences, genetic circuits, data and sbml files for replicating and utilizing our study are provided open-source with the iGEM Foundation (<https://github.com/igem2019>) under GNU GPLv3 licence.

Keywords

Cervical cancer biosensor; miRNA biomarker; synthetic toehold switches; genetic circuit design; reaction network modelling; toehold efficacy modelling; toehold switch grammar; machine learning.

1. Introduction

Cervical cancer is the second most common cancer affecting women worldwide [1–3], with 20% of cases in India [4,5]. Most cervical cancers could be attributed to the HPV16 and HPV18 strains of Human Papilloma Virus (HPV) [1,2]. Lifestyle factors also contribute to the etiology of the disease [2]. Cervical cancer tumorigenesis begins with the metaplastic epithelium being infected by HPV strains at the cervical transitional zone, followed by the persistence of the virus, then the progression of the epithelium cells to pre-cancer stage, and finally the progression to cancer [6]. Pap smear test has been a widely used diagnostic method but an invasive, expensive, and time-consuming one [4,7]. Alternative testing strategies are necessary [8–10], and micro-RNAs (miRNAs) circulating in the blood have emerged as effective biomarkers in many cases [11–14]. Coupling data-driven methods for the identification of reliable early-stage diagnostic / prognostic biomarkers with the design of biosensor offers potential for the development of a new generation of molecular diagnostics [15–18].

Synthetic biology is delivering on its promise of harnessing nature's inherent diversity for the betterment of human health [19]. Bioelectronic diagnostic devices comprise an integrated single-unit reaction chamber housing the sensor element with the necessary reagents for analyte detection, and the corresponding transduction module for providing a reaction readout with optical, piezoelectric, or electrochemical means [20–22]. Toehold switches are a class of synthetic riboregulators with excellent biosensor properties, namely specificity, modularity and orthogonality [23], and such sensors that could sense virtually any RNA sequence [23]. could be freeze-dried on a microfluidic platform for biomarker detection using a cell-free colorimetric assay [24]. Toehold-switch riboregulators are mRNA elements which function by

occluding the ribosome from translating a downstream gene (Fig. 1(a)). Upon base pairing with the trigger RNA sequence, the toehold structure of the switch unfolds, permitting expression of a reporter gene. The toehold-switch concept was extensively utilised in the design of a biosensor for detecting Zika virus infection [25,26]. Takahashi *et al.* designed a toehold switch sensor that could detect the conserved region of *C. difficile* in a paper based platform [27]. In this work, we have formulated an alternative testing strategy for cervical cancer based on data-driven identification of biomarker miRNAs, design of cognate toehold-based sensors, and predictive modelling of circuit performance. This synthetic biology approach is the first integrated pathway connecting data from biomarkers to biosensor design and modelling. Software associated with the study is hosted at <https://github.com/igem2019>, and [Supplementary Data could be found at https://doi.org/10.6084/m9.figshare.14915619.v1](https://doi.org/10.6084/m9.figshare.14915619.v1).

2. Methods

The overall strategy is outlined in Table 1, and discussed in detail below.

2.1 Biomarker Panel Development

The Cancer Genome Atlas (TCGA) provided the miRNA-Seq dataset for normal tissue and cervical adenocarcinoma samples [30]. The dataset contained 312 samples of which 309 were tumour samples and 3 were control samples (tumour-normal matched). The associated clinical data annotated the stage information in the attribute 'patient.stage_event.clinical_stage'. The number of samples in stages I, II, III and IV were found to be 163, 70, 46 and 21 respectively. The patient sample barcode encoded in variable *hybridisation_REF* was parsed, and the samples were correspondingly annotated as cancer and normal.

RSEM-normalized Illumina HiSeq miRNASeq gene expression data [31] was log₂ transformed, and analyzed for differential expression using R (<https://www.r-project.org/>) *limma* package [32]. Linear modelling of stage-annotated gene expression matrix was performed, followed by empirical Bayes adjustment to obtain moderated t-statistics. To account for multiple hypothesis testing and the false discovery rate, the p values of the F-statistic of linear fit were adjusted using the BH method [33]. Differentially expressed miRNAs were considered significant if the absolute log-fold changes > 1.5x and the p-value was <0.05 [34]. A contrast analysis for identifying stage-specific DE miRNAs was then carried out. Finally, the association between the significant early-stage DE miRNAs and the overall survival of a patient was evaluated by univariate Cox proportional hazards regression analysis,

and the significant prognostic miRNA biomarkers were retained. All analysis was done R *survival* package [35].

Table 1. Iterative Design, Build, Test & Learn (DBTL) cycles deployed for developing toehold-switch biosensors specific for a panel of miRNA biomarkers of early-stage cervical cancer.

DBTL phase	Iteration 0: Biomarker signature development	Iteration 1: Optimal toehold switch design	Iteration 2: Reaction network modelling of circuit performance
DESIGN	TCGA data-driven differential expression analysis, survival analysis	Machine learning to design optimal toeholds for identified biomarkers	Modelling the designed circuit using chemical and biochemical kinetics
BUILD	Cancer vs. control linear models, univariate K-M plots, multivariate Cox regression models (using R)	Feature engineering of toehold datasets using ViennaRNA [28] and in-house Python and shell scripts [29]	Generalising the model to second-generation toeholds with anti-miRs, to inform the design of experiments
TEST	Models evaluated using p-values and log fold-change of differential expression, and p-values of prognosis models.	Multivariate regression model of toehold efficacy evaluated using R ² goodness of fit (using R)	Validation of predictive modelling
LEARN	Filtering to four biomarkers, then to three, and finally to two	Optimised toehold switch sequences	Sensitivity, onset of saturation, and specificity of each sensor

2.2 Toehold Switch Design

2.2.1 Toehold switch grammar

A toehold structure consists of two hairpin loops and two base paired stems, that together serve as translation regulator by occluding the RBS in the absence of the cognate trigger molecule ('ground state') [23] (Fig. 1). Upon binding of the cognate miRNA (i.e, the trigger sequence) to the toehold domain, the bottom stem unravels and the double-looped structure of the toehold switch collapses to a linear structure, which exposes the sequestered start codon to expression of the reporter gene. Each toehold switch is a modular entity with distinct sequence domains that could be engineered *de novo* for a particular trigger RNA. Toehold switches hold great potential as sensors for exact sequences of RNA, and are custom designed to hybridise with the cognate molecules. Table 2 describes the regions from 5' to 3' in the generic toehold switch grammar illustrated in Fig. 1(b).

Fig 1: (a) Anatomy of a toehold switch (b) Generic grammar of toehold switch structure for domain parsing, in dot bracket notation.

Table 2. Dot-bracket notation of the generalised toehold switch grammar

Region	Description
d1 dots	Toehold domain - not part of bottom stem
b1 open brackets	Toehold domain - part of bottom stem, based paired with linker
3 dots (first occurrence)	Sequestered nucleotides opposite to start codon, forming secondary loop
b2 open brackets	Ascending top stem
d2 dots	Primary loop containing pre-RBS and RBS
b2 closed brackets	Descending top stem
3 dots (second occurrence)	Start codon
b1 closed brackets	Linker sequence base paired with toehold domain. This leads to reporter gene of interest

The following preliminary criteria were used in the rational design of toehold sequences:

1. To allow effective binding of the trigger to the switch, the horizontal portion of the toehold domain was longer than 10 nucleotides.
2. The descending bottom stem of the toehold switch was fixed at 12 nucleotides, maintaining switch stability in the OFF state as well as the ORF after the start codon.
3. The canonical RBS sequence AGAGGA was used.
4. The primary loop was made adenine-rich to maintain a larger loop structure.
5. The horizontal linker after the bottom stem was retained as AACCUGGCGGCAGCGCAAAG, the 21nt long linker used in Green et al. [23].
6. Mature miRNAs are about 22 nts, and might include a stop-codon trinucleotide towards the 5' region of their sequence. The presence of such stop-codon subsequences could halt the expression of the downstream GFP. Second-generation toehold switches are one solution to this problem [36]. In addition to binding the miRNA, the toehold switch is required to bind another molecule, the anti-miRNA (antimiR) at its 3' end. AntimiRs

are antisense RNA that are complementary and bind to 12 nucleotides at the 3' end of the toehold switch. Essentially, the trigger is represented by a hybrid molecule with a base-paired region and two free ends. The design of the anti-miR sequence could be used in improving the trigger binding event.

2.2.2 Predictive modelling of toehold switch efficacy

The dynamic range describes the sensitivity of the toehold switch, and is the ratio between the maximum and minimum measurable intensities of reporter protein (typically GFP) expression, also known as ON/OFF ratio [23]. The dynamic range of a toehold designed for exact sequences of trigger RNA is the key indicator of its effectiveness, and is generally unknown for new toehold designs. The problem can be addressed using supervised machine learning on toehold-sequence datasets with available dynamic range responses. A previous model developed by CUHK iGEM 2017 offered a web tool for predicting the dynamic range of toehold switches with a modest goodness of fit ($R^2 \sim 0.2$) [37]. To improve on this, we constructed a multivariate linear regression model using the generic grammar of the toehold switch and features engineered from sequence and secondary structure. A validated dataset of the toehold switches with the dynamic range response was available from Green et al. (2014) [23] and Pardee et al (2014, 2016) [24,25]. A total of 230 toehold switch sequences with ON/OFF ratios were thus obtained. The toehold sequence of each instance was parsed into its sub-domains using the generic grammar presented in Fig 1(b). The dot-bracket notation is a representative grammar RNA secondary structure using dots '.' to represent unpaired bases, and matching parentheses '(' and ')' to represent paired bases..additional considerations. 228 toehold instances that conformed to the grammar in Fig 1B were parsed into dot-bracket sequence regions using an in-house regex script (available at <https://github.com/igem2019>) [29].

Some considerations in domain parsing included:

1. Wobble base pair between G-U in the secondary loop containing the start codon.
Wobble base pairs offered better stability in the OFF state and these forward engineered designs could be less leaky
2. Base-paired toehold domains,
3. Base-paired linker regions.
4. Possible unpaired bases in the top stem.

The sequence and structure features used are presented in Table 3. All structural features were calculated using the ViennaRNA suite.

Table 3. Sequence and structural features of interest in the predictive modelling of a toehold switch.

S.No	Feature	Type
1	Toehold domain length	Sequence
2	Number of GC base pairs in the top stem	Sequence
3	Number of GC base pairs in the bottom stem	Sequence
4	Length of the top stem	Sequence
5	Length of the bottom stem	Sequence
6	Length of the primary loop	Sequence
7	Number of A/U rich regions before RBS	Sequence
8	Overall minimum free energy (MFE) of the toehold switch construct estimated using ViennaRNA RNAfold	Structure
9	Minimum free energy of the bottom region of the toehold switch estimated using ViennaRNA RNAfold	Structure
10	Ensemble diversity of the switch and the frequency of the MFE structure in an ensemble estimated using ViennaRNA RNAfold	Structure
11	The solvent accessibility of the entire construct	Structure
12	The net MFE of the second-generation switch-trigger complex	Structure
13	MFE of the top stem	Structure
14	MFE of the bottom stem	Structure
15	MFE of the RBS-linker domain	Structure

Feature selection was performed by removing features correlated with Pearson's $\rho > 0.85$ were removed. Following a 80:20 train:test split, a multivariate linear regression model was built with the selected features on the train data, and evaluated on the test data for significance and goodness-of-fit, using Python scikit-learn (<https://scikit-learn.org/>). The scripts and machine learning model are available in the Supplementary Information [29].

2.2.3 gBlock Design

gBlocks are medium-sized genetic constructs capable of housing one or more genes [38], and deposited in the Registry of Biological Parts [39]. Three independent gBlocks were designed using the standard BioBricks (<https://biobricks.org>), including the constitutive *E. coli* T7 promoter, a 5' prefix sequence with EcoRI and XbaI recognition sites, and a 3' suffix sequence with SpeI and PstI recognition sites (Fig. 2). The reporter gene used was GFP-mut3b, capable of enhanced fluorescence due to three point mutations relative to wild GFP. We found that splitting our parts over multiple gBlocks offered less error-prone and faster synthesis. The gBlocks are available in the Supplementary Material.

Fig 2: Three gBlocks were designed. (i) Toehold switch-GFP construct; (ii) expression system for miR oligo; and (iii) expression system for anti-miR oligo. To allow amplification of the gBlocks, the 5' prefix is concatenated with the VF2 forward primer, and the 3' suffix is concatenated with the VR reverse primer. pLac promoters are specifically inducible with lactose.

2.3 Reaction network modelling

The chemical and biochemical kinetics model was developed by adding the following features to the first generation toehold switch model developed by CLSB-UK iGEM 2017 [40]:

- (i) interaction of transcribed miR and anti-miR oligos to form the miR-anti-miR complex with two free ends, which then binds with the toehold switch
- (ii) maturation kinetics of the expressed GFP. This generalized model can be presented as:

The corresponding ODEs can be given by:

$$\begin{aligned}
 \frac{d[miRNA]}{dt} &= -k_{miR_antimiR_b}[antimiR][miRNA] + k_{miR_antimiR_nb}[miRNA - anti_miRNA] - k_{decay}[miRNA] \\
 \frac{d[anti_miRNA]}{dt} &= -k_{miR_antimiR_b}[antimiR][miRNA] + k_{miR_antimiR_nb}[miRNA - anti_miRNA] - k_{decay}[anti_miRNA] \\
 \frac{d[miRNA - anti_miRNA]}{dt} &= -k_{miR_antimiR_b}[antimiR][miRNA] + k_{miR_antimiR_nb}[miRNA - anti_miRNA] \\
 \frac{d[miRNA - anti_miRNA]}{dt} &= -k_{complex_CTS}[CTS][miRNA - anti_miRNA] \\
 \frac{d[CTS]}{dt} &= k_{transcription}[DNA] - k_{complex_CTS}[CTS][miRNA - anti_miRNA] - k_{decay}[CTS] \\
 \frac{d[OTS]}{dt} &= k_{complex_CTS}[CTS][miRNA - anti_miRNA] - k_{decay}[OTS] \\
 \frac{d[GFP]}{dt} &= k_{translation}[OTS]
 \end{aligned}$$

The kinetic parameters for the transcription, translation and maturation reactions in the genetic circuit were identified from literature [40,41]. The kinetics for the miR-antimiR complex formation were modelled assuming a steady state with negligible backward flux, so that the equilibrium constant approximates the forward rate constant at equilibrium. The equilibrium constant for the complex formation is defined as:

$$k_{eq} = \frac{[hsa_miR_21_5p - anti_hsa_miR_21_5p]}{[hsa_miR_21_5p][anti_hsa_miR_21_5p]} \quad \text{----- Equation (4)}$$

NUPACK was used to obtain the equilibrium concentration values of the miR, antimiR and their complex [42]. The chemical and biochemical kinetics were simulated using MATLAB R2017b (<https://www.mathworks.com/>).

3. Results

3.1 Omics Analysis

The top 100 miRNAs of the linear model were screened for log-fold-change with respect to the control samples, stage-specificity, and significance of the trend, and finally corroborated with

the literature evidence. This analysis yielded four DE miRNAs (Table 4) [12,43]. The full results of the differential expression analysis and linear modelling are presented in Files S1 and S2 respectively. It is seen that has-miR-21-5p is the most significantly differentially expressed miRNA in squamous cell cervical cancer. It is interesting to note that all four miRNAs are upregulated, highlighting their role as oncomiRs (File S3).

Table 4. Putative DE miRNA biomarkers of cervical cancer. LogFC (log-fold change) and adj. p-values with respect to the control as estimated by the linear model are shown, along with the stage-specificity (log-fold change and significance of contrast analysis) of the respective miRNA.

DE miRNAs	logFC	Adj. p value	Stage_1 - Control	Stage_2 - Control	Stage_3 - Control	Stage_4 - Control	Stage-specificity
hsa-miR-21-5p	15.75	3E-151	2.63	2.63	2.75	2.65	4E-08
hsa-miR-20a-5p	6.60	5E-26	2.01	2.02	2.22	2.20	1E-02
hsa-miR-29a-3p	13.06	1E-70	1.99	2.31	2.18	2.64	6E-03
hsa-miR-200a-3p	3.27	9E-07	6.60	6.43	6.32	6.63	3E-17

3.2 Survival Analysis

The differential expression profiling of miRNAs associated with early-stage cervical cancer identified potential biomarkers, which were further investigated using survival analysis with the Cox proportional hazards model and Kaplan-Meier curves. The biomarkers were subjected to a univariate survival analysis and the model significance was estimated using the log rank test [28]. Fig. 3 shows the K-M plots of the univariate survival analysis for the 4 differentially expressed miRNAs. It could be seen that hsa-miR-29a-3p was not significant at 95% (p-value > 0.05), and was consequently dropped from the hazard ratio estimation (Table 5).

Table 5 - Hazard ratio (HR), Z-score and p value for the chosen miRNA biomarkers. A HR > 1 indicates a covariate that is positively associated with the event probability (negatively with survival).

Gene Symbol	HR	Z-score	p-value
hsa-miR-20a-5p	0.523	-2.479	0.013
hsa-miR-200a-5p	1.808	2.076	0.038
hsa-miR-21-5p	0.595	-1.961	0.050

Fig 3. K-M plot for (a) hsa-miR-20a-5p, (b) hsa-miR-200a-5p, (c) hsa-miR-21-5p and (d) hsa-miR-29a-3p

Next, we constructed a survival risk-score using the multivariate Cox regression model [21], and obtained eqn (5).

$$\begin{aligned}
 \text{Survival Risk Score (SRS)} &= (-0.5967 * \text{expression value of hsa-miR-20a-5p}) \\
 &+ (0.5376 * \text{expression value of hsa-miR-200a-3p}) \\
 &+ (-0.4736 * \text{expression value of hsa-miR-21-5p}) \\
 &\text{----- Equation (5)}
 \end{aligned}$$

The coefficients of hsa-miR-20a-5p and hsa-miR-21-5p are negative, indicating a protective effect of these miRNAs. Finally, a biomarker panel of hsa-miR-20a-5p, hsa-miR-21-5p and hsa-miR-200a-3p, where the 307 patients who were considered for our analysis was used to classify patients into high-risk and low-risk groups using an optimal cut point identified with the maxstat statistic of *survminer* R package. Overall survival curves were generated using

the K-M method, and two-sided log rank tests were used to compare the differences in overall survival between the two risk groups. The K-M curve shown in Fig. 4 (a), indicated that the biomarker panel with the three miRNAs was significant (p-value < 0.001). We also evaluated a biomarker panel of just the two miRNAs with protective effect, hsa-miR-20a-5p and hsa-miR-21-5p, considering the significance of hsa-miR-21-5p holds in literature [12,43–48]. The K-M curve shown in Fig. 4 (b), indicated that the significance of this panel was as good as the panel with three miRNAs (p-value < 0.001). From all the above considerations, we settled on the two-marker panel, of hsa-miR-20a-5p and hsa-miR-21-5p, for the prognostic classification of early-stage cervical cancer.

Fig 4. (a) Kaplan-Meier curves of panel with the three chosen biomarkers: hsa-miR-20a-5p, hsa-miR-21-5p and hsa-miR-200a-3p (b) Kaplan-Meier curves of panel with the two chosen biomarkers: hsa-miR-20a-5p and hsa-miR-21-5p

3.3 Toehold Switch Design

Two toehold switches were designed to target the identified miRNA biomarkers, hsa-miR-21-5p and hsa-miR-20a-5p (provide supplementary info on miRNA sequences, toehold sequences, and corresponding DNA sequences). The folding of the transcribed RNA of the designed sequences was predicted using NUPACK [42] and ViennaRNA [28], and the secondary structure was consistent with the toehold conformation (Fig. 5a, b). Neither of the two structures have any rogue base-pairings in the toehold domain. In addition to the secondary structure, the minimum free energy and the dynamic range of the designed toehold switches are key parameters in the effectiveness of the system. The minimum free energies of the hsa-miR-21-5p toehold switch and the hsa-miR-20a-5p toehold switch were estimated to be -21.20 kcal/mol and -19.20 kcal/mol respectively, suggesting a stable toehold conformation, ready for interaction with the cognate miR-antimiR complex.

Fig 5: (a) MFE structure of Second Generation Toehold Switch for hsa-miR-21-5p, (b) MFE structure of Second Generation Toehold Switch for hsa-miR-20a-5p shown in NUPACK

Feature selection for our machine learning model yielded six non-redundant features, all thermodynamic in origin:

1. Overall MFE of the toehold switch construct
2. Frequency of the MFE structure in the ensemble
3. Net MFE (difference in energy between the MFE of the switch-trigger complex and the sum of individual MFE values of the switch and trigger)
4. MFE of the bottom region containing the toehold domain, bottom stem and the horizontal linker
5. MFE of the RBS-linker domain, starting from the primary loop till the end of the construct
6. Specific heat of the toehold switch construct

It is notable that net MFE which is an energetic feature arising from the interaction of the second-generation toehold with the miRNA-antimiR hybrid has emerged as a key feature in the modelling process. The significance of the coefficients of the resulting multivariate linear regression model are noted in Table 6. The model yielded an adj. $R^2 \sim 0.59$ on the test set. With this validation, we applied the model to the designed toehold switches for the two identified miRNA biomarkers. The toehold switch for hsa-miR-21-5p yielded a predicted dynamic range of 52.4, while that of hsa-miR-20a-5p yielded 100.8, both values indicating potential robust

performance of the designed biosensors. The model is freely available for use in toehold-based synthetic biology [29]. All the designed toehold switch sequences and genetic circuits are available in File S4, and the BioBricks submissions are noted in File S5.

Table 6. Significance of the selected features in the constructed multivariate linear regression model (All features were significant)

S. No.	Feature	Coefficient	p-value
1	Overall MFE	16.39	1.7E-03
2	Bottom_region_MFE	-32.81	1.8E-11
3	RBS_linker_MFE	19.21	2.8E-06
4	Net_MFE	5.24	3.3E-03
5	Freq_MFE_struct (frequency)	-327.33	2.4E-02
6	Overall_SH_37 (specific heat)	68.21	6.3E-05

3.4 Reaction Modelling

We employed reaction network mass action kinetics to computationally evaluate the performance of the designed toehold switches as biosensors of miRNA biomarkers (available as executable SBML model in github repository; submitted to EBI BioModels Repository). Before understanding the reaction kinetics of our designed second generation toehold switches, we initially decided to emulate the model proposed by iGEM CLSB-UK (2017) [40] to obtain an insight of their kinetic model on first generation toehold switches and anticipate the plausible errors that could be encountered during our simulations (File S6)

3.4.1 Generalization to Second generation Toehold Switches

We performed NUPACK runs to obtain the structures of hsa-miR-21-5p and anti-hsa-miR-21-5p [42], and analyzed the equilibrium of their interaction at 18 °C. Using the equilibrium concentrations of the complex, hsa-miR-21-5p, and anti-hsa-miR-21-5p (99.97 nM, 0.0264 nM, and 0.0264 nM respectively), the $k_{\text{miR_antimiR_f}}$ was estimated to be in the order of 10^5 nM^{-1} . The concentration profiles for hsa-miR-21-5p, anti-hsa-miR-21-5p and their complex are shown in Fig. 6(a).

Fig 6. (a) Concentration profiles of miR, anti-miR and the complex. It is seen that miR and anti-miR profiles are identical. (b) Concentration profiles of the complex, CTS and OTS in a 2 h time frame with a miR-anti-miR binding rate constant of $10^5 \text{ mol}^{-1}\text{s}^{-1}$. The first 500s are highlighted with an inset over the simulation time of 7200s.

The simulation time was set to 7200 s (2 h) allowing for miRNA-anti-miR complex formation. From the profile in Fig. 6(a), it is clear that the complex formation occurs rather rapidly and both the complex and the miR/anti-miR attain saturation by 500s. We used an incubation time of 8 minutes for the cell-free transcription of the genetic construct to yield significant quantities of the RNA toehold switch [41].

The intensity of GFP fluorescence is the readout from the proposed biosensor. A GFP maturation reaction was included in the model with a maturation factor $\sim 0.2/\text{min}$. To obtain the readout in terms of fluorescence units, we utilized the scaling factor of 79.429 proposed by iGEM Valencia_UPV [49]. The initial DNA concentration was set to 223.5 nM, following established experimental protocols.

From the concentration profile seen in Fig. 6(b), the CTS concentration was initially $\sim 118 \text{ nM}$ due to the transcriptional incubation time. The complex concentration was found to be 98.74 nM upon solving the differential equations. The CTS and complex rapidly hybridize to open the CTS, evident from the significant drop in their concentrations along with a rise in the OTS concentration. At about 500s, the complex gets completely consumed, following which the OTS begins to decay, while CTS continues to rise due to unhindered transcription. Upon

looking into the dynamics of GFP and mature GFP during the initial 500s, translation of GFP occurs with increase in OTS concentration followed by a lazy maturation of the GFP.

The final phase of our modelling included the process of obtaining trends for fluorescence intensity with variations in complex concentration - from 100 pM to 100 nM. We plotted the intensity at 2 h interval against the complex concentration and obtained a linear fit between 10 nM to 100 nM with a goodness-of-fit $R^2 = 0.9993$. This showed that the model is sensitive to changes in the concentration of the trigger miRNA biomarker, satisfying a necessary biosensor property [50–52]. We repeated the process for a time interval of 4h where a similar trend was obtained ($R^2 = 0.9992$) (Fig.7).

Fig 7. The fluorescence intensity varies linearly with complex for a concentrations between 10 nM and 100 nM over (a) 2h and (b) 4h.

Table 7 provides the concentration trajectories of various species and the predicted fluorescence intensity readout for different miRNA-antimiR complex concentrations assuming a 2-hr reaction time. The intensity values for an extended 4-hr reaction time are provided in Supplementary Section S2.

Table 7 - Simulated fluorescence intensity variations with different concentrations of the complex over a 2h period.

miRNA/	100 pM	1 nM	10 nM	25 nM	50 nM	75 nM	100 nM
--------	--------	------	-------	-------	-------	-------	--------

antimiR concentration							
Complex (miR-antimiR- Toehold switch)	98 pM	980 pM	9.87 nM	24.69 nM	49.37 nM	74.05 nM	98.74 nM
CTS	739 nM	738 nM	737 nM	736 nM	733 nM	729 nM	726 nM
OTS	0.018 nM	0.18 nM	1.17 nM	3.01 nM	5.85 nM	9.74 nM	12.1 nM
GFP	0.067 nM	0.67 nM	6.66 nM	16.9 nM	32.32 nM	54.1 nM	65.8 nM
Mature GFP	0.0048 μM	0.048 μM	0.48 μM	1.21 μM	2.43 μM	3.08 μM	4.85 μM
Fluorescence Intensity	1.74	17.40	174.16	432.12	868.03	1348.51	1728.04

4. Discussion

Geraldi *et al.* highlighted the importance of synthetic biology in developing portable *in vitro* based diagnostic kits [53]. Toeholds embedded in genetic circuits enable the measurement of the intensity of expression of a downstream reporter protein – for e.g, GFP – that could yield a precise quantification of the trigger itself. Pardee has reviewed the potential of freeze-dried cell-free (FD-CF) systems in health care for diagnostic and sensing purposes, given their biosafe mode [54], and Tinarfar *et al* reviewed cell free systems and the promising opportunity for the rational design and manipulation of biological systems in relation to cell based systems [55]. Fabricating microfluidic devices with our designed genetic circuits in a cell-free system followed by rigorous testing and validation would yield a point-of-care portable diagnostic tool for the real-time detection of early-stage cervical cancer.

Towards this end, uniting theory and data is key, and Figure 8 encapsulates the methodology we have adopted here, in two parts. The end-to-end computational approach that we have employed in this study is illustrated in Fig. 8(a). It broadly includes four phases, namely identification of biomarker signatures, design of the biosensory element - toehold switch followed by the synthetic genetic circuit design comprising modular constructs and finally validating a proof-of-concept of the designed synthetic circuit using predictive modelling strategies. This approach could be generalised to the design of portable, highly sensitive and specific biosensors for any condition and especially infectious diseases. The current nCoV-

SARS2 pandemic has highlighted the need for constant vigil for outbreak of infections. Such emerging agents require accurate, rapid, and scalable detection platforms, all requirements for which toehold biosensors are well-suited. Fig. 8(b) illustrates a possible scheme for the development and production of such a toehold biosensor. Here the biomarker would be an optimal genomic fragment signature, unique to the pathogen but not in the host genome. The toehold construct serves to signal the presence of the infectious agent that would then release the expression of the reporter gene, yielding a simple visual readout in real-time within minutes. The proposed technique upon execution would generate a simple, reliable, rapid, portable and affordable printed diagnostic tool that could be made available at a global level in a point-of-care setting outside clinical diagnostic laboratory conditions. This simple biosafe product would be an asset in the field, especially in remote locations where it could help in curtailing the transmission and limiting the reproductive ratio of the infectious agent. Such a product would be vital in low-resource settings since it is sterile, abiotic and active for a year at room temperature. Moreover the product is easily scalable, saving the lives of countless frontline corona warriors and resulting in improved patient outcomes. Extension of more than two biomarkers would allow for the simultaneous detection of, say, a family of coronaviruses, an imperative in our times. Thus, end-to-end science is expected to overcome the limitations of ad-hoc diagnostic devices, thereby enabling affordable diagnosis in pandemic situations and other health emergencies.

Fig 8. (a) General scheme representing the end-to-end computational approach for design of toehold switches based biosensors (b) Scheme representing the modus operandus for designing and developing biosensors working on a riboregulator based sensing platform where toehold switches and SARS-CoV-2 viral genomic content act as the biosensory element and analyte, respectively

5. Conclusion

Cervical cancer is a major public health issue, but an addressable one with more effective diagnosis technologies. Early detection results in better treatment and greater compliance. In this work, three DBTL iterations were used to generate a workflow for *in silico* biomarker detection, sensor design and systems dynamics modelling. In the first cycle, we identified miRNAs that were significantly differentially expressed for early-stage cervical cancer using TCGA data, and identified the optimal two miRNA biomarkers. In the second cycle, we designed second generation toehold switches and anti-miR for the two miRNAs. In this course, we developed a machine learning model for toehold dynamic ranges that provided an $\text{adj.}R^2 = 0.59$, a major improvement in efforts in this direction. Finally we simulated the reactions in the cell free system to study their kinetics. From our observations, fluorescence intensity followed a linear range between 10 nM and 100 nM after which we observed a saturation. End-to-end automation of the computational ideas in our workflow would essentially democratize detection of emerging infectious diseases and life-threatening conditions with a readily transportable instrument based on a real-time visual readout.

Author contributions

Conceived and designed the work: AP, SS, PRS. Performed research: PRS, SN, SS, SM, RRS, TS. Analyzed the results: PRS, SS, SN, SM, RRS, AP. Writing- first draft: PRS, SS, SN, SM, RRS. Writing- editing and revision: AP. Supervision & project administration: AP.

Conflict of Interest

The authors declare no competing interests, financial or otherwise.

Acknowledgements

We are grateful to SASTRA Deemed University for financial and infrastructural support to SASTRA iGEM 2019 Team (https://2019.igem.org/Team:SASTRA_Thanjavur). A.P. would like to acknowledge support from DST-SERB EMR/2017/000470.

References

- [1] Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. Lancet 2007. [https://doi.org/10.1016/S0140-6736\(07\)61416-0](https://doi.org/10.1016/S0140-6736(07)61416-0).
- [2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global

- Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021;71:209–49.
<https://doi.org/10.3322/caac.21660>.
- [3] Stelzle D, Tanaka LF, Lee KK, Ibrahim Khalil A, Baussano I, Shah AS V, et al. Estimates of the global burden of cervical cancer associated with HIV. *Lancet Glob Heal* 2021;9:e161–9. [https://doi.org/10.1016/S2214-109X\(20\)30459-9](https://doi.org/10.1016/S2214-109X(20)30459-9).
 - [4] Chandigarh; PGIMER. Health Technology Assessment of Strategies for Cervical Cancer Screening in India. 2019. <https://dhr.gov.in/health-technology-assessment-strategies-cervical-cancer-screening-india> (accessed May 2020)
 - [5] Globocan. Cervix Uteri. <https://gco.iarc.fr/today/data/factsheets/cancers/23-Cervix-uteri-fact-sheet.pdf>; 2020 [accessed January, 2021]
 - [6] Silverberg SG, Loffe OB. Pathology of Cervical Cancer. *Cancer J* 2003;9:335–47. <https://doi.org/10.1097/00130404-200309000-00003>.
 - [7] Ogedegbe G, Cassells AN, Robinson CM, DuHamel K, Tobin JN, Sox CH, et al. Perceptions of barriers and facilitators of cancer early detection among low-income minority women in community health centers. *J Natl Med Assoc* 2005;97:162–70.
 - [8] Jearanaikoon P, Prakrankamanant P, Leelayuwat C, Wanram S, Limpaboon T, Promptmas C. The evaluation of loop-mediated isothermal amplification-quartz crystal microbalance (LAMP-QCM) biosensor as a real-time measurement of HPV16 DNA. *J Virol Methods* 2016;229:8–11. <https://doi.org/10.1016/j.jviromet.2015.12.005>.
 - [9] Wang Y, Chen M, Zhang L, Ding Y, Luo Y, Xu Q, et al. Rapid detection of human papilloma virus using a novel leaky surface acoustic wave peptide nucleic acid biosensor. *Biosens Bioelectron* 2009;24:3455–60. <https://doi.org/10.1016/j.bios.2009.04.034>.
 - [10] Surriabre P, Torrico A, Vargas T, Ugarte F, Rodriguez P, Fontaine V. Assessment of a new low-cost, PCR-based strategy for high-risk human papillomavirus DNA detection for cervical cancer prevention. *BMC Infect Dis* 2019;19:842. <https://doi.org/10.1186/s12879-019-4527-9>.
 - [11] Roy S, Soh JH, Ying JY. A microarray platform for detecting disease-specific circulating miRNA in human serum. *Biosens Bioelectron* 2016. <https://doi.org/10.1016/j.bios.2015.08.039>.
 - [12] Pardini B, De Maria D, Francavilla A, Di Gaetano C, Ronco G, Naccarati A. MicroRNAs as markers of progression in cervical cancer: A systematic review. *BMC*

- Cancer 2018. <https://doi.org/10.1186/s12885-018-4590-4>.
- [13] Filipów S, Łaczmański Ł. Blood Circulating miRNAs as Cancer Biomarkers for Diagnosis and Surgical Treatment Response. *Front Genet* 2019;10:169. <https://doi.org/10.3389/fgene.2019.00169>.
 - [14] Cui M, Wang H, Yao X, Zhang D, Xie Y, Cui R, et al. Circulating MicroRNAs in Cancer: Potential and Challenge. *Front Genet* 2019;10. <https://doi.org/10.3389/fgene.2019.00626>.
 - [16] Dembélé D, Kastner P. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics* 2014;15:14. <https://doi.org/10.1186/1471-2105-15-14>.
 - [18] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6. <https://doi.org/10.1038/415530a>.
 - [19] Slomovic S, Pardee K, Collins JJ. Synthetic biology devices for in vitro and in vivo diagnostics. *Proc Natl Acad Sci* 2015;112:14429–35. <https://doi.org/10.1073/pnas.1508521112>.
 - [20] Vernon SD, Farkas DH, Unger ER, Chan V, Miller DL, Chen YP, et al. Bioelectronic DNA detection of human papillomaviruses using eSensorTM: A model system for detection of multiple pathogens. *BMC Infect Dis* 2003. <https://doi.org/10.1186/1471-2334-3-12>.
 - [21] Ho NRY, Lim GS, Sundah NR, Lim D, Loh TP, Shao H. Visual and modular detection of pathogen nucleic acids with enzyme–DNA molecular complexes. *Nat Commun* 2018;9:3238. <https://doi.org/10.1038/s41467-018-05733-0>.
 - [22] Xu Y, Liu Y, Wu Y, Xia X, Liao Y, Li Q. Fluorescent Probe-Based Lateral Flow Assay for Multiplex Nucleic Acid Detection. *Anal Chem* 2014;86:5611–4. <https://doi.org/10.1021/ac5010458>.
 - [23] Green AA, Silver PA, Collins JJ, Yin P. Toehold switches: De-novo-designed regulators of gene expression. *Cell* 2014. <https://doi.org/10.1016/j.cell.2014.10.002>.
 - [24] Pardee K, Green AA, Ferrante T, Cameron DE, DaleyKeyser A, Yin P, et al. Paper-Based Synthetic Gene Networks. *Cell* 2014;159:940–54. <https://doi.org/10.1016/j.cell.2014.10.004>.
 - [25] Pardee K, Green AA, Takahashi MK, Braff D, Lambert G, Lee JW, et al. Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* 2016;165:1255–66. <https://doi.org/10.1016/j.cell.2016.04.059>.

- [26] Hall RA, Macdonald J. Synthetic Biology Provides a Toehold in the Fight against Zika. *Cell Host Microbe* 2016;19:752–4. <https://doi.org/10.1016/j.chom.2016.05.020>.
- [27] Takahashi MK, Tan X, Dy AJ, Braff D, Akana RT, Furuta Y, et al. A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers. *Nat Commun* 2018;9:3347. <https://doi.org/10.1038/s41467-018-05864-4>.
- [28] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011;6:26. <https://doi.org/10.1186/1748-7188-6-26>.
- [dataset] [29] Srinivasan, S; Palaniappan, A; SASTRA_Thanjavur iGEM Software 2019. https://github.com/igemsoftware2019/SASTRA_Thanjavur.
- [30] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013;45(10):1113-1120. doi:10.1038/ng.2764
- [dataset] [31] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run. Broad Inst MIT Harvard Dataset 2016. https://doi.org/http://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/CESC/20160128.
- [32] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015. <https://doi.org/10.1093/nar/gkv007>.
- [33] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* 1995. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- [34] Sarathi A, Palaniappan A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. *BMC Cancer* 2019. <https://doi.org/10.1186/s12885-019-5838-3>.
- [35] Therneau T. A Package for Survival Analysis in S. R package version. *Survival (Lond)* 2012.
- [36] Green AA, Kim J, Ma D, Silver PA, Collins JJ, Yin P. Complex cellular logic computation using ribocomputing devices. *Nature* 2017. <https://doi.org/10.1038/nature23271>.
- [37] To AC-Y, Chu DH-T, Wang AR, Li FC-Y, Chiu AW-O, Gao DY, et al. A comprehensive web tool for toehold switch design. *Bioinformatics* 2018;34:2862–4.

<https://doi.org/10.1093/bioinformatics/bty216>.

- [38]
- [39] International Genetically Engineered Machines. Registry of Standard Biological Parts http://parts.igem.org/Main_Page. [accessed May 2021]
- [40] CLSB-UK. Project BATMAN <http://2017.igem.org/Team:CLSB-UK>. 2017 [accessed 2019]
- [41] Stögbauer T, Windhager L, Zimmer R, Rädler JO. Experiment and mathematical modeling of gene expression dynamics in a cell-free system. *Integr Biol* 2012. <https://doi.org/10.1039/c2ib00102k>.
- [42] Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, et al. Software News and Updates NUPACK: Analysis and Design of Nucleic Acid Systems. *J Comput Chem* 2011. <https://doi.org/10.1002/jcc.21596>.
- [43] Balacescu O, Balacescu L, Baldasici O, Tudoran O, Achimas-Cadariu P. The Role of miRNAs in Diagnosis, Prognosis and Treatment Prediction in Cervical Cancer. *Colposc. Cerv. Pathol.*, 2017. <https://doi.org/10.5772/68011>.
- [44] Zhao S, Yao D, Chen J, Ding N, Ren F. MiR-20a Promotes Cervical Cancer Proliferation and Metastasis In Vitro and In Vivo. *PLoS One* 2015;10:e0120905. <https://doi.org/10.1371/journal.pone.0120905>.
- [45] Kang H-W, Wang F, Wei Q, Zhao Y-F, Liu M, Li X, et al. miR-20a promotes migration and invasion by regulating TNKS2 in human cervical cancer cells. *FEBS Lett* 2012;586:897–904. <https://doi.org/10.1016/j.febslet.2012.02.020>.
- [46] Zhang Z, Wang J, Wang X, Song W, Shi Y, Zhang L. MicroRNA-21 promotes proliferation, migration, and invasion of cervical cancer through targeting TIMP3. *Arch Gynecol Obstet* 2018;297:433–42. <https://doi.org/10.1007/s00404-017-4598-z>.
- [47] Zamani S, Sohrabi A, Hosseini SM, Rahnamaye-Farzami M, Akbari A. Deregulation of miR-21 and miR-29a in Cervical Cancer Related to HPV Infection. *MicroRNA* 2019;8:110–5. <https://doi.org/10.2174/2211536607666181017124349>.
- [48] Wei W-F, Han L-F, Liu D, Wu L-F, Chen X-J, Yi H-Y, et al. Orthotopic Xenograft Mouse Model of Cervical Cancer for Studying the Role of MicroRNA-21 in Promoting Lymph Node Metastasis. *Int J Gynecol Cancer* 2017;27:1587–95. <https://doi.org/10.1097/IGC.0000000000001059>.
- [49] Valencia_UPV. PRINETERIA. http://2018.igem.org/Team:Valencia_UPV. 2018 [accessed October, 2019].
- [50] Hwang Y, Kim SG, Jang S, Kim J, Jung GY. Toehold Switch-Based Genetic

Modulator for Tuning of Riboswitch Sensor. *Res Sq* 2020:1–16.

<https://doi.org/10.21203/rs.3.rs-23721/v1>.

- [51] Lee K-H, Kim D-M. In Vitro Use of Cellular Synthetic Machinery for Biosensing Applications. *Front Pharmacol* 2019;10. <https://doi.org/10.3389/fphar.2019.01166>.
- [52] Chan SK, Kuzuya A, Choong YS, Lim TS. DNA Switch: Toehold-Mediated DNA Isothermal Amplification for Dengue Serotyping. *SLAS Discov Adv Sci Drug Discov* 2019;24:68–76. <https://doi.org/10.1177/2472555218791743>.
- [53] Geraldi A, Giri-Rachman EA. Synthetic biology-based portable in vitro diagnostic platforms. *Alexandria J Med* 2018;54:423–8. <https://doi.org/10.1016/j.ajme.2018.11.003>.
- [54] Pardee K. Perspective: Solidifying the impact of cell-free synthetic biology through lyophilization. *Biochem Eng J* 2018;138:91–7. <https://doi.org/10.1016/j.bej.2018.07.008>.
- [55] Tinafar A, Jaenes K, Pardee K. Synthetic Biology Goes Cell-Free. *BMC Biol* 2019;17:64. <https://doi.org/10.1186/s12915-019-0685-x>.