Rational RNA Riboswitch Design through a Massive Open Laboratory

Vineet Kosaraju
Abstract

Designed RNA molecules are emerging as promising new, ‘on-demand’ agents for vaccinating against infectious disease, creating gene therapies, and other biomedical applications. However, for RNA to realize its full potential it must be possible to design riboswitches with multiple secondary structures that can either report or act on external conditions as diagnostics or therapeutics, and no such rational process exists. This work aims to both uncover missing design rules and create an algorithm that designs riboswitches. Using a game-based crowdsourcing approach and 200,000 citizen scientists, the EteRNA project successfully created empirically validated design rules and an algorithm, Eternabot, that designs for a single secondary structure. Using a similar approach, crowdsourcing was used to collect human insights into designing riboswitches, and machine learning was used to optimize these insights and derive RNA features. When compared to Eternabot, which designs for several conformations using EteRNA’s originally discovered design rules, the algorithm showed significant improvement both in predicting the stability of RNA riboswitches in silico and in designing perfect riboswitches in vitro. The discovered design rules incorporated into the algorithm along with these results suggest that the algorithm provides a novel, predictive framework for rational RNA nanotechnology for biomedicine.
A. Introduction

1. Background

Next generation research into therapeutics focuses on transforming medicine through the creation of intelligent, responsive RNA molecules, a key component of all cells that is useful in protein translation and gene expression (Clancy and Brown). Whereas current paradigms focus on the selective targeting of proteins’ active sites (Lahti et al.), this process is often difficult due to active site similarity (Nussinov and Tsai). Additionally, the majority of proteins cannot be targeted, as their functions cannot be inhibited via a small molecule approach (Arkin, Tang and Wells). Instead of small molecules, researchers have turned to RNA because they can sense and respond accordingly to a variety of targets, and hold potential in targeting a wide variety of diseases more effectively (Burnett and Rossi). RNA has already been utilized for diseases ranging from macular degeneration (Kaiser et al.) to cancer (Rao et al.). As such, RNA therapeutics hold promise in largely transforming the field of medicine by allowing for targeting of a wider variety of disease causative agents.

Although there have been significant advances in RNA therapeutics, RNAs until now have been engineered through trial and error in vitro selection processes, which are both expensive and time-consuming, and no rational method exists for accurately designing RNA molecules with a specific function. This research focuses on taking a rational approach to the problem of designing RNA molecules that encapsulate a specific biomedical function, via a unique citizen science approach piloted by the EteRNA project (Lee et al.). Ultimately, this research expands scientists’ basic knowledge of RNA secondary structure design through community-derived insights and removes a bottleneck in the path to creating complex RNA diagnostics, therapeutics, sensors, and nanotechnology.

2. RNA Secondary Structure

Similar to DNA, RNAs are composed of a chain of aromatic ring bases attached to a sugar-phosphate backbone. RNA forms a highly modular polymer composed of four nucleotide monomers: adenine, guanine, cytosine, and uracil, which characterize the RNA’s primary structure (Pietzsch). However, unlike DNA molecules, which are often double-stranded and form a standard double helix structure, RNA molecules can form complex structures due to intricate interactions. These complex folds are characterized in two dimensions by RNA’s secondary structure, where hydrogen bonding between two nucleotides form base pairs that fold into
helices, and unpaired nucleotides form loops. In the secondary structure, bases may interact either through the formation of strong Watson-Crick base pairs, between guanine and cytosine or adenine and uracil, or weaker base pairs, for example between guanine and uracil (Varani and McClain). Additionally, under changes in physiological conditions, such as temperature, or surrounding molecular concentrations, RNA sequences may change their structure, allowing them to hold multiple conformational states and reactions (Figure 1). These multi-state RNAs that are able to change structure in response to their environment are known as riboswitches, and are especially crucial in designing new diagnostics and therapeutics.

![RNA secondary structure](image)

**Figure 1:** RNA molecules can change conformations due to stimuli, including temperature and external molecules. The RNA portrayed has a single sequence, displayed at the bottom, which can switch between the two displayed secondary structure states, at the top, with the addition or removal of a small molecule called flavin mononucleotide. Base paired nucleotides are connected with gray lines, representing bonds, and unpaired nucleotides are unconnected and form circular loops.

### 3. Practical Applications of Riboswitch Design

Riboswitches have been recently utilized in therapeutics for a wide variety of applications. Apart from therapeutics, riboswitches also hold large potential for the creation of diagnostics that sense biomarkers and nanocomputers that handle basic logic gates. Furthermore, RNA riboswitches are an especially powerful targeting molecule as they are cheap and easy to manufacture at the primary structure level, and are highly modular in their ability to form complex folds. Any molecule that RNA can detect in its environment can also be detected by a
riboswitch, including a wide variety of small molecule ligands and other nucleic acid based molecules, thereby providing flexibility to RNA riboswitch diagnostics. For these reasons, the ability to design riboswitches is one that will aid researchers greatly in the future in a variety of fields. Unfortunately, current methods for designing riboswitches are mostly based on trial-and-error, where a starting sequence is mutated several thousands of billions of times. These sequences are synthesized, and the top sequences are selected and re-amplified, for up to 10-15 repetitions (Szeto et al.). As such, the overall process of designing a single accurate riboswitch is both overly expensive and time-consuming. Additionally, selection methods can only filter for RNAs with a binding affinity to external molecules, and cannot select for a specific RNA function. Therefore, no rational methods exist for accurately designing riboswitches with a specific medical function, providing a bottleneck for research into next-generation therapeutics and nanotechnology.

4. Crowdsourcing the RNA Design Problem

Two major challenges currently exist in the realm of RNA structural research: a prediction problem where a secondary or tertiary structure is to be derived from a specific sequence, and a design problem where an optimal RNA sequence is to be calculated from a specific secondary or tertiary structure. Due to the accurate measurements of free energy change parameters, contributing to the Turner Nearest Neighbors rules (Jaeger, Turner and Zuker), the former problem has been mostly solved for secondary structure prediction. However, the latter problem remains difficult due to the sheer amount of exponential computational space for solving a single secondary structure. In order to tackle the RNA design problem, researchers released an Internet-scale online game called EteRNA in which over 200,000 citizen scientist participants design RNA sequences for a variety of target secondary structures (Lee et al.). These sequences are synthesized in vitro and these citizen science gamers receive experimental results, allowing them to improve their designs. After a few years and several rounds of experimental testing, the researchers were able to collate insights from these participants into a machine-learning algorithm called Eternabot, which outperforms other state-of-the-art RNA design algorithms, such as NUPACK (Zadeh et al.) and RNAInverse (Gruber et al.). These results from EteRNA have demonstrated the ability of crowdsourcing to use the collective insights of a large community of people to tackle difficult problems such as designing a single RNA secondary
structure. However, no such effort to rationally design riboswitches in silico has previously existed.

5. Research Goals

This research has three specific goals that when combined hope to further research in next generation RNA engineering. Building on the success of Eternabot for single state designs, **Specific Goal 1** is to determine features of accurate RNA riboswitches using insights derived from participants of the EteRNA platform, who have recently begun investigating multi-state RNA molecules on several design challenges. These insights should accurately encapsulate several important features of accurate, sensitive, and responsive riboswitches. **Specific Goal 2** involves creating an algorithm that predicts the success of an arbitrary riboswitch, and comparing its accuracy to that of Eternabot, which designs for several conformations using EteRNA’s originally discovered design rules. Although other RNA design tools exist, they cannot create riboswitches that respond to external stimuli, and thus are not suitable for comparison. This goal tests the hypothesis that specific insights derived by non-expert participants will be able to accurately determine the ability of a RNA riboswitch with multiple secondary structures to change structural conformations. **Specific Goal 3** involves using the algorithm to design RNA riboswitches that respond to a variety of external stimuli, namely ligands and microRNAs, which were chosen due to their combined relevance in medical diagnostics (see Methods, Stimuli Response Modeling for Riboswitch Design). These riboswitches will be tested rigorously in vitro in order to confirm the algorithm’s ability to match a set of constraints to create RNA designs for biomedically relevant challenges. Additionally, these riboswitches will be further analyzed to improve the algorithm with specific in vitro results. Specifically, riboswitches that perform efficiently and are highly accurate will be analyzed to determine specific aspects of their primary or secondary structures that enable and aid conformational state switching. Similarly, riboswitches that are highly inaccurate will be analyzed to discover RNA features that the algorithm should exclude in future designs, allowing for future improvements. Ultimately, these three goals will determine if insights provided by a community of citizen scientists can enable the efficient and accurate development of rationally designed RNA riboswitches, enabling further research into RNA diagnostics and therapeutics.
B. Methods

1. Significance of RNA Medicine and Crowdsourcing

Prior to this work, participants in the RNA crowdsourcing platform EteRNA have begun investigating and creating well-functioning riboswitches, both in silico, and in vitro. However, this process occurred over several rounds of testing and is highly time-consuming. Rather than depend on a variable stream of solutions from non-experts, we can greatly reduce the time required to create new riboswitches by distilling insights and RNA features from these citizen scientists and incorporating them into a machine-learning algorithm. A crowdsourcing approach was utilized where citizen scientists from EteRNA submitted strategies that predicted an RNA sequence's success in responding to stimuli on a scale of 0 to 100 using features determined through previous experimental experience using the gamified massive open lab (Figure 2).

![Figure 2](image-url)

**Figure 2:** The interface used by EteRNA citizen scientists to determine specific features of RNA riboswitches that they hypothesized were linked to a riboswitches' ability to sense external molecules. At the top is a listing of all submitted designs, allowing citizen scientists to browse through the history of synthesized RNA molecules, and at the bottom is the gamified 'puzzle view' of a specific RNA design, allowing for a deeper look at the features of a RNA.
These strategies were collected from the participants to determine specific features of an RNA sequence that allow it to accurately switch states. In this pilot study, a total of 42 strategies were received, encapsulating more than 300 RNA features, and were subsequently implemented in Python, enabling the prediction of the accuracy of an arbitrary riboswitch. This study is thus especially significant as it is the first to both take a rational approach to creating biomedically relevant riboswitches and distill qualities of accurate riboswitches into design features that will aid in future research.

2. Experimental Validation

In order to validate features of riboswitches chosen by citizen scientists, previous RNA riboswitch data from EteRNA were utilized. Each month, in the EteRNA experimental pipeline, around 12,000 RNA molecules are placed on a sequencing chip and various concentrations of fluorescent bacteriophage MS2 coat protein are flowed onto the chip (Peabody). Each of these RNAs has been designed to bind to MS2 in one state, and not bind to it in the other. Because the protein is tagged with a fluorescent agent, RNA that are bound to it will appear as green, whereas unbound RNA will appear as red, allowing for determination of which state an RNA is in (Figure 3).

The RNA's binding affinity for MS2 can subsequently be easily determined by fitting a binding curve based on the specific concentrations of molecules that evoke state-change responses. This binding affinity represents the RNA's ability to both switch and respond to its environment, and is converted to a score between 0 to 100. On this scoring basis, any RNA riboswitch with a score above 80 is of high quality, and any riboswitch with a score of 100 is

![Figure 3](image-url)
considered perfect. This score has been shown to be reproducible, demonstrating its ability to correctly encapsulate the level of accuracy of an arbitrary riboswitch (Figure 4). Previous experimental data consists of several rounds of these microarray experiments that were carried out in the past year; these sequences and results are used for both helping the players learn and for training the algorithm using machine learning. The experimental data were split into both a training set and a validation set to perform cross validation on the implemented design algorithm.

3. Riboswitch Accuracy Predictions

After the strategies were collected and implemented, each was optimized over the training set using machine learning to reduce potential error. The strategies encapsulate specific RNA features that either aid or harm the responsiveness of riboswitches, and each feature is based off of parameters defined by the submissions created by citizen scientists. As these parameters are sometimes arbitrary, or inaccurate, it was first required that each parameter was modified to best predict the accuracy of RNA riboswitches. In order to do so, each strategy was run on a training set of RNA sequences and the average squared error between the strategy's predicted score and the actual synthesis score from the training set was calculated. The strategy's parameters were then modified using a Downhill Simplex algorithm (Nelder and Mead) to minimize the average squared error between the training set data and each strategy's predicted score (Figure 5).

Figure 4: Each riboswitch is validated experimentally to determine its ability to respond to external stimuli. This response is measured as a signal representing a binding affinity to the external stimuli, and is then converted to a score between 0 to 100. As the linearity of the plot demonstrates, this score has been shown to be highly reproducible across test experiments.
In order to create an accurate combination of these strategies to predict and design riboswitches, the RNA strategies were each assigned weights using a Lasso (L1) regularization, creating a sparse model (Figure 6). These weights allow for the creation of an algorithm that would proportionally use each strategy based on its accuracy on predicted experimental data from the training set. With this algorithm, researchers can successfully predict the accuracy of an RNA riboswitch in silico, which aids in reducing the complexity of designing riboswitches for in vitro use in nanotechnology or therapeutics.

4. Stimuli Response Modeling for Riboswitch Design

In order to design riboswitches, stimuli commonly used in therapeutics must first be modeled in silico. Two such stimuli were chosen for this research: small binding molecules called ligands, and microRNAs. Ligands, specifically flavin mononucleotide (FMN), were
chosen as a research subject due to the wide amount of research previously performed on identifying RNA aptamers sequences that can sense ligands at low concentrations; for example, FMN aptamers exist that sense the FMN molecule at concentrations as low as 0.5 μM (Fan et al.). MicroRNAs, which are short sequences of RNA molecules that are typically less than 25 nucleotides long, are also well studied, and additionally are functionally well conserved across different species of plants (Zeng et al.), animals, and microbes (Chen and Rajewsky). Further, microRNAs are involved in core tasks in the cell such as gene silencing (Huntzinger and Izaurralde), and can often be used as biomarkers for diseases (Briasoulis et al.). In this regard, miRNAs have been shown to be down-regulated in several diseases, including cancer, and thus serve as strong biomarkers for infectious agents and other illnesses (Mar-Aguilar et al.). As such, the ability to sense miRNAs, a task suited for RNA riboswitches, enables the creation of accurate diagnostics for diseases.

In order to determine if a FMN molecule successfully binds to a specific RNA, a free energy calculation was used. To perform these energy calculations, the algorithm uses ViennaRNA, a package for the prediction of RNA folds through energy minimization. With ViennaRNA, the energy of the molecule alone and the energy of the unbound RNA are calculated and compared to the energy of the bound RNA molecule. If the energy difference between the unbound RNA system and the bound RNA-molecule system can be satisfied by the energy bonus provided by the FMN molecule alone, then the FMN molecule binds. A similar approach is utilized to determine if a miRNA successfully binds to another RNA, where the free energy of the bound RNA duplex along with an energy bonus must be greater than the free energy of the RNA and the miRNA in independent systems. These calculations allow for rapid determination of stimuli modeling and aid accurate modeling of RNA riboswitches.

5. Implementation of Riboswitch Designer Algorithm

To design new RNA riboswitch sequences, a Monte Carlo simulation optimized over the two unique measures of a riboswitch’s accuracy: the predicted accuracy of a riboswitch based on crowdsourced RNA features, and the in silico modeling of a riboswitches’ response to stimuli. The simulation begins with an RNA sequence composed solely of Adenine nucleotides in unpaired regions, and strong Guanine-Cytosine bonds in base pairing regions. Based on this sequence, random mutations are made, and depending on the two scoring metrics, the new sequence is either accepted, or removed in the simulation. The first scoring metric calculates the
current RNA design's base pair distance between a variety of constraints on the riboswitch such as including required secondary structure features, meeting required amounts of specific nucleotide contents, excluding specific secondary structure features, excluding repetitious nucleotide sequences, and responding to stimuli in solely one state in an *in silico* model. This scoring metric must be fully satisfied to create an accurate RNA riboswitch. An additional scoring metric based on the predictive algorithm previously created using player submitted strategies was used to optimize the designed RNA riboswitches. In this manner, the algorithm uses the first scoring metric to ensure that the constraints of the RNA are met and that it accurately folds *in silico*, while optimizing the second scoring metric to increase the predicted *in vitro* score. With the combination of these two scoring metrics, an accurate RNA riboswitch design algorithm was created.

6. Importance of Riboswitch Design Puzzles

The design algorithm was evaluated on several separate RNA riboswitch puzzles, testing a riboswitch’s ability to bind to FMN and a miRNA at arbitrary locations in the RNA sequence. The specific miRNA chosen, microRNA-208a, has been linked to several cardiac disorders, including cardiac hypertrophy (Callis et al.), while FMN was chosen to test the riboswitches’ ability to respond to small molecules of similar low concentrations found in the body. Thus, using a riboswitch to detect these molecules could aid in creating cheap and feasible diagnostics for use in biomedicine (Figure 7). Each of these riboswitch puzzles required that a MS2 hairpin structure that binds to the fluorescent MS2 protein be present in one of two possible states, along with several other constraints. Several variations of these puzzles were created that moved the location of the MS2 hairpin to specified spots in the riboswitch design. Additionally, two unique types of variations were also created, to test if the algorithm could design riboswitches that applied to several situations. The first variation required that the MS2 hairpin must form in the same state as the external stimuli and thus the presence of the stimuli allows for its formation. Alternatively, the second variation required that the MS2 hairpin must form in the opposite state as the external stimuli and thus the absence of the stimuli allows for its formation.
C. Data Analysis and Results

1. Specific Goal 1: Strategy Collection and Riboswitch Feature Distillation

Specific Goal 1 focused on determining features of accurate RNA riboswitches through crowdsourcing and was satisfied by giving citizen science participants two weeks to formulate hypotheses and ideas regarding multi-state RNA molecules in a pilot study. At the end of this time period, citizen scientists had created 42 strategies, each of which was composed of several RNA features, in total encapsulating over 300 RNA features. Notably, some of the top strategies submitted by citizen scientists contain features that can be used as riboswitch design rules that determine the accuracy of an in vitro riboswitch through quick and efficient in silico modeling (Figure 8). These design rules include constraints to promote a response to external stimuli, limits for the percentage of types of nucleotides and base pairs, and principles that dictate the placement of RNA stacks, hairpins, and other secondary structure features. Additionally, several
other rules not proposed by citizen scientists were also included in the algorithm, to incorporate other tactics that are both commonly used and are available through in silico modeling. These rules include a calculation of the probability of pairing on a specific hairpin that indicates the success of a riboswitch design, a calculation of the free energy difference between the several conformational states, and a calculation of the change in base pairs in between states. Through combining the strategies from citizen scientists with common computational techniques, and using machine learning to optimize these features, the algorithm was successfully implemented, and thus Specific Goal 1 was accomplished.

1. Call to Action
2. Player Forum Posts
3. Collected Strategies
4. Successful Features

<table>
<thead>
<tr>
<th>Elements: MS2 Gates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encapsulated features: 6</td>
</tr>
</tbody>
</table>

When binding a microRNA, if it is a ‘turn-on’ RNA, reward if the MS2 gate is 5 bps, but if it is a ‘turn-off’, reward if the gate is 6-9 bps. […] An MS2 gate is defined as a stack that forms close to the MS2 hairpin. A ‘turn-on’ RNA binds the MS2 protein only in the presence of external stimuli, whereas a ‘turn-off’ RNA binds the MS2 protein in the absence of external stimuli.

**Figure 8:** After receiving a call to action (1), citizen scientists began formulating hypotheses and strategies about successful riboswitches in the form of forum posts (2). Afterwards, they submitted strategies to be implemented in the algorithm (3), which successfully distilled a variety of features to use in designing stimuli-responsive riboswitches (4). In this format, citizen scientists contributed by creating strategies with encapsulated features that could successfully predict the accuracy of an in vitro riboswitch.

**2. Specific Goal 2: Accuracy Comparison to Eternabot**

One aspect of determining whether the created algorithm can accurately synthesize in vitro riboswitches is analyzing its ability to predict and rationalize a large compendium of prior results. Because the predicted scores of the algorithm are directly used as a metric in the riboswitch generation simulation, if the predicted scores are highly inaccurate, it will be difficult for the algorithm to generate accurate riboswitches. Similarly, if the algorithm’s accuracy in predicting experimental scores is better than or worse than Eternabot’s accuracy, it will be easier or harder, respectively, to generate a perfect riboswitch. As such, Specific Goal 2 focused on testing the algorithm’s fidelity through large amounts of experimental data, in order to ascertain whether citizen scientist rules could accurately characterize a multi-state RNA’s ability to change its conformational shape and if the new algorithm improved over Eternabot.
In order to evaluate the algorithm’s accuracy, a validation set that encapsulated riboswitches was used for cross validation on both the new algorithm, and Eternabot, for comparison. This validation set was formed by randomly taking half of the RNA synthesis results from previous *in vitro* experiments (see Methods, Experimental Validation). When Eternabot was evaluated on the validation set to determine the accuracy of its predictive ability, it was determined that Eternabot’s predictions had a correlation of around -0.07 with the experimental riboswitch results, and had an average error of around 27.7% (Figure 9).

![Figure 9](image)

*Figure 9: When evaluated on the validation set, Eternabot has a low negative correlation of -0.07 and a high error of 27.7%, indicating its inability to accurately predict riboswitch success.*

This negative correlation that is close to 0 indicates that Eternabot has a very weak predictive ability for riboswitches, and that in general those predictions are inaccurate. Additionally, as the error is relatively large, Eternabot’s predictions are usually not indicative of the experimental results, and thus the predictive ability is weak. This error can further be seen as Eternabot’s predictions in general follow a distribution that is skewed to the left and centered at a score of 70, whereas experimental results indicate a bimodal distribution with centers at 30 and 60. Ultimately, because Eternabot’s prediction distribution’s center is higher than that of the distribution of actual experimental data, and its skew indicates that more solutions are predicted to have a high score, rather than their true lower experimental score, Eternabot’s predictions are
overly optimistic. These optimistic predictions are highly inaccurate, reducing its ability to be applied to riboswitches.

Alternatively, when the new algorithm was evaluated on the validation set, results were far more promising, indicating that it can successfully be applied to riboswitches (Figure 10). In terms of comparing the new algorithm’s predicted scores to actual experimental results, there is a stronger positive correlation of 0.207, indicating more powerful and accurate predictions for riboswitches. Additionally, the new algorithm has a lower error of 18.72% and in general predicted scores are closer to the experimental results. The new algorithm’s distribution of predicted scores effectively portrays this trend: while the distribution is relatively normal, it is now centered in between the two modes of the experimental distribution, making it more accurate. Ultimately, this increase in accuracy occurs across the board, in terms of correlation, average error, and predicting the spread of the data. As such, even though the algorithm’s predictions are not completely accurate, it has significantly improved from Eternabot. This improvement in the algorithm also demonstrates how citizen scientists can contribute meaningfully to scientific research and specifically to the realm of RNA engineering.

![Figure 10](image_url) **Figure 10:** When evaluated on the validation set, the new algorithm has a higher positive correlation of 0.207 and a lower error of 18.72%, indicating improvement in predicting riboswitch success.
3. Specific Goal 3: *In-Vitro* Synthesis of Original Designed Riboswitches

In order to determine if the new algorithm is feasible for biomedically relevant applications, we first considered whether its generated riboswitches would respond to external stimuli, such as ligands and miRNAs. As such, over six hundred RNA solutions were generated using the presented algorithm over the course of a few hours, and then synthesized *in vitro*. At the same time riboswitches generated by the Eternabot algorithm were also synthesized, to determine the extent to which the created algorithm generated improvement. Results reveal that while both Eternabot and the new algorithm created perfect riboswitches that can accurately change between multiple secondary structure conformations, the new algorithm was able to generate more of these riboswitches (Figure 11).

![Top Scoring RNA Riboswitches](image.png)

**Figure 11:** When analyzing biomedically relevant riboswitches with a score above 80 that were generated by each of the algorithms, the new algorithm outperforms Eternabot in the majority of the cases.

More specifically, the new algorithm successfully created more riboswitches in the score range of 80 to 100, indicating its improved ability to design responsive riboswitches. Although Eternabot outperforms the new algorithm in designing sequences with a score range of 60 to 80, and the new algorithm outperforms Eternabot in creating sequences with a score from 40-60 (Figure 12), these riboswitches, with a score from 40 to 80, are meaningless, as they are not
accurate enough to be used in medical applications. Further, for each of the top 11 scoring sequences of each algorithm that are above 80 and thus biomedically relevant, in a direct one to one comparison between the algorithms, the new algorithm outperforms Eternabot in 9 of the cases. As such, the new algorithm improved over Eternabot by creating more riboswitches, which also scored more highly, and generating more perfect riboswitches with a score of 100.

Figure 12: A one to one comparison of all of the generated riboswitches for each of the algorithms reveals that Eternabot outperformed the new algorithm in generating riboswitches with scores from 60 to 80. However, because these riboswitches have low experimental scores, they ultimately have little value in biomedicine as diagnostics.

4. Additional Riboswitch Design Rules

In all RNA riboswitch designs that accurately switched states and received high experimental in vitro scores, nucleotides that were involved in binding to the external stimuli formed areas of high positive free energy (Figure 13). This indicates that an unstable binding of the stimuli molecules can easily switch into a state with a more stable negative free energy. As such, a high positive free energy in binding areas is indicative of aiding riboswitch switching of conformational states, and thus aids responsiveness to stimuli.
Additionally, nucleotides that switched base pairs also formed loops of high positive free energy. It can thus be concluded using these free energy comparisons that to aid the formation of multiple states, riboswitches should be relatively unstable in specific regions containing hairpins or multiloops. Interestingly enough, regions with highly positive or negative free energies changed folds while regions with a slightly positive free energy remained conserved throughout conformational states. This observation indicates that when designing riboswitches, the algorithm can use free energy as a predictor of state change. These features of top RNA designs can be incorporated within the algorithm to rationally design riboswitches with higher free energies in nucleotides that either form loops or bind to stimuli so as to promote switching between two or more states. Thus, these additional design features can increase the algorithm’s accuracy and open up a variety of options for researchers to investigate.

D. Conclusions and Discussion

1. Conclusions

This study has for the first time demonstrated the ability of a game-based crowdsourcing approach combined with a machine-learning algorithm to successfully design RNA molecules
that can occupy several conformational states by responding to external stimuli. Using crowdsourcing to collate potential features that define accurate RNA riboswitches, and machine learning to optimize and select over these features, the newly created algorithm improves in silico methods in terms of prediction and riboswitch design. At a pilot stage, the algorithm has also successfully created in vitro RNA riboswitches that respond to biomedically relevant stimuli. Additionally, the algorithm improves over existing techniques in several ways, including time and accuracy. More specifically, whereas current crowdsourcing techniques require a month to generate solutions, the algorithm can generate solutions within minutes. Further, the algorithm’s predictive power greatly increases from previous computational techniques, with a significant increase in correlation, and it creates more riboswitches that are also more accurate at responding to stimuli. More notably, several design rules have been proposed by citizen scientists for secondary structure riboswitch design, including one dictating placement of specific RNA features, one that limits the percentage of specific nucleotide bases and base pairs, and one that places constraints on miRNA-responsive riboswitches. As these design rules have been experimentally validated, it has been shown that insights into multiple secondary structure RNA design created by citizen science participants can aid the creation of accurate riboswitches.

2. Future Work

Although the current results are promising and have several applications to biomedicine, other interesting areas of investigation and research can be explored. As participants in the EteRNA crowdsourcing platform do not fully understand RNA molecules that switch structures, it is valuable to incrementally include additional crowdsourced strategies as their understanding progresses. With more rounds of experimental results directly related to RNA riboswitches, citizen scientists will learn from their mistakes in previous RNA designs, and correct for those mistakes, eventually improving over time. After a period of several months, once citizen scientists have experimented with several rounds of RNA riboswitch synthesis, their new insights should be re-incorporated into the algorithm, allowing for the algorithm to select additional features, and greatly improving its accuracy. Additionally, apart from using a Lasso regularization to generate a sparse network for combining the strategy features, other methods such as L2 could be explored.
3. Overall Implications

Ultimately, this study has made several crucial advances to the field of RNA engineering for use in medical applications, by integrating the RNA crowdsourcing platform EteRNA with a novel machine-learning algorithm for designing riboswitches. *In vitro* results show that human intuition derived from thousands of citizen scientists combined with optimizations performed by computational techniques can be used to accurately predict the ability of an RNA sequence to switch states and to efficiently design stable, stimuli-sensitive RNA riboswitches. These results further reveal improvements from previous computational algorithms, and are on par with current time consuming techniques that use citizen science alone. Additionally, the combination of citizen science with machine learning has created several previously unestablished RNA features that govern the ability of a riboswitch molecule to respond to its environment. These features aid the rapid development of riboswitches by researchers with a simple *in silico* approach, saving time and money wasted on lengthy selection based efforts. Even at a pilot stage, the citizen scientist distilled features and the algorithm have already proven successful at designing perfectly responsive riboswitches to biomedically important design challenges, such as miRNA-208a. As such, the algorithm development and feature generation of this work will usher in a new wave of RNA therapeutics and diagnostics by enabling researchers to rapidly create RNA molecules that respond to their environment.
REFERENCES


