**Title:** Sequence constraints for efficient strand invasion for co-transcriptional gene regulation by an adenine sensing riboswitch

**Introduction:**

Riboswitches are regulatory regions of RNA that exert genetic control by coupling metabolite sensing with folding. These elements are located in the 5’ untranslated region of many bacterial genes, though also found in Archaea and few Eukarya (31128223). They are able to function in the absence of proteins to regulate gene output in a cis-fashion (15173824). These elements have two principal components: a receptor domain, and a regulatory domain. The receptor domain, also called the aptamer domain, is a highly structured and conserved region that binds a ligand with high specificity (21925376, 14523920, 15173824). The regulatory domain, or expression platform, is a dynamic region that creates a genetic output by folding co-transcriptionally based on the ligand input in the receptor domain (20943759). These genetic elements are of great interest for synthetic biology or clinical applications as potential tools because of their ability to function as sensors that can produce a genetic output. To leverage these genetic elements as tools, investigation must be done to understand underlying principles that govern the sequence landscape of important regions within the riboswitch.

While much work has been done in the past two decades to characterize riboswitches, the governing principles of the expression platform remain elusive. The unifying feature of this region between riboswitches is its ability to fold and exert allosteric control in response to aptamer occupancy, but that can be achieved in many ways and control gene expression at various levels (15173824). Current knowledge about riboswitch design comes from studies surveying natural riboswitches (25015992, 36150954) or evolving novel aptamers through techniques like SELEX that often modify existing riboswitch scaffolds (36617976, 28092358). Limitations to these approaches are three-fold. First, the focus is placed on the aptamer domain, usually with the goal of evolving a novel sensor for a chosen effector ligand. Second, many of these analyses produce answers without providing insight into the principals governing these sensors. Third, many of these aptamers are developed *in vitro,* potentially limiting their portability into other systems. This has created a lack of understanding both about overarching strategies to design nucleic acid-based sensors, and more specifically how to design and manipulate a regulatory domain.

The *pbuE* adenine responsive riboswitch from *B. subtilis* is a simple and small riboswitch. It is an ideal model system for the study of riboswitches, as it is well characterized and folds with high fidelity (25573585, 16201765, 25550163, 16931335, 24590258). It is a transcriptional switch that regulates expression of a purine efflux pump in *B. subtilis*, turning ON expression when excess adenine is present in the local environment (16931335, 25550163). Transcriptional riboswitches are under immense temporal constraints as they must survey the environment and either form the ON or OFF conformation before the RNA polymerase escapes the switch. Although they are referred to as riboswitches, the conformation with the terminator helix is far more thermodynamically stable, making the system work more as a co-transcriptional “fuse” than a switch (25573585,19595806, 15173824). The formation of the stable OFF conformation is only accomplished when the 3’ end of the riboswitch successfully invades into P1, displacing the initial duplex. This strand invasion is toehold-mediated by the duplex region P4, which stabilizes the 3’ end to invade into the intact P1 helix ().

Strand invasion is a prolific phenomenon in nucleic acid biology, relevant in various contexts from riboswitches to homologous recombination to CRISPR/Cas9 gene editing (37095744). For a system with a single stranded region invading into a duplex each migration step, that is the formation of a basepair for one duplex at the expense of the other duplex, takes between 10 and 20 µs (592403, 6264399, 37095744). During the invasion, the two strands will step forward and backward in a random walk process until the invading duplex is fully formed, ending the exchange process (37095744). A bias exists for the terminator helix over the aptamer domain because P(T) is inherently longer than P1, and P3 due to the toehold (P4) and lack of ssRNA (J3/1). It has also been shown that the stability of P4, the toehold, will determine the rate at which invasion proceeds (37095744, https://doi.org/10.1023/A:1023928811651, 19894722, 24019238). This bias reinforces the switch being more of a fuse and emphasizes the importance of ligand binding and understanding how the sequence identity of each strand in this regulatory element either allows or disallows the ability of the ON and OFF switch to form.

While techniques exist to evolve novel aptamers (), the ability to optimize and understand principles behind the regulatory domain remains a bottleneck in the ability to utilize riboswitches are tools. To that end, Drogalis and Batey completed a mutagenic analysis of regions of the expression platform of the *pbuE* riboswitch to glean information about how modifying the expression platform would impact the signal amplitude between the ON- and OFF-states of the switch (33259551).



The resulting sequence, schematized in Figure X, greatly outperformed the 5.6-fold switching of pbuE with a near 100-fold induction in cells. Modifications were made under the assumption that Watson-Crick (WC) base pairing in both the receptor and regulatory domains would create the best switch (33259551). Therefore, the switch became highly symmetrical, save the aptamer domain, and regions thought to be key to nucleation of the terminator helix (such as P4) were enriched with GC pairs. In addition to changing the primary sequence, the pre-aptamer sequence was shortened by 27 nucleotides to reduce misfolding and an AA mismatch was introduced in P3, proximal to L3, to promote strand invasion into the aptamer domain and destabilize the tertiary structure of the binding pocket. The mutagenic analysis, while yielding a high preforming variant did not aim elucidate the design principles behind this improved switching ability.

In this study, we aim to understand the design principles governing a transcriptional riboswitch’s ability to fold co-transcriptionally and attenuate gene expression (19595806). This study, through multiple genetic screens, offers insights into primary and secondary structural requirements for the processes of strand-invasion and nucleation that allow for the dynamic properties of this unique regulatory element. The regulatory domain of the riboswitch was evaluated through multiple discrete libraries, each randomizing a different region of interest. Those libraries were analyzed through qualitative and quantitative methods to assess the riboswitch functionality, and functional variants were extracted and sequenced. Through the analysis of regions in the terminator helix, we find that strand invasion nucleation residues are the most essential and highly conserved. When looking at paired regions or regions in competition between the aptamer and terminator helix, balance between the two paired regions must be tuned to optimize functionality. From these data it seems that the development of novel biosensors that use this scaffold can be easily designed with limited manipulation to key areas of the switch based on the parameters of the system.

**Methods:**

*Molecular Cloning of Riboswitch Libraries into Reporter Vector*

Riboswitches were amplified by PCR with consistent outer primers and a varied inner primer depending on library (Table #). PCR products were amplified for 12-20 rounds to minimize bias, product was confirmed on 2% agarose gel. Riboswitch libraries were then cloned into a reporter vector that was designed from the low-copy pBR327 plasmid (23654267, 25550163) using standard molecular cloning techniques (https://doi.org/10.1002/abio.370050118). The riboswitches were ligated upstream of *gfpUV* to control expression of fluorescent protein.

*Primary and Secondary Screening of Libraries*

Libraries were then transformed into *E. coli* Keio parental cells (BW25113) () with standard protocol. For those libraries that destabilized the expression platform (LIST THEM), transformants were plates onto CSB plates supplemented with 100 µg/mL carbenicillin. For the remaining libraries (LIST), which destabilized the aptamer domain were plated on CSB plates supplemented with 100 µg/mL carbenicillin and 0.5 mM 2AP. After overnight growth at 37 °C, plates were illuminated with a 3XX nm light to visualize *gfpUV* expression. These plates were photographed and the number of colonies on each plate was counted with the software Opencfu. Based on the number of colonies counted on each plate, likelihood of observing any one sequence was calculated using the second equation in *Reetz, MT et al* (18567049). When illuminated, those colonies that were dim on liganded plates or bright on unliganded plates were transferred onto a pair of gridded plates (one with ligand and one without) and again incubated overnight at 37 °C. Gridded plates were again illuminated with the same light after overnight growth and those colonies with were dim on the unliganded plates and bright on the liganded plates were selected to be screened further as they appeared to be functional riboswitches.

*Cell-based Activity Assay*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Library | nt varied | observed1 | functional2 | prob.3 |
| Anchor | 5 | 5,118 | 17 | 99.3 |
| Anchor\_exchange (P1\_gateway) | 5 | 5,069 | 97 | 99.3 |
| Gateway | 6 | 22,799 | 141 | 99.6 |
| L4 | 6 | 202 | 77 | 4.81 |
| P4 | 6 | 14,145 | 127 | 96.8 |
| P1\_exchange | 6 | 13,573 | 102 | 96.4 |
| P3 | 5 | 9,029 | 15 | >99.9 |

**Results and Discussion**

**TO DO:**

* **RENAME LIBRARIES**
* **SCHEMATIC WITH ALL LIBS**
* **ZOOMED IN SCHEMATICS WITH NUMBERING SYSTEMS**

**A genetic screen to elucidate design principles.**

To learn broadly about folding principles and sequence preferences of this nucleic acid element, a genetic screen was performed. The screen, while laborious, provides invaluable information about how those riboswitches that functionally optimally differ in sequence features from underperforming variants. As work is done to elucidate principles behind nucleic acids to port them into other contexts, screening over a more stringent protocol like selection was favorable. In screening multiple libraries, each highlighting a different feature of the regulatory domain, an artificial phylogeny of over 500 sequences was assembled with known activity.

**P3 and Gateway: Invading strand**

To understand the constraints involved with invading an intact duplex, two single stranded libraries were designed 3’ of the P4 toehold. Both libraries are invading a duplex region, P3 (P3 library, table S#) or P1 and P3 (Gateway library, table S#), but all of them randomized nucleotides were on the invading strand. The P3 library had 5 randomized nucleotides, spanning the length of P3. ~ 9,000 colonies were observed, 15 unique functional sequences were produced, and there was over a 99.9% chance of observing any given sequence. The Gateway library was larger, 6 randomized nucleotides, and included 2 bases in competition with P1, two paired with J3/1, and two in competition with P3. ~ 23,000 colonies were observed in the gateway library yielding 141 unique functional sequences, resulting in ~ 99.6% chance of observing any given sequence.

Chart

Description automatically generated with medium confidence

The gateway library had few high preforming variants, with only 3 that had FI >0.5 of NH5. None of them surpassed NH5 in FI, but one was about to repress more robustly than NH5. That switch had a point mutation at the 6th position, changing the pairing with the complement stand from an AU pair to a GU wobble. The position with the highest conservation was the 4th randomized nucleotide, which participates in the basepair directly before the 3’ strand invades into P3. Across all bins, U is favored at that position to make a WC pair with A in J2/3, the top bin has 100% conservation (n=3), the mid bin has 95% (n=19) and the bottom has 70%. The other positions are much more variable with nucleotide identity save those that form pairs with U, which allow for A or G in all bins (positions 2, 3, and 6). While, both A and G were represented at those positions, replacing A with G diminished switches ability to repress, with one switch that had a G at position 2 preforming the same in the presence of ligand but allowing about 4 times more leaky expression in the absence of ligand, greatly impacting the FI value of the switch.

Chart, scatter chart

Description automatically generated

While the P3 library was screened thoroughly, very few (n=18) unique sequences were identified. No sequence preformed comparably to *NH5*, which supports that initial hypothesis that canonical base pairing lends to the best switch in regions of competition between two stems. The one sequence that was a top performer had the sequence GAGGA, which only differs from the *NH5* sequence at the first position, resulting in a GG mismatch rather than an GC pair with the complementary strand. All of the sequences in the middling bin were also single point mutations, all four of which change an AU pair to a GU wobble pair. The consensus sequence from all 18 variants is NDBDN, though A is disfavored at position 1 (n=1) and C is disfavored at position 5 (n=1). For those positions that pair with uracil (positions 2 and 5), G and A seem to be equally tolerated. While those sequences were found, one variant with the sequence CGGGG, which changes both AU pairs to GU pairs, had a marked decreased ability to switch and repress compared to *NH5*.

**P1/P(T) balance**

One library was designed to better understand the requirements of regions at equilibrium in P1 and P(T) as strand invasion occurs or is disallowed in the case of expression proceeding. This library had 6 randomized nucleotides spanning two base pairs in P1 and two base pairs in P(T). During screening, ~14,000 colonies were observed, and 102 functional sequences were extracted, resulting in an ~96% change that any one sequence was observed (table S#).

**Chart, scatter chart

Description automatically generated**

**L4 and P4: Site of nucleation**

The toehold libraries reside in L4 and P4. These regions have been highlighted due to their role in stabilizing the expression platform enough to allow for invasion into the aptamer domain and creation of the terminator helix (). To assess the sequence requirements of the two regions, two libraries were designed, each with 6 nucleotides randomized. Each library was screened with a different strategy based on performance. The L4 library produced numerous highly functional switches which rivaled NH5, so screening was halted due to the lack of insight and large burden of screening that region. Only ~200 sequences were observed of this library, which yielded 77 functional sequences, emphasizing the abundance of robust switchers within that library (table S#). Though only 4.8% of observing any potential sequence were screened in the L4 library, the data produced indicated that continued screening via this approach was unlikely to be informative. The requirements of the P4 library, which includes the three base pairs in P4 distal to L4, were much more stringent based on colonies screened, ~14,000, and functional sequences extracted, 127 (table S#). The P4 library was screened until ~97% of potential sequences had been observed (table S#).

The L4 library was of interest because a the hexauridine tract in the NH5 parental mutant is hypothesized to be a “programmed pause” site for the RNA polymerase in *E. coli* and showed marked improvements for both expression and repression compared to mutants without this feature (33259551). But, when tested against a variant with 5’UUCCUU rather than 5’UUUUUU, activity was very similar, so the sequence requirements of this region were unclear following mutagenesis (33259551).

The screen of L4 yielded no sequence preference for any position, indicating that the identity of the nucleotides in this loop does not impact the ability of the riboswitch to function. 11 sequences had FI equal to or greater than NH5, reaching 1.4-fold of NH5, and 10 of those (as well as 7 sequences with lower FI than NH5) were able to repress more robustly than NH5. One sequence of note had a repression value ~0.6 compared NH5 and an FI of 1.2 compared to NH5 was 5’UAAUUG. This sequence has the potential to lengthen P4 with a GU pair, which could assist in the nucleation of the terminator without being overtly deleterious to the switch’s ability to express in the presence of ligand. Of those 17 sequences that repress the same or better than NH5, 9 of them could potentially form a WC or GU wobble pair between the first and last nucleotide. In the previous study a GAAA tetraloop was inserted as L4 with an accompanied lengthened P4, but it caused a slight reduction in activity while retaining robust repression.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

The randomized region in P4 relied heavily on ability to base pair between the two strands. Those that preformed best had a change in the third base pair, which may have been to destabilize the terminal base pair in P1. Further studies should be done to understand the impact of removing that base pair on the performance of the sequences that did well in this library. All those in the top bin had productive base pairing in the second base pair (n-1 of invasion for the extended P1) and the majority (83%) had productive pairing in the other two pairs. This indicates pairing is important in this region, but the strength/identity of each pair may change depending on the sequence of the competing P1 strand.

Of note, one sequence which switched each of the 3 base pairs to AU rather than the parental GC had large decrease in ability to repress. While it expressed FP slightly more than *NH5*, *NH5* repressed nearly 9 times as much expression in the absence of 2AP, resulting in a marked change in fold induction. Many colonies in the bottom bin also formed 3 productive base pairs, so the formation of any WC or wobble pair is not sufficient for robust switching. For example, one colony that also forms 3 AU pairs and differs only in the directionality of basepair 2 and 3. That sequence lacks the ability to express or repress near the parental or the other sequence with 3 AU pairs.

One sequence, with a fold induction less than a tenth of *NH5*, still maintains 3 GC base pairs. The only difference is that the first base pair has been changed from GC to CG. The deleterious impact of this change may be related to stacking energies or interactions with the pre-aptamer sequence (???), but overall suggests that directly analyzing the strength of each base pair will not directly translate to the in vivo activity of the system.

When looking at the three bins, it is most informative to look at secondary structure rather than primary sequence. When comparing percentage of productive base pairs across bins the top bin has a higher percentage of productive base pairs. The middle and bottom bin have similar proportions of pairing at the position distal to L4, but increased pairing in the other two base pairs.

**Anchor and Anchor exchange: Boundary**

Two libraries were designed to understand the boundary between the toehold and the initiation of invasion. One region probed was a single stranded library spanning P4, excluding the nucleotide proximal to L4 and extending two base pairs into the strand invading P1 for a total library size of 5 nucleotides. The anchor library was screened until there was a ~99% chance of observing any one sequence, which involved observing ~5,100 colonies and yielded 17 functional and unique sequences. The other library accompanying the anchor library in understanding the boundary of P4 includes nucleotides in all 3 of the strands involved in the branch migration. The anchor exchange library includes the additional P1 basepair at the base of P1, which was added during the design of NH5 as well as two base pairs in P4, the pair distal from L4 being in direct competition with the randomized base pair in P1.

Chart, scatter chart

Description automatically generated

This library located on the 3’ end of the P4 stem near the site of nucleation yielded 5 sequences deemed top switches, and one middling switcher, with all others falling in the bottom bin. Those sequences in the top two bins were all single point mutations, usually converting an GC pair to a GU pair, except for one sequence which introduced a CC mismatch at position 1 and an AG mismatch at position 4. One sequence, which introduced a GU wobble pair at position 3, had a fold induction higher than *NH5* (*NH5* corrected FI=1.2) but was unable to repress as well as *NH5*, suggesting that the stability of a GC pair at that position likely promotes higher fidelity in the formation of the terminator helix. The lack of deviation from the parental sequence, especially amongst sequences that preformed similar to *NH5*, supports the hypothesis that WC base pairing in both P1 and P4 creates a highly functional riboswitch.

All sequences pulled out from the screen maintained a guanine at position 2 and most sequences had a G at position 1 and C at position 3. The conservation at these positions may indicate the stability requirements of nucleation of P4. Position 2 lies at the n-1 position to invasion based on the *NH5* pre-aptamer sequence, and position 3 is the first basepair that must participate in invasion.

A picture containing plot, line, text, screenshot

Description automatically generated

From the sequence alignments and associated activities, it seems that there are many pairings that yield highly functional switches, both in their ability to repress robustly as well as their ability to have differential expression based on ligand state. What is clear is that canonical base pairing at any of the surveyed base pairs in neither necessary nor sufficient to create a functional switch. Only 4 of the functional sequences form 0 productive base pairs if the positions interact as they do in *NH5*. All of those sequences are in the bottom bin, indicating suboptimal levels of repression and fold induction. Based on the distribution of base pairing amongst the three bins, the formation of P1 seems to be unimportant if even disfavored. While the mid and bottom bins follow similar trends, there is a noticeable enrichment in productive pairing in P4 for the top bin, hinting at the importance of stable P4 formation for robust activity.