Boston College

Doctoral Thesis

On the use of Coarse Grained Thermodynamic Landscapes to Efficiently Estimate Kinetic Pathways for RNA Molecules

---

Author:
Evan Senter

Supervisor:
Dr. Peter Clote

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Clote Lab
Department of Biology

Thursday 18th June, 2015

Chapter 1

Ribofinder

Riboswitches are regulatory mRNA elements that modulate gene expression via structural changes induced by the direct sensing of a small-molecule metabolite. Most often found in bacteria, riboswitches regulate diverse pathways including the metabolism and transport of purines, methionine, and thiamin amongst others. The structure of a riboswitch includes an aptamer domain—involved in the direct sensing of the small-molecule—and an expression platform whose structure changes upon the aptamer binding the metabolite. Because of the discriminatory nature of metabolite sensing, groups have had great success in finding representative examples of aptamers across a diverse collection of bacterial species; RFam 12.0 currently contains 26 different families of aptamers involved in different metabolic pathways. Whereas there exists strong sequence and structural similarity within the aptamer of a riboswitch family, the expression platform is

highly variable, and thus challenging to capture using traditional SCFG-based approaches. For this reason databases such as RFam only contain the aptamer portion of the riboswitch, and there exists no database providing sequences including expression platforms, necessary for capturing the 'on' and 'off' conformations of this regulatory element. We have developed a new pipeline—called `Ribofinder`—which can detect putative riboswitches including their expression platforms and likely conformational structures across a wide collection of genomic sequences.

1.1 Pipeline

At the time of our retrieval (2014-11-25 at 09:14), the RefSeq database hosted by NCBI comprised 5,121 complete bacterial genomes with corresponding genomic annotations. In order to both detect putative full riboswitches across this collection of data as well as filter the candidates down to a number tractable for experimental validation, we developed a novel pipeline which takes a three-tiered approach to candidate selection. Our approach is to *a*) identify a pool of candidate riboswitches across genomic data; *b*) perform a coarse-grained filtering of the candidate pool based on structural characteristics; and finally *c*) fine-grained curation of the candidates based on a collection of measures and pairwise similarity.

In the following discussion, we describe the application of `Ribofinder` to identify unannotated G-box purine riboswitches; guanine-sensing cis-regulatory elements which modulate the expression of genes involved in purine biosynthesis.

1.1.1 Candidate Selection

The RefSeq data we used for analysis contains 5,121 annotated bacterial genomes across 2,732 different organisms, totaling over $9.5*10^9$ bases. We used the program `Infernal` to determine the coordinates of putative aptamer structures within the RefSeq genomes, and `TransTermHP` to locate candidate rho-independent transcription terminators.

1.1.1.1 Detecting Aptamers with `Infernal`

`Infernal` uses a stochastic context-free grammar (SCFG) with a user-provided multiple sequence alignment (MSA) to efficiently scan genomic data for RNA homologs, taking into consideration both sequence and structural conservation. Using the purine aptamer MSA from RFam 12.0 (RF00167), `Infernal` (v1.1.1, default options) detects 1,537 significant hits having E-value <= 0.01. Because `Infernal` leverages the concept of a 'local end'—a large insertion or deletion in the alignment at reduced cost—it is possible for the software to return a significant hit whose aligned structure does not have the canonical three-way junction observed in all purine riboswitches. `Ribofinder` prunes these truncated `Infernal` hits by converting the alignment structure into a parse tree, and only permitting trees of sufficient complexity to contain a multiloop (described further in 1.1.2). The pyrimidine residue abutted next to the P1 stem in the J3-1 junction differentiates between guanine and adenine-sensing riboswitches by binding the complimentary

purine ligand; for our interest in G-box riboswitches exclusively we require the presence of a cytosine at this residue. In total, using `Infernal` with these additional filters yields 1,280 G-box aptamers across 555 unique organisms (note: here and elsewhere I define a 'unique organism' as having a unique taxonomy ID).

1.1.1.2 Detecting Expression Platforms with `TransTermHP`

`TransTermHP` detects rho-independent terminators in bacterial genomes in a context-sensitive fashion by leveraging the protein annotations available in PTT data. These terminator sequences canonically have a stable hairpin loop structure immediately preceding a run of 5+ uracil residues, the combination of which causes the ribosomal machinery to stall and dissociate from the transcript. `TransTermHP` performs a genomic scan to determine candidate loci with this motif, and returns scored hits. The scoring system considers both structural homology and the genomic contextual information available in the PTT file. Across our collection of bacterial genomes acquired from NCBI RefSeq data, `TransTermHP` identified 2,752,469 rho-independent terminators using the default filters.

Due to the spatially-mediated structural regulation of purine riboswitches, whereby ligand interaction with the aptamer domain induces local structural rearrangement in the expression platform, we paired aptamers with corresponding terminators by minimizing the genomic distance, with an upper bound of 200 nucleotides between the end of the aptamer domain and start of the terminator. This approach yields

577 candidate riboswitches, 81 of which have multiple rho-independent terminators within range of a putative aptamer produced by `Infernal`. For these, we simply pair the closest `TransTermHP` hit with the aptamer domain.

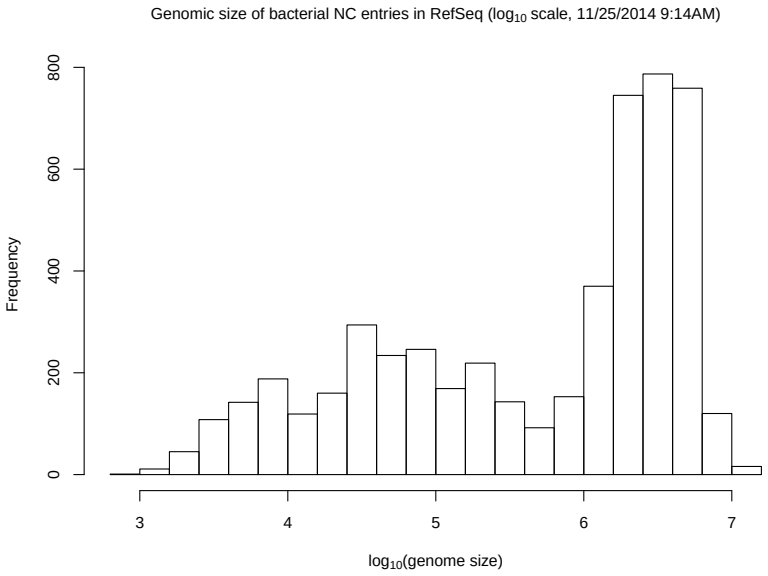

FIGURE 1.1: Histogram displaying the distribution of genome sizes across the RefSeq data analyzed, comprising 5,172 bacterial genomes. Genome size is shown using a $\log_{10}$ scale.

1.1.2 Notation for Representing Abstract RNA Shapes

Given an RNA sequence $\mathbf{s} = a_1, a_2, \cdots, a_n$, where positions a_i are drawn from the collection of single-letter nucleotide codes, i.e. $\{\mathrm{A}, \mathrm{U}, \mathrm{G}, \mathrm{C}\}$, it is possible to describe a corresponding secondary structure $\mathcal{S}$ compatible with $\mathbf{s}$ using the dot-bracket notation. In this notation, each nucleotide a_i has a corresponding state s_i, where s_i is denoted as a '.' if unpaired and a '(' [resp. ')'] if the left [resp. right] base in a basepair. Given any two basepairs (i, j) and (k, l) in $\mathcal{S}$, then

$i < k < j \iff i < l < j$; pseudoknots are not permitted in the structure. A secondary structure taking this form is said to have balanced parentheses, and can additionally be represented using a context-free grammar such as:

$$S \to S\,. \mid .\,S \mid (S) \mid SS \mid \epsilon \tag{1.1}$$

The grammar from (1.1) can be used to generate a parse tree $\mathcal{T}$ for $\mathcal{S}$. The benefit of working with $\mathcal{T}$ over $\mathcal{S}$ is that the parse tree offers an abstract representation of secondary structure shape independent of sequence length, permitting us to classify and eventually constrain a large collection of sequences having variable length which are all expected to have the same abstract tree shape. This is analogous to what the Giegerich lab refers to as their 'type 5' structural abstraction using the `RNAshapes` tool. Every node in $\mathcal{T}$ represents a helix in $\mathcal{S}$, and internally tracks the indices of both its beginning (i, j) and closing (k, l) basepair. We use a level-order naming convention to refer to helices within the parse tree, whereby a position $\mathtt{p}_1$ references the first child of the root node, $\mathtt{p}_{1,2}$ references the second child of $\mathtt{p}_1$, and generally $\mathtt{p}_{i_1, i_2, \cdots, i_n}$ refers to the ${i_n}^{\text{th}}$ child of $\mathtt{p}_{i_1, i_2, \cdots, i_{n-1}}$. To reference specific nucleotides in the context of their location relative to a helix, we use the opening and closing basepairs (i, j) and (k, l) as landmarks. Thus, $\mathtt{p}_1(l)$ is the index in $\mathcal{S}$ of the right-hand side closing basepair of $\mathtt{p}_1$. We use the notation $\mathtt{t}_i$ to refer to the subtree of $\mathcal{T}$ whose root is $\mathtt{p}_i$.

Finally, we introduce the concept of a tree signature. The tree signature for a tree

$\mathcal{T}$ is a list of the node depths when traversed in a depth-first pre-order fashion. To provide a concrete example, consider the following experimentally validated xpt G-box riboswitch from Bacillus subtilis subsp. subtilis str. 168 (NC_000964.3 2320197-2320054) with corresponding gene-off structure:

```
ACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGCUUGCUCUUUAUUUGAGCGGGCA
.(((((((((.....(((((.......)))))..........((((((.......))))))..)))))))).((((.(((((((........)))))))))).....((((((((((((((.......))))))))))
```

The `RNAshapes` 'type 5' representation for this structure is `[[][]][][]` (note the coalesced left bulge in the hairpin immediately downstream the closing multiloop stem, at helix $\mathtt{p}_2$) and the tree signature for this parse tree of the structure is `[0, 1, 2, 2, 1, 1]`.

1.1.3 Coarse-Grained Filtering

We leverage the notion of abstract structural filtering initially to ensure that all `Infernal` aptamer hits have a tree signature of `[0, 1, 2, 2]`, which represents a three-way junction, and that the binding site for the guanine ligand $\mathtt{p}_1(l-1) = \mathrm{C}$. These filters, in combination with the proximal terminator hairpins produced by `TransTermHP` yield the aforementioned 577 candidate guanine riboswitches for which we then try to produce reasonable gene-on and off structures.

1.1.3.1 Constrained Folding to Predict Switch Structures

To restrict our search to unannotated G-box riboswitches, and further ensure that we are not re-detecting sequences based off the RFam covariance model provided to `Infernal`, we constrain our search to those RefSeq organisms not represented in the RFam seed alignment. 503 of the 577 candidates, or 87.18% represent putative unannotated riboswitches not represented by RF00167.

The gene-off structure $\mathcal{S}_{\text{off}}$ for a G-box riboswitch is the easier of the two to find computationally, since the terminator loop is exceptionally thermodynamically stable. In the gene-on conformation $\mathcal{S}_{\text{on}}$, the P1 stem of the multiloop partially dissociates and an anti-terminator loop forms between the region immediately 3' of the P1 stem and what was the left-hand side of the terminator loop. This truncated P1 stem, which closes the three-way junction in the aptamer, is exceptionally unstable based on present energy models available for structural folding, and requires special treatment to reconstitute in our final structures.

The software `RNAfold` (v2.1.8) allows for the folding of RNA molecules with 'loose' constraints. In this model of constrained folding, the resulting structure produced by the software guarantees not to explicitly invalidate any user-provided constraints, but does not guarantee all constraints will be satisfied in the resulting structure. For each of the candidate guanine riboswitches, having $\mathcal{T}_{\texttt{Infernal}}$ and $\mathcal{T}_{\texttt{TransTermHP}}$, we build the following constraint masks:

<table>
<tr><th>G-box gene-off constraint mask</th><th>G-box gene-on constraint mask</th></tr>
<tr><td colspan="2">Prohibit basepairing upstream of $\mathtt{p}_1(i)$ and downstream of $\mathtt{p}_2(l)$.
Force basepairs and unpaired regions in $\mathtt{t}_1$, with the exception of $\mathtt{p}_1$.
Prohibit formation of $\mathtt{p}_1$ stem, which closes the three-way junction.</td></tr>
<tr><td>Force basepairs and unpaired regions in $\mathtt{t}_2$.</td><td>Require m nucleotides starting from $\mathtt{p}_1(l+3)$ to pair to the right, where $m = len(\mathtt{p}_2)$, and require the left-hand side of the $\mathtt{p}_2$ helix to pair to the left.
Disallow pairing downstream of $\mathtt{p}_2(j)$.</td></tr>
</table>

These constraint masks are run using the command-line flags `-d 0 -P rna_turner1999.par` to disable dangles and use the Turner 1999 energies respectively. Experimental evidence using inline probing suggests that the 'on' conformation of the G-box riboswitch has a reduced P1 stem length of 3 base pairs; in practice we were unable to force `RNAfold` to respect this constraint regardless of command-line options specified. For this reason we reconstitute the P1 stem in both structures after constrained folding, having length equivalent to it the `Infernal` P1 stem (resp. 3 basepairs) in the gene-off (resp. gene-on) structure.

This difficulty with `RNAfold` can be shown by using the constraint-produced structures as exhaustive constraints themselves. All unpaired nucleotides in $\mathcal{S}_{\text{off}}$ and

$\mathcal{S}_{\text{on}}$ are notated by a '`x`' and all base pairs by '`()`' for the 5' and 3' side of the pair respectively to form new constraints mask $\mathcal{C}_{\text{off}}$ and $\mathcal{C}_{\text{on}}$, having all bases' state explicitly specified. By refolding all 577 candidate sequences with $\mathcal{C}_{\text{off}}$ and $\mathcal{C}_{\text{on}}$ using the same options as before, only 463 (or 80.24%) of the resulting structures from $\mathcal{C}_{\text{off}}$ have the tree signature prefix `[0, 1, 2, 2, 1]`, and just 21 (or 3.64%) of the $\mathcal{C}_{\text{on}}$ structures correctly re-fold their multiloop.

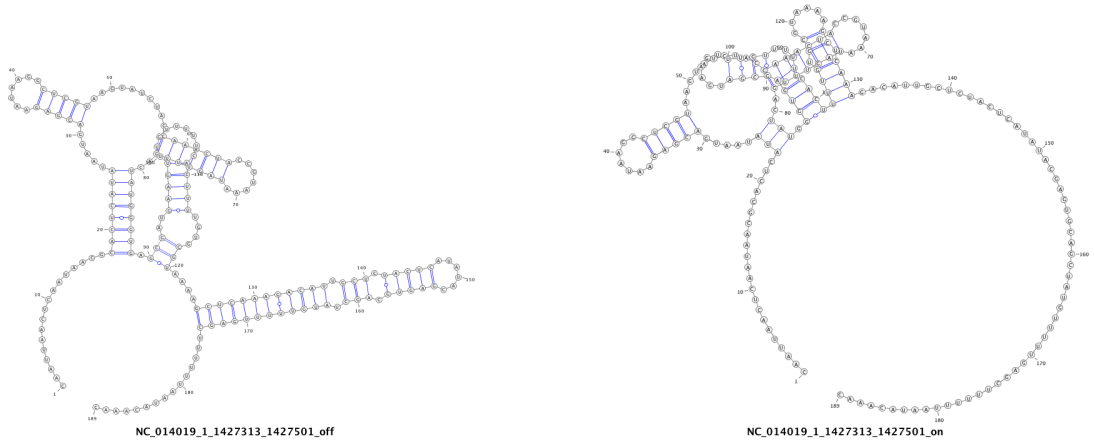

FIGURE 1.2: *Left:* the computationally predicted gene-off conformation of sequence NC_014019.1 1427313-1427501, using `RNAfold` from the ViennaRNA 2.1.8 suite, with dangles disabled and the Turner 1999 energies. This sequence is located upstream of the xpt gene in Bacillus megaterium QM B1551. *Right:* the gene-on conformation.

1.1.4 Fine-Grained Filtering

Until now, we have described our approach for generating the 503 guanine riboswitch candidates in RefSeq, alongside their gene-on and off structures. Unfortunately the experimental validation of all 503 candidates is not tractable, so it was necessary to reduce this collection again to a more manageable size, while only keeping the most promising candidates.

1.1.4.1 Using `FoldAlign` to Rank Candidates