Sequence design for RNA-RNA interactions

Maria Waldl, Hua-Ting Yao, Ivo Hofacker

Sequence design for RNA-RNA interactions

Maria Waldl[†][0000−0001−7098−5712] and
Hua-Ting Yao[†][0000−0002−1720−5737] and
Ivo L. Hofacker[0000−0001−7132−0800]

Abstract The design of RNA sequences with desired structural properties presents a challenging computational problem with promising applications in biotechnology and biomedicine. Most regulatory RNAs function by forming RNA-RNA interactions, e.g., in order to regulate mRNA expression. It is therefore natural to consider problems where a sequence is designed to form a desired RNA-RNA interaction and switch between structures upon binding. This contribution demonstrates the use of the `Infrared` framework to design interacting sequences. Specifically, we consider the regulation of the rpoS mRNA by the sRNA DsrA and design artificial 5'UTRs that place a downstream protein coding gene under control of DsrA. The design process is explained step-by-step in a Jupyter notebook, accompanied by Python code. The text discusses setting up design constraints for sampling sequences in `Infrared`, computing quality measures, constructing a suitable cost function, as well as the optimization procedure. We show that not only thermodynamic, but also kinetic folding features can be relevant. Kinetics of interaction formation can be estimated efficiently using the `RRIkinDP` tool, and the chapter explains how to include

Maria Waldl
Department of Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria
Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, 1090 Vienna, Austria
Institute of Computer Science and Interdisciplinary Center for Bioinformatics, Leipzig University, D-04107 Leipzig, Germany
e-mail: `maria@tbi.univie.ac.at`

Hua-Ting Yao
Department of Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria
e-mail: `htyao@tbi.univie.ac.at`

Ivo L. Hofacker
Department of Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria
Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, 1090 Vienna, Austria
e-mail: `ivo@tbi.univie.ac.at`

[†]Shared first authors

kinetic folding features from `RRIkinDP` directly in the cost function. The protocol implemented in our Jupyter notebook can easily be extended to consider additional requirements or adapted to novel design scenarios.

Key words: RNA sequence design, RNA-RNA interactions, RNA structure, RNA folding kinetics

1 Introduction

In recent years, there has been notable advancement in RNA design tools, addressing various design challenges such as negative design problems, multi-structure design, and complex constraints. Tools like `RNAblueprint` [1], `Redprint` [2], and `Infrared` [3] exemplify this progress, offering sophisticated algorithms and user-friendly interfaces for designing RNA molecules with precise control over their structural and functional properties.

RNA-RNA interactions play critical roles in regulating gene expression, RNA processing, and other fundamental biological processes. These interactions involve base pairing between complementary regions of two RNA molecules, resulting in the formation of complex secondary and tertiary structures. Understanding the thermodynamic and kinetic features governing RNA-RNA interactions is crucial for unraveling their functional implications and engineering RNA molecules with tailored properties. Since RNA-RNA interactions can be treated as an extension of RNA secondary structures, they are arguably more amenable to rational design than interactions with proteins or small molecules and may thus provide an ideal avenue for engineering artificial regulators of gene expression.

In this chapter we also present the integration of `RRIkinDP` [4], a computational tool that allows to evaluate not only thermodynamic, but also kinetic features of RNA-RNA interactions, into RNA design schemes such as the `Infrared` framework. `RRIkinDP` offers efficient evaluation of interaction features, making it fast enough to directly integrate into the design process. This allows researchers to engineer RNA sequences with functional characteristics beyond equilibrium properties.

`Infrared` provides a versatile framework for the rapid development of task-specific RNA design tools. Leveraging the capabilities of `Infrared`, our approach builds upon existing design methodologies to optimize various aspects of RNA sequences, including accessibility, seed stability, barrier optimization, alternative structures, and dynamic switching through interaction formation. Integration into the constraint framework within `Infrared` further enhances computational efficiency, enabling the design of complex RNA molecules with reduced computational costs.

In contrast to earlier tools for multi-structure design [5, 6] that allowed only to specify a list of target structures, but provided little choice of cost function, optimization strategy and additional design constraints, `Infrared` provides a framework in which arbitrary design constraints can be formulated, making it equally suitable for interaction design and classical multi-structure design.

The design approach outlined in this chapter is presented in the form of a Jupyter notebook, available at `https://github.com/ViennaRNA/RRIdesign`. This notebook provides a comprehensive guide to implementing our design methodology and offers insights into the rational engineering of RNA sequences for specific applications.

By elucidating the integration of `RRIkinDP` into RNA design frameworks and showcasing the capabilities of `Infrared`, this chapter aims to provide researchers with valuable insights and methodologies for engineering RNA molecules with enhanced functionality and versatility.

1.1 Theory

There are several common approaches to predicting RNA-RNA interactions as extensions of RNA secondary structure prediction. The `ViennaRNA` package, which we're using in this contribution, provides the programs `RNAcofold` and `RNAup`. `RNAcofold` performs the prediction by concatenating the two RNA strands and then runs a standard folding algorithm except that loops containing the break-point between the two strands are treated specially. This results in a pseudo-knot free joint structure and therefore excludes structures containing, e.g., kissing hairpins. `RNAup` [7] views interaction formation as a two step process consisting of freeing up the interaction sites on both molecules and then forming the interaction between single stranded regions. The binding free energy can then be expressed as

$$\Delta G_{bind} = \Delta G_{open} + \Delta G_{interact}, \tag{1}$$

where $\Delta G_{open} \geq 0$ is the free energy required for making the binding site accessible and $\Delta G_{interact} \leq 0$ the free energy gained from forming interaction helices. The algorithm proceeds by precomputing ΔG_{open} for every possible region $[i \dots j]$ before then finding the interaction minimizing ΔG_{bind}. `IntaRNA` in addition requires a perfectly matching seed interaction, which can speed up the interaction search. The main restriction of `RNAup` and `IntaRNA` is that there can only be one interaction region. Note that the unrestricted RNA-RNA interaction problem is NP-hard similarly to the problem of predicting secondary structures with pseudo-knots.

A putative interaction may be thermodynamically stable, but hard to reach by the dynamic folding process due to high energy barriers. `RRIkinDP` is a tool to analyze the energy landscape for structure formation and can extract useful features, such as the highest energy barrier. By restricting itself to direct trajectories, i.e., shortest paths from initial contact to full interaction, the problem can be solved efficiently using dynamic programming. In addition to the evaluation of folding paths, the energy landscape provides information on possible short initial seed interactions, that can stabilize the first contact between the two RNAs and thereby initiate the interaction formation. Further information on the feasibility of an interaction is provided by the accessibility of the interaction site, or the seed interaction site,

where a high accessibility indicates that it is easy (not much energy is needed) to make the interaction site unpaired in the intramolecular structure and therefor available to form interaction base pairs. Furthermore, the structure of the landscape can reveal shorter suboptimal interactions that are kinetically more favorable than the full interaction predicted by thermodynamic criteria.

The perhaps most important ingredient for any successful design is the cost function that will be optimized. Thermodynamic features in the cost function should typically be based on partition function calculations. A simple, commonly used, cost function for designing bistable sequences folding into two target structures S_1, S_2 might read

$$C(x) = E(x, S_1) - G(x) + E(x, S_2) - G(x), \tag{2}$$

where $G(x) = -RT \ln Z$ is the ensemble free energy derived from the partition function Z over all possible structures of sequence x and $E(x, S)$ is the free energy of structure S on sequence x. Minimizing the cost function above is equivalent to maximizing the product of the Boltzmann probabilities $p(S_1) \cdot p(S_2)$ of the two structures.

The above example assumes that we want to precisely specify both target structures down to each individual base pair. However, in a typical application we may want to only specify structures in part, thus giving the design process more freedom. Partial specification can easily be incorporated by using constrained folding. Given a partial structure P that specifies some base pairs as well as some unpaired positions (but leaves other positions free), we can easily compute a constrained partition function Z^C, which sums only over those structures that fulfill the constraints, and the corresponding free energy $E^C(x, P) = -RT \ln Z^C$. Simply substituting $E(x, S)$ with the constrained version yields a suitable cost function for partially specified structures. The opening energies used by `RNAup` for interaction prediction (see above) are another example of features that can be easily computed using a constrained partition function.

In prokaryotes, translation initiation starts with binding of the ribosome via an RNA-RNA interaction between the Shine-Delgarno sequence of the mRNA and a CCUCC element near the end of the 16S rRNA. The opening energy of the ribosome binding site (RBS) is therefore an important determinant of translation efficiency and can be used to predict the effect of small RNA (sRNA) regulation [8]. In our application example, we will start from the well known interaction between the bacterial sRNA DsrA and its target mRNA rpoS. The rpoS mRNA is normally translated poorly, since the RBS, which we assume as a 30nt region starting with the Shine-Dalgarno motif, is sequestered in a stable secondary structure. Binding of DsrA further upstream leads to a refolding of the untranslated region (UTR) making the RBS accessible, thus upregulating translation. In this chapter, we will as a showcase design the 5' UTR of a messenger RNA coding sequence, e.g., GFP to put it under the control of the DsrA sRNA. We note that one could analogously design an artificial sRNA to target a given (fixed) mRNA.

Our design criteria will therefore be (i) sRNA and mRNA should bind strongly at reasonable concentrations (ii) the mRNA should exhibit poor RBS accessibility

without the sRNA (iii) in the bound state the RBS is highly accessible. Since the RBS reaches into the coding sequence, we will include the first codons in designed sequence, while requiring a fixed encoded amino acid sequence. In a second step, as a showcase for the design of more complex interaction features, we will also add (iv) a stable interaction seed as well as (v) a low folding barrier along the interaction formation path as design criteria.

In practice, a design problem will often have numerous additional constraints, e.g. to allow for handling of the sequences (such as primer binding sites, restriction sites, etc.), or to ensure that the designed sequence is compatible with the expression system. The `Infrared` framework allows to easily incorporate such constraints into the design process. Many sRNA-mRNA interactions, including the DsrA-rpoS, are facilitated by Hfq [9]. Hfq has distinct binding sites for sRNAs and mRNAs and while we keep the DsrA sequence (and thus its Hfq binding site) fixed, a corresponding binding motif might be designed into the mRNA. Finally, since not all requirements of sRNA regulation are fully understood, designing a fully functional sequence will likely require multiple rounds of computational design and experimental testing.

2 Materials

Our RNA-RNA interaction design approach relies on several key components: the `ViennaRNA` library for free energy evaluations, `RNAup` and `IntaRNA` for interaction prediction, `RRIkinDP` for computing kinetic and thermodynamic RNA-RNA interaction features, and the design framework `Infrared`. The design approach and examples presented in this book chapter are provided as Jupyter notebooks. This section outlines the steps to install these tools within a Conda environment.

2.1 Installing Conda Package Manager

We recommend installing the necessary dependencies using the Conda package manager. Instructions for installing Conda, in the form of Miniconda, on MacOS and Linux can be found here: `https://docs.anaconda.com/free/miniconda/miniconda-install`.

2.2 Installation of RRIkinDP, ViennaRNA, and Infrared

After installing and activating Conda, the required dependencies can be easily installed using the `conda install` command. We recommend setting up a new environment and installing the dependencies at the same time to avoid compatibility issues between different package versions, e.g., with the following commands:

```
$ conda create --name rridesign \
      conda-forge::'infrared>=1.2' \
      bioconda::'viennarna>=2.6.2' \
      bioconda::intarna \
      bioconda::'rrikindp>=0.0.2'
$ conda activate rridesign
```

For additional installation instructions, see the Notes section at the end of the chapter.

2.3 Design Approach and Example Scripts

The Jupyter notebook with the design approach and example files can be downloaded from `https://github.com/ViennaRNA/RRIdesign`, e.g., with `git clone`:

```
$ git clone git@github.com:mwaldl/RRIdesign.git
```

To run the Jupyter notebook, ensure a notebook interface is installed, e.g., JupyterLab via Conda:

```
$ conda install conda-forge::jupyterlab
```

Navigate to the downloaded git repository and start JupyterLab:

```
$ cd RRIdesign
$ jupyter lab
```

This will open JupyterLab in your web browser, where you can access the notebook `RRIdesign.ipynb` containing all the scripts described in the Methods section.

2.4 RNA Sequence Data

For illustration, we use the DsrA-rpoS interaction from Escherichia coli. The corresponding sequences can be obtained from the RefSeq Genomes data base entry of the Escherichia coli strain K-12 substr. MG1655 at NCBI [10]:

- DsrA: NC_000913.3(2025313-2025225)
- rpoS 5'UTR fragment (TSS -150nts +27nts): NC_000913.3(2867701-2867525)

As an example of a coding sequence that can be controlled by DsrA via a designed mRNA 5'UTR we use the GFP coding sequence from Aequorea victoria, GenBank: M62654.1.

3 Methods

In this section we will briefly describe functions included in the Jupyter notebook and accompanying python module.

3.1 The DsrA-rpoS Example

The first section of the notebook examines the interaction between the DsrA and rpoS sequences as a natural example of translational control via sRNA–mRNA interaction. We use a piece of the rpoS mRNA comprising 150nt upstream of the start codon as well as the first 9 codons (27nt) of the coding sequence after the start codon AUG.

Note that results will vary depending on the extent of the mRNA region. In the first step we predict the minimum fee energy (MFE) secondary structures of dsrA and rpoS. For convenience, the `rri` module contains the `prediction_unbound` function which returns the MFE structures of both sequences. A second helper function `draw_RNAplot` calls the `RNAplot` program of the `ViennaRNA` package, highlighting the Shine-Delgarno motif, the AUG, as well as the 30nt RBS. The resulting plot should look as shown in Fig. 1, where the region downstream of the RBS is sequestered by a stem structure. It has been shown that this stem structure inhibits the translation of rpoS [11].

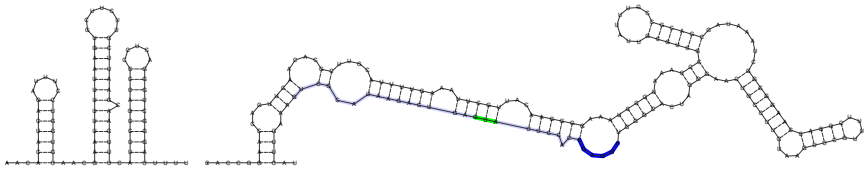


Fig. 1: Predicted MFE structure for DsrA (left) and rpoS (right).The Shine-Delgarno sequence is highlighted in blue, the AUG in green, and the RBS in gray.

The `rri.prediction_bound` helper function computes the rpoS structure in the bound state, by first predicting the interaction using `RNAup`, then performing a structure prediction for rpoS under the constraint, that the region of interaction with DsrA cannot form intra-molecular structure. Therefore, the intramolecular stem that blocks the RBS has to open upon DsrA binding [12]. The resulting interactions structure of both molecules is shown in Fig. 2. Note that for simplicity we do not include intramolecular DsrA structure. The interaction between DsrA and the rpoS 5'UTR is predicted to range from position 10 to position 40 in the DsrA and from position 25 to position 54 in the rpoS fragment with a free energy of about -12kcal/mol.

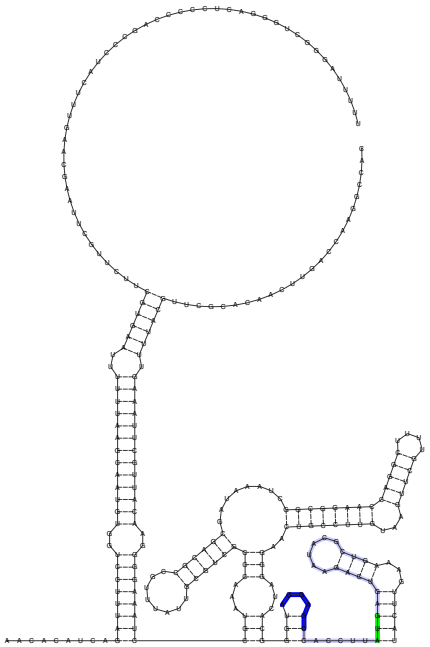


Fig. 2: Predicted MFE structure for rpoS after binding to DsrA.The Shine-Delgarno sequence is highlighted in blue, the AUG in green, and the RBS in gray.

Binding of DsrA has not resulted in a completely single-stranded RBS. This is expected since 30nt long regions are almost never completely free of base pairs. However, we can quantify the RBS accessibility by computing the opening energies ΔG_{open} before and after binding. ΔG_{open} is the difference between two ensemble free energies

$$\Delta G_{open} = \Delta G^{RBS} - \Delta G \geq 0 \tag{3}$$

where ΔG denotes the ensemble free energy of the mRNA over all possible structures, while ΔG^{RBS} is the corresponding free energy restricted to structures with unpaired RBS. To compute the binding energy after binding, we use the additional constraint that positions in the binding region cannot form (intra-moleular) base pairs. Using the helper function `rri.dGopen`, we obtain $\Delta G_{open}^{free} = 11.23$ kcal/mol before and $\Delta G_{open}^{bound} = 6.34$ kcal/mol after binding.

In addition to the binding energy and the RBS accessibility, we computed kinetic features of the native DsrA-rpoS interaction. Fig. 3 represents the energy landscape of all intermediate interaction. From this we can, e.g., search for the optimal starting seed interaction as well as a direct folding path that involves the lowest possible energy barrier. Assuming a seed length of 5 nucleotides for the first stable interaction, the landscape suggests that the optimal start point for the DsrA-rpoS interaction is

region from position 12 to 16 with a free energy of 0.68 kcal/mol. From this seed, the full interaction can be reached without crossing any barrier states with a less favorable energy than the seed interaction.

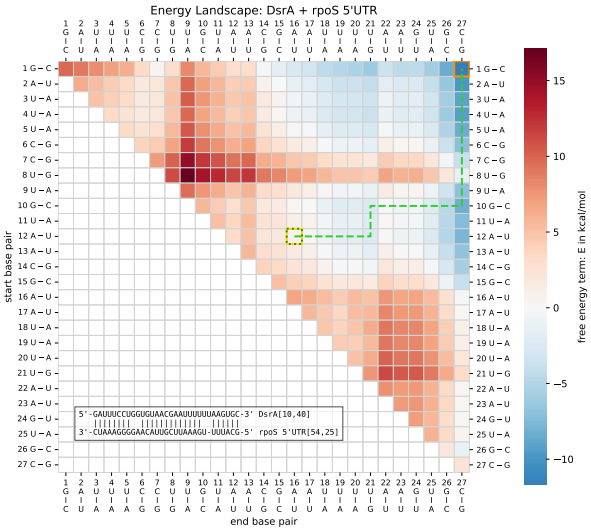


Fig. 3: Energy landscape of the predicted minimum free energy interaction between DsrA and the 5' UTR of rpoS in E. coli. The energy landscape represents all possible sub-interactions $\mathcal{I}_{k,l}$ from the kth to the lth base pair with in the full DsrA-rpoS interaction, which is shown in the box at the bottom left corner. The cell in the kth row and lth column represents the binding energy of sub-interaction $\mathcal{I}_{k,l}$ in kcal/mol indicated by the red to blue color scale, where red (blue) signifies a high (low) energy. Assuming a seed interaction of length m, the folding path starts with a cell located on the mth diagonal, the diagonal starting with $(1, m)$. The possible moves are either to the right or to the top. Moving one cell to the right (top) in this matrix corresponds to extending the interaction by one base pair on the right (left) side of the sub-interaction. An ideal path should encode a sequence of decreasing free energy from the start to the end. The final interaction is the sub-interaction corresponding to the end of the path. A minimum barrier folding path with seed length $m = 5$ is shown as a green dashed line, starting with $\mathcal{I}_{12,16}$ (yellow) and ending at the full interaction (orange). The barrier state (dark green) is same as the starting state.

3.2 Design Model

`Infrared` provides different solvers for user-defined problems described as the object `Model`. In the context of RNA design, a model typically contains a set of variables, each with 4 possible values $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$ and representing one position in the RNA; a set of constraints, each restricts nucleotide composition for a subset of variables; and a set of functions defining different (pseudo-) energies of sequences. From the model, valid RNA sequences satisfying the constraints are sampled with a Boltzmann weight defined by the energies. The constraints can overlap, allowing us, e.g., to specify multiple target structures. Without an energy function the sampling will be uniform over all valid sequences. `Infrared` also allows to target a specific value for each energy in the output samples.

In this section, we will explain step-by-step the `Model` object construction to design a similar sRNA-mRNA system. First, we need to import the framework and its RNA extension.

```
import infrared as ir
import infrared.rna as irrna
```

To simplify the task, we focus only on designing 150nt upstream of the start codon. The sRNA and coding sequences are fixed to the DsrA sequence and the first 9 codons of GFP. We also insert a Shine-Dalgarno motif at the same location as in the rpoS example. Let `seq` be the concatenation of the described sRNA and mRNA sequence constraints expressed as a IUPAC sequence, we create an `Infrared` design model as

```
model = ir.Model(len(seq), 4)
for i, x in enumerate(seq):
    model.add_constraints(ir.ValueIn(i, irrna.iupacvalues(x)))
```

The first line initiates a model of n variables with 4 values where n is the total length of sRNA and mRNA. The last two lines restrict the possible values for each variable given by the IUPAC sequence.

Next, we extract two structure constraints `mRNA_str` and `inter_str` from DsrA-rpoS system to guide the conformation before and after binding. The first constraint `mRNA_str` imposes the structure of RBS region as shown in Fig. 1. The second one `inter_str` describes the binding region and overlaps with `mRNA_str` to destabilize RBS conformation after binding (Fig. 2). The constraints are expressed in well-balanced dot-bracket notation and added to the model using `add_constraints`.

```
for name, dbn in [('M', mRNA_str), ('I', inter_str)]
    bps = irrna.parse(dbn)
    model.add_constraints(irrna.BPComp(i, j) for (i, j) in bps)
```

We further define the energy of folding to each structure with a simple base pair energy model using `add_functions`

```
model.add_functions([irrna.BPEnergy(i, j, (i-1, j+1) not in bps)
    for (i,j) in bps], f'energy{name}')
```

and define a more complex Turner energy with `add_feature`.

```
model.add_feature(f'Energy{name}', f'energy{name}',
    lambda sample, ss=ss:
        RNA.energy_of_struct(irrna.ass_to_seq(sample), ss))
```

Here the base pair energy is used as weight to sample an RNA sequence from the Boltzmann distribution.

```
sampler = ir.Sampler(model)
sampler.sample()
```

The additional features allow us to control efficiently the final structure energy of the sampled sequences. For instance, the code below returns a sequence forming the interaction `inter_str` with an energy of approximately -15 kcal/mol and the `mRNA_str` structure with -7 kcal/mol with a tolerance of 1.5 kcal/mol under Turner energy model. We invite readers to read `Infrared` RNA design tutorial for more details [13].

```
sampler = ir.Sampler(model)
sampler.set_target(-15, 1.5, 'EnergyI')
sampler.set_target(-7, 1.5, 'EnergyM')
sampler.targeted_sample()
```

From the sampling procedure, we obtain RNA sequences that satisfy the imposed sequence and structure constraints, thus already conforming to the criteria of *positive* design. However, this does not guarantee that candidate sequences actually form the desired structures and fulfill other design objectives (binding conformation, accessibility) while avoiding alternative structures. Avoiding unwanted features is known as *negative* design. To address it, we deploy local optimization to search the neighborhood of sampled sequences. This requires a move set that defines the local neighborhood of sequences that still fulfill the constraints, as well as a cost function to minimize, which will be discussed in the next section.

At each step, a connected component (a set of related positions) defined by structure constraints is randomly and uniformly selected and the corresponding subsequence is resampled. This could result in a point mutation or larger change, depending on the size of the connected component. In addition, we ensure the new sequence is distinct to the previous one. The new sequence is accepted if its cost is lower or, otherwise, with an acceptance probability controlled by temperature. We provide two optimization routines in the notebook: `mc_optimize` performs a Markov Chain Monte Carlo (MCMC) simulation at constant temperature, while `sa_optimize` performs a simulated annealing, i.e., an MCMC with a cooling schedule.

3.3 Cost Function

The last piece to complete the sRNA-mRNA system design is the cost function to minimize during optimization. We consider two cost functions to address 5 design criteria listed in Sec. 1:

(i) sRNA and mRNA should bind strongly;
(ii) the mRNA should exhibit poor RBS accessibility without the sRNA;
(iii) in the bound state the RBS is highly accessible;
(iv) there is a suitable, accessible, seed interaction;
(v) the energy barrier for interaction formation is low.

An equilibrium thermodynamics cost function will employ only the first three design criteria, while a cost function with kinetics will additionally cover the last two. In this section, we present in detail these two cost functions and show the designs obtained at the end of optimization. Note that the designed sequences will vary depending on the sampled starting sequences and the partial resampling while running the Jupyter notebook.

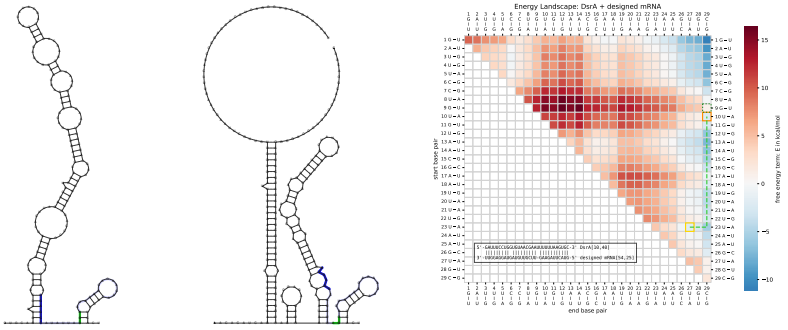
3.3.1 Thermodynamics Cost Function with RBS Accessibility Optimization

Our first cost function will be based on 3 free energy values: (i) the binding energy as computed by `RNAup` / `IntaRNA` ΔG_{bind}, (ii) the RBS opening energy for the unbound mRNA ΔG_{open}^{free} as well as (iii) for the sRNA bound mRNA ΔG_{open}^{bound}. We can find suitable target values θ for these free energies by comparing to known sRNA – mRNA interactions. Combining several optimization criteria can be tricky, since the optimization procedure may end up over-optimizing some (easy) criteria, while ignoring hard to optimize ones. To avoid this, we make use of a `LogSumExp` function, which computes a smooth maximum over a list of values, $\mathrm{LogSumExp}(x_1, \ldots, x_n) = \ln \sum \exp x_i$. Note that the feature values x_i may need to be scaled and shifted first. As cost function we can use, e.g.,

$$C(x) = \mathrm{LogSumExp}\Big((\Delta G_{bind} - \theta_{bind}), (-\Delta G_{open}^{free} + \theta_{open}^{free}), (\Delta G_{open}^{bound} - \theta_{open}^{bound})\Big) \quad (4)$$

The second term has opposite sign since we want to minimize ΔG_{bind} and ΔG_{open}^{bound}, while maximizing ΔG_{open}^{free}. See the notebook for chosen target values.

Fig 4 visualizes one mRNA design obtained after 1 000 steps using the thermodynamic cost function. The predicted values are −12.97 kcal/mol for binding energy, 12.29 kcal/mol for opening energy before binding, and 5.01 kcal/mol for opening energy after binding. The difference of opening energy can also be observed on MFE predictions. The Shine-Dalgarno is sequestered in an almost perfect helix before binding, after binding it is part of three consecutive loops separated by an isolated base pair.



(a) (b) (c)

Fig. 4: Predicted MFE structure for designed mRNA with thermodynamics cost function before (a) and after (b) binding to DsrA and its energy landscape (c). The predicted RBS opening free energy is 12.29 kcal/mol before and 5.01 kcal/mol after binding.

3.3.2 Adding Kinetic Features to the Cost Function

The mRNA UTRs designed with the thermodynamic cost function feature the targeted binding and opening energies. However, when evaluating kinetic features, we find that the designed sequences might often not fold into the full interaction within reasonable time, e.g., due to high barriers along the folding path. For instance, the minimum barrier energy of the design shown in Fig. 4(c) is about 2.7 kcal/mol higher than the binding energy of seed interaction. To include kinetic features into the design optimization, we convert the design criteria (iv), a stable seed interaction, and (v), a low barrier state energy, into two energies.

Given a fixed seed interaction length m, we define the first energy as the minimum barrier energy $\Delta G_{barrier}$ among all barriers that folding paths reach starting with a seed interaction of length m (design criteria (v)). The second energy is the binding energy ΔG_{seed} of seed interaction from which the path has a barrier energy at $\Delta G_{barrier}$ to measure the stability of seed interaction (design criteria (iv)). Both these energies are available after calling `rri.EnergyLandscape` from the `RRIkinDP` module to analyze the interaction formation landscape. Combining with Eq. 4, we obtain the kinetic cost function as

$$C(x) = \text{LogSumExp}\Big((\Delta G_{bind} - \theta_{bind}), (-\Delta G_{open}^{free} + \theta G_{open}^{free}), (\Delta G_{open}^{bound} - \theta_{open}^{bound}), (\Delta G_{seed} - \theta_{seed}), (\Delta G_{barrier} - \theta_{barrier})\Big) \tag{5}$$

An example sequence, obtained after 1 000 optimization steps, is shown in Fig. 5. As with the thermodynamic cost function, we obtain a sequence where the RBS

region is sequestered in the unbound state and weakly structured after binding with DsrA. The RBS opening energies are 12.58/4.88 kcal/mol before/after binding. The landscape, however, has notably lower energy barriers than the example from Fig. 4 as well as the native DsrA-ropS interaction (Fig. 3). The interaction is predicted to have a free energy around −17 kcal/mol, the optimal 5nt seed region consists of the 17th to 21th base pairs (out of 36 base pairs in the interaction).

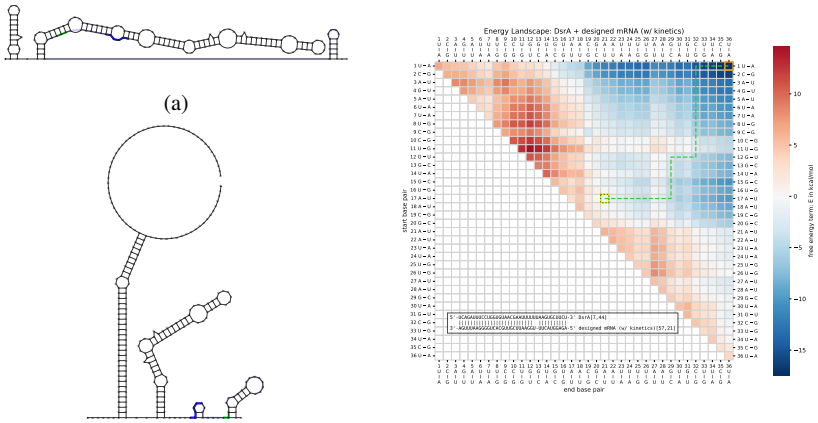


Fig. 5: Predicted MFE structure for designed mRNA with kinetic cost function before (a) and after (b) binding to DsrA and its energy landscape (c). The predicted RBS opening free energy is 12.58 kcal/mol before and 4.88 kcal/mol after binding. `RRIkinDP` predicts the interaction starts with the seed region from 17th to 21th base pairs.

4 Notes

By default, Conda is set up with the newest Python version, which may lead to compatibility issues with certain packages. To ensure compatibility, you can create a new Conda environment with a specific Python version and install the required packages:

```
$ conda create --name rridesign python=3.10
$ conda activate rridesign
$ conda install conda-forge::'infrared>=1.2' \
      bioconda::'viennarna>=2.6.2' \
      bioconda::intarna \
      bioconda::'rrikindp>=0.0.2'
```

Installing packages using Conda may sometimes be slow or fail because Conda is unable to resolve dependencies. Mamba, a drop in replacement for Conda, is in general faster and better at resolving dependencies. Installation instructions can be found here: `https://mamba.readthedocs.io/en/latest/installation/mamba-installation.html`. We highly recommend using Mamba, especially if you anticipate combining this tutorial with other tools.

References

[1] Hammer S, Tschiatschek B, Flamm C, Hofacker IL, Findeiß S (2017) RNAblueprint: flexible multiple target nucleic acid sequence design. Bioinformatics 33(18):2850–2858, DOI 10.1093/bioinformatics/btx263

[2] Hammer S, Wang W, Will S, Ponty Y (2019) Fixed-parameter tractable sampling for RNA design with multiple target structures. BMC Bioinformatics 20(1):209, DOI 10.1186/s12859-019-2784-7

[3] Yao HT, Marchand B, Berkemer SJ, Ponty Y, Will S (2024) Infrared: a declarative tree decomposition-powered framework for bioinformatics. Allgorithms for Molecular Biology DOI 10.21203/rs.3.rs-3366298/v1, preprint

[4] Waldl M, Beckmann IK, Will S, Hofacker IL (2023) Modeling kinetics of RNA-RNA interactions on direct paths. bioRxiv DOI 10.1101/2023.07.28.548983, URL `https://github.com/mwaldl/RRIkinDP`

[5] Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M (2001) Design of multistable RNA molecules. RNA 7(2):254–65, DOI 10.1017/s1355838201000863

[6] Höner zu Siederdissen C, Hammer S, Abfalter I, Hofacker IL, Flamm C, Stadler PF (2013) Computational design of RNAs with complex energy landscapes. Biopolymers 99(12):1124–36, DOI 10.1002/bip.22337

[7] Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL (2006) Thermodynamics of RNA-RNA binding. Bioinformatics 22(10):1177–82, DOI 10.1093/bioinformatics/btl024

[8] Amman F, Flamm C, Hofacker IL (2012) Modelling translation initiation under the influence of sRNA. Int J Mol Sci 13:16223–16240, DOI 10.3390/ijms131216223

[9] Soper T, Woodson SA (2008) The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. RNA 15:1907–1917, DOI 10.1261/rna.1110608

[10] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts

P, Murphy TD, Pruitt KD (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44(D1):D733–D745, DOI 10.1093/nar/gkv1189, URL https://doi.org/10.1093/nar/gkv1189, https://academic.oup.com/nar/article-pdf/44/D1/D733/9482930/gkv1189.pdf

[11] Brown L, Elliott T (1997) Mutations that increase expression of the rpoS gene and decrease its dependence on hfq function in Salmonella typhimurium. Journal of Bacteriology 179(3):656–662, DOI 10.1128/jb.179.3.656-662.1997, URL https://journals.asm.org/doi/abs/10.1128/jb.179.3.656-662.1997, https://journals.asm.org/doi/pdf/10.1128/jb.179.3.656-662.1997

[12] Lease RA, Cusick ME, Belfort M (1998) Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. Proceedings of the National Academy of Sciences 95(21):12456–12461, DOI 10.1073/pnas.95.21.12456, URL https://www.pnas.org/doi/abs/10.1073/pnas.95.21.12456, https://www.pnas.org/doi/pdf/10.1073/pnas.95.21.12456

[13] Yao HT, Ponty Y, Will S (2022) Developing complex RNA design applications in the Infrared framework. RNA Folding-Methods and Protocols