Flexible and universal multiple target nucleic acid sequence design

Stefan Hammer[1,2], Birgit Tschiatschek[2], Christoph Flamm[1], Ivo L. Hofacker[1,2,3] and Sven Findeiß[1,2]

1 University of Vienna, Department of Theoretical Chemistry, Vienna, A-1090, Austria
2 University of Vienna, Bioinformatics and Computational Biology Research Group, Vienna, A-1090, Austria
3 University of Copenhagen, Center for Non-coding RNA in Technology and Health, Copenhagen, DK-1870, Denmark

Abstract

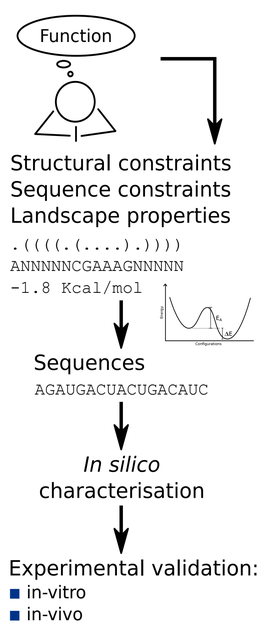
Motivation: Due to the close structure to function relationship, the availability of good structure prediction methods, and energy models, RNA is perfectly suited to design molecules with predefined properties. Currently available RNA design tools implement very specialized use-cases which cannot be easily adapted to new applications. Often, complicated sampling and optimization methods were developed to suit a specific design goal.

Results: We developed RNAblueprint, a C++ library implementing a **graph coloring approach** to uniformly sample sequences compatible to structural and sequence constraints from the typically huge solution space. **Fair sampling** from the solution space makes optimization runs much more performant and raises the probability to find better solutions. Scripting interfaces allow to easily adapt existing code to new scenarios which makes the whole design process very **universal and flexible**. We implemented novel design approaches such as a multistable thermoswitch or a sRNA mediated translation regulation system to show the advantages of our software.

Introduction to RNA design

To be able to *de novo* design RNA molecules it was necessary to develop methods to solve the inverse folding problem and approaches for an in silico characterization. A design task can be generally split into three components:

(1) generate valid sequences compatible to sequence and structural constraints, **(2)** formulate an **optimization problem** where the sampled sequences are optimized towards a optimal score calculated by the objective function and **(3)** use *in silico* characterization methods such as **filtering and clustering** with respect to problem specific features not included in the objective.



Fair sampling of RNA sequences

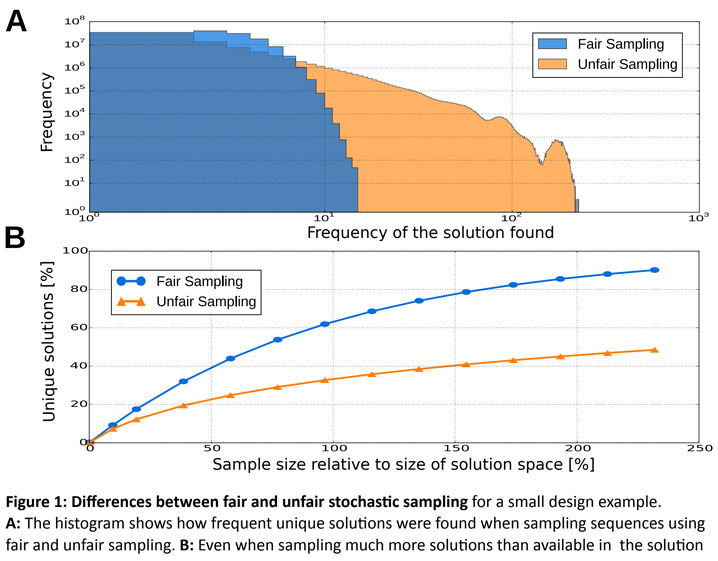


Figure 1: Differences between fair and unfair stochastic sampling for a small design example. **A:** The histogram shows how frequent unique solutions were found when sampling sequences using fair and unfair sampling. **B:** Even when sampling much more solutions than available in the solution space (∼230%), we still get only 50% of all possible solutions with unfair sampling, while we are able to sample about 90% with the fair method.

The developed software [1] implements a graph theoretical approach [2] to be able to handle **multiple structural constraints**. In addition, it takes any **sequence constraint** in IUPAC annotation as input and **generates sequences** compatible to all constraints.

RNAblueprint offers functions to get **properties of the solution space** and current search spaces, such as the number of solutions. This feature can, for example, be used for **mutational studies**. For efficient optimizations we guarantee to sample every solution with the same probability from the whole solution space as shown in Figure 1.

Multi-state thermoswitch

To show the advantages of our software, namely the flexibility, universality and efficiency, we implemented serveral designs including a multistate thermoswitch. This RNA molecule is able to fold into three distinct states at specific temperatures. The objective function not only includes a term to favour the given structures at their temperatures, but also to counter select for the structures at all other temperatures.

Objetive Function

$$f(x) = \underbrace{\sum_{i}^{M} E_{Ti}(x,\Omega_i) - G_{Ti}(x))}_{\text{favor structure at temperature}} + \xi \times \underbrace{\sum_{i}^{M}[E_{Ti}(x,\Omega_i) - \sum_{j\neq i}^{M} E_{Ti}(x,\Omega_j)]}_{\text{penalize other states at temperature}}$$

x ... sequence
Θ ... structure
M ... number of structures
$E_T(x,\Theta_i)$ ... energy of structure at temperature T
$G_T(x)$ ... gibbs free energy at temperature T

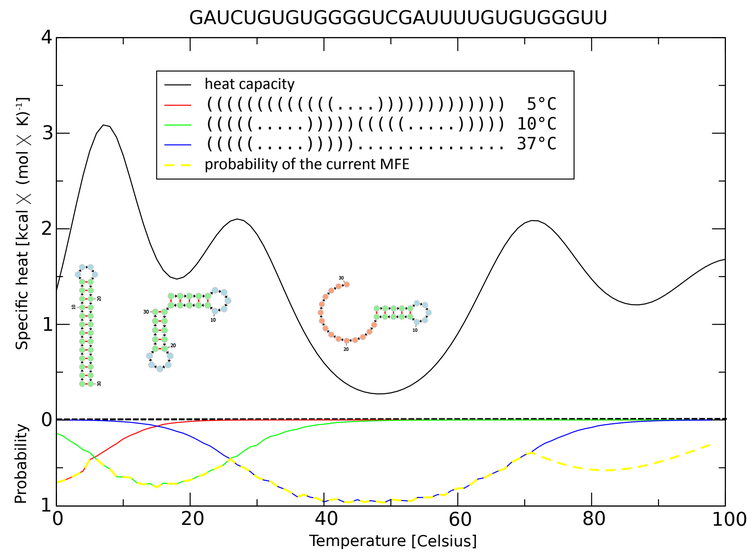


Figure 3: Heat capacity curve and thermodynamic probabilities of the target structures evidence the sucessful design. The molecule folds into the desired states at the given temperatures with high probabilities and shows distinct transitions between the states. The RNA completely unfolds above 70 degrees to then exhibit the open chain formation as minimum free energy structure.

sRNA regulated gene expression

We furthermore implemented a regulatory 5'UTR that retulates expression of its downstream gene by responding to the presence of a small RNA. We therefore came up with a novel objective function which not only contains the accessibility of the RBS and the sRNA binding site, but also the concentration of the formed sRNA/5'UTR complex for efficient binding. An additional term minimizes the crosstalk of the 5'UTR with the downstream reporter gene.

Objetive Function

$$f(x) = \underbrace{1 - [S_{complex}]/[mRNA_0]}_{\text{Efficient duplex formation}} + \xi_1 \times \underbrace{(1 - P(5'UTR_{unpaired}))}_{\text{RBS and AUG are accessible}} + \xi_2 \times \underbrace{(1 - P(sRNA_{unpaired}))}_{\text{sRNA binding site accessible}} + \xi_3 \times \underbrace{(1 - P(mRNA_{context}))}_{\text{5'UTR folds independent of the reporter gene}}$$

$[S_{complex}]$ ... concentration of the 5'UTR:sRNA complex in a specific structure S
$[mRNA_0]$ ... start concentration of the mRNA set to 10^{-5} mol/l
$P(x_{unpaired})$ ... probability of a specified sub sequence being unpaired in sequence x
$P(mRNA_{context})$ ... probability of the 5'UTR folding independently of the coding region

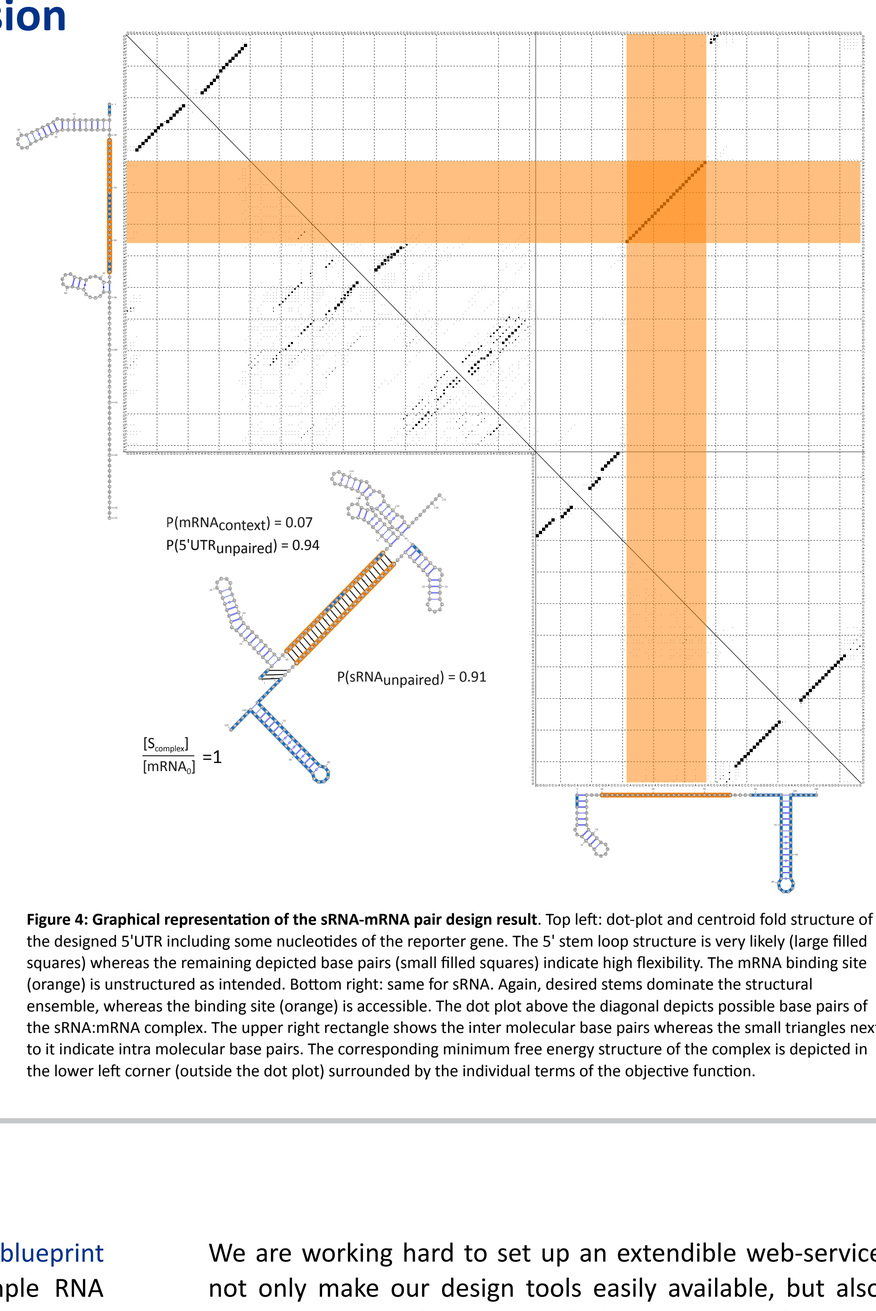
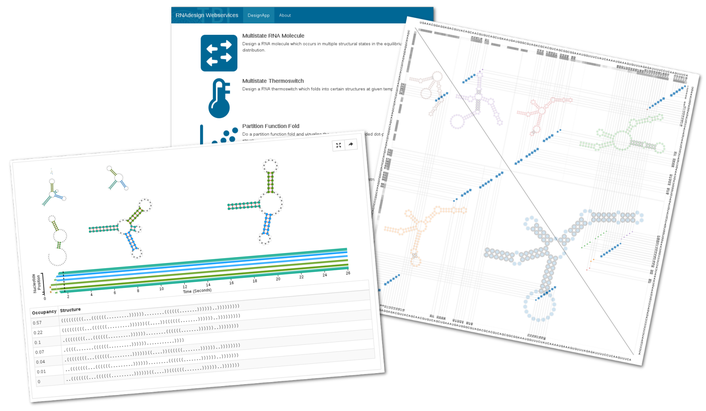


Figure 4: Graphical representation of the sRNA-mRNA pair design result. Top left: dot-plot and centroid fold structure of the designed 5'UTR including some nucleotides of the reporter gene. The 5' stem loop structure is very likely (large filled squares) whereas the remaining depicted base pairs (small filled squares) indicate high flexibility. The mRNA binding site (orange) is unstructured as intended. Bottom right: same for sRNA. Again, desired stems dominate the structural ensemble, whereas the binding site (orange) is accessible. The dot plot above the diagonal depicts possible base pairs of the sRNA:mRNA complex. The upper right rectangle shows the inter molecular base pairs whereas the small triangles next to it indicate intra molecular base pairs. The corresponding minimum free energy structure of the complex is depicted in the lower left corner (outside the dot plot) surrounded by the individual terms of the objective function.

Conclusion

We provide a software solution in form of the RNAblueprint library which makes it possible to uniformly sample RNA sequences compatible to structural and sequence constraints. With our framework it is possible to incorporate almost any software like NUPACK [4] or ViennaRNA [5] in the calculation of the objective function and it is now feasible to explore a much broader range of objectives. We showed examples on how to use the developed software to implement a design method using *de novo* objective functions that incorporate terms which were not available in other existing design frameworks.

We are working hard to set up an extendible web-service to not only make our design tools easily available, but also to provide a range of analysis tools for in silico verification and characterization of the generated designs or any other RNA of interest.



http://nibiru.tbi.univie.ac.at/blueprint/

References

[1] **Stefan Hammer**, Birgit Tschiatschek, Christoph Flamm, Ivo L. Hofacker and Sven Findeiß. RNAblueprint: Flexible and universal multiple target nucleic acid sequence design. Bioinformatics, submitted 2016.
[2] Christian Höner zu Siederdissen, **Stefan Hammer**, Ingrid Abfalter, Ivo L. Hofacker, Christoph Flamm, and Peter F. Stadler. Computational design of RNAs with complex energy landscapes. Biopolymers, 99(12):1124–1136, 2013.
[3] Christoph Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multistable RNA molecules. RNA, 7(2):254 265, February 2001.
[4] Joseph N. Zadeh, Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. NUPACK: Analysis and design of nucleic acid systems. Journal of Computational Chemistry, 32(1):170–173, 2011.
[5] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1):26, November 2011.

This work has been supported by the European Commission under the Environment Theme of the 7th Framework Program for Research and Technological Development (Grant agreement number 323987).