Computational RNA Design

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,
University of Leipzig
Max Planck Institute for Mathematics in the Sciences
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology
Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
Center for non-coding RNA in Technology and Health, U. Copenhagen
The Santa Fe Institute (external faculty)

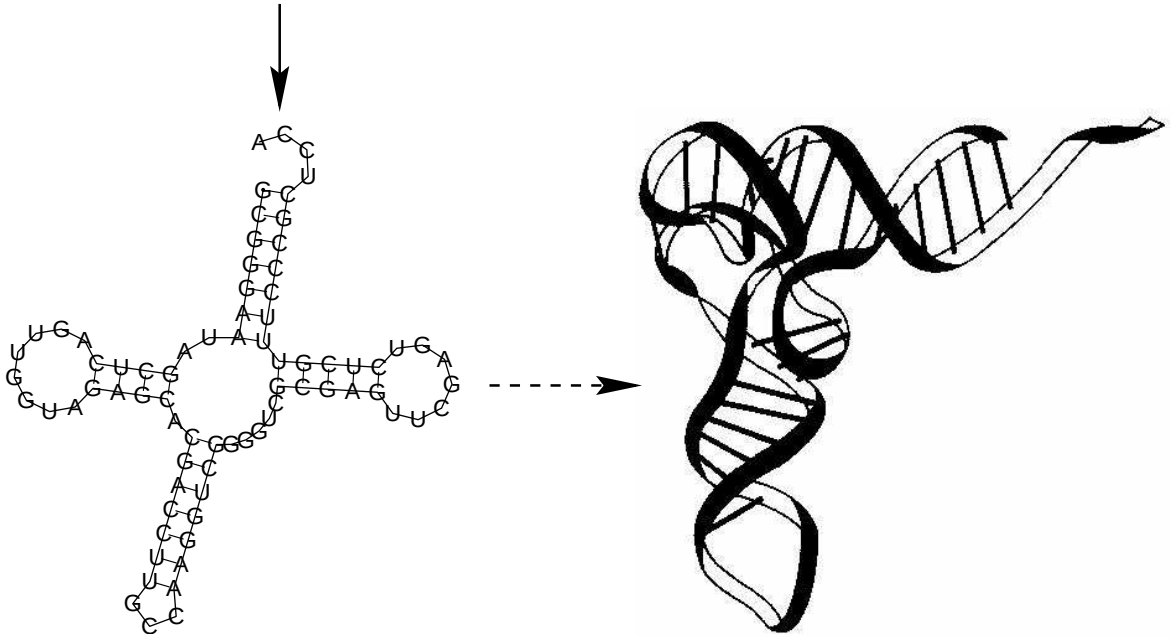
Boston, Jul 07 2014

From Sequence to Shape and Back

A historical note: why inverse folding is easy in practice ...

The starting point

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA



Sequence-structure relationships in RNA

Sequence-Structure Map of RNA

1. *Redundancy:* Many more sequences than structures
2. *Sensitivity:* Small changes in the sequences may lead to large changes in the structure
3. *Neutrality:* A substantial fraction of mutations does not alter the structure.

Implications:

1. *Neutral Networks:* $S(\psi)$ forms a connected "percolating" network in sequence space for all "common" structures.
2. *Mutual Accessibility:* The neutral networks of any two structures almost touch each other somewhere in sequence space.

- RNA sequence evolve drift-like while maintaining secondary structure
- Substitution pattern reflects selective constraints on the structure

Proc.Roy.Soc.B **255** 279-284 (1994), Proc. Natl. Acad. Sci. USA **93**, 397-401 (1996),

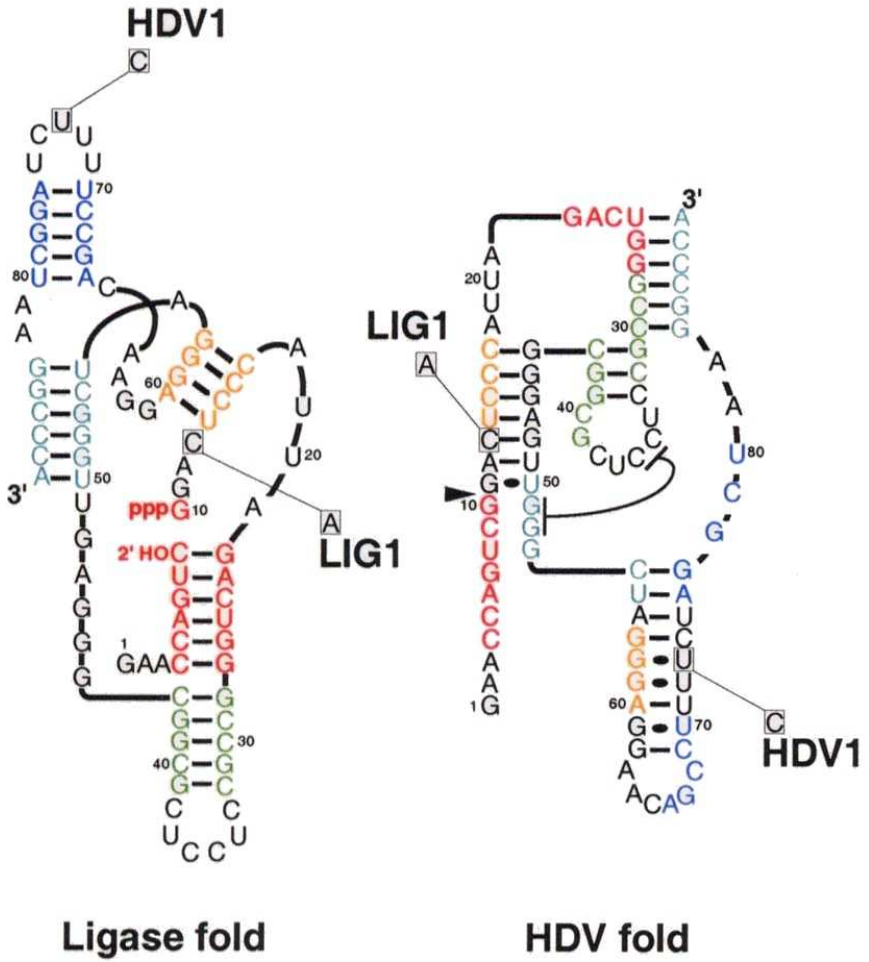
Bull. Math. Biol. **59**, 339-397 (1997), RNA **7**: 254-265 (2000).

Implementations

Several programs are available that implement simple inverse folding:

- `RNAinverse` (Hofacker *et al.* 1994)
  unbiased search using adaptive walks and a simple hierarchical problem decomposition
- `RNAdesigner` (Andronescu 2004)
  Uses sequence bias in paired/unpaired regions, more sophisticated decomposition to speed up search
- `INFO-RNA` (Busch & Back 2006)
  Uses a sequence with minimal energy for the target structure as starting point (strong sequence bias)
- `NUPACK` (Zadeh, Wolf, Pierce 2011)
  Allows design of multi-strand structures
- `RNAiFOLD` (Garcia-Martin, Clote, Dotu 2013)
  uses constraint programming or a large neighborhood search heuristic

One Sequence, Two target Structures



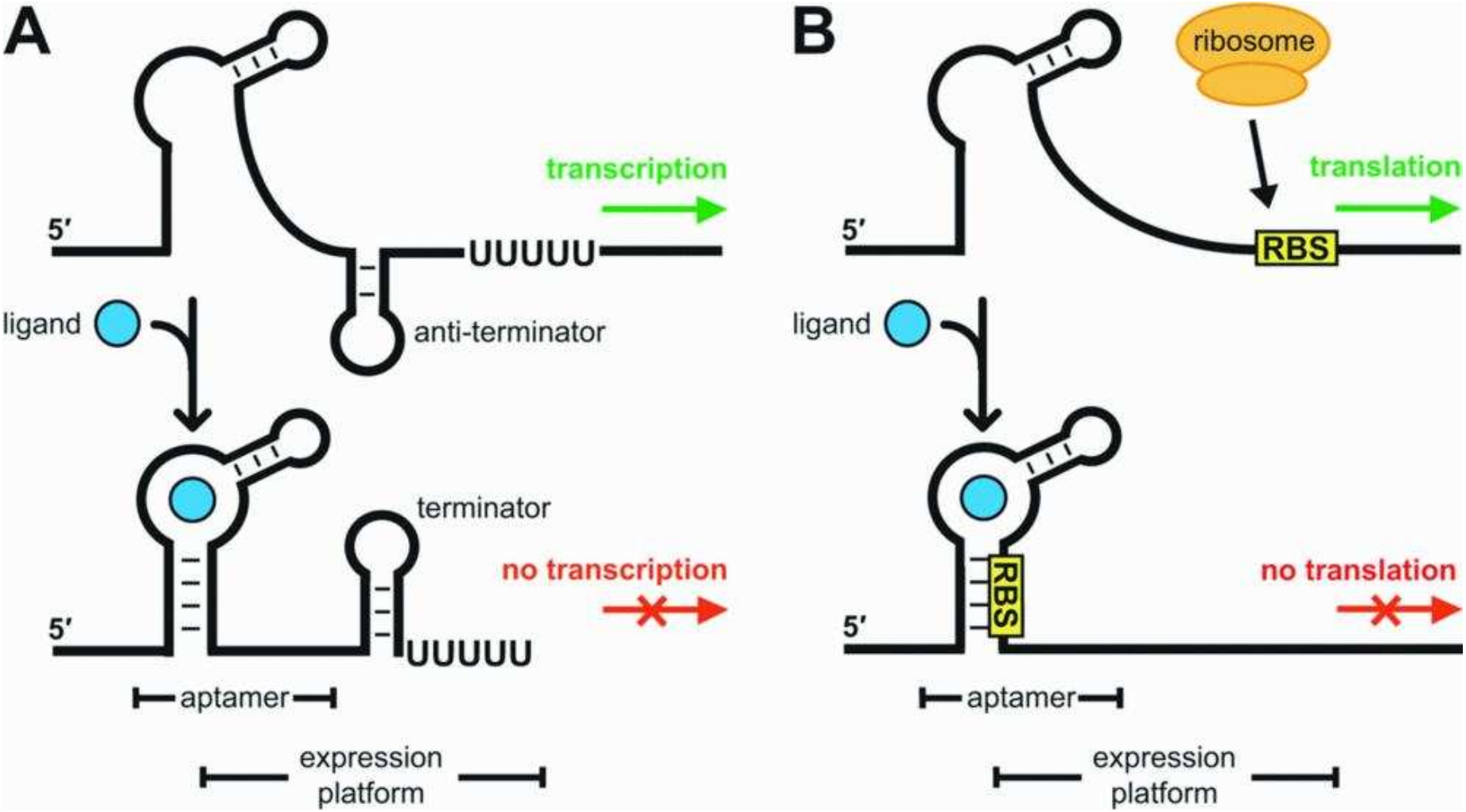
Schultes, EA & Bartel, DP; Science (2000), **289**:448-452

Design of Artificial Riboswitches

- Riboswitches are a convenient gadget in synthetic biology
- Task: combine ligand-specific sensor with an effector (i.e., some form of a regulatory element)
- Question: to what extent is this really modular?
- Idea: use RNA structure prediction to model the interplay of sensor and effector

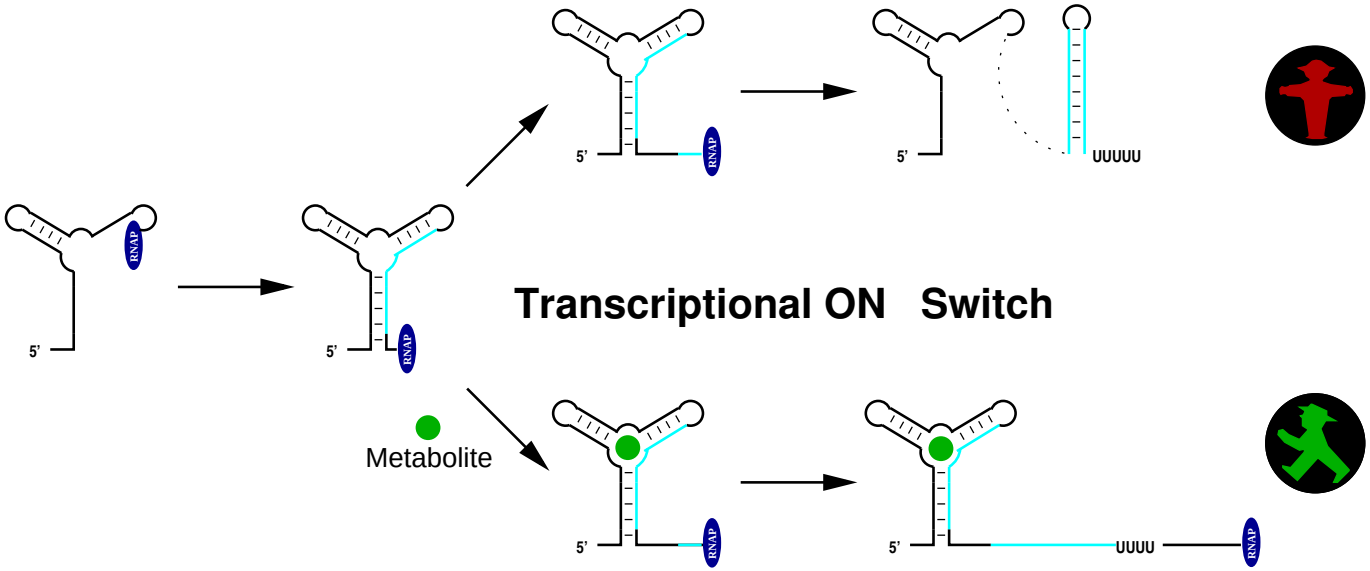
Riboswitches: Regulators of Gene Expression

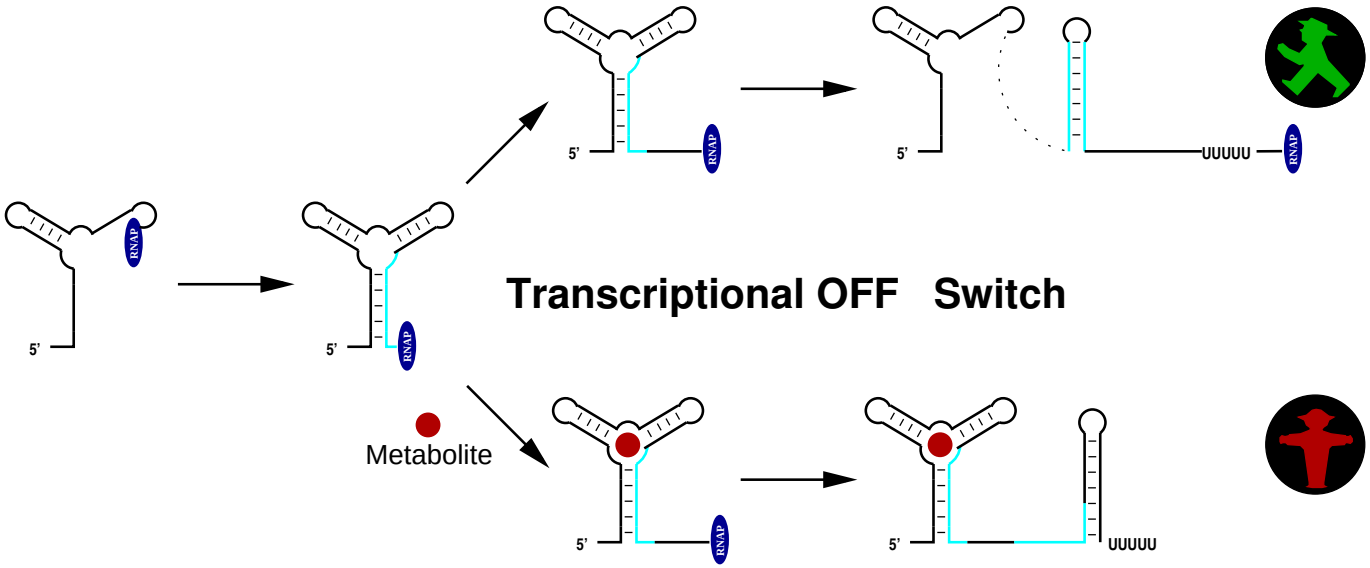
Transcriptional *versus* translational riboswitch



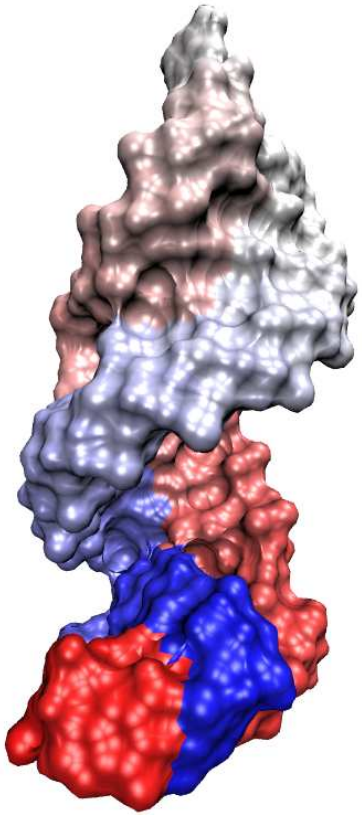
Kim & Breaker, *Biol. Cell* (2008)

Ribo-Switching of Transcription

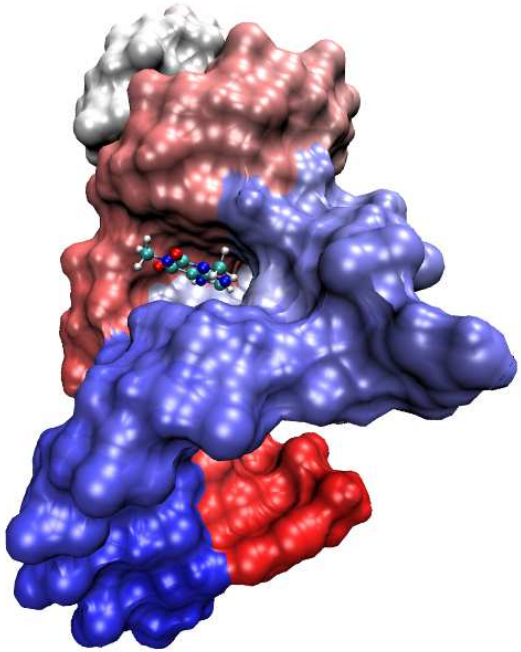




Theophylline Aptamer

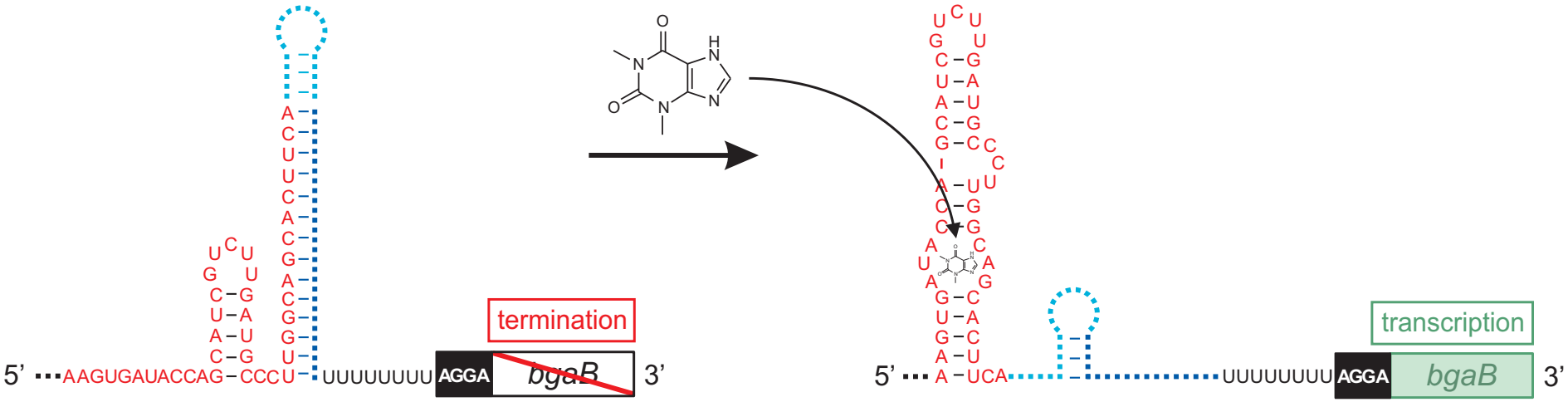


Unbound aptamer
Model predicted using Rosetta



Theophylline bound aptamer
Crystal Structure
(PDB-ID 1O15)

Design Idea



Goal: a theophylline triggered on-switch

Essence of the Multistable Design Problem

- Design a sequence that *compatible* with not just one but *several* target structures
- Each target should be almost a ground state
- **Questions:**
  - When can this be solved?
  - How can we include ligang specificity
- First step: generate sequences that are compatible with all design goals.
- 2nd step: optimize the sequences toward the design goal(s)

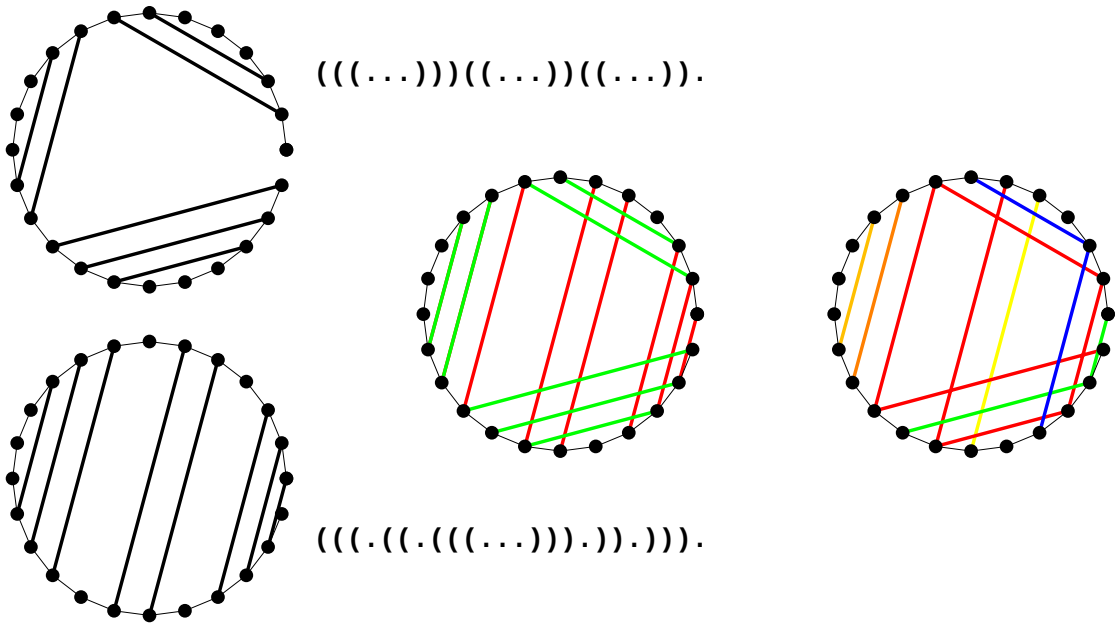
Bi-Stable Structures

Given two structures $\mathcal{S}_1$ $\mathcal{S}_2$, are there sequences compatible to both?

intersection theorem:

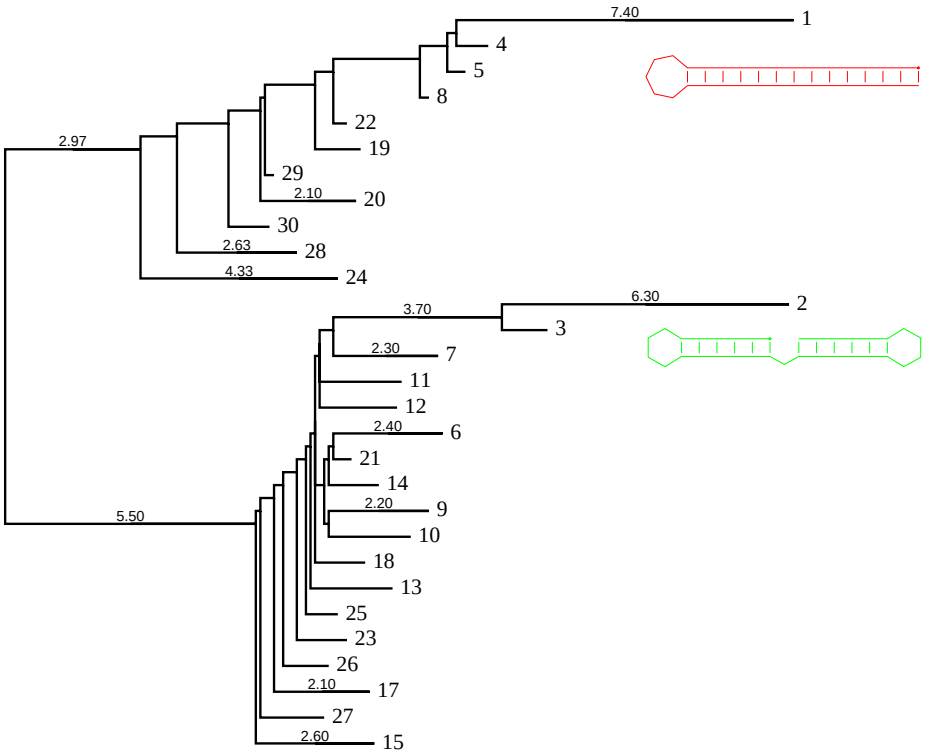
$$\mathbf{C}[\mathcal{S}_1] \cap \mathbf{C}[\mathcal{S}_2] \neq \emptyset$$

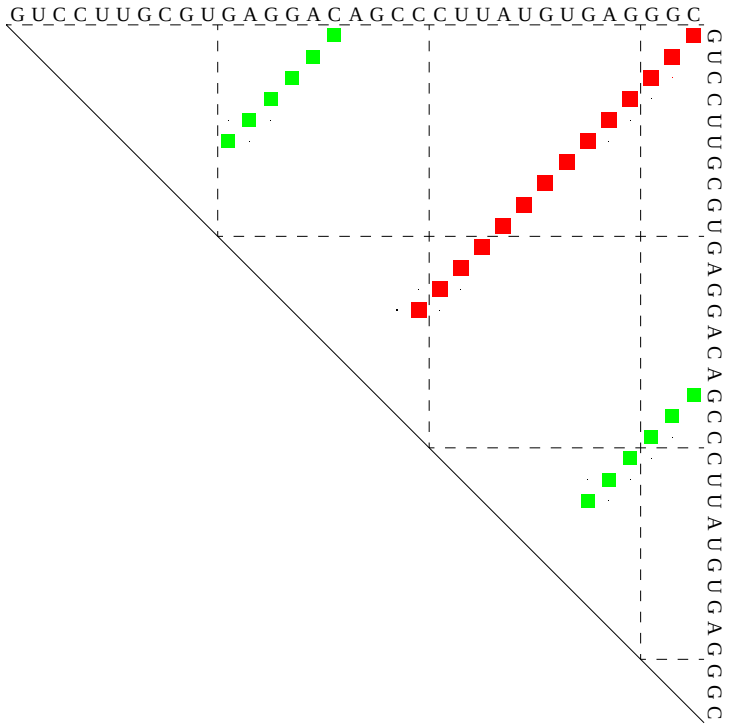
Proof: Dependency graph decomposes into paths and cycles of even length



the alternating sequence AUAUAU... is compatible with each path and cycle.

Examples of bistable structures





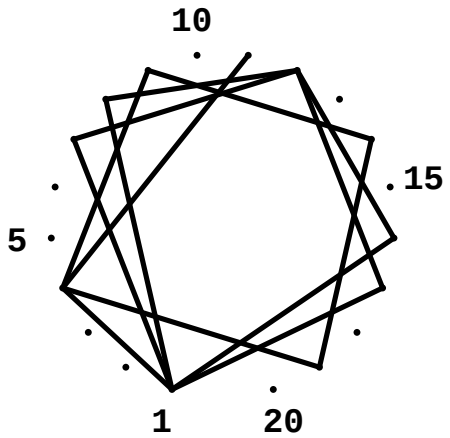
$$\Xi(x) = E(x, \Omega_1) + E(x, \Omega_2) - 2G(x) + \xi \left(E(x, \Omega_1) - E(x, \Omega_2)\right)^2$$

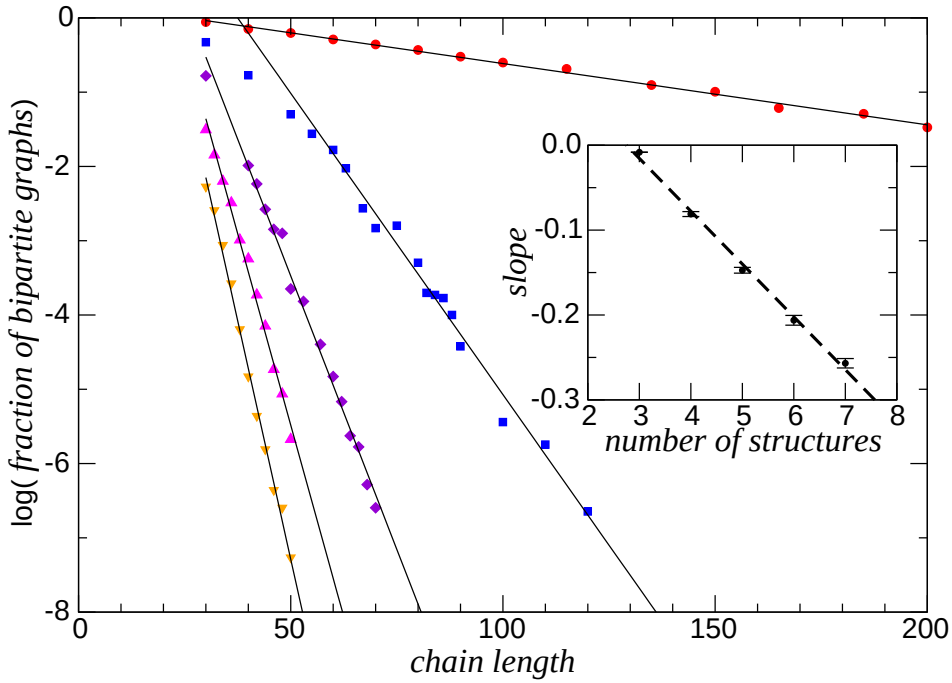
Multi-Stable Structures

Generalization to multiple Targets:

Theorem. There is a sequence satisfying each secondary structure constraints $S_1, S_2, \ldots, S_M$ if and only if the overlap graph $S_1 \cup S_2 \cup \cdots \cup S_M$ is bipartite.

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





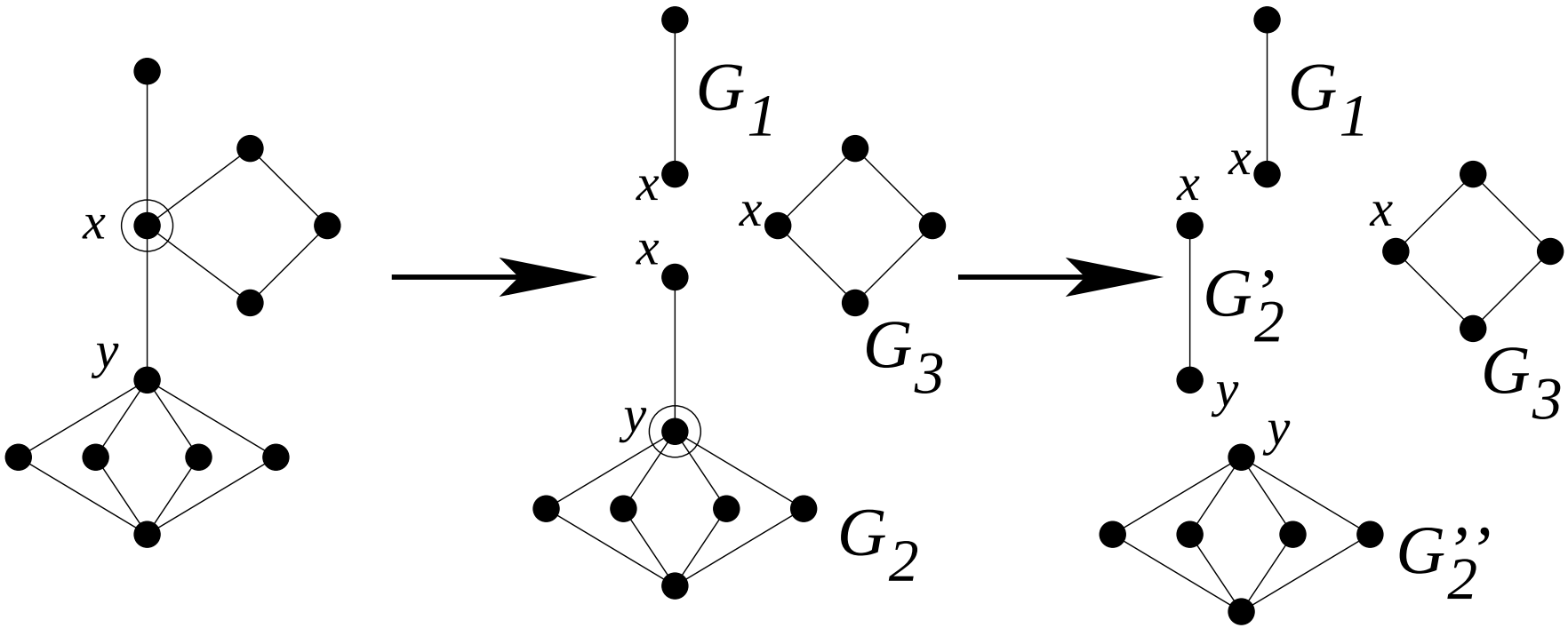
Solving Multi-Constraint Design Problems

- one possibility: constraint programming [Dotu's work]
- stochastic heutristics
  - Complex search space. Only $\mathbf{C} := \bigcap_{i=1}^{M} \mathbf{C}(\Omega_i)$ allowed
  - How to choose a good (fair) starting position?
    simple for $M = 2$: constraints are path and cycles. Simple recursions to sample uniformly from $\mathbf{C}$
  - Difficult for $M > 2$: need more complex descompositions of graphs

How to sample uniformly?

Use the overlap graph: it defines the dependence structure of nucleoides.

- Zeroth Step: connected components are indepenently sampled.
- First Step: block decomposition of the overlap graph.
  Color every block separately with fixed colors at the cut points

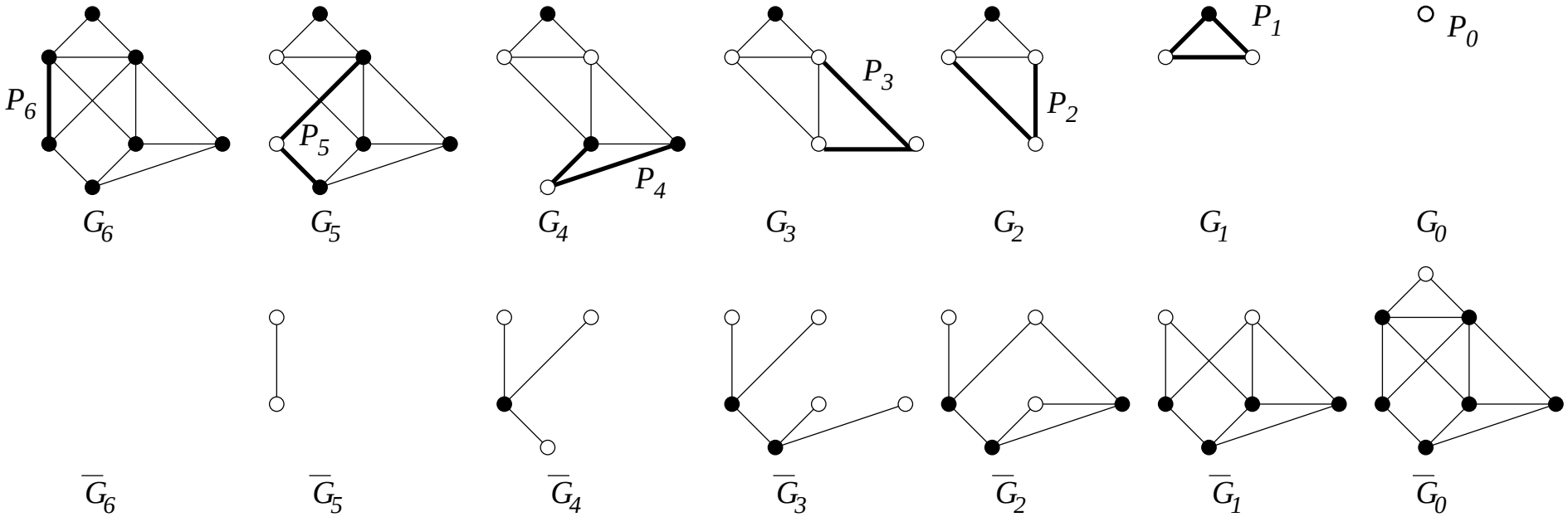


- Second Step: Color Blocks by Dynamic Programming

Biopolymers **99**: 1124-1136 (2013)

Coloring dense blocks: Ear decomposition

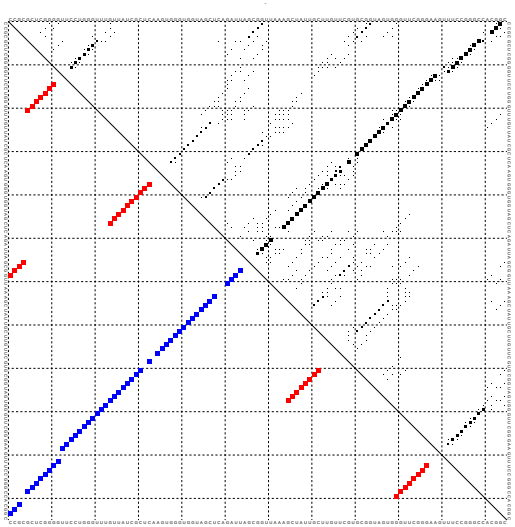
ear decomposition

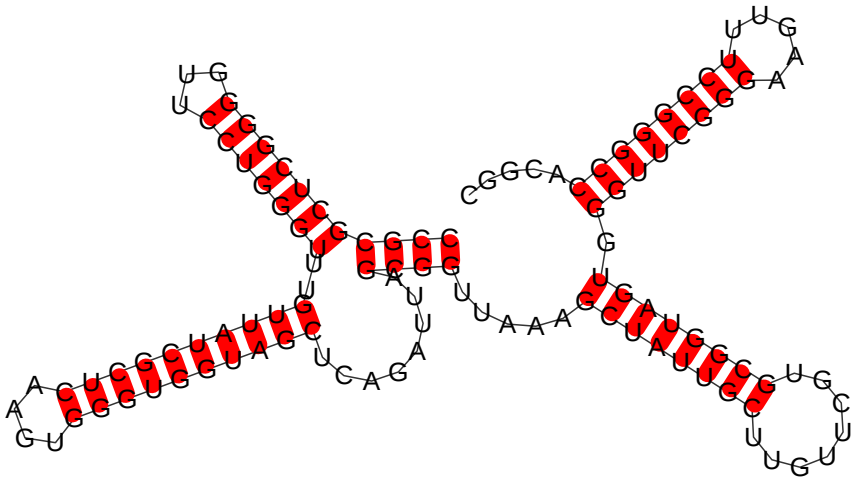


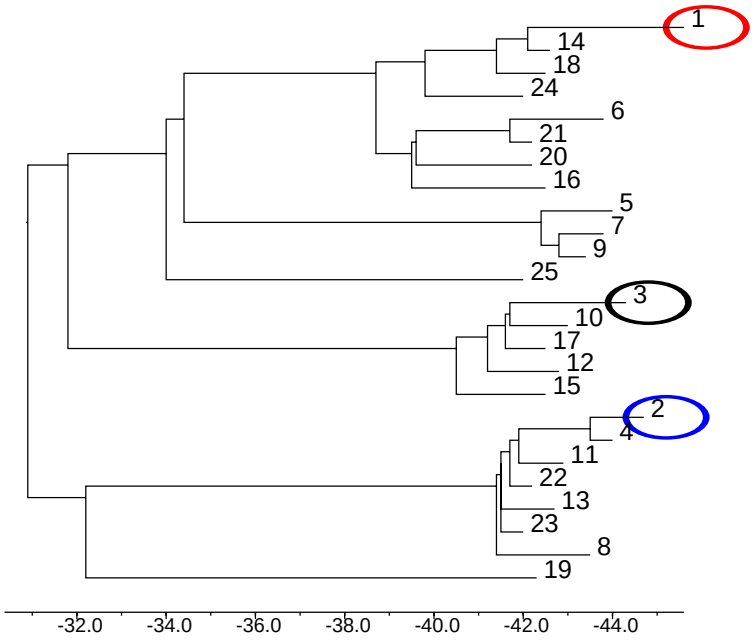
complement graph with attachment vertices

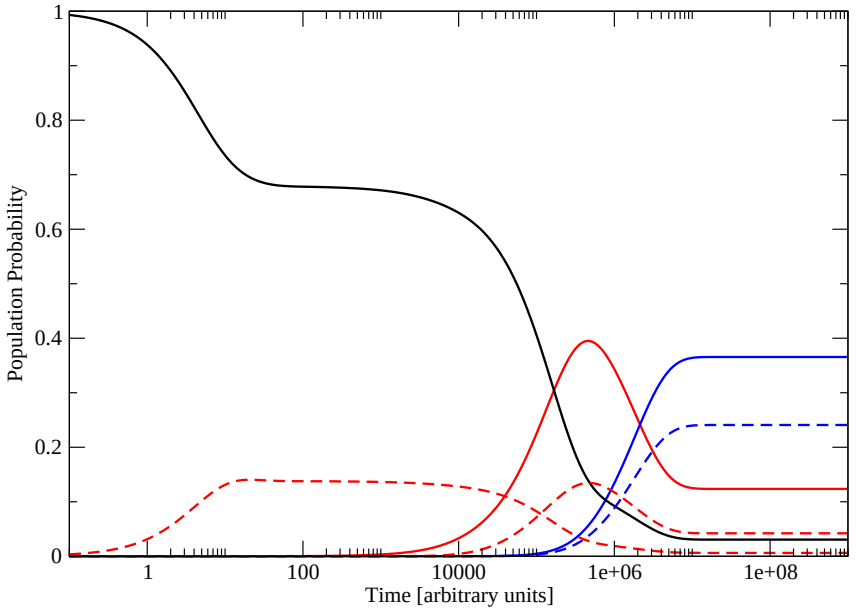
- dynamic programming approach to count colorings with given color combinations at the attachment vertices.
- memory exponential in the maximum number of attachment vertices α, CPU time in the maximum size of the union of attachment vertices in consecutive steps β

A design for the SV11 switch

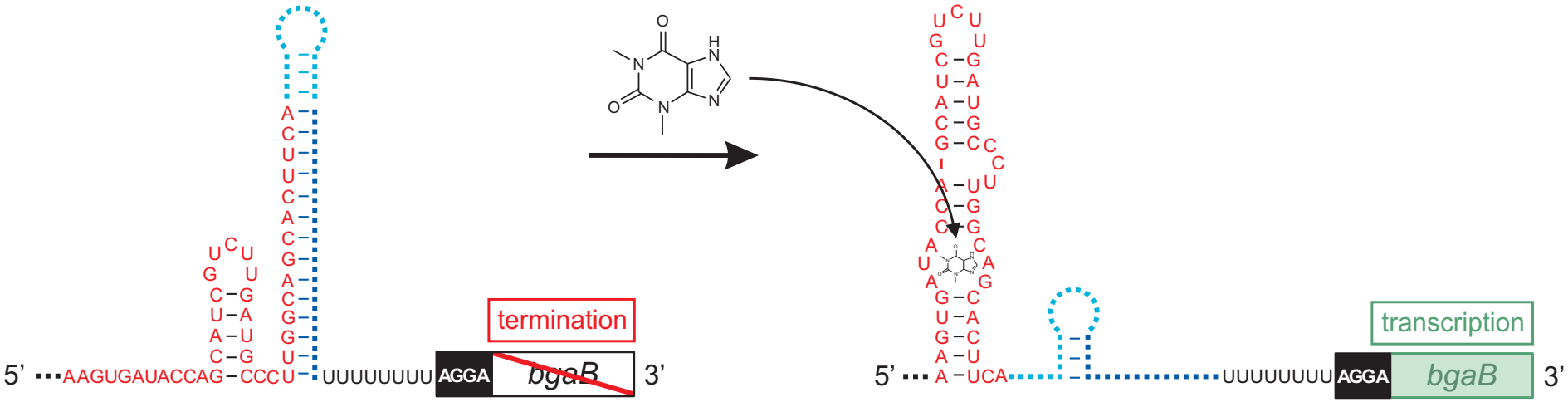






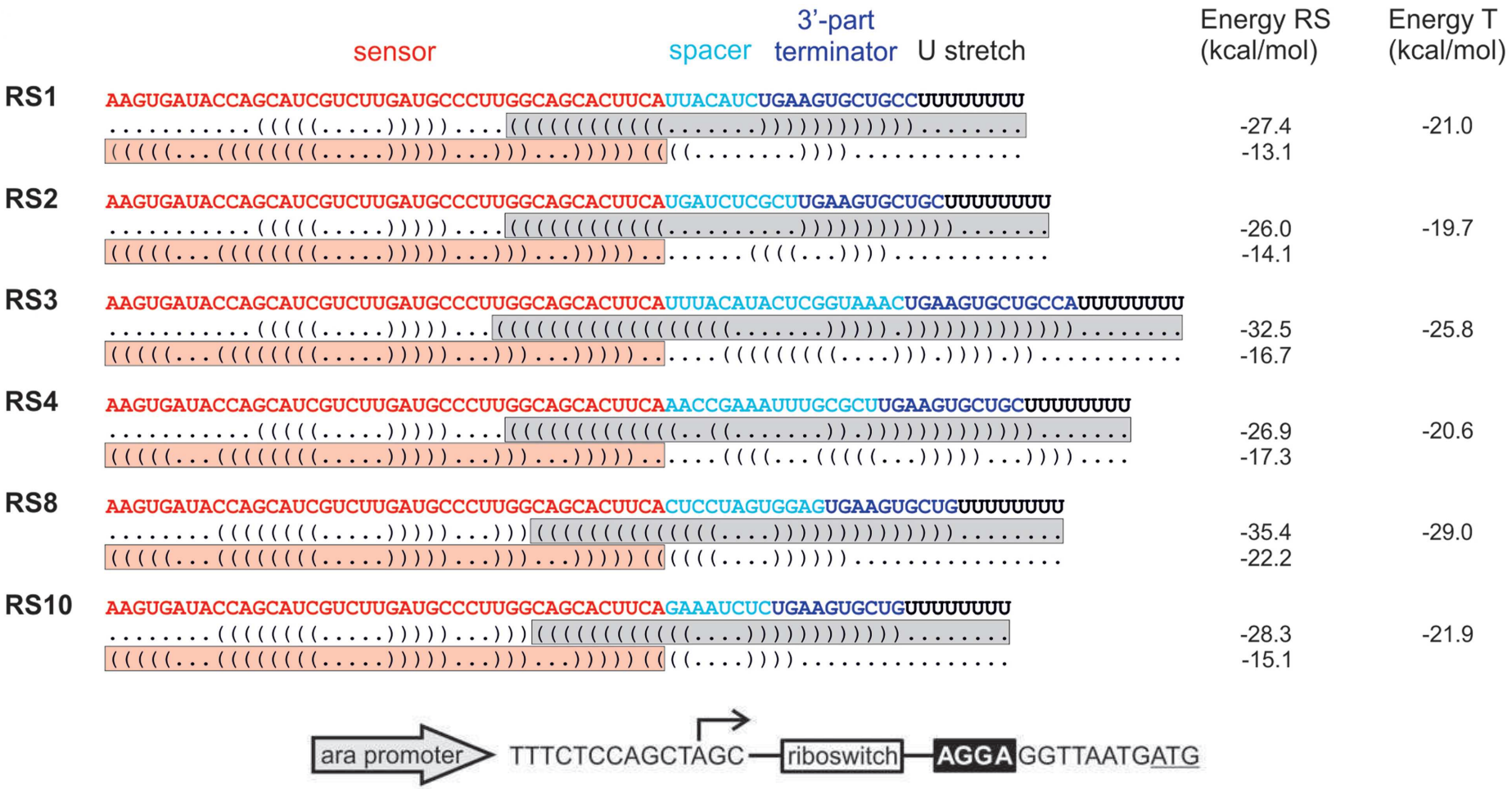


Back to the Theophylline Switch



Goal: a theophylline triggered on-switch

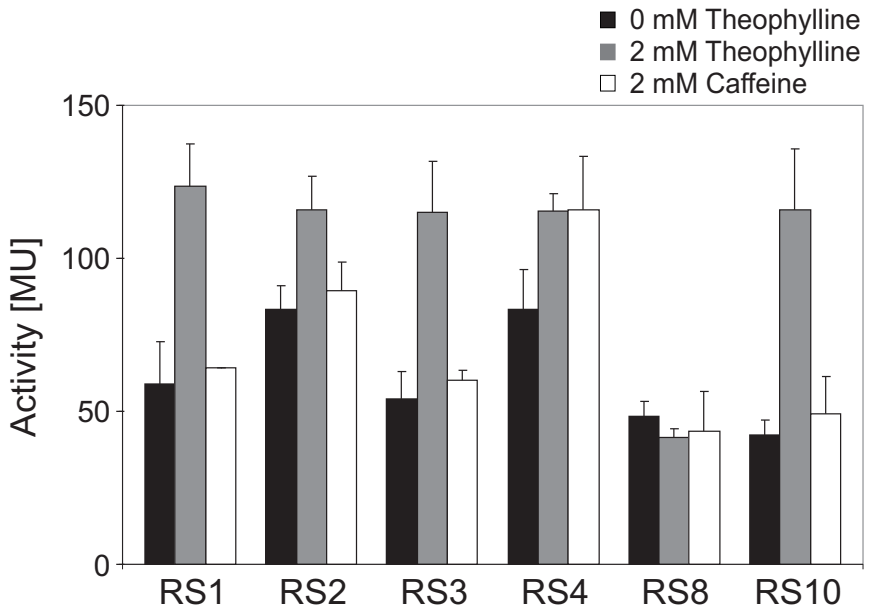
Designed Theophylline Switches



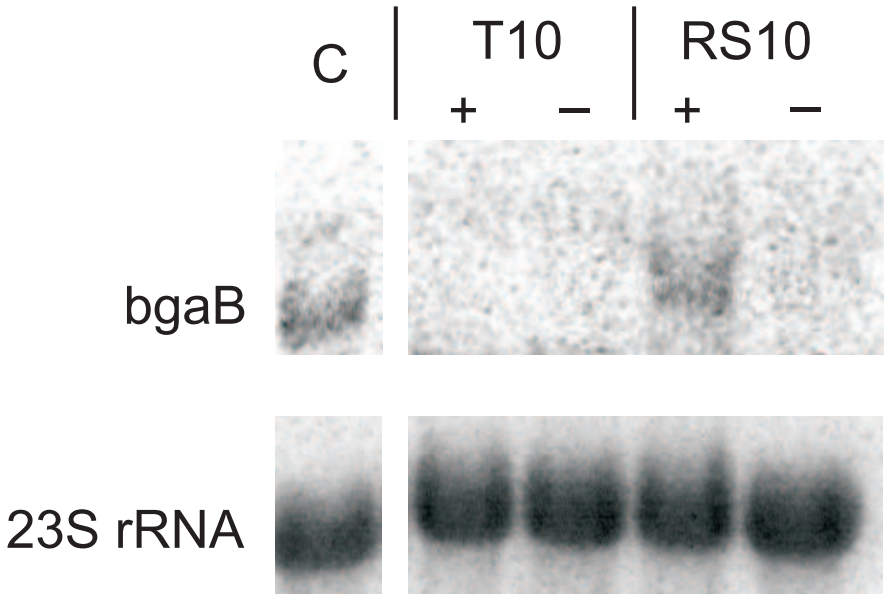
Transcribed from arabinose promoter of plasmid pBAD, using β-galactosidase as reporter gene.

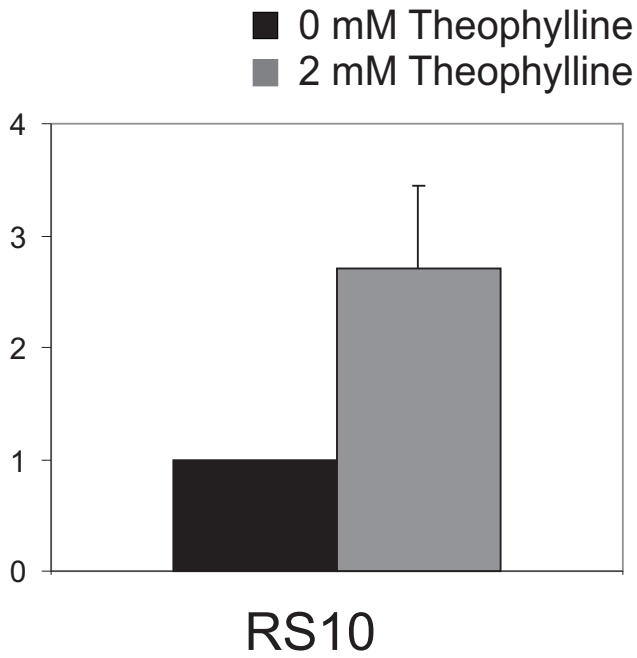
Construct Expression

	sensor	spacer	3'-part terminator	poly(U)
RS1:	aptamer-	UUACAUC----------	-UGAAGUGCUGCC	--UUUUUUUU
RS2:	aptamer-	UGAUCUCGCU--------	-UGAAGUGCUGC	---UUUUUUUU
RS3:	aptamer-	UUUACAUACUCGGUAAAC	-UGAAGUGCUGCCA	-UUUUUUUU
RS4:	aptamer-	AACCGAAAUUUGCGCU--	-UGAAGUGCUGC	---UUUUUUUU
RS8:	aptamer-	CUCCUAGUGGAG------	-UGAAGUGCUG	----UUUUUUUU
RS10:	aptamer-	GAAAUCUC----------	-UGAAGUGCUG	----UUUUUUUU



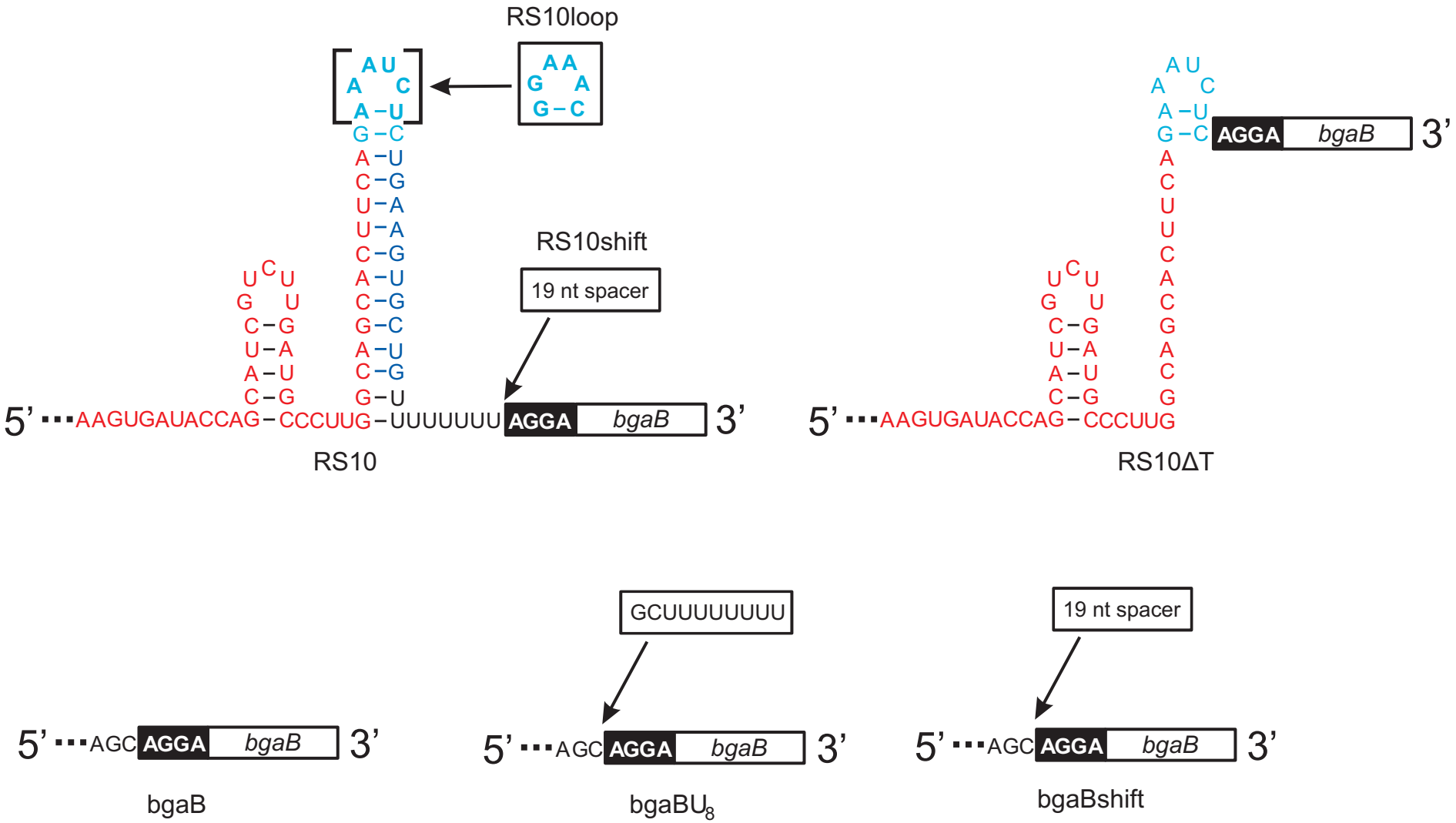
Transcriptional Switching



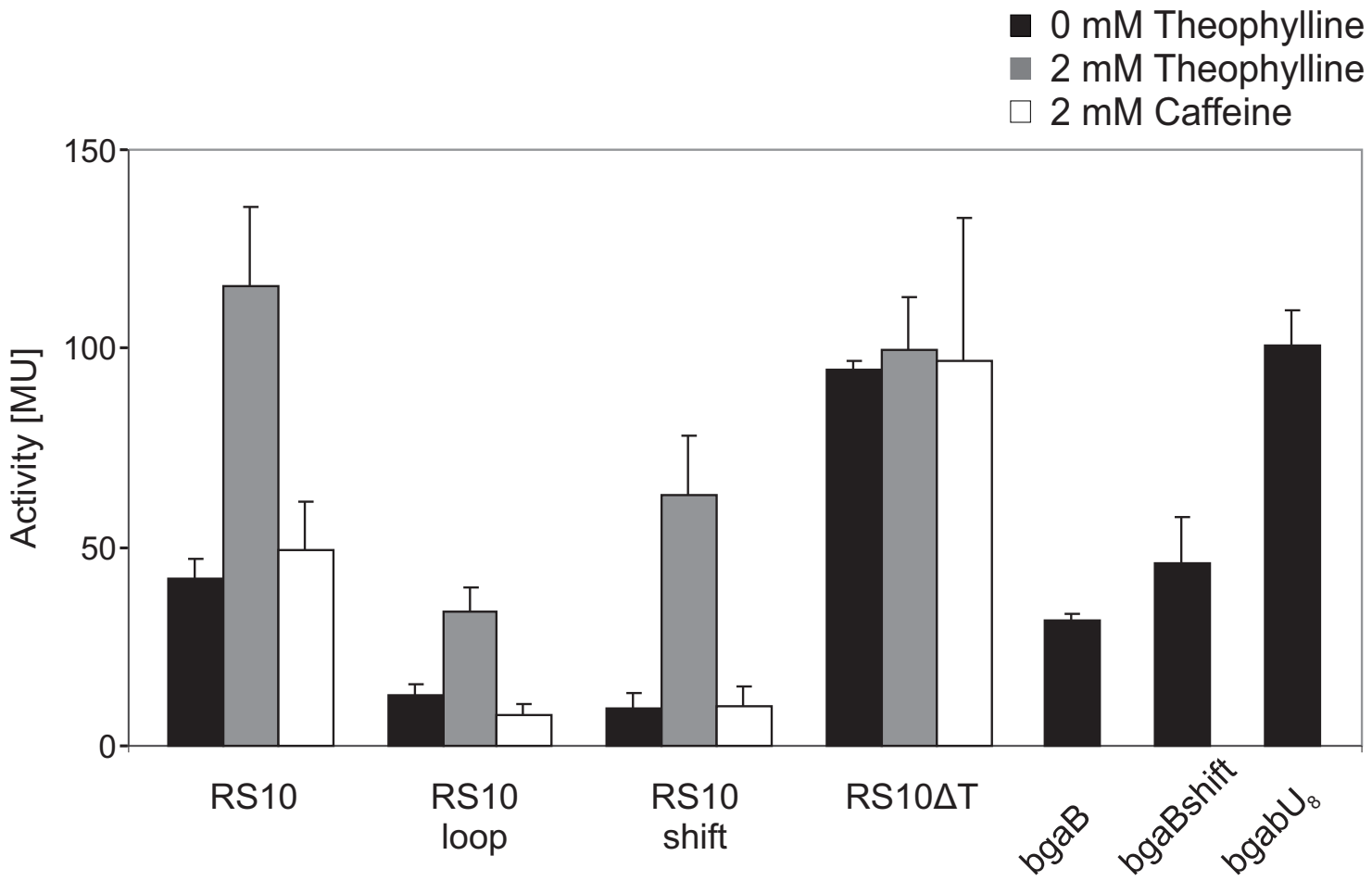


Northern blot of RS10 and terminator T10

Optimizing the Switch

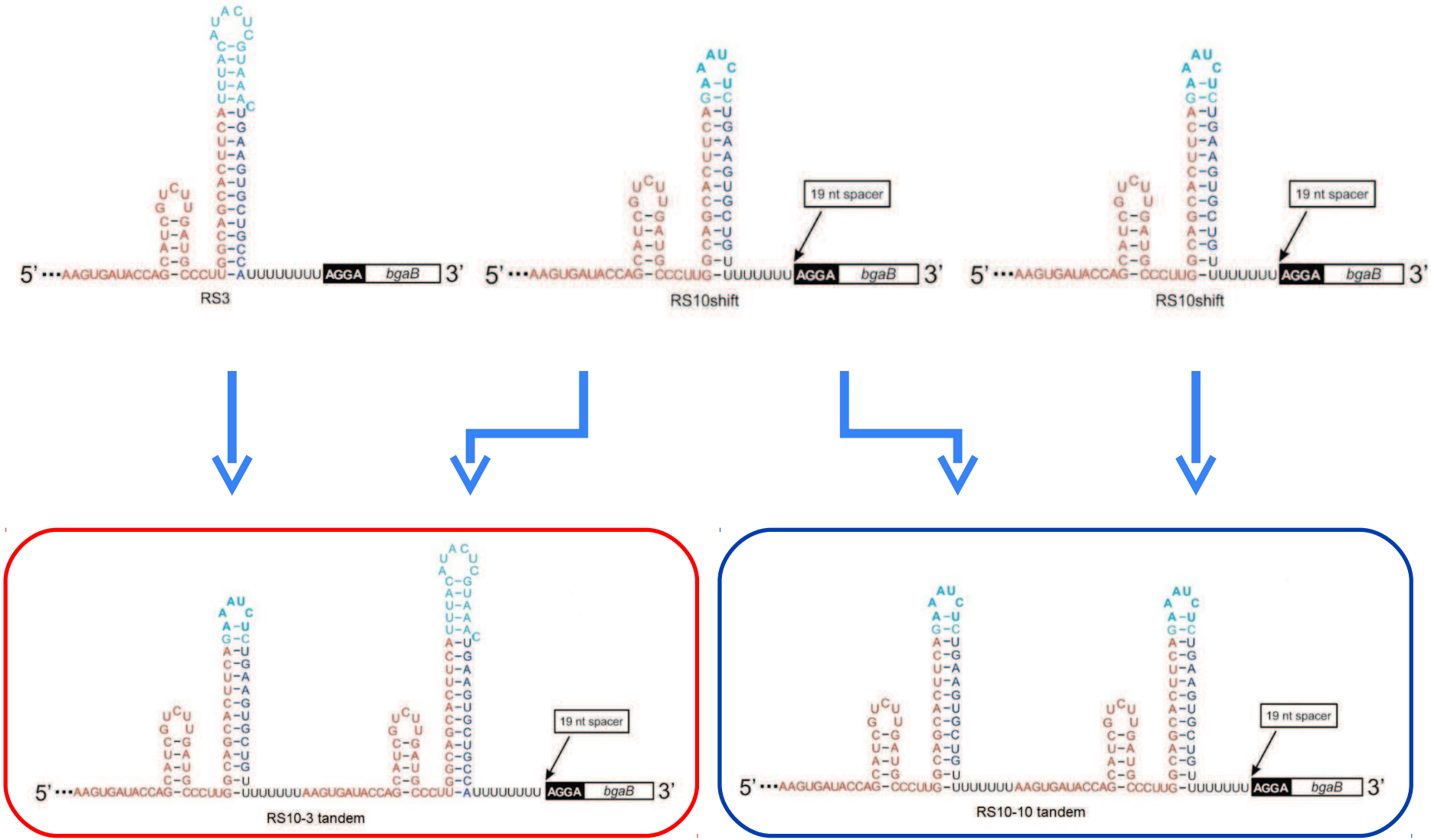


Optimizing the Switch

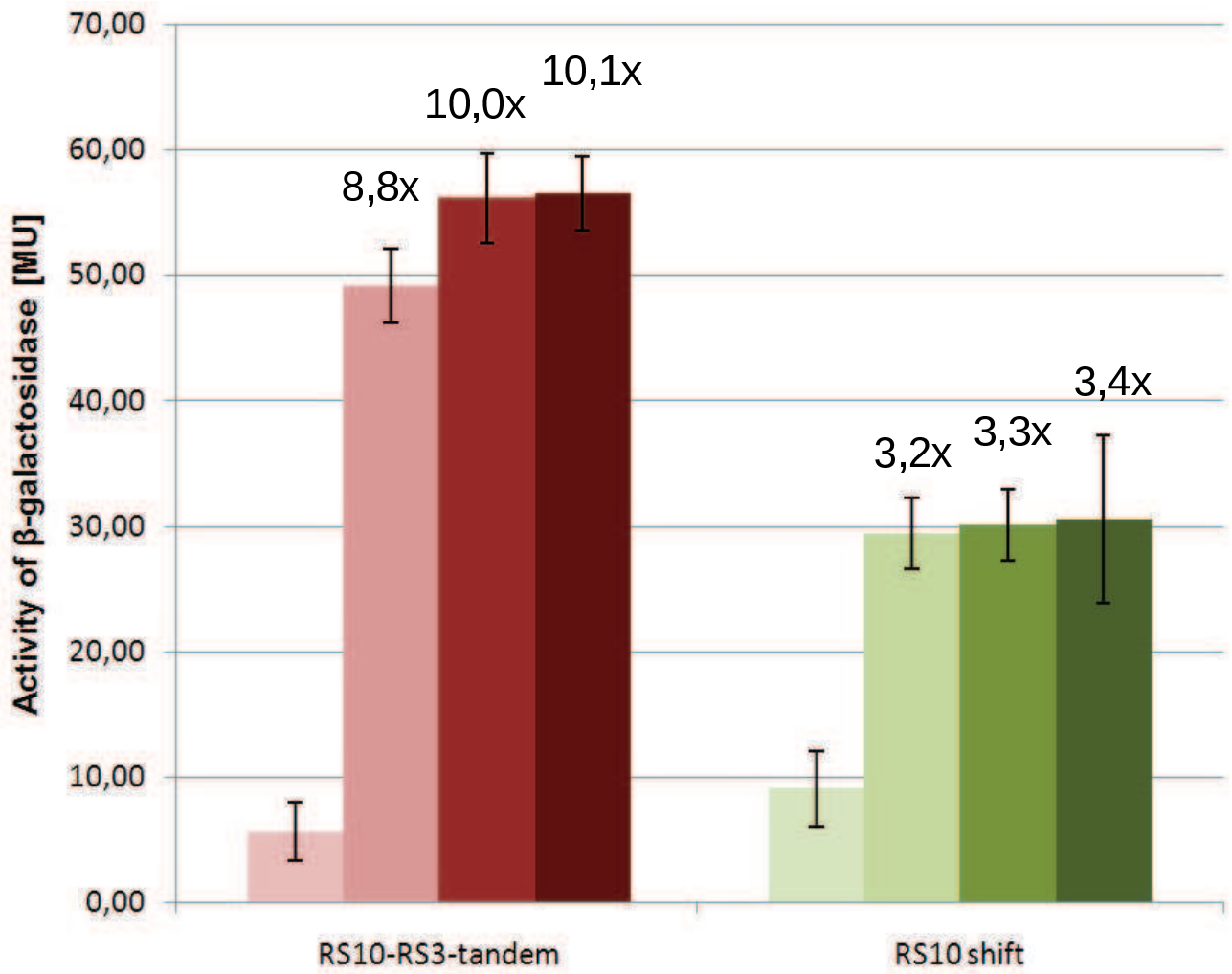


Activities of optimized constructs

Working in Tandem ...



... works even better



A more principled way to include ligand binding

- Known/measured binding energy $-\varepsilon$ of the ligand to a particular structural motif Ψ necessary for binding.
- without ligand: we want structural feature Ω
- with ligand: we want structural feature Ψ
- Compute the partition function $Z[\Psi]$ over all structures with feature Ψ.

  Partition function over structures without feature Ψ is $Z[\neg\Psi] := (Z - Z[\Psi])/Z$.
- Binding distorts the ensembl of structure when the ligand is present: $Z_L = Z[\neg\Psi] + Z[\Psi]\exp(-\varepsilon)$
- Objectives
  - without ligand: $p_0(\Omega) := Z[\Omega]/Z \to max$ and $p_0(\Psi)$ should be small
  - with ligand: $p_L(\Psi) := Z[\Psi]\exp(-\varepsilon)/Z_L \to$ max and $p_L(\Omega)$ should be small.

    ...easy if Ψ and Ω are mutually exclusive, otherwise we also need the partition function $Z(\Omega \wedge \Psi)$.

Using Constraints ...

Modified folding algorithms that scores certain structures differently
$Z\{\psi; e\}$ scores a (set of) pattern(s) ψ (in practice e.g. the loop forming the binding pocket) with bonus energies e
Key relationship:

$$\frac{[RNA \cdot L]}{[RNA]\,[L]} = K = \frac{Z\{\psi; e\}}{Z z_L}$$

Set $z_L = 1$ for a small molecular ligand and gauge the binding energies accordingly.

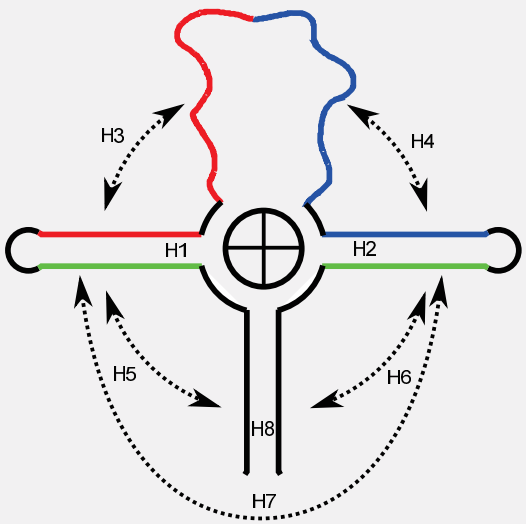
$$Z\{\psi; e\} \approx Z[\neg\Psi] + Z[\Psi]\exp(-\varepsilon)$$

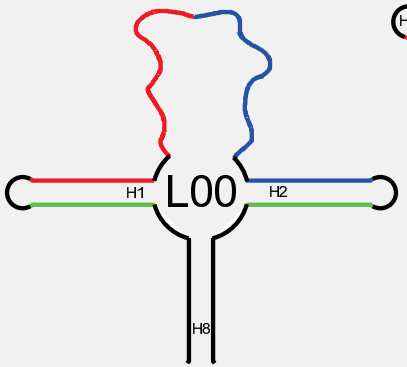
in the more general model with (soft) constraints we may include a more elaborate parametrization that includes e.g. a set of variant binding site structures ...
Details of the theory (and implementations) are still being developed ... see Ronny's talk

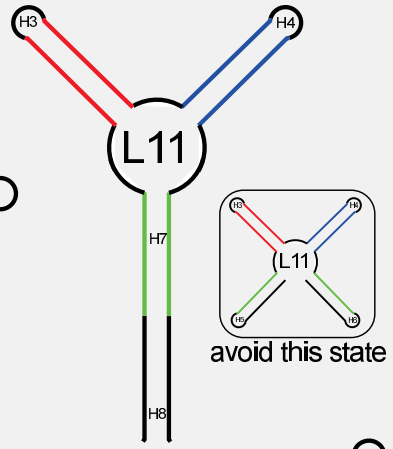
Designing an Integrated XOR Switch

Idea: Combination of two adptamers interacting non-linearly

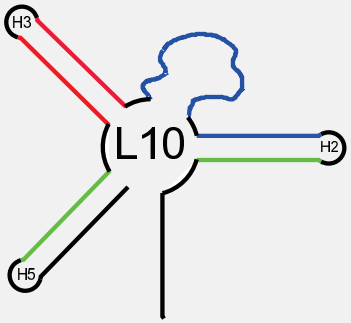
	Lig1	Lig2	AT	T	
L00	0	0	0	1	OFF
L10	1	0	1	0	ON
L01	0	1	1	0	ON
L11	1	1	0	1	OFF

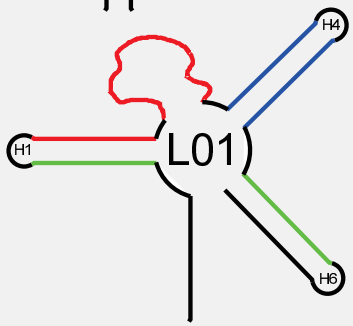


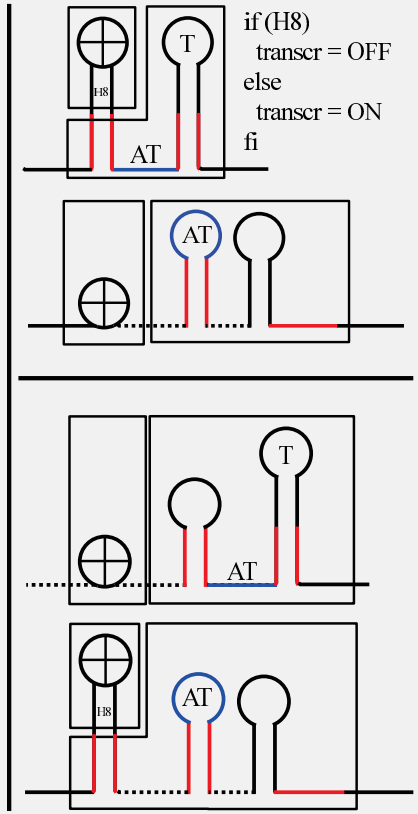




avoid this state

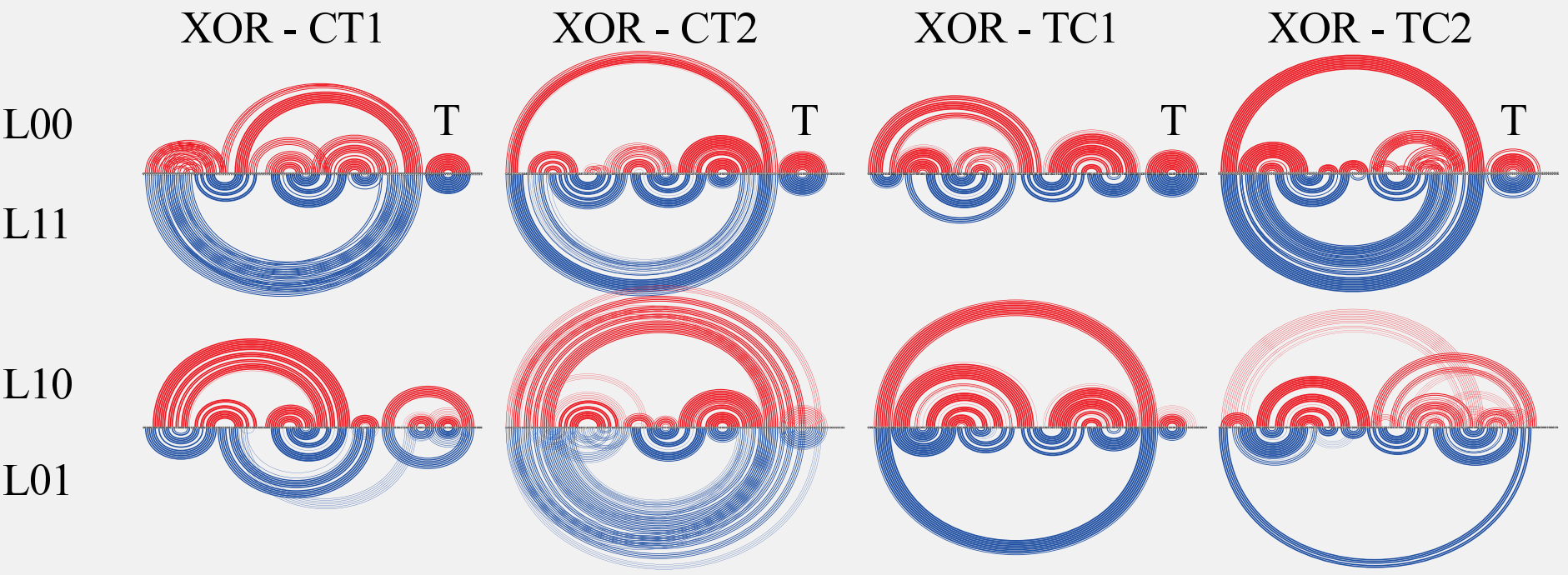




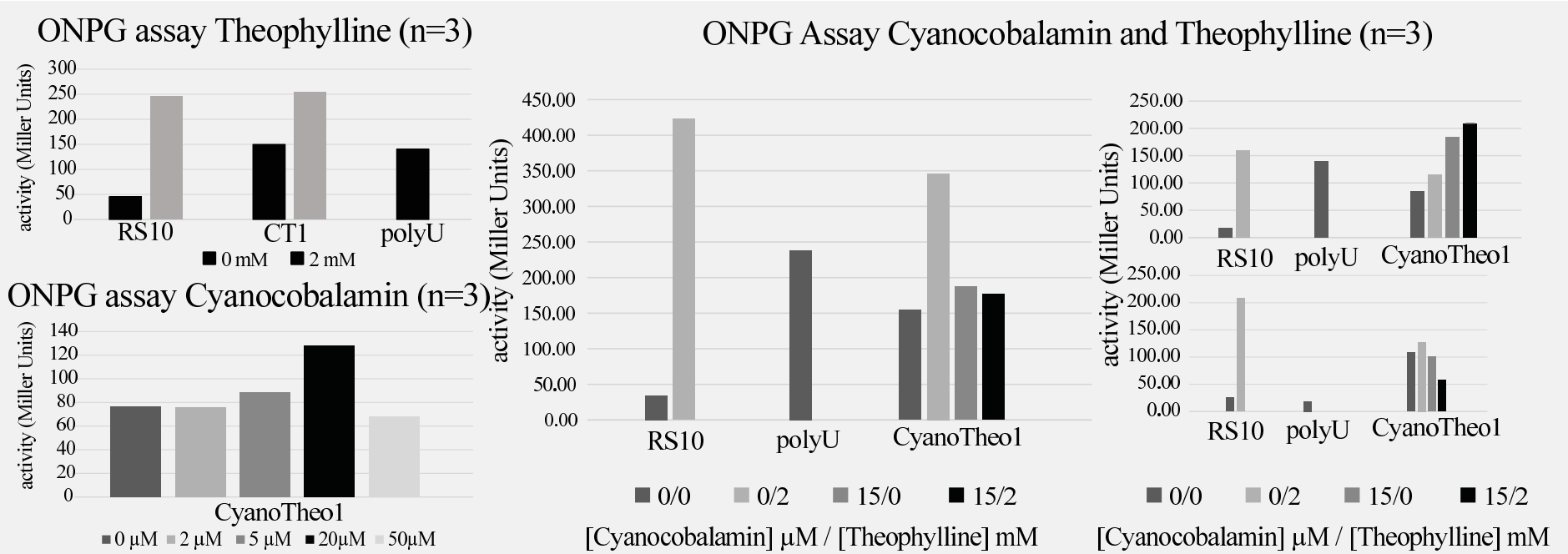


Designing an Integrated XOR Switch

4 computational designs for the cobalamin and theophyllin aptamers

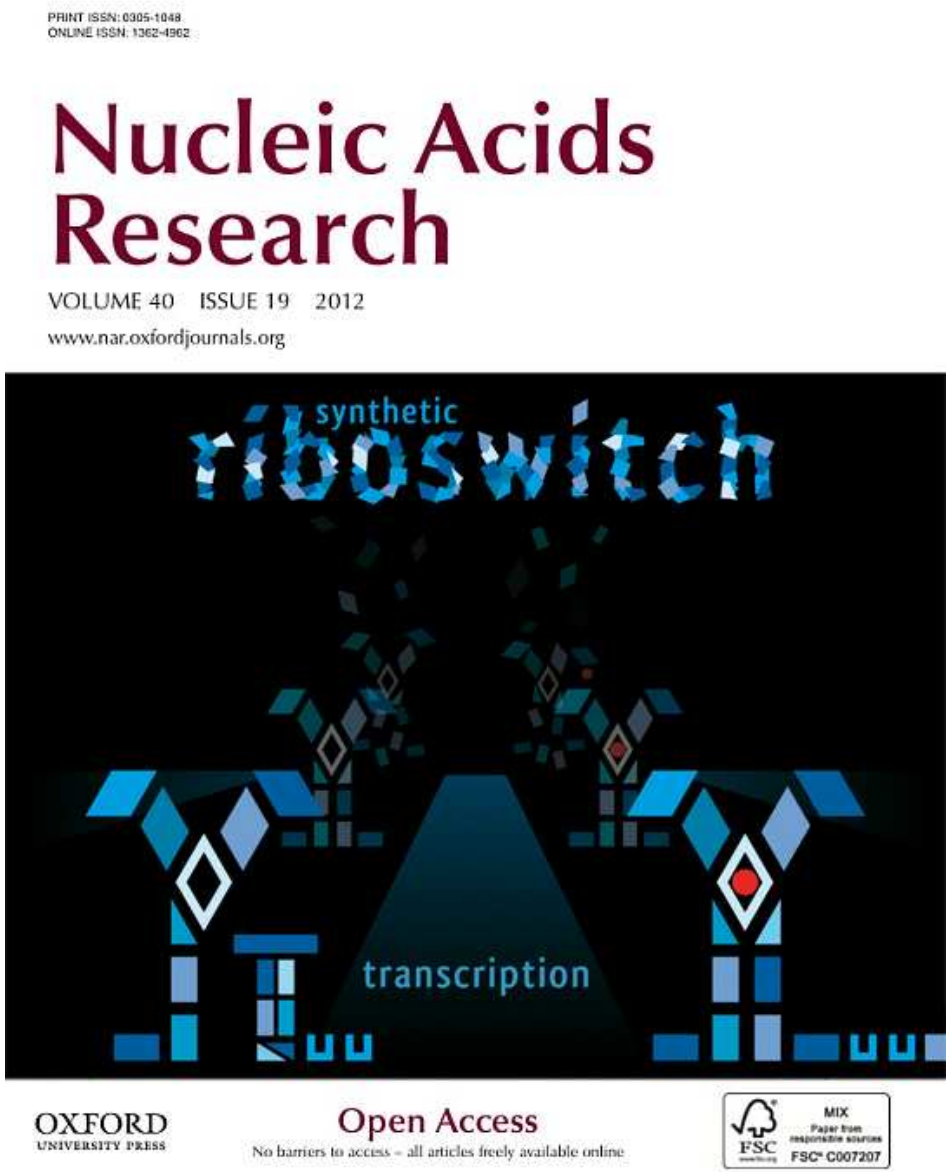


Initial experimental tests



... are promising ... but a lot of optimization will still be necessary

Acknowledgements



Funding: DFG

Collaborators:

Ronny Lorenz

Mörl Lab (Leipzig):

Manja Wachsmuth

Nadine Weissheimer

Hofacker Lab (Vienna):

Sven Findeiß

Christoph Flamm

Christian Höner zu Siederdissen