Design of Multi-Stable Nucleic Acid Sequences

Ingrid G. Abfalter[1], Christoph Flamm[1], Peter F. Stadler[1,2]

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien,
Währingerstraße 17, A-1090 Wien, Austria
Phone: ++43 1 4277 52731; Fax: ++43 1 4277 52793;
Email: ingrid@tbi.univie.ac.at.
[2]Bioinformatik, Institut für Informatik, Universität Leipzig,
Kreuzstraße 7b, D-04103 Leipzig, Germany

RNA molecules that can fold into two or more predefined alternative metastable structures can be designed rationally. We outline an algorithm for this task that reduces the problem to vertex coloring the union of all prescribed outerplanar secondary structure graphs. Starting from an ear decomposition of this composite graph colorings are produced by a dynamic programming procedure. Sequences can then be optimized for particular properties by means of standard optimization heuristics.

1. Molecular Switches

RNAs play a central role within living cells performing a variety of tasks [2]. The function of a biopolymer is predominantly determinded by its three-dimensional structure. The folding process of a single-stranded nucleic acid-molecule is of a hierarchic nature: Initially, stable secondary structure elements are formed which then act as a scaffold for tertiary contacts that finally determine the assembly of the native structure. The secondary structure of RNA, see Fig. 1 for different representations, contributes most to the free energy of the tertiary structure, hence its computation is used as a simplified model of the real nucleic acid-structure [6].

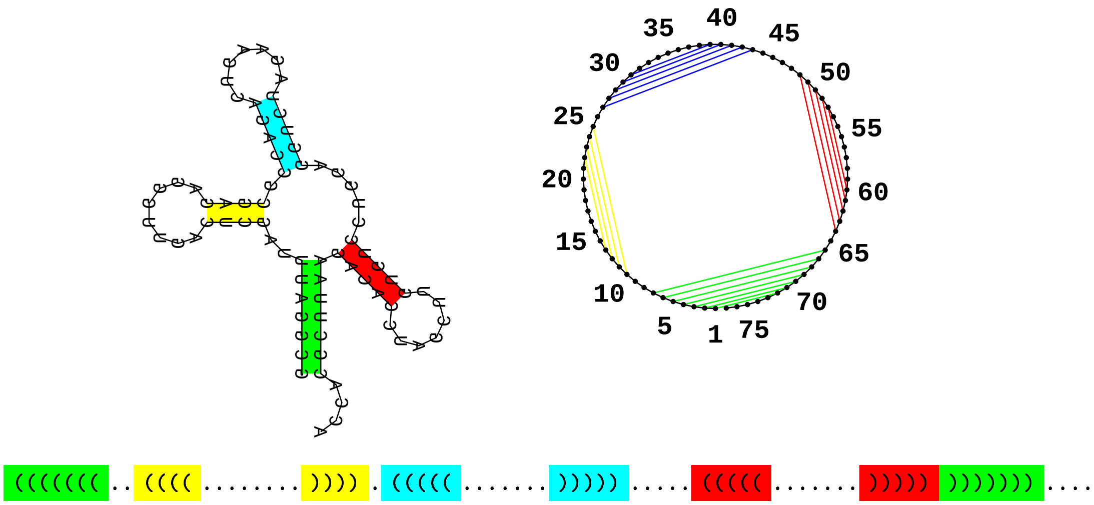

Figure 1. RNA secondary structure: (left) conventional representation. (right) circle representation. (below) bracket-dot-representation.

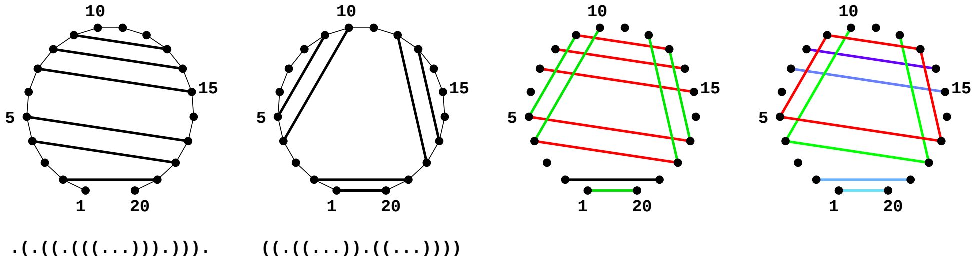

Figure 2. Dependency graph Ψ. (left) Circle representation of secondary structures 1 and 2. (middle) The dependency graph is constructed by super-imposing the circle representations of the two structures. Edges that can only be found in structure 1 are green, those only in structure 2 red, edges contained in both structures are black. (right) Paths are coloured blue and green, cycles red [3].

Nucleic acid molecules may have alternative non-native conformations with energy levels comparable to the ground state and high energy barriers that separate them. These meta-stable conformations can fulfill functions completely different from that of the native structure. In nature this important feature of RNA is used to implement *molecular switches* that regulate a variety of biological processes, see e.g. [1]. Here we describe a computational approach for the rational design of multi-stable nucleic acid-switches with two or more pre-defined secondary structures.

2. The Computer Model

Our approach is based on the following theorem that characterizes the realizability of a collection of M distinct secondary structures by a single RNA sequence. Let $\mathcal{A}$ denote the alphabet of monomers and let $\mathcal{B} \subset \mathcal{A} \times \mathcal{A}$ be the set of allowed basepairs. In particular, we have the alphabet of ribonucleotides $\mathcal{A} = \{A, U, G, C\}$ and the alphabet of RNA base pairs $\mathcal{B} = \{AU, UA, GC, CG, GU, UG\}$. For a secondary structure Θ (described as the set of base paired sequence positions) we write $\mathbf{C}[\Theta]$ for the set of all sequences that are compatible with Θ in the sense that every basepair $(i, j) \in \Theta$ is realized by a pair $(x_i, x_j) \in \mathcal{B}$ of pairing nucleotides.

The *dependency graph* Ψ of a collection of secondary structures $\{\Theta_i\}$ with n nucleotides consists of n vertices and edges connecting k with l if and only if (k, l) is a basepair in at least one of the secondary structures Θ_i, see Fig. 2. With this terminology we showed in [3]:

Theorem 1. *(Generalized Intersection Theorem) Suppose $\mathcal{B} \subseteq \mathcal{A} \times \mathcal{A}$ contains at least one symmetric base-pair then:*

(1) $\boldsymbol{C}[\Theta_1] \cap \boldsymbol{C}[\Theta_2] \cap \cdots \cap \boldsymbol{C}[\Theta_k] \neq \emptyset$ *if the dependency graph* $\Psi = \Theta_1 \cup \Theta_2 ... \cup \Theta_k$ *is bipartite.*

(2) *There are* $\displaystyle\prod_{\text{components } \psi \text{ of } \Psi} F(\psi)$ *sequences in* $\displaystyle\bigcap_j \boldsymbol{C}[\Theta_j]$.

(3) *For the biophysical alphabet holds:* $\bigcap_j \boldsymbol{C}[\Theta_j] \neq \emptyset$ *if and only if* Ψ *is a bipartite graph.*

The design algorithm uses this result to decide whether a sequence that concurrently forms all structures $\{\Theta_i\}$ exists. If so, we uniformly sample from these sequences and use a heuristic optimization procedure, in the simplest case an adaptive walk, to obtain compatible sequences that optimize a cost function Ξ that encodes additional constraints, for instance the design goal to have similar energies on all prescribed structures Θ_i.

In more detail, the algorithm consists of the following steps:

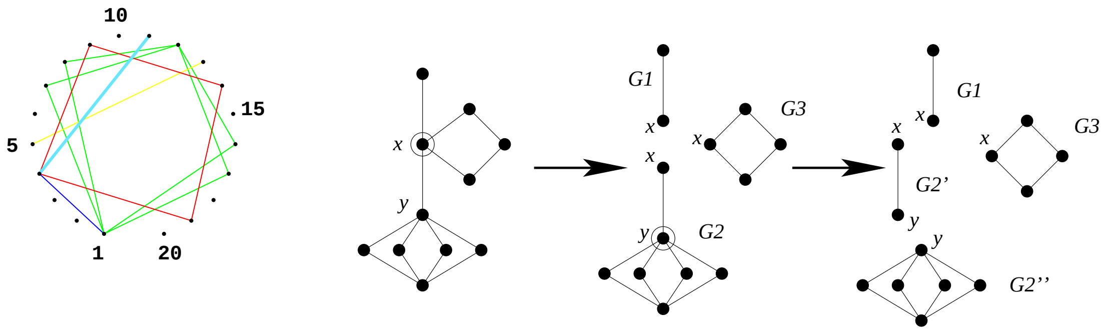

Figure 3. The dependency graph Ψ of more than two structures may consist of more complicated connected components that need further decomposition. The left part shows two connected components, the complex one is depicted in blue, green and red according to the biconnected components it is later decomposed into. The second connected component conists only of a path (yellow). The complex component (right) is decomposed at the cut-vertices x, y into two paths of the length 1, a circle of the length 4 and a block.

(1) INPUT: Predefine a set of secondary structures.
(2) Draw the dependency graph.
(3) Test for the bipartite property of the graph Ψ using a simple breadth-first-search coloring algorithm. Stop if Ψ is not bipartite.
(4) Decompose the graph into its connected components, then further into the biconnected components and finally decompose also the blocks by Whitney's Ear Decomposition.
(5) Count the number of compatible sequences.
(6) Generate sequences with uniform distribution on the set of compatible sequences.
(7) Optimize the sequence according to an energy criteria.
(8) OUTPUT: Optimized nucleic acid sequence compatible with all predefined structures.

Bipartiteness is tested using a simple breadth-first-search coloring algorithm.

In order to be able to design sequences with a uniform distribution we have to count the number of sequences compatible with a set of structures. According to assertion 2 of the Intersection Theorem we count them by splitting the graph into its connected components. Each connected component is computed independently. All their combinations of possible assignments are taken into account at the very last step of the computation of the cardinality of the intersection.

We use a depth-first-search algorithm to find *cut-vertices* and decompose the connected components into their biconnected components. The classes of biconnected components are paths, cycles and complex blocks. The blocks are further decomposed by *Whitney's Ear-Decomposition* [5], see Fig. 4.

3. Sequence Design as Graph Coloring

The computation of all possible sequence assignments of the connected components is nontrivial, since we have to satisfy the constraint that cut-vertices are occupied by a given nucleotide when we concatenate the partial sequences from the individual biconnected components. Furthermore, the attachment points of the ears must be assigned with the same bases in each concatenation step of the blocks. Further progress can be made by looking at the design problem a bit more abstractly.

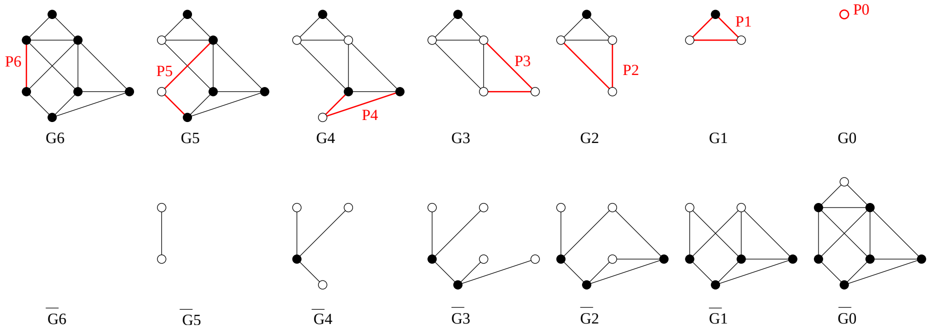

Figure 4. Graphs associated with an ear-decomposition: (top) Ear-decomposition of a block: In each step from G_6 to G_0 a path (ear) is removed until a central cycle is left. (bottom) The corresponding $\overline{G}_k$ of each step is shown. The attachment points of the ears are depicted by unfilled vertices.

A sequence that is compatible with all secondary structures can be viewed as a coloring of the vertices of Ψ such that adjacent vertices have colors $(a,b) \in \mathcal{B}$. In this section we briefly outline how this problem can be solved by a dynamic programming algorithm.

A vertex coloring of a graph G is a mapping $\mathbf{c}$ of the vertices $V(G)$ onto a set of colors. The important observation is that colorings can be obtained by putting together partial colorings: Let H be a subgraph of G and consider $U \subseteq W \subseteq V(H)$. The partial coloring $\mathbf{c}_U$ of H is simply a map from U into the color set. Let Ω be an abstract evaluation function, e.g. the number of "legal" colorings or the list of all colorings. Then we can write

$$\begin{aligned} \Omega(V(H),\mathbf{c}_U) &= \bigvee_{\mathbf{c}_{W\setminus U}} \Omega(V(H),\mathbf{c}_U \circ \mathbf{c}_{W\setminus U}) \qquad \text{for all } U \subseteq W \\ \Omega(V(H),\mathbf{c}_U) &= \Omega(V(H_1)\cup U,\mathbf{c}_U) \circ \Omega(V(H_2)\cup U,\mathbf{c}_U) \\ &\qquad \text{for all } H_1, H_2 \subseteq H \text{ and all } U \text{ such that } V(H_1 \cap H_2) = U \end{aligned} \tag{1}$$

$\vee$ and $\wedge$ are associative and commutative operators. In our example of counting conflict free colorings $\vee$ is addition $\wedge$ is multiplication.

Graph coloring is a well known NP-complete problem [4]. Of course our approach cannot overcome this in general. We can, however, search for a decomposition of G that allows us to apply equ.(1) with acceptable resource requirements. To this end we consider an ear decomposition $\mathcal{E} = (P_0, P_1, \dots)$ of G into paths and the associated series of subgraphs of G, Fig. 4 given by

$$G_k = \bigcup_{i=0}^{k} P_i \qquad \overline{G}_k = \bigcup_{i=k+1}^{\mu} P_i \qquad A_k = G_k \cap \overline{G}_k \tag{2}$$

We observe that G_k is biconnected for all $k > 0$. By definition $G_0 = P_0$, $G_1 = P_1$, $G_\mu = G$, $\overline{G}_0 = G$ and $\overline{G}_\mu = \varnothing$, the empty graph. We have therefore

$$\overline{G}_k = P_{k+1} \cup \overline{G}_{k+1} \tag{3}$$

The graphs A_k are disconnected and consist of the *attachment points* of $\overline{G}_k$ on G_k. Starting from the outer-most paths $\overline{G}_{\mu-1}$ and proceeding inwards until we reach $\overline{G}_0 = G$ we evaluate the number (or list) of colorings given fixed colors at the attachment points of $\overline{G}_k$. Explicitly, we have

$$\Omega(\overline{G}_k;\mathbf{c}_{A_k}) = \bigvee_{\mathbf{c}_{A_{k+1}\setminus A_k}} \left[\Omega(\overline{G}_{k+1};\mathbf{c}_{A_{k+1}\setminus A_k} \circ \mathbf{c}_{A_{k+1}\cap A_k}) \wedge \Omega(P_{k+1};\mathbf{c}_{A_{k+1}\setminus A_k} \circ \mathbf{c}_{A'_k}) \right] \tag{4}$$

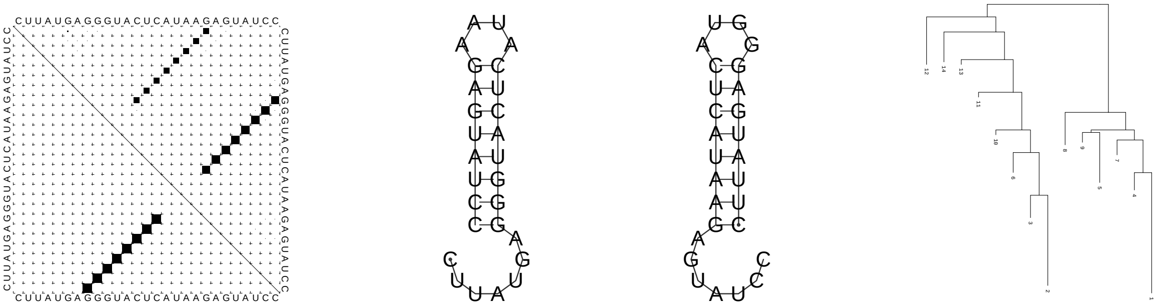

Figure 5. Example of a bistable switch molecule. From left to right; Dotplot: both structures have approximately the same statistical weight within the thermodynamic equilibrium; MFE structure; Metastable structure; note that the two structures have no base pair in common; Tree of local minima: Shown are the two (meta-)stable conformations, that are seperated by an energy barrier of ~ 11.2kcal/mol.

Here A'_k denotes the end-points of P_{k+1}. $A_{k+1} \setminus A_k$ is the set of the attachment points of $\overline{G}_{k+1}$ that are buried in the interior of P_{k+1} and play no role in further steps. The path P_{k+1} is subdivided by the interior attachment points into $|A_{k+1} \setminus A_k| + 1$ sub-paths for which the weights can be computed recursively. We need to evaluate all possible colourings of the interior attachement vertices for all possible colorings of the exterior attachement vertices $(x \in A_k)$. The performance of this way of coloring the graph is therefore determined by the maximal number of attachment vertices for $0 \leq k < \mu$.

4. Optimization

Uniform samples of coloring (i.e., sequences that are compatible with all secondary structures) can be obtained by means of a standard backtracking procedure when the numbers of coloring with given colors on the vertices A_k are tabulated for each k.

These sequences are used to initialize optimization heuristics, such as Adaptive Walks, that search the set of compatible sequences for those that optimize a cost-function Ξ. The function Ξ encapsulates desired properties of the molecule such as (nearly) equal energies for all prescribed secondary structures and constraints on the energy barriers between these metastable states. An example of such a designed RNA switch is shown in Fig. 5.

Acknowlegements. This work is supported in part by the *DFG* Bioinformatics Initiative and by a scholarship of the *Österreichische Akademie der Wissenschaften* to I.A.

References

[1] T. Baumstark, A. R. Schroder, and D. Riesner. Viroid processing: Switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. *EMBO J.*, 16:599–610, 1997.

[2] T. R. Cech and B. L. Bass. Biological catalysis by RNA. *Annu. Rev. Biochem.*, 55:599–630, 1986.

[3] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multi-stable RNA molecules. *RNA*, 7:254–265, 2001.

[4] T. R. Jensen and B. Toft. *Graph Coloring Problems.* John Wiley & Sons, New York, 1994.

[5] H. Whitney. Non-separable and planar graphs. *Trans. Am. Math. Soc.*, 34:339–362, 1932.

[6] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.