Design of Multi-Stable RNA Molecules

Christoph Flamm[a], Ivo L. Hofacker[a,*], Sebastian Maurer-Stroh[a], Peter F. Stadler[a,b], Martin Zehl[a]

[a]Institut für Theoretische Chemie und Molekulare Strukturbiologie
Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52742 Fax: **43 1 4277 52793

[b]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
Phone: (505) 984 8800 Fax: (505) 982 0565

[*]Address for correspondence:
Ivo L. Hofacker
Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52738 Fax: **43 1 4277 52793 Email: ivo@tbi.univie.ac.at

Abstract

We show that the problem of designing RNA sequences that can fold into multiple stable secondary structures can be transformed into a combinatorial optimization problem that can be solved by means of simple heuristics. Hence it is feasible to design RNA switches with prescribed structural alternatives. We discuss the theoretical background and present an efficient tool that allows the design of various types of switches. We argue that both the general properties of the sequence-structure map of RNA secondary structures and the ease with which our design tool finds bi-stable RNAs, strongly indicates that RNA switches are easily accessible in evolution. Thus conformational switches are yet another function for which RNA can be employed.

Keywords: RNA switches, sequence design, RNA folding, RNA secondary structure prediction

1. Introduction

Over the last ten years it became evident that RNA plays a central role within living cells and actively performs a variety of tasks in many different biological contexts. These functions are often intimately related to the three-dimensional structure of the molecules. The process of RNA folding is thought to be of a hierarchical nature (Brion and Westhof, 1997; Tinoco Jr and Bustamante, 1999). Stable RNA secondary structure elements fold fast, on a microsecond time scale, and determine the subsequent assembly of the tertiary fold. The energies involved in secondary structure formation are large compared to those of the tertiary contacts, hence the basic properties of the conformational energy landscape of an RNA molecule can be understood at the level of secondary structures (Flamm et al., 2000). One of its most important features is the fact that non-native conformations can have energies comparable to the ground state and they can be separated from the native state by very high energy barriers. Stable alternative conformations have been observed experimentally for a variety of RNA molecules (Emerick and Woodson, 1993; Fresco et al., 1966; Hawkins et al., 1977).

When the formation of non-native-like secondary structure happens at early stages of the folding process, major structural reorganizations of the folding chain become necessary to reach the native state. This involves breaking a large number of base pairs and hence may be very costly in terms of energy. Mis-folded conformations therefore often constitute folding traps that can dramatically slow down the RNA folding process (Pan et al., 1998, 1997; Treiber et al., 1998).

Alternative conformations of the same RNA can determine completely different functions (Baumstark et al., 1997; Perrotta and Been, 1998). SV11, for instance, is a relatively small molecule that is replicated by Qβ replicase (Biebricher et al., 1982; Biebricher and Luce, 1992). It exists in two major conformations, a meta-stable multi-component structure and a rod-like conformation, constituting the native state, separated by a huge energy barrier. While the meta-stable conformation is a template for Qβ replicase, the ground state is not. By melting and rapid quenching the molecule can be re-converted from the inactive stable to the active meta-stable form (Zamora et al., 1995). The capability of RNA molecules to form multiple (meta)-stable conformations with different function is used in nature to implement so called *molecular switches* that regulate and control the flow of a number of biological processes.

Alternative foldings are probably involved in the viroid replication process (Gultyaev et al., 1998; Hecker et al., 1988; Loss et al., 1991). The terminator and anti-terminator, two alternative RNA hairpins, regulate gene expression in *E. coli* and *B. subtilis* by attenuation (Babitzke and Yanofsky, 1993; Fayat et al., 1983; Putzer et al., 1992). A complex series of mRNA rearrangements regulates the plasmid maintenance in the *hok/soc* system of plasmid R1 (Nagel et al., 1999). The binding of the elongation factors EF-Tu and EF-G, respectively, to alternative conformations of 28S rRNA during the elongation cycle in protein biosynthesis has been proposed in (Wool et al., 1992). There is convincing evidence that the codon-anticodon arrangement and the proper recognition of the tRNA at the ribosomal A site are controlled by an RNA switch (Lodmell and Dahlberg, 1997; von Ahsen, 1998). Artificial RNA switches have been designed as well. For instance, Soukup and Breaker (1999) have engineered a molecule that is triggered by ligand binding using a switching mechanism similar the one proposed for the ribosomal `A` site.

A particularly impressive example has been described by Schultes and Bartel (2000), who designed a sequence that can satisfy the base-pairing requirements of both the hepatitis delta virus self-cleaving ribozyme and an artificially selected self-ligating ribozyme, which have no base pairs in common. This *"intersection sequence"* displays catalytic activity for both cleavage and ligation reactions. Structural probing demonstrated that the new molecule indeed adopts the tertiary folds of both ribozymes.

Giegerich et al. (1999) presented a software tool that can be used to investigate the possibility of structural switching in a given sequence. Their program `paRNAss` clusters suboptimal structures by structural similarity and energy barriers. An RNA switch is then a sequence where clusters of suboptimal structure are clearly separated from each other with a significant energy barrier between them. In this contribution we consider the complementary problem, namely the design of a switching sequence when the structural alternatives are given.

We show that bi-stable, and more generally, multi-stable RNA molecules with a variety of additional properties can be found rather easily. We present a computational method that allows the design of RNA sequences that fold into prescribed alternative conformations. In the case of bi-stable molecules the two alternative conformations can be chosen at will. The freedom of choice is limited only when three or more alternative structures are required. We give examples of small RNAs that are designed to change their preferred conformation in a desired temperature range, and that have energy barriers of a desired height. In a more sophisticated application an artificial analog of SV11 RNA is designed in mere ten minutes using a `Perl` program based upon the `Vienna RNA Package` (Hofacker et al., 1994). The ease with which RNA sequences with properties of switches can be found suggests that this mechanism is readily available in evolution. The known features of the sequence-structure map of RNA secondary structure folding can be used to derive the same conclusion.

2. Theory

2.1. RNA Structures and Compatible Sequences. An RNA secondary structure can be understood as a set Ω of base pairs. For simplicity we assume that the sequence positions are numbered consecutively from 1 to n, the set of unpaired positions will be denoted by Υ.

Base pairs in secondary structures satisfy two constraints:

- A base may participate in at most one base pair.
- Base pairs must not cross, i.e., we cannot have two base pairs (i, j) and (k, l) with $i < k < j < l$. This condition excludes pseudo-knots.

The base pairing rules of RNA allow only 6 types of base pairs out the 16 possible combinations. Given a secondary structure Ω, this restricts the choice of sequences that are *compatible* with Ω, since for each pair $\{i, j\} \in \Omega$ and each compatible sequence x, $x_i x_j$ must be either one of the four Watson-Crick pairs AU, UA, GC, CG, or one of the two "wobble" pairs GU and UG. Much of the discussion below remains valid for arbitrary alphabets $\mathcal{A}$ of nucleic acids and general pairing rules $\mathcal{B}$. For the biophysical alphabet we have, of course,

$$\mathcal{A} = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{U}\} \qquad \mathcal{B} = \{\mathsf{AU}, \mathsf{UA}, \mathsf{CG}, \mathsf{GC}, \mathsf{GU}, \mathsf{UG}\} \tag{1}$$

We denote the set of all sequences that are compatible with a structure Ω by $\mathbf{C}[\Omega]$. Clearly, for each $i \in \Upsilon$ we may choose an arbitrary letter from the nucleic acid alphabet, while for each pair we may choose one of the possible base pairs. Using the notation $|X|$ for the number of elements in the set X (e.g. $|\Omega|$ denotes the number of base pairs), we have

$$|\mathbf{C}[\Omega]| = |\mathcal{A}|^{|\Upsilon|}|\mathcal{B}|^{|\Omega|} \tag{2}$$

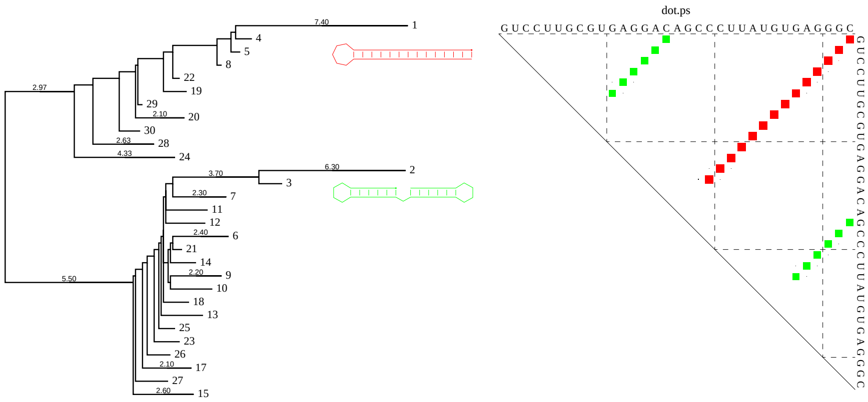

Figure 1. Equilibrium base pair probabilities (right) and energy barriers between the 30 lowest local minima (left) for the designed sequence GUCCUUGCGUGAGGACAGCC-CUUAUGUGAGGGC. The sequence has two dominating conformations, a rod like one (red) and a two component structure (green), all other possible base pairs have very low probability. The two conformations have energies of -17.1 and -17.0 kcal/mol, respectively, and are separated by an energy barrier of 17.2kcal/mol (indicated by the height of the saddle point connecting the two states in the tree).

sequences that are compatible with the secondary structure Ω. For the biophysical alphabet we have explicitly $4^{|\Upsilon|}6^{|\Omega|}$, while for the restricted alphabet $\{\mathsf{G},\mathsf{C}\}$ with pairing rule $\{\mathsf{GC},\mathsf{CG}\}$ we have $2^{|\Upsilon|+|\Omega|} = 2^{n-|\Omega|}$ compatible sequences.

2.2. Design as an Optimization Problem. The energy of an RNA sequence in a particular secondary structure can be evaluated in a "nearest-neighbor" model, for which most energy parameters have been carefully measured (Jaeger et al., 1989; Walter et al., 1994; Mathews et al., 1999). Within this energy model the *RNA folding problem* of finding the (near) optimal secondary structures of a given sequence can be solved efficiently by means of dynamic programming (Zuker and Stiegler, 1981; Zuker, 1989; McCaskill, 1990). We use the implementation `Vienna RNA Package`, version 1.3.1, to evaluate the "folding function" Φ, that is, to compute the secondary structure $\Phi(x)$ of a given sequence x.

The structural dissimilarity $D(\Omega_1, \Omega_2)$ of two RNA secondary structures Ω_1, Ω_2 can be quantified by a variety of distance measures, see e.g. (Shapiro and Zhang, 1990; Hofacker et al., 1994; Reidys and Stadler, 1996). In the simplest case we count the number of base pairs that are either in Ω_1 or in Ω_2 but not in both. In set notation this is the symmetric difference metric

$$D(\Omega_1, \Omega_2) = \big|(\Omega_1 \cup \Omega_2) \setminus (\Omega_1 \cap \Omega_2)\big|. \tag{3}$$

Sequence x folds into structure Ω, i.e., $\Phi(x) = \Omega$, if and only if $D(\Omega, \Phi(x)) = 0$. Hence, the *inverse folding problem* of finding a sequence x that folds into a prescribed secondary structure Ω can be rephrased as the following combinatorial optimization problem:

$$\text{Find } x \in \mathbf{C}[\Omega] \text{ such that } \Xi(x) = D(\Omega, \Phi(x)) \to \min\,. \tag{4}$$

The program `RNAinverse`[1] is based on this idea (Hofacker et al., 1994).

It is straight forward to modify equ.(4) to search, for instance, for sequences in which the ground state is much more stable than any structural alternative (Hofacker et al., 1994): Let $E(x;\Omega)$ be the energy of structure Ω for sequence x, and let $G(x)$ be the ensemble free energy of sequence x, which can be computed

[1] A web interface for designing sequences with `RNAinverse` can be found at `http://www.tbi.univie.ac.at/cgi-bin/RNAinverse.cgi`.

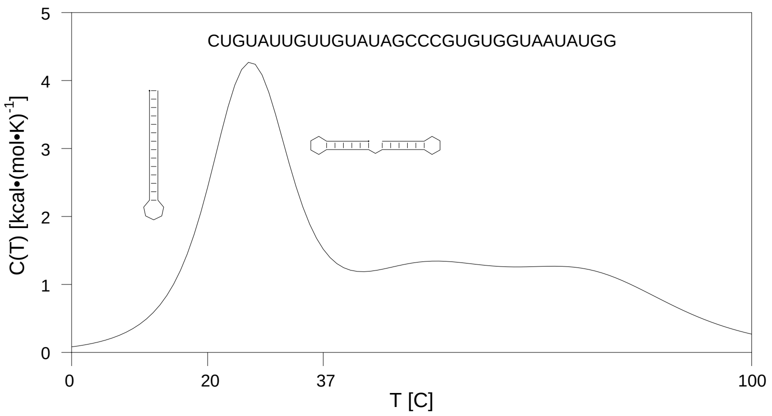

Figure 2. Specific heat of a designed RNA sequence, calculated using the `RNAheat` program of the `ViennaRNA` package. The sequence switches from a V-shaped to a rod-like structure between 20° and 37° Celsius.

by McCaskill's (McCaskill, 1990) algorithm. Sequences with the desired property minimize

$$\Xi(x) = E(x,\Omega) - G(x) = -RT \ln p \tag{5}$$

where p is the probability of structure Ω in the Boltzmann ensemble of sequence x.

We found that the combinatorial optimization problems (4,5) is easily solvable by means of adaptive walks. Starting from a randomly chosen initial sequence x_0 we produce mutants by exchanging a nucleotide at the unpaired positions Υ or by replacing one of the six pairing combinations by another one in a pair in Ω. A mutant is accepted if the cost function $\Xi(x)$ decreases.

It is the purpose of this contribution to demonstrate that the inverse folding approach can be generalized to more complicated design problems involving two or more structural constraints on the sequences. Below we give a few examples of design schemes.

Example 1. Given two distinct secondary structures Ω_1 and Ω_2 (with the same sequence length n) we want to design a sequence x that has a Boltzmann ensemble consisting almost exclusively of Ω_1 and Ω_2 such that these two structural alternatives occur with roughly equal frequencies. A suitable cost function for this design problem is

$$\Xi(x) = E(x,\Omega_1) + E(x,\Omega_2) - 2G(x) + \xi\left(E(x,\Omega_1) - E(x,\Omega_2)\right)^2 \tag{6}$$

where $\xi > 0$ is a constant that weights the relative importance of thermodynamic stability and equal frequencies. An example is shown in Figure 1.

Example 2. A "switch" that changes its preferred structure from Ω_1 to Ω_2 when the temperature changes from T_1 to T_2 can be obtained with a cost function such as the following

$$\begin{aligned}\Xi(x) = &\left(E_{T_1}(x,\Omega_1) - G_{T_1}(x)\right) + \left(E_{T_2}(x,\Omega_2) - G_{T_2}(x)\right)\\ &+ \xi\left\{\left(E_{T_1}(x,\Omega_1) - E_{T_1}(x,\Omega_2)\right) + \left(E_{T_2}(x,\Omega_2) - E_{T_2}(x,\Omega_1)\right)\right\}\end{aligned} \tag{7}$$

The first term favors Ω_1 at temperature T_1 and Ω_2 at T_2. The second term explicitly penalizes the wrong structure relative to the correct one. Such a design is shown in figure 2.

The "design by optimization" approach is by no means limited to thermodynamic properties of the RNA molecule. Kinetic properties can be prescribed as well.

Example 3. Given two distinct secondary structures Ω_1 and Ω_2 we wish to design a sequence that has Ω_1 and Ω_2 as stable local energy minima with roughly equal

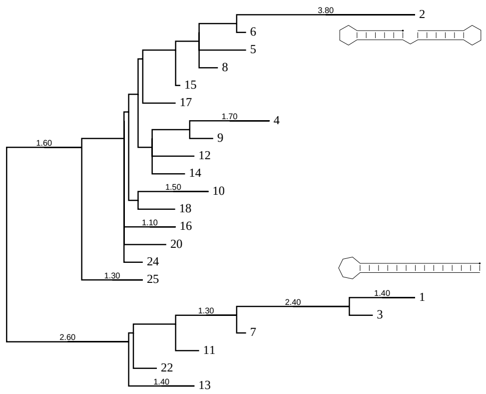

Figure 3. Barrier tree for the sequence GUGUUUGAGAGGAUAUGGCGUUUUUUUG-GAUGC. The sequence has the same two dominating conformations as the one in fig 1, but was designed to have a small energy barrier of about 8kcal/mol. The two conformations have the same energy of -7.7 kcal/mol and the exact energy barrier is 8.7 kcal/mol.

energy, and for which the energy barrier between these two minima is roughly ΔE. An appropriate cost function is for instance

$$\begin{aligned}\Xi(x) = E(x,\Omega_1) + E(x,\Omega_2) - 2G(x) + \xi\left(E(x,\Omega_1) - E(x,\Omega_2)\right)^2 \\ + \zeta\left(B(x,\Omega_1,\Omega_2)) - \Delta E\right)^2\end{aligned} \tag{8}$$

where $B(x,\Omega_1,\Omega_2)$ is the height of the energy barrier between the two structures and $\zeta > 0$ is a weighting factor. The computation of $B(x,\Omega_1,\Omega_2)$, which in itself is a non-trivial problem, is discussed in section 2.7. An example is shown in figure 3.

The cost functions (6,7,8) are only defined for sequences that are compatible with both Ω_1 and Ω_2. The optimization can — and should — therefore be restricted to the *intersection* $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_1]$. The structure of this set, which is crucial for the design algorithm, is described in detail in the following sections. In particular we need to solve two mathematical problems on our way: (1) how to *fairly* choose a starting point for the optimization procedure in the intersection, and (2) how two mutate the sequence such that (a) all mutants are compatible with both structures and (b) there is as little sequence bias as possible. If these conditions are satisfied, repeatedly running the algorithm will produce a fair sample of possible solutions.

2.3. The Intersection Theorem.

Theorem 1. *(Intersection Theorem) If the nucleic acid alphabet admits at least one type of complementary base pairs, then, for any two secondary structures Ω_1 and Ω_2 there exists at least one sequence that is compatible with both structures, in symbols*

$$\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2] \neq \emptyset \tag{9}$$

Proof. An abstract group-theoretical proof can be found in (Reidys et al., 1997). Here we give a different, purely combinatorial version.
Consider two secondary structures Ω_1 and Ω_2. To construct the *dependency graph* Ψ, we use the sequence positions $\{1,\ldots,n\}$ as vertices, and draw edges connecting i and j for each base pair (i,j) in Ω_1 and Ω_2.

Since sequence constraints can arise only from base pairs, i.e. edges in Ψ, it is clear that each connected component of Ψ is independent from all others. To construct a sequence in $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$ we can therefore assign each component separately. Since any vertex is incident with at most two edges, the connected components of Ψ are only paths, cycles and isolated vertices, see figure 4.

We may distinguish three types of positions and associated components in Ψ:

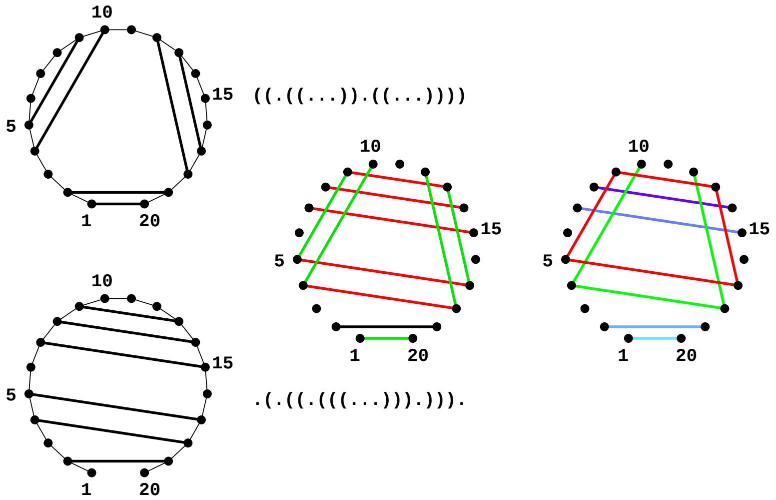

Figure 4. Construction of the dependency-graph Ψ of two secondary structures. L.h.s.: circle representations of the two structures; middle: The edge colored dependency graph Ψ can be viewed as the superposition of the circle representations. R.h.s.: The connected components of Ψ are shown in different colors.

(i) Positions that are unpaired in both structures form isolated vertices. For these we may select an arbitrary letter from $\mathcal{A}$.
(ii) Positions that are paired with the same pairing partner in both structures form paths of length 2. We may assign any one of the possible base pairs in $\mathcal{B}$ to such a path.
(iii) The remaining positions are paired differently in the two structures and can belong to cycles or paths. They are discussed below.

Let us color the graph Ψ such that each pair from Ω_1 (Ω_2) is drawn in red (green), while leaving pairs that occur in both structures black. Each sequence position in class (iii) is now incident with at most one red and one green edge. Furthermore, red and green edges alternate along paths and cycles. Cycles therefore must have even number of edges and vertices. If XY and YX are base pairs, we may associate the alternating sequence XYXYX... with the vertices of each path and cycle.

Independence of the cycles and paths of Ψ implies that there are indeed sequences that are compatible with both Ω_1 and Ω_2. □

The intersection theorem does not directly generalize to more than two sequences. However, using the idea of the edge-colored dependency graph we obtain the following

Theorem 2. [Generalized Intersection Theorem] *Suppose $\mathcal{B} \subseteq \mathcal{A} \times \mathcal{A}$ contains at least one symmetric pair, i.e.,* XY $\in \mathcal{B}$ *implies* YX $\in \mathcal{B}$*. Then*

(i) $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2] \cap \cdots \cap \mathbf{C}[\Omega_k] \neq \emptyset$ *if Ψ is bipartite.*
(ii) The number of sequences in $\bigcap_j \mathbf{C}[\Omega_j]$ *can be written in the form of equ.(10).*
(iii) For the biophysical alphabet (1) holds: $\bigcap_j \mathbf{C}[\Omega_j] \neq \emptyset$ *if* **and only if** *Ψ is a bipartite graph.*

Proof. The first part is easy to prove. If Ψ is bipartite than we assign X to one partition and Y to the other. The base pairs that are present in any one of the secondary structures Ω_j are the edges of Ψ. Hence to each edge we have a base pair XY.
Clearly, sequence positions that are not contained in the same connected component of Ψ are independent, hence equ.(10) is correct.
A graph is bipartite if and only if it contains no cycles of odd length. Hence we

have to show that no odd cycle can be realized by the biophysical RNA alphabet. Consider the graph

$$\mathsf{A} - \mathsf{U} - \mathsf{G} - \mathsf{C}$$

When producing a valid sequence of letters for a cycle C_k we have to follow the edges in this graph. Thus, if we start with a particular letter X, all other occurrences of the same letter X must appear after an even number of steps along the cycle. This includes encountering X after having gone around the cycle. Odd cycles therefore cannot be associated with a valid sequence and the theorem follows $\square$

2.4. The Size of the Intersection. The edge-colored graph Ψ introduced in the proof of the Intersection Theorem 1 can be used to enumerate the size of $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$. It will be convenient to add the pairs of Ω_{12} as paths of length 1 and the unpaired positions in Υ_{12} as isolated vertices to Ψ. With this definition we may write

$$\left|\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]\right| = \prod_{\text{components } \psi \text{ of } \Psi} F(\psi) \tag{10}$$

where $F(\psi)$ is the number of sequences that are compatible with a connected component ψ of Ψ. For an isolated vertex (unpaired base) $F(\{i\}) = |\mathcal{A}|$, the number of different nucleotides. For a base pair $F(\{i,j\}) = |\mathcal{B}|$, the number of possible base pairs. The values $F(\psi)$ for larger components depend on the details of the base pairing rules.

In the proof of the Intersection Theorem we have used that the components of the dependency graph can be only isolated points, paths, and cycles in the case of two structures. Let us write P_n and C_n for a path and cycle with n vertices, respectively. For $\mathcal{A} = \{\mathsf{G},\mathsf{C}\}$ and $\mathcal{B} = \{\mathsf{GC},\mathsf{CG}\}$ we have $F(P_n) = F(C_n) = 2$. For $\mathcal{A} = \{\mathsf{G},\mathsf{C},\mathsf{A},\mathsf{T}\}$ and $\mathcal{B} = \{\mathsf{GC},\mathsf{CG},\mathsf{AT},\mathsf{TA}\}$ we have $F(P_n) = F(C_n) = 4$, independent of n. In both cases the sequence along a path P_n or cycle C_n is uniquely determined by the first letter. In the case of the biophysical RNA alphabet

$$\mathcal{A} = \{\mathsf{G},\mathsf{C},\mathsf{A},\mathsf{U}\} \qquad \mathcal{B} = \{\mathsf{GC},\mathsf{CG},\mathsf{AU},\mathsf{UA},\mathsf{GU},\mathsf{UG}\} \tag{11}$$

we have a much more complicated situation because of the GU-pairs. We find

$$\begin{aligned} F(P_n) &= 2\big(\mathrm{Fib}(n) + \mathrm{Fib}(n+1)\big) = 2\mathrm{Fib}(n+2) \\ F(C_n) &= 2\big(\mathrm{Fib}(n-1) + \mathrm{Fib}(n+1)\big) \end{aligned} \tag{12}$$

where $\mathrm{Fib}(n)$ is the n-th Fibonacci number. For the derivation of these formulae we refer to the appendix.

Note that the size of the intersection is always large for the biophysical alphabet, which should facilitate the design problem. For other alphabets the intersection is small, if the dependency graph consists of few large components. In this case design of switching sequences will be infeasible.

2.5. Random Sequences in $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$. In order to avoid a bias towards particular sequence motifs we need to find sequences in $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$ that are "as random as possible". The combinatorial results in the previous section can be used in a straight forward way to generate sequences in $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$ with a uniform distribution.

Clearly, sequence positions in different connected components ψ of Ψ are independent. The problem hence reduces to generating sequences for a connected component ψ. From the recursions equ.(12, see appendix) it is clear that the probabilities $p_Q(k;\mathsf{X}|\mathsf{Y})$ of finding a particular letter X in the k-th position of a cycle or path depends only on the letter Y in the previous position and whether a cycle C_n or a path P_n is considered. We first note that

$$p_{Q_n}(k;\mathsf{G}|\mathsf{C}) = 1 \qquad p_{Q_n}(k;\mathsf{U}|\mathsf{A}) = 1 \tag{13}$$

where Q_n denotes either a cycle or a path of length n. Furthermore, $p_Q(k;\mathsf{X}|\mathsf{Y}) = 0$ if XY is not a valid base pair.

In the case of paths the situation is simple. For the first letter of a path we have

$$p_P(1;\mathsf{X}|\emptyset) = \frac{F(Q_n^{\mathsf{X}})}{F(Q_n)} \tag{14}$$

The recursions (16) immediately imply

$$\begin{aligned} p_P(k;\mathsf{G}|\mathsf{U}) = p_P(k;\mathsf{U}|\mathsf{G}) &= \frac{F(P_{n-k}^{\mathsf{G}})}{F(P_{n-k+1}^{\mathsf{U}})} = \frac{F_2^p(n-k)}{F_2^p(n-k+1)} = \\ &= \frac{\mathrm{Fib}(n-k+1)}{\mathrm{Fib}(n-k+2)} \\ p_P(k;\mathsf{A}|\mathsf{U}) = p_P(k;\mathsf{C}|\mathsf{G}) &= \frac{\mathrm{Fib}(n-k)}{\mathrm{Fib}(n-k+2)} \end{aligned} \tag{15}$$

for $2 \le k \le n$.

In the case of cycles only the initialization is different, since any even length path starting with G or U is also a valid cycle. Cycles starting with C or A can be constructed by appending a path of length $n-1$ starting with G or U, respectively. The procedure is summarized as algorithm 1 in the appendix.

2.6. Consistent Mutations. We distinguish two different types of "mutations" in cycles: (a) local mutations that conserve the purine-pyrimidine pattern and lead to sequences that have Hamming distances of at most 3, and (b) non-local mutations with exchange R$\leftrightarrow$Y at each position of the cycle.

Let us first consider local mutations. We select a position k in the cycle at random and mutate according to the rule G$\leftrightarrow$A or C$\leftrightarrow$U. Then we have to perform the required "repairs" in the cycle. For instance AUG$\rightarrow$ACG$\rightarrow$GCG. Note that the letter before A is necessarily U, hence no further repair in necessary. It is not hard to check that in the worst case a "repair" of the previous and the following position is necessary. The Hamming distances between local mutants are therefore never larger than 3. Furthermore, the "repairs" are obviously unique; i.e., there is a single mutant for each position in the cycle.

```
GC  GU  AU  UA  UG  CG
↓   ↓   ↓   ↓   ↓   ↓
AU  AU  GU  CG  CG  UG

GC  GU  AU  UA  UG  CG
 ↓   ↓   ↓   ↓   ↓   ↓
GU  GC  GC  UG  UA  UA
```

Figure 5. Mutations at consecutive positions of a cycle yield distinct mutants.

In order to see that this scheme leads to a uniform distribution on the intersection we observe that any cycle of length ℓ has exactly ℓ local neighbors, namely exactly one for each position. It remains to verify that the mutants obtained from changing different positions are indeed distinct. Since a mutation at position k affects at most the two neighboring positions it is sufficient to show that the mutants arising from mutations in adjacent positions are always distinct. Figure 5 lists all possible cases, showing that such mutants are indeed distinct.

Non-local mutations are generated by replacing a cycle with a new, randomly generated sequence as described in section 2.5. This is necessary since the local moves discussed above preserve the purine-pyrimidine pattern.

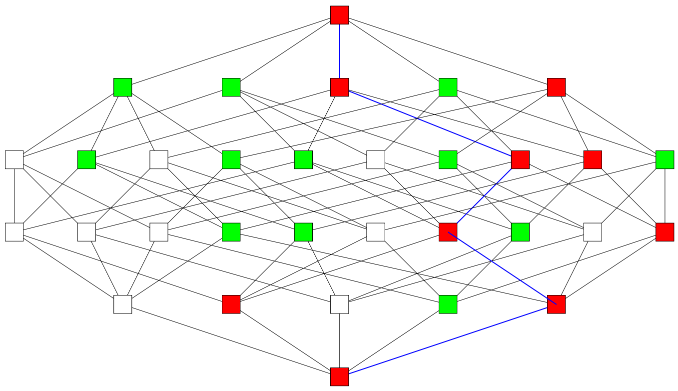

Figure 6. Direct paths connecting two structures with $d = 5$. With $m = 2$ only the colored structures are evaluated and only red structures are used to generate conformations in the next distance class.

2.7. Estimating Barrier Height. The energy barriers separating local minima are the most important factor influencing the folding kinetics of an RNA molecule (Flamm et al., 2000). For short sequences these barrier heights and the structures at the saddle points (transition states) can be determined exactly with the help of complete suboptimal folding (Wuchty et al., 1999) in the following manner:

We assign each suboptimal structure to a basin corresponding to the lowest local minimum that can be reached along a path that visits only structures with lower energy. The saddle point between two basins is then the lowest energy structure that has neighbors belonging to each of the two basins (Vertechi and Virasoro, 1989). The definition of neighbors and local minima of course depends on the choice of the move set, in the simplest case insertion and deletion of individual base pairs (Flamm et al., 2000).

In general the exact determination of barrier heights is too costly to be used in each evaluation of the cost function. Estimates, however, can be calculated relatively cheap. If we consider only opening and closing of single base pairs as the move set by which secondary structures can re-fold, then the base pair distance $D(\Omega_1, \Omega_2)$, (equ. 3) gives us the minimum number of moves needed to transform Ω_1 into Ω_2. Morgan and Higgs (1998) have introduced the notion of *direct paths*, which consist of exactly $d = D(\Omega_1, \Omega_2)$ moves. Because evaluation of all possible direct paths is still too costly, they used a simple greedy algorithm to derive upper bounds on the height of a barrier.

To improve the greedy estimate, we use the follwoing procedure: Starting at the first structures we generate all conformations that are one step closer to the second structure. Of the resulting partial paths we keep the best m, these candidates are then extended by one step in the next iteration. Thus, we perform a breadth first search of the possible paths and bound the search by keeping only the best m candidates at each step, see figure 6.

If we already know an upper bound for the barrier height we can reduce the search space further by terminating each path as soon as its energy becomes higher than our bound. It can therefore be useful to repeat the above procedure a few times with increasing values of m. Note that for $m = 1$ we recover the greedy algorithm of Morgan and Higgs (1998).

2.8. Implementation. The procedure defined above was implemented using the scripting language `Perl`. This allows easy modifications, such as variations of the cost functions. On the other hand, the program makes use of the `C` routines of the `Vienna RNA Package` (via a `Perl` extension module), and thus has access to fast

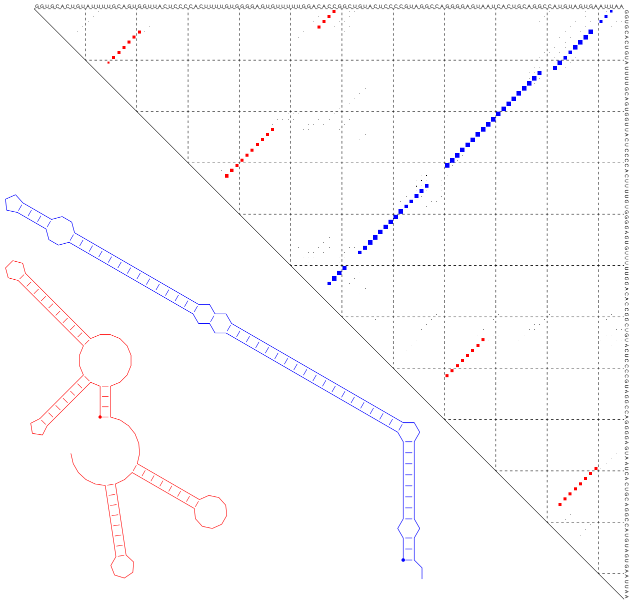

Figure 7. Equilibrium base pair probabilities for a sequence designed to have the same two meta-stable conformations as the SV11 sequence: a rod like one (blue) and a three component structure (red). Both states have an energy of -56.2kcal/mol.

routines for computation of RNA secondary structures and base pairing probabilities.

For the small example of figure 1 the program takes about 1.3s per sequence on a 333 MHz Pentium II. While manual design is not too hard for such an example, the optimization procedure yields significantly better sequences. For the 115nt SV11 example, figure 7, it designs one sequence in about 10min.

3. Discussion

In the introduction we have briefly reviewed the experimental evidence for a functional role of bi-stable RNAs in a variety of different contexts. We have shown how the RNA design problem can be transformed into an easily solvable combinatorial optimization problem on the set of RNA sequences that are compatible with all desired structures. The *intersection theorem* guarantees that for any two prescribed secondary structures there is always a non-empty set of compatible sequences.

The computational procedure for finding RNA switches, including switches that can be triggered by external stimuli such as temperature changes, work surprisingly efficiently for (nearly) arbitrary pairs of structures. In particular, it does not require sophisticated optimization procedures. In fact, a simple local optimization scheme such as an adaptive walk is sufficient.

The ease with which switches can be designed suggests that RNA switches are also readily accessible in evolution. Hence the known cases are probably not exceptional instances of unusual RNA behavior, but represent another class of functions for which nature can employ RNA. This view is strongly supported by the large body of experimental, computational, and theoretical evidence that has been accumulated on the sequence-structure maps of nucleic acids.

Thus RNA switches are likely not exceptional instances of unusual RNA behavior but another class of functions for which nature employs RNA.

(1) The additivity of the energy parameters for nucleic acid secondary structures implies that the energetic effects of point mutations on the ground

state are bounded by a constant[2] independent of the chain length n, see (Fontana et al., 1993b). If x' is a point mutant of a sequence x with ground state structure $\Phi(x) = \Omega$ and if x' is still compatible with Ω, then Ω must also appear as a low-energy suboptimal structure of the sequence x', at most a few kcal/mol above the energy of the mutants ground state $\Omega' = \Phi(x')$.

(2) The *neutral set* $\Phi^{-1}(\Omega)$ consists of all sequences whose ground state (under fixed environmental conditions) is the secondary structure Ω. Extensive computational studies (Schuster et al., 1994; Grüner et al., 1996a,b) showed that $\Phi^{-1}(\Omega)$ is approximately uniformly embedded in the set $\mathbf{C}[\Omega]$ of sequences that are compatible with Ω. In the case of common secondary structures (i.e., those with a typical distribution of stack and loop sizes (Fontana et al., 1993a)) the neutral sets $\Phi^{-1}(\Omega)$ form connected networks that are densely embedded in $\mathbf{C}[\Omega]$. Therefore, the neutral networks of two structures Ω_1 and Ω_2 come very close together on the set $\mathbf{C}[\Omega_1] \cap \mathbf{C}[\Omega_2]$ of sequences that are compatible with both structures (Reidys et al., 1997; Reidys, 1997).

Suppose both Ω and Ω' are common structures. The denseness of their neutral networks on $\mathbf{C}[\Omega]$ and $\mathbf{C}[\Omega']$, and therefore also on $\mathbf{C}[\Omega] \cap \mathbf{C}[\Omega']$ now implies that the energies of both Ω and Ω' are similar to each other and close to the ground state energy for every sequences in the intersection $\mathbf{C}[\Omega] \cap \mathbf{C}[\Omega']$ (see Ancel and Fontana, 2000). Sequences that can fold into both structures therefore should be frequent in $\mathbf{C}[\Omega] \cap \mathbf{C}[\Omega']$, which is exactly what we find.

Acknowledgements. Stimulating discussions with Robert Giegerich are gratefully acknowledged. This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj. 13545-INF, and by the *European Commission* within the framework of the Biotechnology Program (BIO-4-98-0189).

Appendix

Derivation of Equ.(12). Let us first consider paths. A path that starts, say, with a G can be extended in exactly two ways: (1) with path starting with C or (2) with a path starting with U. A path that starts with C, on the other hand, must be continued with a path starting with G. The number of paths of length n starting with a particular letter X are therefore given by the following recursions:

$$\begin{aligned} F(P_n^{\mathsf{G}}) &= F(P_{n-1}^{\mathsf{U}}) + F(P_{n-1}^{\mathsf{C}}) \\ F(P_n^{\mathsf{U}}) &= F(P_{n-1}^{\mathsf{A}}) + F(P_{n-1}^{\mathsf{G}}) \\ F(P_n^{\mathsf{A}}) &= F(P_{n-1}^{\mathsf{U}}) \\ F(P_n^{\mathsf{C}}) &= F(P_{n-1}^{\mathsf{G}}) \end{aligned} \tag{16}$$

These recursions are of course started with $F(P_1^{\mathsf{X}}) = 1$ for all $\mathsf{X} \in \mathcal{A}$.

For cycles the situation appears more complicated at first glance, since we have to make sure that start and end can pair. For the biophysical pairing rules, however, we observe that pairs always consist of one purine, A,G, and one pyrimidine U,C, so that purines and pyrimidines alternate along a path. Thus in a path of even length, starting with the letter G the last letter is guaranteed to pair the G at the beginning. The number of cycles starting with G is therefore equal to the number of paths $F(C_n^{\mathsf{G}}) = F(P_n^{\mathsf{G}})$. Similarly, we have $F(C_n^{\mathsf{U}}) = F(P_n^{\mathsf{U}})$ Now consider a cycle starting with C. Clearly, the position before and after the C must both be G. Hence

[2]To be precise, this is true only if the structure in question does not contain very long loops in which case we can only prove a bound of the order $\ln n$ arising from joining two long loops to an even longer one by eliminating the separating base pair. Such structures, however, are not common and hence need not concern us here.

Table 1. Values of $F(\psi)$ for small components sizes n.

n	$F_1^p(n)$	$F_2^p(n)$	$F_1^c(n)$	$F_2^c(n)$	$\mathrm{Fib}(n)$
0					0
1	1	1	-	-	1
2	1	2	1	2	1
3	2	3	-	-	2
4	3	5	2	5	3
5	5	8	-	-	5
6	8	13	5	13	8
7	13	21	-	-	13
8	21	34	13	34	21

C_n^{C} can be thought of as constructed from a cycle C_{n-2}^{G} by inserting CG immediately after the start. Hence we obtain the following recursions:

$$\begin{aligned} F(C_n^{\mathsf{A}}) &= F(C_{n-2}^{\mathsf{U}}) = F(P_{n-2}^{\mathsf{U}}) \\ F(C_n^{\mathsf{C}}) &= F(C_{n-2}^{\mathsf{G}}) = F(P_{n-2}^{\mathsf{G}}) \end{aligned} \tag{17}$$

Since A and C, and G and U, respectively behave in the same way we introduce the abbreviations

$$\begin{aligned} F_1^p(n) &= F(P_n^{\mathsf{A}}) = F(P_n^{\mathsf{C}}) \qquad & F_2^p(n) &= F(P_n^{\mathsf{G}}) = F(P_n^{\mathsf{U}}) \\ F_1^c(2m) &= F(C_{2m}^{\mathsf{A}}) = F(C_{2m}^{\mathsf{C}}) \qquad & F_2^c(2m) &= F(C_{2m}^{\mathsf{G}}) = F(C_{2m}^{\mathsf{U}}) \end{aligned} \tag{18}$$

This yields

$$F_2^p(n) = F_2^p(n-1) + F_1^p(n-1) = F_2^p(n-1) + F_2^p(n-2) \tag{19}$$

with the initial conditions

$$F_2^p(0) = 1 \qquad F_2^p(1) = 2 \tag{20}$$

The recursions (19) are the same as the recursions for the Fibonacci numbers

$$\mathrm{Fib}(n) = \mathrm{Fib}(n-1) + \mathrm{Fib}(n-2) \qquad \mathrm{Fib}(0) = 0,\ \mathrm{Fib}(1) = 1 \tag{21}$$

except for the initial conditions.

Taking the initializations into account, see Table 1, we obtain

$$\begin{aligned} F_1^p(n) &= \mathrm{Fib}(n) \qquad & F_2^p(n) &= \mathrm{Fib}(n+1) \\ F_1^c(2m) &= \mathrm{Fib}(2m-1) \qquad & F_2^c(2m) &= \mathrm{Fib}(2m+1) \end{aligned} \tag{22}$$

Of course we have $F(P_n) = 2F_1^p(n) + 2F_2^p(n)$ and $F(C_n) = 2F_1^c(n) + 2F_2^c(n)$. Thus

$$\begin{aligned} F(P_n) &= 2\big(\mathrm{Fib}(n) + \mathrm{Fib}(n+1)\big) = 2\mathrm{Fib}(n+2) \\ F(C_n) &= 2\big(\mathrm{Fib}(n-1) + \mathrm{Fib}(n+1)\big) = 2\mathrm{Lucas}(n) \end{aligned} \tag{23}$$

The Lucas numbers $\mathrm{Lucas}(n)$ satisfy the same recursion as $\mathrm{Fib}(n)$. However, the initialization is different: $\mathrm{Lucas}(0) = 2$ and $\mathrm{Lucas}(1) = 1$.

Generating random compatible sequences. The discussion is Section 2.5 leads to a recursive algorithm for filling a cycle by a random sequence. We use the symbol $a \lll\!\!\!+ b$ to mean the concatenation ab of the strings a and b. In addition algorithm 1 requires a function `fillUpath` analogous to `fillGpath`.

algorithm 1 A recursive algorithm to fill a cycle. We use the symbol $a \lll\!\!+ b$ to mean the concatenation ab of the strings a and b. In addition a function `fillUpath` analogous to `fillGpath` is needed.

Function `fillcycle`(n):
if $n = 0$ **then**
    return $\emptyset$;
$\xi \leftarrow$ uniformly distributed random number in $[0, 1]$;
if $\xi < F_1^c(n)/F^c(n) = \mathrm{Fib}(n-1)/(2\mathrm{Lucas}(n))$ **then**
    return $'\mathsf{AU}' \lll\!\!+$ `fillUpath`$(n-2)$;
else
    if $\xi < \mathrm{Fib}(n-1)/\mathrm{Lucas}(n)$ **then**
        return $'\mathsf{CG}' \lll\!\!+$ `fillGpath`$(n-2)$;
    else
        $\xi \leftarrow \xi - 2F_1^c(n)/F^c(n)$;
        if $\xi < F_2^c(n)/F^c(n) = \mathrm{Fib}(n+1)/(2\mathrm{Lucas}(n))$ **then**
            return $'\mathsf{U}' \lll\!\!+$ `fillUpath`$(n-1)$;
        else
            return $'\mathsf{G}' \lll\!\!+$ `fillGpath`$(n-1)$;

Function `fillGpath`(n):
if $n = 0$ **then**
    return $\emptyset$;
$\xi \leftarrow$ uniformly distributed random number in $[0, 1]$;
if $\xi < F_2^p(n-2)/F_2^p(n) = \mathrm{Fib}(n-1)/\mathrm{Fib}(n+1)$ **then**
    return $'\mathsf{CG}' \lll\!\!+$ `fillGpath`$(n-2)$;
else
    return $'\mathsf{U}' \lll\!\!+$ `fillUpath`$(n-1)$;

References

Ancel LW, Fontana W. 2000. Plasticity, evolvability and modularity in RNA. *J. Exp. Zool. (Mol. Dev. Evol.).* in press.

Babitzke P, Yanofsky C. 1993. Reconstitution of Bacillus subtilis Trp attenuation in vitro with TRAP, the Trp RNA-binding attenuation protein. *Proc. Natl. Acad. Sci. USA*. 90:133–137.

Baumstark T, Schroder AR, Riesner D. 1997. 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.

Biebricher CK, Diekmann S, Luce R. 1982. Structural analysis of self-replicating RNA synthesized by Qβ replicase. *J. Mol. Biol.* 154:629–648.

Biebricher CK, Luce R. 1992. In vitro recombination and terminal elongation of RNA by Qβ replicase. *EMBO J.* 11:5129–5135.

Brion P, Westhof E. 1997. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* 26:113–137.

Emerick VL, Woodson SA. 1993. Selfsplicing of the Tetrahymena pre-rRNA is decreased by misfolding during transcription. *Biochemistry.* 32:14062–14067.

Fayat G, Mayaux FJ, Sacerdot C, Fromant M, Springer M, Grunberg-Manago M, Blanquet S. 1983. Escherichia coli phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identification of the gene for the ribosomal protein L20. *J. Mol. Biol.* 171:239–261.

Flamm C, Fontana W, Hofacker I, Schuster P. 2000. RNA folding kinetics at elementary step resolution. *RNA.* 6:325–338.

Fontana W, Konings DAM, Stadler PF, Schuster P. 1993a. Statistics of RNA secondary structures. *Biopolymers.* 33:1389–1404.

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P. 1993b. RNA folding and combinatory landscapes. *Phys. Rev. E.* 47:2083–2099.

Fresco JR, Adains A, Ascione R, Henley D, Lindahl T. 1966. Tertiary structure in transfer ribonucleic acids. *Cold Spring Habor Symp. Quant. Biol.* 31:527–539.

Giegerich R, Haase D, Rehmsmeier M. Prediction and visualization of structural switches in RNA. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 4, pages 126–137, Singapore, 1999. World Scientific Press.

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996a. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.* 127:355–374.

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996b. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath. Chem.* 127:375–389.

Gultyaev AP, Batenburg FH, Pleij CW. 1998. Dynamic competition between alternative structures in viroid RNAs simulated by an RNA folding algorithm. *J. Mol. Biol.* 276:43–55.

Hawkins ER, Chang SH, Mattice WL. 1977. Kinetics of the renaturation of yeast tRNA$_3^{Leu}$. *Biopolymers.* 16:1557–1566.

Hecker R, Wang ZM, Steger G, Riesner D. 1988. Analysis of RNA structures by temperarure-gradient gel electrophoresis: viroid replication and processing. *Gene.* 72:59–74.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie.* 125:167–188.

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*. 86:7706–7710.

Lodmell JS, Dahlberg AE. 1997. A conformational switch in Escherichia coli 16S ribosomal RNA during decoding of messenger RNA. *Science*. 277:1262–1267.

Loss P, Schmitz M, Steger G, Riesner D. 1991. Formation of a thermodynamically metastable structure containing hairpin II is critical for infectivity of potato spindle tuber viroid RNA. *EMBO J.* 10:719–727.

Mathews D, Sabina J, Zucker M, Turner H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. 29:1105–1119.

Morgan SR, Higgs PG. 1998. Barrier heights between groundstates in a model of RNA secondary structure. *J. Phys. A*. 31:3153–3170.

Nagel JHA, Gultyaev AP, Derdes K, Pleij CWA. 1999. Metastable structures and refolding kinetics in *hok* mRNA of plasmid R1. *RNA*. 5:1408–1419.

Pan T, Fang X, Sosnick T. 1998. Pathway modulation, circular permutation and rapid RNA folding under kinetic control. *J. Mol. Biol.* 286:721–731.

Pan T, Thirumalai D, Woodson SA. 1997. Folding of RNA involves parallel pathways. *J. Mol. Biol.* 273:7–13.

Perrotta AT, Been MD. 1998. A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation. *J. Mol. Biol.* 279:361–373.

Putzer H, Gendron N, Grunberg-Manago M. 1992. Co-ordinate expression of the two threonyl-tRNA synthetase genes in Bacillus subtilis: Control by transcriptional antitermination involving a conserved regulatory sequence. *EMBO J.* 11: 3117–3127.

Reidys C, Stadler PF. 1996. Bio-molecular shapes and algebraic structures. *Computers & Chem.* 20:85–94.

Reidys CM. 1997. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.* 19:360–377.

Reidys CM, Stadler PF, Schuster P. 1997. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.* 59:339–397.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*. 289:448–452.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. (London) B*. 255:279–284.

Shapiro BA, Zhang K. 1990. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*. 6:309–318.

Soukup GA, Breaker RR. 1999. Engeneering precision RNA molecular switches. *Proc. Natl. Acad. Sci. USA*. 96:3584–3589.

Tinoco Jr I, Bustamante C. 1999. How RNA folds. *J. Mol. Biol.* 293:271–281.

Treiber DK, Rook MS, Zarrinkar PP, Williamson JR. 1998. Kinetic intermediates trapped by native interactions in RNA folding. *Science*. 279:1943–1946.

Vertechi AM, Virasoro MA. 1989. Energy barriers in SK spin glass models. *J. Phys. France*. 50:2325–2332.

Ahsen Uvon. 1998. Translational fidelity: Error-prone versus hyper-accurate ribosomes. *Chem. Biol.* 5:R3–R6.

Walter AE, Turner DH, Kim J, Lyttle MH, Müller P, Mathews DH, Zuker M. 1994. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*. 91:9218–9222.

Wool IG, Glück A, Endo Y. 1992. Ribotoxin recognition of ribosomal RNA and a proposal for the mechanism of translocation. *Trends Biochem.* 17:266–269.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*. 49:145–165.

Zamora H, Luce R, Biebricher CK. 1995. Design of artificial short-chained RNA species that are replicated by Qβ replicase. *Biochemistry*. 34:1261–1266.

Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science*. 244:48–52.

Zuker M, Stiegler P. 1981. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*. 9:133–148.