Exploring the Connection Between Synthetic and Natural RNAs in Genomes: A Novel Computational Approach

Uri Laserson[1,2], Hin Hark Gan[1], and Tamar Schlick[1,2]

[1] Department of Chemistry
[2] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA

Abstract. The central dogma of biology—that DNA makes RNA makes protein—was recently expanded yet again with the discovery of RNAs that carry important regulatory functions (e.g., metabolite-binding RNAs, transcription regulation, chromosome replication). Thus, rather than only serving as mediators between the hereditary material and the cell's workhorses (proteins), RNAs have essential regulatory roles. This finding has stimulated a search for small functional RNA motifs, either embedded in mRNA molecules or as separate molecules in the cell. The existence of such simple RNA motifs in Nature suggests that the results from experimental *in vitro* selection of functional RNA molecules may shed light on the scope and functional diversity of these simple RNA structural motifs *in vivo*. Here we develop a computational method for extracting structural information from laboratory selection experiments and searching the genomes of various organisms for sequences that may fold into similar structures (if transcribed), as well as techniques for evaluating the structural stability of such potential candidate sequences. Applications of our algorithm to several aptamer motifs (that bind either antibiotics or ATP) produce a number of promising candidates in the genomes of selected bacterial and archaeal species. More generally, our approach offers a promising avenue for enhancing current knowledge of RNA's structural repertoire in the cell.

1 Introduction: Importance of RNA Structure and Function

RNA molecules play essential roles in the cellular processes of all living organisms. The wonderful capacity of RNA to form complex, stable tertiary structures has been exploited by evolution. RNA molecules are integral components of the cellular machinery for transcription regulation, chromosome replication, RNA processing and modification, and other essential biological functions [9, 41, 48]. Recent discoveries that noncoding RNAs (ncRNAs—RNA molecules that have specific functions other than directing protein synthesis through translation) make up a significant portion of the transcriptome (entire set of expressed RNA and protein transcripts of an organism) further suggest the prominent role of RNA in cellular function [16, 34]. In particular, cells employ many small ncRNAs such as microRNAs (21–23 nt) to regulate

gene expression; small interfering RNAs (siRNAs), often complementary to messenger RNAs, to mediate mRNA degradation; and small nucleolar RNAs, important in post-transcriptional modification of ribosomal and other RNAs [3, 23, 24]. Thus, ncRNAs may hold the key for understanding genetic control of cell development and growth [31, 32].

With rapidly growing interest in RNA structure and function, as well as emerging technological applications of RNA to biomedicine, new scientific challenges regarding RNA are at the forefront. One important objective is to increase the functional repertoire of RNA. This can be tackled by systematically identifying and characterizing the ncRNAs in genomes, or by creating novel synthetic functional RNAs using *in vitro* selection methods. Although *in vitro* selection is a proven tool for finding novel functional RNAs, it is limited in scope to relatively small RNAs ($<$250 nt) [53].

Here we report development of a theoretical/experimental approach for expanding the functional repertoire of RNA using a combination of *in vitro* selection and computational methods. Alternatively, the analysis of RNA motifs as mathematical graphs, as developed recently ([14, 26, 35], and see the Appendix), may suggest candidate novel RNA topologies to direct computational/experimental searches for ncRNAs in genomes and synthetic functional RNAs in the laboratory. Before we present our approach, we describe the motivation for exploring the connection between synthetic and natural RNAs.

2 Exploring the Connection between Synthetic and Natural RNAs

In recent years, numerous target-binding nucleic acid molecules (known as *aptamers*) have been identified; the targets include organic molecules, antibiotics, peptides, proteins, and whole viruses [21, 53]. In addition, *in vitro* selection experiments have produced novel RNA enzymes (ribozymes) and led to applications in biomolecular engineering, e.g., allosteric ribozymes and biosensors [44–46].

The process of *in vitro* selection simulates evolution in the laboratory [10, 47, 51, 53]. Starting with a large pool of small, random-sequence RNA molecules, the sequences are iteratively selected for a physical or chemical property (e.g., binding affinity or catalysis) by amplifying the enriched sequence pool using the polymerase chain reaction (PCR). After $\sim$8–15 cycles have been completed, the molecules are cloned and sequenced. The resulting artificial molecules may be target-binding RNAs (called *aptamers*) or novel catalytic RNAs. Since the process of selecting functional molecules is similar to evolution, and the process takes advantage of the PCR, common motifs among the sequences with a significant amount of sequence conservation often emerge.

The binding and catalytic properties of synthetic and natural RNAs are mediated by specific sequence and structural motifs. In 2001, Szostak's group demonstrated that the motif of the natural functional hammerhead ribozyme can be selected from random sequence pools [39], suggesting multiple origins for the ribozyme. Because a synthetic aptamer discovered by Wallace and Schroeder [52] and crystallized by

the Patel group [49] binds streptomycin (which is naturally produced by bacteria) tightly and specifically, Piganeau and Schroeder reinforced the likely connection between natural and synthetic RNAs [36]. Indeed, motifs similar to aptamers may exist in natural RNAs because aptamers bind to certain targets that are prevalent in Nature and recent findings also show that metabolite-induced RNA conformational changes control gene expression in bacteria and other organisms [33, 54, 55]. As Piganeau and Schroeder conclude in their commentary on the recent article by Patel and collaborators [49] on the structure of a 40-nt aptamer binding the antibiotic streptomycin, *"We can now predict that many biosynthetic pathways will be regulated by metabolite binding 'natural aptamers,' and we might even find a structure similar to the streptomycin aptamer in a bacterium producing streptomycin"* [36].

Since *in vitro* selection is a technique that simulates an evolutionary process, it is reasonable that structures discovered through *in vitro* selection may have also evolved in the cell. The methods we develop here are meant to explore these intriguing connections between natural and synthetic RNAs in a general and systematic manner.

3 Methods

The components used in our method are standard, but the combination is novel. Namely, we combine the following techniques and tools: output from *in vitro* selection of functional molecules; RNAMotif [29], a computational tool that allows searches for RNA sequences in genomic databases that might fold into specified secondary conformations; and the Vienna RNA Package [22], which can predict the secondary structures of RNA molecules from their sequence, as well as calculate other properties, such as "sub-optimal foldings" and heat-capacity curves.

3.1 Aptamer Search Algorithm

Our aptamer search method has three major steps:

1. *Create motif descriptors* by extracting the critical structural features from the experimental aptamer molecule that confer onto the molecule its special physical or chemical property (e.g., binding affinity);
2. *Search the genomes* of selected organisms using the RNAMotif tool for the aptamer structure specified by the descriptor we have created;
3. *Assess the quality* of the candidate sequences (e.g., nature and stability of fold, energetic stability, statistical significance, etc.).

The first step involves the analysis of motif data from *in vitro* selection experiments and, if available, any structural studies of specific motifs. Important structural information includes overall qualitative structure (e.g., loops, bulges, hairpins, lengths of stems, etc.) as well as any specific sequence information that may be critical to the hydrogen-bonding scheme (Fig. 1).

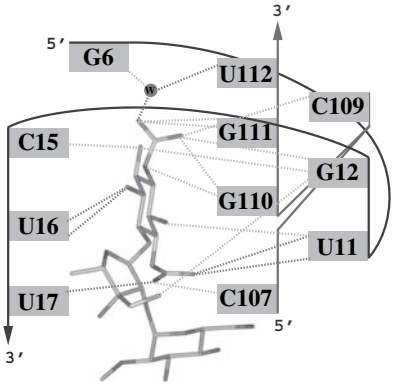

Figure 1. The hydrogen-bonding scheme of the streptomycin aptamer developed by the Schroeder group [52]. The dotted lines represent hydrogen bonds between the streptomycin molecule (center) and the nucleotides in the binding pocket. This type of information is ideal for pinpointing the specific nucleotides that account for the binding specificity. The secondary structure of the molecule is shown in Fig. 2d. (Adapted from [49])

Figure 1 illustrates this process for the streptomycin aptamer. The hydrogen-bonding scheme in the streptomycin-aptamer complex (discovered by the Schroeder group [52]) was crystallized and structurally characterized by the Patel group [49]. The corresponding secondary structure is shown in Fig. 2d. This motif has three helices (each of variable length) and two asymmetric bulges. Because the bases shown in Fig. 1 are known experimentally to be important to the binding specificity, we define the target sequence motif in the first asymmetric bulge to be `GNANNUG`. Likewise, we retain bases associated with other bulges since they contribute to the binding affinity.

After collecting the relevant structural information, we search the genomes of numerous organisms using RNAMotif as described below. This yields for each aptamer a pool of sequences that could potentially fold into the specified experimental aptamer structure. To filter this pool further, we "fold" each candidate sequence using the 2D prediction tool in the Vienna RNA Package. We retain the predicted structure if it is similar to the experimental aptamer structure but discard it otherwise; the similarity is determined by inspection: the sequence is retained if the general topology of the predicted structure is similar to the experimental structure. This process greatly reduces the size of the candidate sequence pool.

Finally, we subject the remaining candidates (sequences that fold as desired) to further tests that allow us to assess their potential as significant matches based on energetic and statistical measurements for significance and stability.

3.2 RNAMotif Scanning Tool: Searching for Secondary Structural Motifs

Sequence and secondary structural motifs can be searched for in genomes using the scanning tool RNAMotif [29]. The qualitative topological and secondary structural

elements are specified in a "descriptor" (specifying the structural connectivity and the length of helices, loops, bulges, etc.) as well as any specific sequence information (such as a GNRA loop). The program searches for sequences in genomes that could potentially fold into the specified secondary structure based on Watson-Crick base pairing rules. Additionally, the sequences are ranked based on the energy of folding them into the specified secondary structure.

3.3 RNA Folding Algorithms: Secondary Structure Prediction

Available 2D RNA folding algorithms can predict the patterns of base pairing, the presence of base pair mismatches, and regions with unpaired bases (e.g., loops, bulges, and junctions). For RNA tree structures, the 2D folding algorithm bundled with the Vienna RNA package (P. Schuster and coworkers [22]), is widely used. (One of the limitations of the 2D prediction algorithms is that pseudoknots[3] are not accounted for since they are difficult to predict, but efforts include pseudoknot prediction [37].) This shortcoming in 2D RNA folding algorithms limits our RNA folding applications to tree structures, which nonetheless represent a large set of structures to explore.[4] The current algorithms also do not account for the effect of magnesium ions, which have been shown to be critical for RNA folding. However, the Vienna RNA package is reasonably accurate for short sequences ($<$100 nt), and in addition to predicting the minimum energy 2D structure, it can calculate suboptimal structures and melting curves (specific heat curves, as described next).

3.4 Melting Curve Analysis of Candidate Sequences

Biological molecules possess stable structures at physiological temperatures. The stability of RNA secondary structures with respect to temperature change can be analyzed using heat-capacity or melting curves. These curves represent the amount of energy absorbed per unit change in temperature (e.g. $\partial H/\partial T$ versus T, where H is the RNA's conformational enthalpy). For each candidate sequence, we compute the theoretical melting curve using tools available in the Vienna RNA Package. Additionally, to discriminate random from functional RNA sequences, we shuffle each candidate sequence 1,000 times and compute the corresponding melting curves. The melting temperature, T_m, is defined as the temperature of the highest peak; it may also be interpreted as the transition temperature at which the RNA molecule's secondary structure experiences significant disruption.

[3] A *pseudoknot* forms when consecutive single-stranded regions **a**, x, **b**, y, **c**, z, **d** (where x, y, and z are the connecting regions) fold on each other such that **a** hydrogen-bonds with **c** and **b** with **d**. A pseudoknot is technically part of a molecule's tertiary structure though it is often convenient to consider it a secondary structural element since it involves base-pairing interactions [41].

[4] Note that general 3D prediction algorithms for RNAs are not yet available and constitute a significant challenge. This gap between 2D and 3D structure prediction is best addressed at present by subjecting any predictions based on 2D folds to experimental tests.

Additionally, for each sequence (the candidate and the randomly shuffled ones), we plot the melting temperature (T_m) against the free energy (F) as a further indicator of stability. The T_m versus F plot maps the global physical characteristics of RNA secondary folds. Using principal component analysis, we compute a 90% confidence ellipse, and check to see whether the candidate sequence falls outside the ellipse (i.e., is significantly stable).

3.5 Conformational Energy Landscape Analysis

We also calculate the conformational energy landscape of the secondary RNA fold of each candidate sequence. The shape of this energy landscape offers an alternative approach for discriminating RNA-like from non-RNA-like molecules. To assess different secondary structures, we plot for each suboptimal structure (up to 10 kcal/mol above the global minimum) its energy (E) against a "distance" (D) from the minimum energy structure. We define D to be the base pair dissimilarity between the minimum energy structure and each suboptimal structure. This E versus D plot for optimal and suboptimal structures defines the conformational energy landscape of an RNA sequence.

Furthermore, we quantify the conformational energy landscape of each sequence by computing the "Valley index," V, as described in [27]. The Valley index is a Boltzmann-weighted average of the distance between all pairs of structures (minimum-energy and sub-optimal). A large Valley index implies many low energy competing structures that have very different conformations.

Similar to the melting curve analysis, each sequence is randomly shuffled 1,000 times, and the Valley index is plotted against the free energy of the sequence. A 90% confidence ellipse is computed, and the candidate sequence is tested to see whether it falls outside of the ellipse.

3.6 Statistical Analysis

Each secondary structure motif can be considered as a "word" in the 4-letter nucleotide alphabet. For example, the probability of finding any given letter in a random sequence of nucleotides is $1/4$, so the probability of finding any specific sequence of length N is $1/4^N$. Likewise, a complicated descriptor has an associated expected frequency. However, given the complexity of secondary structure descriptors, this number is difficult to compute analytically. Therefore, these frequencies are estimated using a Monte Carlo method, in which the number of matches to a given descriptor in a random 1 Mb sequence is averaged over 1 million tries. Provided the genomes of interest have a nucleotide distribution that is close to uniform,[5]

[5]The assumption that that a genome is uniformly distributed in the four bases is strong. Many genomes deviate from uniform distributions, and often it is more fruitful to consider di- or trinucleotide distributions. However, our uniform distribution assumption makes the analyses simple and can later be refined. We also performed tests in which we biased the base distribution to mimic the *Streptomyces avermitilis* genome (the genome we used with

our computation provides an estimate of the number of matches to the descriptor that we expect by probability alone. With this information, we can estimate whether a secondary structure motif is over or under-represented in a given genome.

3.7 Computational Performance

Creating the descriptor definition in terms of skeletal motifs requires biological intuition and experimentation. Once defined, we employ the simple "programming language" of Macke and colleagues [29] to describe an RNA secondary structure of any complexity.

The genome searches using RNAMotif are very rapid. The search involving the most complicated descriptor (streptomycin) and largest genome (*Streptomyces avermitilis*) can be completed in less than one minute. In general, the computational speed depends on genome size and complexity of the descriptor.

Subsequently, the candidate sequences are folded. We use the RNAfold program as part of the Vienna RNA package. For a sequence of less than 100 nucleotides, the predicted secondary structure is computed in a matter of seconds. Generating sub-optimal structures, however, is relatively lengthy, because sequences of similar lengths can have drastically different numbers of sub-optimal structures (e.g., ranging from $\sim$150 for some sequences to $\sim$200,000 for others). However, the generation of each suboptimal structure is extremely fast (even the case where over 200,000 structures were generated took only 2 minutes to finish).

The energetic and statistical analyses are the most computationally expensive. For the scatter plots, we compute the Valley index and melting curve for 1,000 random permutations of a given sequence. For the statistical evaluation, we search a 1 Mb sequence 1 million times. The combined calculations take several weeks, with the majority of the time in the Monte Carlo evaluation.

All computations were performed on an SGI 300 MHz MIPS R12000 IP27 processor with 4 GB of memory.

4 Results

4.1 Candidate Sequences in Bacterial and Archaeal Genomes: Initial Search Results

We began searches for three aptamers that showed binding affinity to antibiotics (chloramphenicol [6], streptomycin [49, 52], and neomycin B [25]) in bacterial genomes, as well as for an aptamer that showed affinity for ATP [40] in archaeal genomes. The descriptors are shown in Fig. 2. We use representative species for searching: since the *Streptomyces* family accounts for many of the known antibiotics (including those used here), we use all available *Streptomyces* genomes. Additionally, we choose *E. coli* genomes since they are among the best characterized

the largest deviation from uniformity) and computed the expected frequency of matches; the results were similar to the case of uniform distribution.

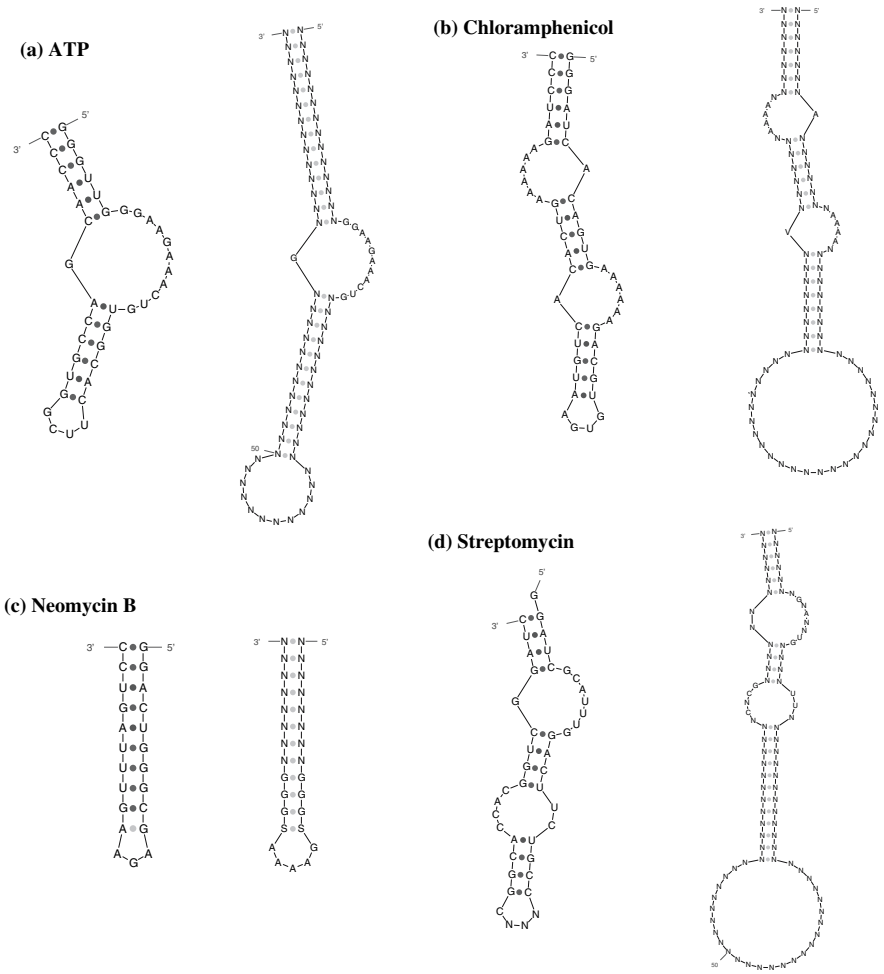

Figure 2. The four experimental aptamers (left side of each pair) used in this study along with our constructed motif descriptors (right). Regions with many 'N's are variable in length. The nucleic acid base symbols are defined as follows: N: any base; V: A, C, or G; S: C or G. We search the genome of an organism for any sequence that fits into the descriptor consensus (allowing for slight mismatches). The experimental papers associated with the aptamers are: ATP [40], Chloramphenicol [6], Neomycin B [25], and Streptomycin [49, 52]

bacterial species (and a staple of many biological studies). Other bacterial genomes were selected randomly. For the searches involving the ATP aptamer, we used all the available archaeal genomes at the time of the study from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). The search results are displayed in Tables 1 and 2.

Table 1. Number of initial candidate sequences for the ATP-binding aptamer in selected archaeal genomes. Bold entries exhibit significant deviations from the expected number of matches

Genome	Size (Mb)	ATP	
		Observed	Expected
Aeropyrum pernix	1.7	**6**	**1.6**
Archaeoglobus fulgidis DSM 4304	2.2	**6**	**2.1**
Halobacterium sp. NRC-1	2.6	4	2.5
Methanobacterium thermoautotrophicum str. ΔH	1.8	**6**	**1.7**
Methanococcus jannaschii	1.7	4	1.6
Methanopyrus kandleri AV19	1.7	3	1.6
Methanosarcina acetivorans C2A	5.8	**12**	**5.5**
Methanosarcina mazei Goe1	4.1	5	3.9
Pyrobaculum aerophilum	2.3	2	2.2
Pyrococcus abyssi	1.8	3	1.7
Pyrococcus furiosus DSM 3638	1.9	**8**	**1.8**
Pyrococcus horikoshii	1.8	3	1.7
Sulfolobus solfataricus	3.0	2	2.8
Sulfolobus tokodaii	2.7	4	2.5
Thermoplasma acidophilum	1.6	2	1.5
Thermoplasma volcanium	1.6	2	1.5
	Total	72	

Table 2. Number of initial candidate sequences for the three antibiotic-binding aptamers in selected bacterial genomes. Bold entries exhibit significant deviations from the expected number of matches

Genome	Size (Mb)	Chloramph.		Streptomycin		Neomycin B		**Tot.**
		Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
Streptomyces avermitilis	9.2	0	3.1	2	0.7	**6**	**19.8**	8
Streptomyces coelicolor	8.8	0	2.9	1	0.7	**1**	**18.9**	2
E. coli K12	4.7	**7**	**1.6**	1	0.4	11	10.1	19
E. coli O157:H7	5.6	**7**	**1.9**	1	0.4	**17**	**12.0**	25
E. coli O157:H7 EDL933	5.6	**7**	**1.9**	1	0.4	**19**	**12.0**	27
E. coli CFT073	5.3	**7**	**1.8**	0	0.4	**17**	**11.4**	24
Neisseria meningitidis MC58	2.3	1	0.8	0	0.2	9	4.9	10
Neisseria meningitidis Z2498	2.2	**5**	**0.7**	0	0.2	9	4.7	14
Sinorhizobium meliloti	3.7	1	1.2	1	0.3	**3**	**7.9**	5
Chlamydia trachomatis	1.1	**4**	**0.4**	0	0.1	1	2.4	5
	Tot.	39		7		93		139

Tables 1 and 2 describe our results for these four aptamer targets. First, we note that the neomycin B candidate pool is much larger than both the chloramphenicol and streptomycin pools, because it is simpler (12–17 nt, with a simple loop motif; see Fig. 2c).

Second, we see that the average number of matches to neomycin B in *Streptomyces* genomes is four, while the average number in *E. coli* genomes is four times greater. Since the average *Streptomyces* genome size is almost twice that of the average *E. coli* genome ($\sim$9 Mb versus $\sim$5 Mb), these trends are significant because probability alone would predict the number of hits in *Streptomyces* genomes to be about *twice* as many matches as *E. coli*, not *one fourth* as we obtain. This suggests a preference for or discrimination against this aptamer, i.e., either *Streptomyces* is significantly missing this structure, or *E. coli* has a significantly large occurrence of it.

More rigorously, the Monte Carlo method described above estimates the expected frequency of each of the descriptors per 1 Mb of uniformly distributed random sequence. The Monte Carlo results are listed in Table 3 and the expected number of matches are listed in Tables 1 and 2. This information reveals several interesting trends. For streptomycin, which exhibits a number of matches similar to the computed expected number of matches, no significant findings of aptamer hits can be claimed; however, pools for neomycin B, chloramphenicol, and ATP show significant deviations from the expected behavior (e.g., much smaller for the *Streptomyces* genomes and much larger for *E. coli* genomes for neomycin B). The archaea genomes tend to agree with the expected number of matches, except for several genomes marked in bold in Table 1.

Table 3. Computed expected frequencies of the descriptors per 1 Mb of uniformly distributed random sequence

Descriptor	Frequency ± 1 Standard Error
Chloramphenicol	0.3320 ± 0.0007
Streptomycin	0.0778 ± 0.0003
Neomycin B	2.1471 ± 0.0015
ATP	0.9424 ± 0.0009

4.2 Structurally Filtered Search Results

Next, we further filter the pools of matches to eliminate candidate sequences that may not fold as intended. This step reduces our total pool of candidate sequences greatly, from 139 to 32 matches in bacterial genomes and from 72 to 5 matches in archaeal genomes. In particular, only one candidate sequence remains for the streptomycin aptamer.

4.3 Energetic Analysis Discriminates Natural from Random RNA Sequences

With the significantly smaller pool of filtered candidate sequences, we now proceed to evaluate the candidates based on energetic considerations. Because biological

RNA molecules form stable structures at physiological temperatures, specific heat curves and conformational energy landscapes have expected characteristics, as elaborated below.

First, we compute the melting curves and melting temperatures of the candidate sequences and compare them to those of random permutations of the sequence. We plot the melting temperatures of our candidate and its permuted pool against the free energy. A promising candidate should have distinct characteristics from the random pool. In certain cases, we indeed find the candidate sequence to be apart from the bulk of the random sequences; Fig. 3 shows how the sequence of interest falls outside a computed 90% confidence ellipse.

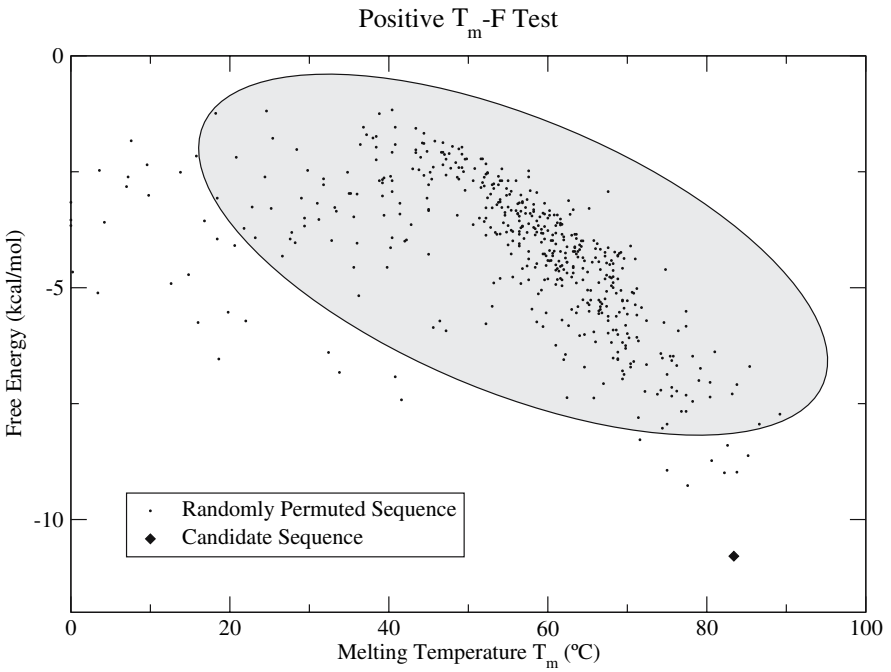

Figure 3. Scatter plot of the melting temperatures (T_m) of the candidate sequence and its 1,000 random permutations versus the free energy of their structure ensembles. Notice that the candidate sequence lies outside the 90% confidence ellipse, signifying an especially stable structure

Secondly, we analyze the conformational energy landscape of each candidate sequence, as shown in Fig. 4. That is, we plot the energy of each suboptimal structure versus the distance of the suboptimal structure to the minimum energy structure. We observe that some candidates possess conformational energy landscapes with multiple low minima, while others have very steep single-minima landscapes. A stable aptamer structure should have a sharp, deep minimum and a funnel-like landscape (Fig. 4, circles). This means that for a given sequence, the more different a fold

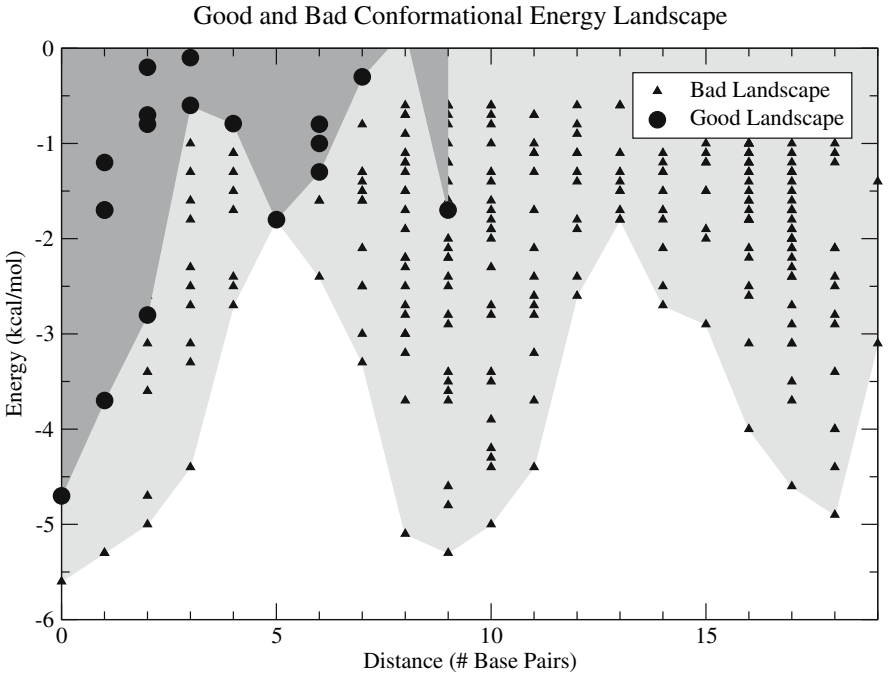

Figure 4. Conformational energy landscapes of two candidate sequences for the neomycin B antibiotic. The "good" candidate, marked with circles, exhibits a sharp, steep slope, while the "bad" candidate, marked with triangles, is more random, with multiple isoenergetic minima

looks from the minimum energy fold, the higher energy it is. A random landscape, or a shallow one with multiple minima (as in Fig. 4, triangles) likely has other different and stable structures for the same sequence that are almost isoenergetic to the minimum energy structure. This information is quantified in the computation of the Valley index. Structures with unfavorable conformational energy landscapes have larger Valley indices. This provides for a quantitative test of the energetic stability of the candidate sequence's minimum energy structure. As described above, each sequence is shuffled 1,000 times and the Valley indices of all the sequences are plotted against their free energy (see Fig. 5). A sequence that falls outside the bulk of the random sequences is probably significantly stable. Using the analyses above, we evaluate the candidate sequences that are most likely to have the physical properties of biological molecules. Our tests are stringent indicators of stability, as only 24% of the candidate sequences pass the melting temperature test, while only 14% pass the Valley index test. Any sequence that passes the Valley index test also passes the melting temperature test. Furthermore, the tests were conducted on several known biological RNAs (5S, U5, U6, U7, and Gln tRNA), and all of them passed the tests with the exception of one sequence which did not pass the Valley index test.

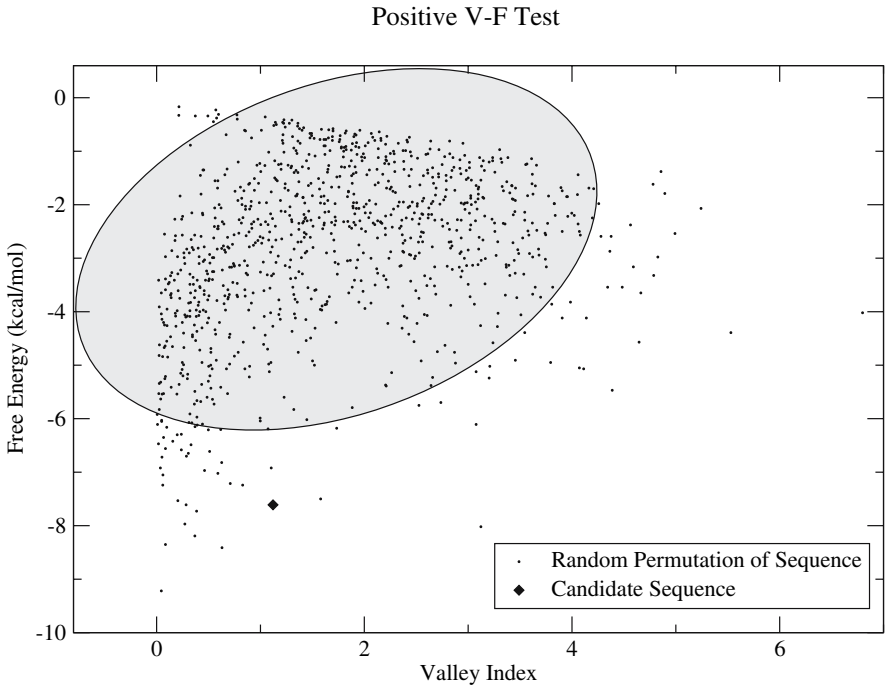

Figure 5. Plot of the Valley index of a candidate sequence versus its free energy, as well as 1,000 random permutations of the sequence. Note how the candidate sequence lies outside of the 90% confidence ellipse, suggesting an especially stable structure

Candidate Sequences

Selected candidate sequences are shown in Table 4, including the results of energetic tests, the computed physical quantities, and their locations in the genome. Some of the sequences occur in non-coding regions, while others occur in genes of known and unknown functions. Many of the sequences occur multiple times in different genomes, and even in the same genome. For example, the two candidates from the ATP pool pass both the melting temperature and the Valley index tests. They occur inside the plasmids of *Halobacterium* in non-coding regions. On the other hand, the streptomycin pool candidate sequence passes neither of the energetic tests, yet shows very positive physical qualities (e.g., high melting temperature) and is also located in a non-coding region. The first two neomycin B sequences pass both energetic tests with the first one in a non-coding sequence while the second is located in a hypothetical protein with a function that is not currently understood. Finally, the chloramphenicol pool sequence passes neither of the tests, and does not exhibit especially stable characteristics. However, it is located in the ECs1492 gene, which encodes for a transcription-repair coupling factor that is responsible for a mutation frequency decline. It is possible that this may explain another mode of action for chloramphenicol, namely that it increases transcription mutation frequency. Finally,

based on the data in Tables 1 and 2, it is possible to attribute increased significance to certain matches based on the statistical representation of the motif in a given genome. For example, it is easily seen that *E. coli* accounts for a significantly large number of neomycin B motifs, compared with what is expected. Therefore, the neomycin B matches located in the *E. coli* genomes may have a higher likelihood of existing *in vivo*.

5 Conclusions and Future Directions

The method presented here for searching for artificial aptamer structures in the genomes of various organisms has produced promising RNA sequences that may be functionally important. In response to the assertion of Piganeau and Schroeder [36], it is indeed likely that aptamer structures will be found *in vivo*.

Further work is required to search all genomes comprehensively, investigate other aptamers, develop ways to reliably distinguish biological RNAs from random noise and spurious matches, and most importantly, verify the findings experimentally.

More broadly, the integration of mathematical RNA modeling (such as by graph theory, see Appendix) and experimental methods has the potential to greatly expand our knowledge of RNA's repertoire through the development of new tools for analyzing RNA motifs, genome analysis, and improvement of the *in vitro* selection technology. These tools and technologies will likely broaden the scope of RNA-based methods for biomedical applications including selection of complex synthetic RNAs and identification of ncRNAs associated with various physiological functions. Ultimately, it is prudent to address the challenging problem of connecting and better characterizing the relationships between 2D and 3D RNA folds.

Acknowledgments

We thank Marco Avellaneda, Dinshaw Patel, Renée Schroeder, and Mike Waterman for constructive discussions and many suggestions related to this work. We thank Tom Macke and Ivo Hofacker for technical assistance with RNAMotif and the Vienna RNA package, respectively, and Yanli Wang for preparing Fig. 1. We also thank Dave Scicchitano for supporting this study through the Department of Biology. UL thanks the entire Schlick Lab for many helpful discussions and for helping sort through a myriad of technical problems, and the Howard Hughes Medical Institute for an undergraduate summer research fellowship through the Honors Summer Institute at NYU. We also gratefully acknowledge the support of NSF (DMS-0201160), NIH (GM055164 and ES012692), and the Human Frontier Science Program (RGP0076) for supporting this research.

Table 4. Selected candidate sequences from the candidate pools for streptomycin, chloramphenicol, neomycin B, and ATP (after filtering by looking at predicted 2D structures). Included are the computed energetic data, test results, locations of the start sites of the sequences, and GenBank annotations. The corresponding RNA sequences have T replaced by U

Sequence Genome	Location	Gene	F (kcal/mol)	T_m (°C)	T_m-F Test	V	V-F Test
ATP							
5′-GCTGGTCGAA GACACTGGCT GTCGCTGTCG ACGGCGATCA GC							
Halobacterium sp. NRC1 plasmid	pNRC200 92229	non-coding	−19.43	71.6	+	4.54	+
Halobacterium sp. NRC1 plasmid	pNRC100 92229	non-coding					
STREPTOMYCIN							
5′-GTACCCGGAC GTGCCCTTCC AGGCGTCCAT GGAGGCCTGG CTCGGGGCGG TGC							
S. avermitilis	7323209	non-coding	−28.64	111.4	−	0.84	−
NEOMYCIN B							
5′-TGCGGGCGAA CAGTTTGCA							
E. coli O157:H7 EDL933	5503670	non-coding	−7.61	81.0	+	1.12	+
5′-TTGAGCAGGG GCGTGAAGTT TTTGCTTTG							
E. coli CFT073	3864206	Smf	−10.79	83.4	+	1.44	+
5′-AGTCTGGTGG GCGATATGTT TATTATGAT							
E. coli K12	1324039	yciQ	−6.10	57.8	+	2.24	−
E. coli O157:H7	1826735	ECs1840					
E. coli O157:H7 EDL933	2260441	Z2542					
E. coli CFT073	1569619	yciQ					
CHLORAMPHENICOL							
5′-TCAGAGCTGA AAAACTGGCC CCGAGTGCAG CTAAAAACTG A							
E. coli O157:H7	1532105	ECs1492	−10.58	72.6	−	2.40	−
E. coli O157:H7 EDL933	1617201	Mfd					

Appendix: Use of Graph Theory to analyze RNA 2D Structure and Function

RNA Genomics and Graph Theory

This article focused on using RNA motifs from *in vitro* selection experiments to discover novel functional RNA molecules in genomes. An alternate approach to RNA

genomics developed in our group is the use of graph theory for course-grained secondary structure modeling (see, for instance, [11, 14, 26, 35]). Indeed, the utility of the graph theory approach to RNA secondary structure has been known as early as the 1980's with work done on tree edit distances [43], RNA structure comparison [5, 28], and RNA structure statistics [12]. Graph theory analysis of genomes is promising because all RNA structures can be schematically represented as two-dimensional graphs and thus novel graph topologies from graphical enumeration can be used to drive discovery of novel RNA motifs in genomes via methods and analyses similar to those described here. Below, we outline the essentials and advantages of graph theory for describing, cataloguing, and predicting RNA structures in the hope that this will stimulate mathematicians to work in this area.

RNA Structural Motifs and Graph Theory

RNA molecules are hierarchical in nature since their secondary structures are known to be stable independently of their tertiary structures [50]. Thus, many groups approach RNA by focusing on 2D RNA structures [7, 13, 15, 17, 30, 38, 56]. RNA secondary motifs have a network-like topology with stems linking loops, bulges, and junctions (Fig. 6). Such a topological RNA representation allows exploration of RNA

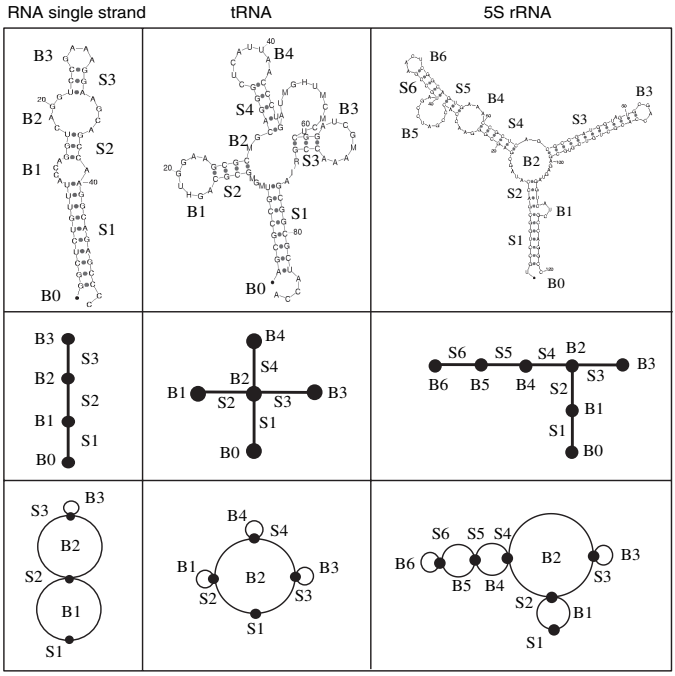

Figure 6. Graphical representations of RNA secondary structures (*top*) as tree (*middle*) and dual (*bottom*) graphs

topologies using graph theory, a field in mathematics widely used for analyzing networks and enumerating structural possibilities, including hydrocarbons, genetic and biochemical networks, ecology, transportation, and the Internet [4, 8, 18].

Figure 6 shows three RNA secondary (tree) motifs represented as tree graphs: the vertices (•) are RNA loops, bulges or junctions, and the edges (lines, —) are RNA stems (precise rules are detailed in [14]). Thus, the schematic tree graphs represent the connectivity between the RNA secondary elements (e.g., stems, loops, bulges, junctions). The tree graphs provide intuitive representations of RNA structures, but they cannot represent other important RNA types, such as pseudoknots. For completeness, we developed another class of RNA graphs called *dual graphs* (third row of Fig. 6; [14]); dual graphs can represent all RNA trees and pseudoknots and can be generalized to represent unusual RNA structures with triple, quadruple, and higher-order helices (e.g., occurring in RNA frameshift signal of HIV-1 and RNAs interacting with antibiotic neomycin [1, 2]).

Since the "RNA graphs" are discrete, they allow us to enumerate all possible 2D RNA motifs using enumeration methods of graphs. Graphical enumeration of RNA topologies can be performed analytically or computationally depending on the complexity of the structures. For example, for unlabeled trees, the number of possible graphs with V vertices is obtained from the coefficients c_i associated with the x_i term of the counting polynomial derived by Harary and Prins [20]:

$$t = \sum_i c_i x^i = x + x^2 + x^3 + 2x^4 + 3x^5 + 6x^6 + 11x^7 + 23x^8 + 47x^9 + 106x^{10} + \cdots$$

For example, there is only 1 distinct graph each for $V = 1, 2, 3$ vertices (since $c_1 = c_2 = c_3 = 1$) and 2 distinct 4-vertex graphs ($c_4 = 2$), 3 distinct 5-vertex graphs ($c_5 = 3$), and so on.

These sets of distinct graphs represent libraries of theoretically possible RNA topologies, which include naturally occurring, candidate, and hypothetical RNA motifs theory (see Schlick lab's RNA-As-Graphs (RAG) web resource at `monod.biomath.nyu.edu/rna/` and [11]). Known RNAs in public databases (NDB and others) can thus be matched to the topologies we describe (see Fig. 6). Significantly, because we found that the known 2D RNA motifs represent only a small subset of all possible topologies, we hypothesize that some of the missing motifs may represent undiscovered *naturally occurring* RNAs while others may be designed and then synthesized in the laboratory.

Sequence Space Versus Topology Space

Current theoretical [42] and experimental [53] approaches to RNA structure explore RNA's sequence space. Experimental *in vitro* selection techniques exploit random sequence pools for comprehensive searches for novel RNAs. In the search for RNA genes in genomes, scanning algorithms require sequence and structural motifs as input. In contrast, our RNA analysis focuses on structural motifs rather than sequences

per se. A critical advantage of RNA graph analysis is that the space of topologically distinct structures is vastly smaller than the nucleotide sequence space. In fact, we estimate, based on Harary-Prins enumeration formula for tree graphs (above) [19, 20], that the number of distinct RNA tree topologies can be parameterized as $\sim 2.5^{(N/20)-3}$ for $N > 60$ compared with 4^N for the nucleotide sequence space! The markedly smaller RNA topology space implies great potential for the search for novel RNA structures. Once a novel target topology/motif is identified, the corresponding RNA sequences can be found in two ways: for natural RNAs, the selected motif can be found by scanning the genomes; and, for synthetic RNAs, they can be designed using modular assembly of existing RNA fragments (i.e., using a library of

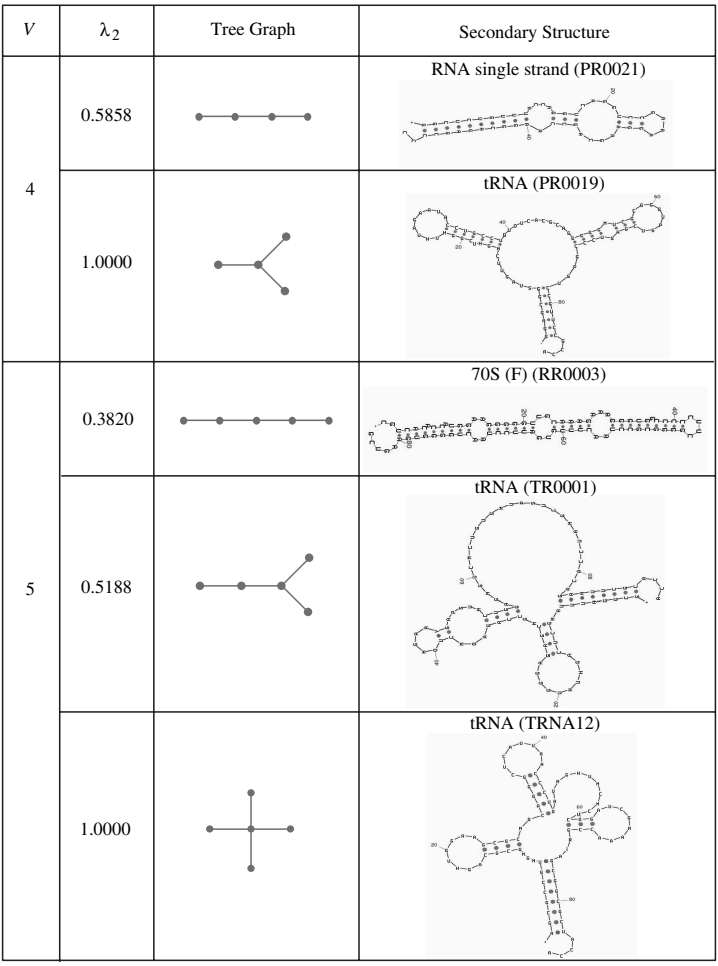

Figure 7. RNA tree motif libraries for $V = 4, 5$. V is vertex number and λ_2 is the second smallest eigenvalue of the Laplacian matrix

sequence/motif building blocks and application of 2D folding algorithms). Both of these research directions are currently being pursued in our laboratory.

RAG: RNA-As-Graphs Web Resource

Our RNA graphical representations present an opportunity for cataloguing of RNA structures based on their topological properties (Fig. 7). Cataloguing RNA's structural diversity, including hypothetical motifs, is vital for identifying novel RNA structures and for pursuing RNA genomics initiatives. Our RNA-As-Graphs (RAG; `monod.biomath.nyu.edu/rna`) web resource catalogues and ranks all mathematically possible (including existing and candidate) RNA secondary motifs on the basis of graphical enumeration results. We archive RNA tree motifs as "tree graphs" and other RNAs, including pseudoknots, as general "dual graphs." All RNA motifs are catalogued by graph vertex number (a measure of sequence length) and ranked by topological complexity (second smallest eigenvalue (λ_2) corresponding to the graph's *Laplacian* matrix). RAG's inventory immediately suggests candidates for novel RNA motifs, either naturally occurring or synthetic. Through RAG, we hope to pursue and further stimulate efforts to predict and design novel RNA motifs and thereby contribute to RNA genomics initiatives.

References

[1] D. P. Arya, R. L. Coffee, Jr., and I. Charles. Neomycin-induced hybrid triplex formation. *J. Amer. Chem. Soc.*, 123:11093–11094, 2001.

[2] D. P. Arya, R. L. Coffee, Jr., B. Willis, and A. I. Abramovitch. Aminoglycoside-nucleic acid interactions: Remarkable stabilization of DNA and RNA triple helices by neomycin. *J. Amer. Chem. Soc.*, 123:5385–5395, 2001.

[3] J. P. Bachellerie, J. Cavaille, and A. Huttenhofer. The expanding snoRNA world. *Biochimie*, 84:775–90, 2002.

[4] A. L. Barabási and E. Bonabeau. Scale-free networks. *Sci. Amer.*, 288:60–69, 2003.

[5] G. Benedetti and S. Morosetti. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biol. Chem.*, 59:179–184, 1996.

[6] D. H. Burke, D. C. Hoffman, A. Brown, M. Hansen, A. Pardi, and L. Gold. RNA aptamers to the peptidyl transferase inhibitor chloramphenicol. *Chem. Biol.*, 4:833–843, 1997.

[7] J. H. Chen, S. Y. Le, and J. V. Maizel. Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucl. Acids Res.*, 28:991–999, 2000.

[8] K. J. Devlin. *Mathematics: the New Golden Age*. Penguin, London, 1988.

[9] J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418:222–228, 2002.

[10] A. D. Ellington and J. W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.
[11] D. Fera, N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick. RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, 5:88, 2004.
[12] W. Fontana, D. A. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.
[13] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083–2099, 1993.
[14] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucl. Acids Res.*, 31:2926–2943, 2003.
[15] C. Gaspin and E. Westhof. An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J. Mol. Biol.*, 254:163–174, 1995.
[16] W. W. Gibbs. The unseen genome: Gems among the junk. *Sci. Amer.*, 289:46–53, 2003.
[17] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucl. Acids Res.*, 31:439–441, 2003.
[18] J. Gross and J. Yellen. *Graph Theory and its Applications*. CRC Press, Boca Raton, FL, 1999.
[19] F. Harary. The number of homeomorphically irreducible trees and other species. *Acta Math.*, 101:141–162, 1959.
[20] F. Harary. *Graph Theory*. Addison-Wesley, Reading, MA, 1969.
[21] T. Hermann and D. J. Patel. Adaptive recognition by nucleic acid aptamers. *Science*, 287:820–825, 2000.
[22] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994. `www.tbi.univie.ac.-at/~ivo/RNA/`
[23] A. Huttenhofer, J. Brosius, and J. P. Bachellerie. RNomics: Identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.*, 6:835–843, 2002.
[24] A. Huttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J. P. Bachellerie, and J. Brosius. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, 20:2943–2953, 2001.
[25] L. Jiang, A. Majumdar, W. Hu, T. J. Jaishree, W. Xu, and D. J. Patel. Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure Fold Des.*, 7:817–827, 1999.
[26] N. Kim, N. Shiffeldrim, H. H. Gan, and T. Schlick. Candidates for novel RNA topologies. *J. Mol. Biol.*, 341:1129–1144, 2004.

[27] J. Kitagawa, Y. Futamura, and K. Yamamoto. Analysis of the conformational energy landscape of human snRNA with a metric based on tree representation of RNA structures. *Nucl. Acids Res.*, 31:2006–2013, 2004.

[28] S. Y. Le, R. Nussinov, and J. V. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, 22:461–473, 1989.

[29] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucl. Acids Res.*, 29:4724–4735, 2001.

[30] H. Margalit, B. A. Shapiro, A. B. Oppenheim, and J. V. Maizel, Jr. Detection of common motifs in RNA secondary structure. *Nucl. Acids Res.*, 17:4829–4845, 1989.

[31] J. S. Mattick. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, 2:986–991, 2001.

[32] J. S. Mattick and M. J. Gagen. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.*, 18:1611–1630, 2001.

[33] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, and R. R. Breaker. Genetic control by a metabolite binding mRNA. *Chem. Biol.*, 9:1043–1049, 2002.

[34] Y. Okazaki, M. Furuno, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–573, 2002.

[35] S. Pasquali, H. H. Gan, and T. Schlick. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucl. Acids Res.*, 2005. In Press.

[36] N. Piganeau and R. Schroeder. Aptamer structures: A preview into regulatory pathways? *Chem. Biol.*, 10:103–104, 2003.

[37] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[38] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16:583–605, 2000.

[39] K. Salehi-Ashtiani and J. W. Szostak. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414:82–84, 2001.

[40] M. Sassanfar and J. W. Szostak. An RNA motif that binds ATP. *Nature*, 364:550–553, 1993.

[41] T. Schlick. *Molecular Modeling: An Interdisciplinary Guide*. Springer-Verlag, New York, NY, 2002.

[42] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B. Biol. Sci.*, 255:279–284, 1994.

[43] B. A. Shapiro and K. Z. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, 6:309–318, 1990.

[44] G. A. Soukup and R. R. Breaker. Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci. USA*, 96:3584–3589, 1999.

[45] G. A. Soukup and R. R. Breaker. Nucleic acid molecular switches. *Trends Biotechnol.*, 17:469–476, 1999.

[46] G. A. Soukup and R. R. Breaker. Allosteric nucleic acid catalysts. *Curr. Opin. Struct. Biol.*, 10:318–325, 2000.

[47] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Q. Rev. Biophys.*, 4:213–253, 1971.

[48] G. Storz. An expanding universe of noncoding RNAs. *Science*, 296:1260–1263, 2002.

[49] V. Tereshko, E. Skripkin, and D. J. Patel. Encapsulating streptomycin within a small 40-mer RNA. *Chem. Biol.*, 10:175–187, 2003.

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

[51] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505–510, 1990.

[52] S. T. Wallace and R. Schroeder. In vitro selection and characterization of streptomycin-binding RNAs: Recognition discrimination between antibiotics. *RNA*, 4:112–123, 1998.

[53] D. S. Wilson and J. W. Szostak. In vitro selection of functional nucleic acids. *Ann. Rev. Biochem.*, 68:611–647, 1999.

[54] W. Winkler, A. Nahvi, and R. R. Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial expression. *Nature*, 419:952–956, 2002.

[55] W. C. Winkler, S. Cohen-Chalamish, and R. R. Breaker. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. USA*, 99:15908–15913, 2002.

[56] M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In J. Barciszewski and B. F. C. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43. Klewer Academic Publishers, Dordrecht, NL, 1999.