Contributed chapter to *Computational Studies of DNA and RNA*, J. Sponer and F. Lankas editors, Springer-Verlag.

August 17, 2005

Chapter 1

RNA: The Cousin Left Behind Becomes a Star

Tamar Schlick
Department of Chemistry and Courant Institute of Mathematical Sciences,
New York University, 251 Mercer Street, New York, New York 10012.
Phone: (212) 998-3116 *Fax: (212) 995-4152* *Email: schlick@nyu.edu*

Abstract A brief introduction to RNA structure and function is offered, including recent exciting discoveries concerning RNA's starring roles in gene regulation. Challenges in RNA research are outlined, and the role of modeling and bioinformatics approaches to these problems is suggested. Applications of a graph-theory representation of RNA secondary structures to RNA analysis and design are illustrated.

> *Even after the completion of many genome sequences, both the number and diversity of ncRNA genes remain largely unknown.*[1]
>
> —Eddy, 2001

> *New evidence suggests, however, that this junk DNA may encode RNA molecules that perform a variety of regulatory functions. This new theory may explain why the structural and developmental complexity of organisms does not parallel their numbers of protein-coding genes.*[2]
>
> —Mattick, 2004

1. Introduction

While proteins are household words and DNA is an icon — in science as well as art (e.g., Ref.[3]) — their biomolecular cousin, RNA, has largely been left behind until recently. Indeed, RNA's starring role in the cell is emerging with new discoveries concerning RNA's vital regulatory roles. We now appreciate that RNA molecules are integral components of the cellular machinery for protein synthesis and transport, RNA editing, chromosome replication and regulation, catalysis and many other functions (see Table 1 for RNA's diverse roles). In this chapter, some of the current excitement in the field of RNA biology and chemistry are described, along with pressing challenges concerning RNA structure and function. Novel computational approaches, including molecular modeling and simulation and mathematical representations of secondary RNA structures, have great potential for impacting the field of RNA research, including genome-wide initiatives concerning RNA structure and function, or *ribonomics*. Such applications are illustrated for RNA structure analysis and design using graph theory representation of RNA secondary structures.

Table 0.1. Some classes of non-coding RNA (ncRNA).

RNA	Function
transfer RNA (tRNA)	protein synthesis
ribosomal RNA (rRNA)	protein synthesis
small nucleolar RNA (snoRNA)	rRNA modification
micro RNA (miRNA)	translation regulation
transfer-messenger RNA (tmRNA)	protein stability in ribosome
telomerase RNA	replication
guide RNA (gRNA)	mRNA editing
spliced leader RNA (SL RNA)	mRNA trans-splicing
small nuclear RNA (snRNA)	RNA splicing
hammerhead ribozyme	self-cleavage
hepatitis delta virus ribozyme	self-cleavage
Group I intron	self-splicing
Group II intron	self-splicing
RNase P	pre-tRNA processing
23S rRNA	peptide bond formation

2. RNA at Atomic Resolution

RNA is a single-stranded polynucleotide chain which can fold upon itself to form double-stranded segments stabilized by complementary hydrogen bonds such as adenine with uracil, cytosine with guanine, as well as thermodynamically stable guanine-uracil wobble pairs. These folded structures are imperfect due to non-complementary base pairs and unpaired bases and thus form bulges, loops, junctions, and other motifs, as shown in Figure 1.1, stabilized by various stacking interactions, hydrogen bonding, and intramolecular networks between distant regions in the linear sequence[4].

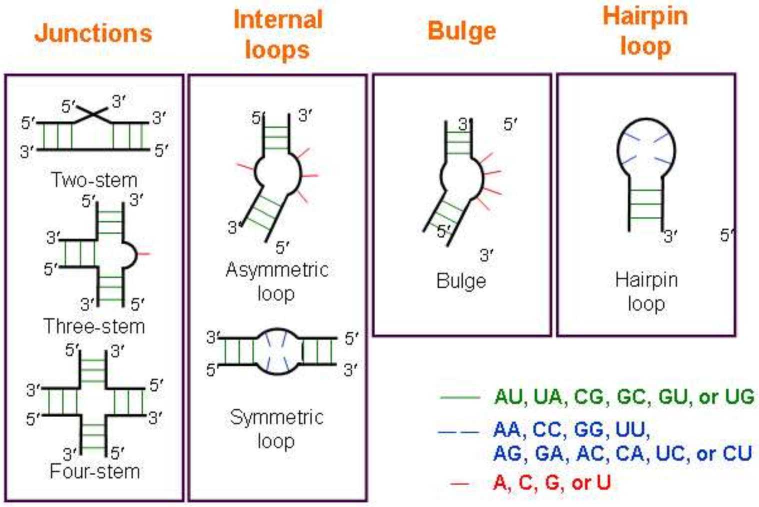


Figure 1.1. RNA Secondary Structural Motifs.

The motifs shown in Figure 1.1 are known as *secondary structures*; these can lead to compact and complex *tertiary interactions*, as shown in Figure 1.2 for the hammerhead ribozyme[5] and Figure 1.3 for the hepatitis delta virus (HDV) ribozyme[6]. The latter RNA reveals a common feature of RNA that can be considered as a super-secondary structural element: a *pseudoknot*. RNA pseudoknots have a stretch of nucleotides within a hairpin loop that pairs with nucleotides external to that loop. In other words, hydrogen bonding occurs between alternating regions (e.g., **a** with **c** and **b** with **d** to produce an intertwined geometry), as illustrated in Figure 1.4.

The clover-leaf structure of the tRNA molecule has been known for over 25 years, and for a long time was the only well-characterized major structure of an RNA molecule; see Ref.[7] for a perspective following the high-resolution determination of yeast phenylalanine tRNA in 2000. However, with vast improvements in crystallization procedures for RNA (e.g., RNA structure determination through crystallization with a protein that would not interfere with the enzyme's activity[8]), as well as alternative approaches for studying RNAs such as high-resolution NMR,

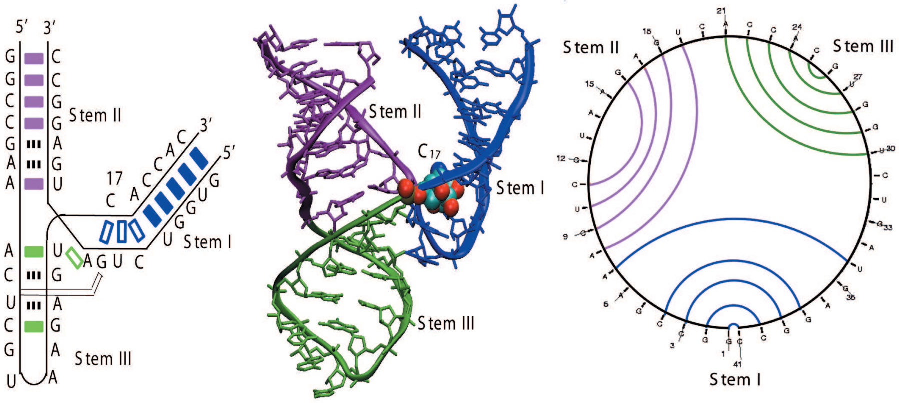


Figure 1.2. Hammerhead Ribozyme Secondary and Tertiary Structures: two types of base-pairing diagrams (left and right) illustrate the secondary structure, and the middle illustration shows the folded, three-dimensional configuration.

spectroscopy, cross-linking relations and phylogenic data analysis, our knowledge of RNA structure has increased dramatically.

The high-resolution ribosome structures have added dramatically to our library of known RNA structures[9,10,11,12,13,14,15,16,17]. The ribosome is the cell's protein synthesis factory, a complex of many proteins and several RNA molecules, which are folded as many stable secondary motifs; the ribosome's small and large subunits cooperate tightly to coordinate the interplay between tRNA, mRNA, and proteins in the process of protein synthesis and catalyze the peptide bond formation. As of Summer 2005, there are about 825 known structures of RNA in the public databases (see http://www.rcsb.org), but many entries are duplicates of the same molecule or motif.

3. RNA's Diversity

The wonderful capacity of RNA to form complex, stable tertiary structures has been exploited by evolution.

Traditionally, RNA is known for its key role as mediator between the agent of heredity — DNA — and the cell's workhorses — proteins (see Figure 1.5). For example, tRNA molecules carry amino acids and deposit them in the correct order, mRNAs mediate translation of the hereditary information from DNA into protein, rRNAs are involved in protein biosynthesis (within a complex of ribosomal RNA and numerous proteins). However, work in the 1980s established that RNA, like protein, can act as a catalyst in living cells. Thomas Cech and Sidney Altman received the 1989 Nobel Prize for chemistry for this discovery of

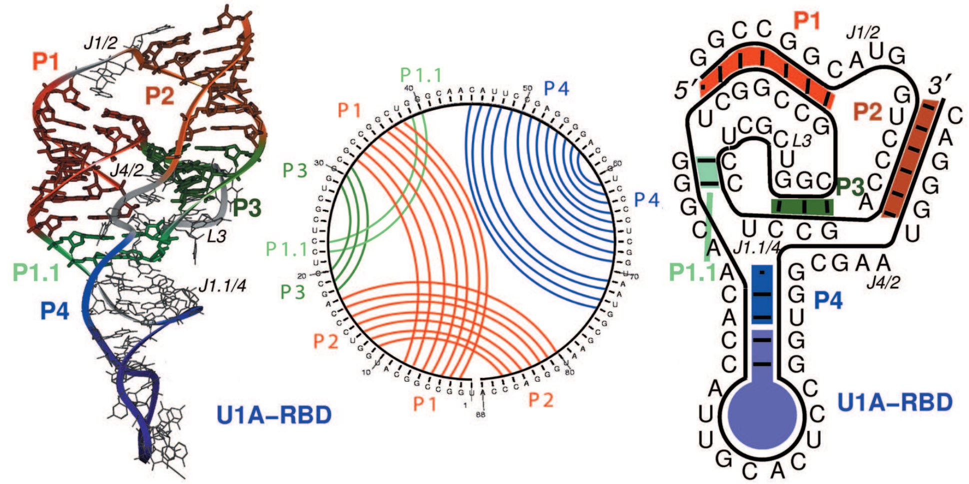


Figure 1.3. Hepatitis Delta Virus (HDV) ribozyme structure. The secondary and tertiary structures are shown, along with the pseudoknot details. Two pseudoknots are present: between regions **P1** and **P2** (red) and **P1** and **P3** (green). This intertwined base-pairing is evident from the middle image, which shows crossings in the paired bases (compare to the hammerhead ribozyme in Figure 1.2, which shows no crossings).

RNA enzymes or *ribozymes*. More than 500 ribozyme types have been found in a diverse range of organisms. Many ribozymes make or break phosphodiester bonds in nucleic acid backbones, but other biological and chemical functions are continuously being discovered.

Ribozymes have also been designed (e.g., Refs.[18,19,20]), as spare in composition as two base building-block units (rather than four)[21]. In fact, the 83-nucleotide ribozyme composed only of two different building blocks — uracil and 2,6-diaminopurine — was shown to catalyze the ligation of two RNA molecules with a rate 36,000 times faster than the uncatalyzed reaction[21]. That RNA's genetic code may be simpler than today's four bases lends further support to the "RNA world" hypothesis.

4. Recent Discoveries Concerning RNA's Starring Role

Exciting recent discoveries concerning RNA came in the end of 2002, when a high-quality draft sequence of the mouse genome was published and analyzed (see the 5 December 2002 issue of *Nature*, volume 420). The 2.5-billion size of the mouse genome is slightly smaller than the human genome (3 billion base pairs in length), and the number of estimated mouse genes, around 30,000, is roughly similar to the number approx-

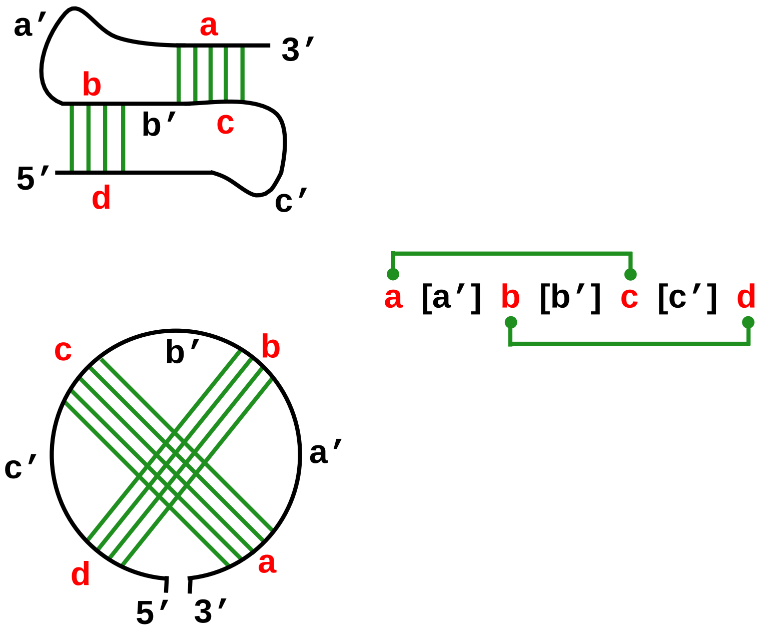


Figure 1.4. RNA 3D Folds Involve Pseudoknots: RNA Pseudoknots have an intertwined form of base pairing, which can be evident from a circular representation of base pairing, as shown in Figure 1.2 (no pseudoknot) and Figure 1.3 (2 pseudoknots).

imated for humans. Intriguingly, the various analyses reported in December 2002 revealed that only a small percentage (1%) of the mouse's genes has no obvious human counterparts. This similarity makes the mouse genome an excellent organism for studying human diseases and novel treatments. But the obvious dissimilarity between mice and men and women also begs for further comparative investigations: why are we not more like mice? Part of this question may be explained through an understanding of how mouse and human genes might be regulated differently.

Aside from complex *networks* of genes rather than *single* genes that are responsible for controlling phenotypes, another factor for this difference between humans and mice traits is related to control of gene activation and function by a novel class of genes called *RNA genes* — RNA transcripts that do not code for proteins (*ncRNA* for non-coding RNAs). These genes have essential regulatory functions that may play significant roles in each organism's survival[22,23,24,25]. In fact, scientists are amazed at how RNA's critical activities have eluded them so long. As early as 1961, suggestions that RNA can control gene activity were mentioned in Jacob and Monod's classic paper[26]; but only recently have some of these mechanisms and immense applications been discovered[2].

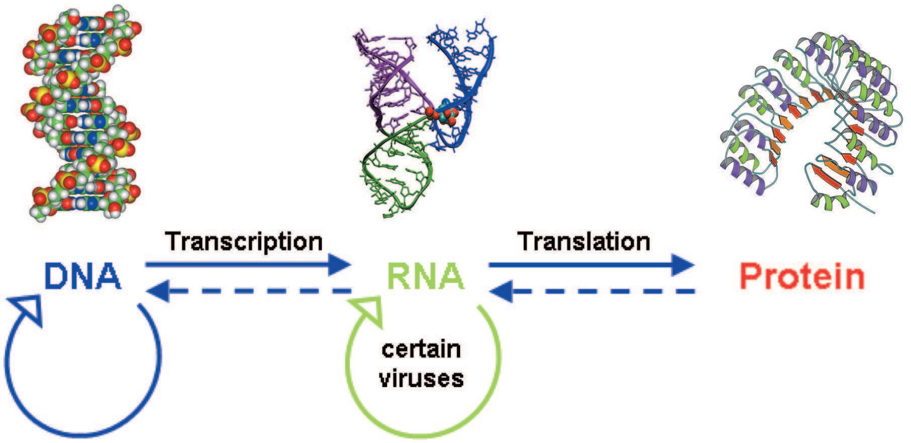


Figure 1.5. The Central Dogma of Biology. In the traditional flow of information, DNA makes RNA through transcription and RNA makes protein in translation. However, discoveries that RNA can act as an enzyme, and can carry actual instructions for protein synthesis (other than via the RNA code) enrich RNA's cellular roles dramatically.

Such newly found roles for RNAs, especially concerning tiny RNAs that do not encode proteins (e.g., micro-RNAs) but can influence gene action, won DNA's cousin the venerable trophy of "Breakthrough of The Year" by *Science* editors in 2002 (see the 20 December 2002 issue of *Science*, volume 298). Non-protein coding stretches of mRNAs range in size from only 20 nucleotides to over 10000 nucleotides[27]; they are required to control the translation from the mRNA transcript into protein.

The 2002 award recognized a large group of papers that unraveled various fascinating features of small RNAs (affectionately termed *nanoRNAs*). Micro-RNAs (miRNAs), generally 21 to 25 nucleotides long, form a regulatory class of ncRNAs[28]. These RNAs, encoded in genomes, control gene expression by repressing translation of target genes through, for example, binding to $3'$ untranslated regions of the messenger-RNA targets or by destroying the messenger-RNA targets. Thus, regions in the genome which were previously denoted as "junk-DNA" may actually define a gold-mine of biological information.

Such small RNAs in animals, plants, and fungi are collectively termed RNAi (for RNA interference). The agents that initiate RNAi in a sequence-specific manner are double-stranded RNA segments (siRNAs, for small interfering RNAs) that silence gene expression. For example, they may seek out messenger RNA (corresponding to them) and destroy it, or they may bind to chromatin and/or modify chromatin structure[29,30,31]. Though initially regarded as anomalies, work is revealing that such siRNAs regulate gene expression in a variety of organisms. SiRNAs can also be synthetic, designed to target specific genes; they have therapeutic applications.

Such interference mechanisms by RNA silencing through chromatin structure can provide an organism a natural defense against invading viruses and transposons (DNA segments that migrate within and across organisms and are associated with bacterial pathogenicity)[32]. Consequently, this natural protection is being exploited by scientists using siRNAs to target viral genes that can inhibit the replication of HIV-1, polio, or other viruses (e.g., Ref.[33]). Moreover, RNA interference mechanisms are being investigated by several companies who are applying them to discover the functions of genes by turning them off to determine the effect on the plant or the animal. For example, a landmark study on obesity employed RNA interference[34] to inactivate about 85% of the roundworm's predicted 19,757 genes that code for proteins in a single experiment[35]. Results of the experiment helped identify the genes that play a role in an organism's tendency for obesity.

These fascinating discoveries regarding RNA's interference with gene activity are associated with many *epigenetic* phenomena — changes in gene expression that do not involve alterations in the genome and persist across at least one generation.

A different kind of epigenetic control was also discovered by Breaker and coworkers[36,37] in bacterial messenger RNAs containing sequences that sense small molecules directly to control translation of mRNA into protein. Namely, specific control regions of mRNA can bind directly to metabolites, such as associated with vitamin B synthesis and import, and induce a conformational change in RNA's folding state; this metabolite-triggered conformational change acts as part of the signal transduction pathway that senses vitamin level and controls enzyme production. Conformational switches have also shown to be important for the catalytic activity of the hepatitis delta virus (HDV) ribozyme[38].

Such a switch of RNA conformation between two states in a ligand-dependent manner (a *riboswitch*) also opens new avenues for thinking about RNA design in a variety of contexts[39]. Like the *Paracelsus challenge* for proteins[40], one can formulate a similar challenge for RNA design: describe minimal changes in the nucleotide sequence to trigger a conformational rearrangement in the folding of a given RNA molecule. Such a challenge in the RNA world will be appropriately approached by a combination of computational and experimental wizardry. In fact, *Science* editor Jennifer Couzin[41] exclaims: *"Having exposed RNAs' hidden talents, scientists now hope to put them to work."*.

Already, numerous applications can be envisioned for RNAs as regulators of gene expression, therapeutic agents, molecular switches, and molecular sensors. This is because therapeutic agents could be designed to exploit RNA's functional sites as potential drug targets[42]. The construction of ligands that bind RNA and interfere with protein synthesis, transcription, or viral replication may lead to new antibiotic/antiviral drugs, for example. Together with the design of novel RNAs and of RNA sequences called *aptamers* — RNAs that are *apt* to bind to specific molecular targets or perform desired catalytic functions by design (see Refs.[43,44] for example) — RNA offers a great molecular machinery with potential benefits to biomedicine and nanotechnology. Many such novel synthetic RNAs have been found by random RNA sequence pool ("in vitro") experiments[27,45,46,47,48]. With rapidly growing interest in RNA structure and function and its applications in biomedical research, enhancing the repertoire of both natural and synthetic RNA is a central goal.

5. Major Challenges in RNA Research

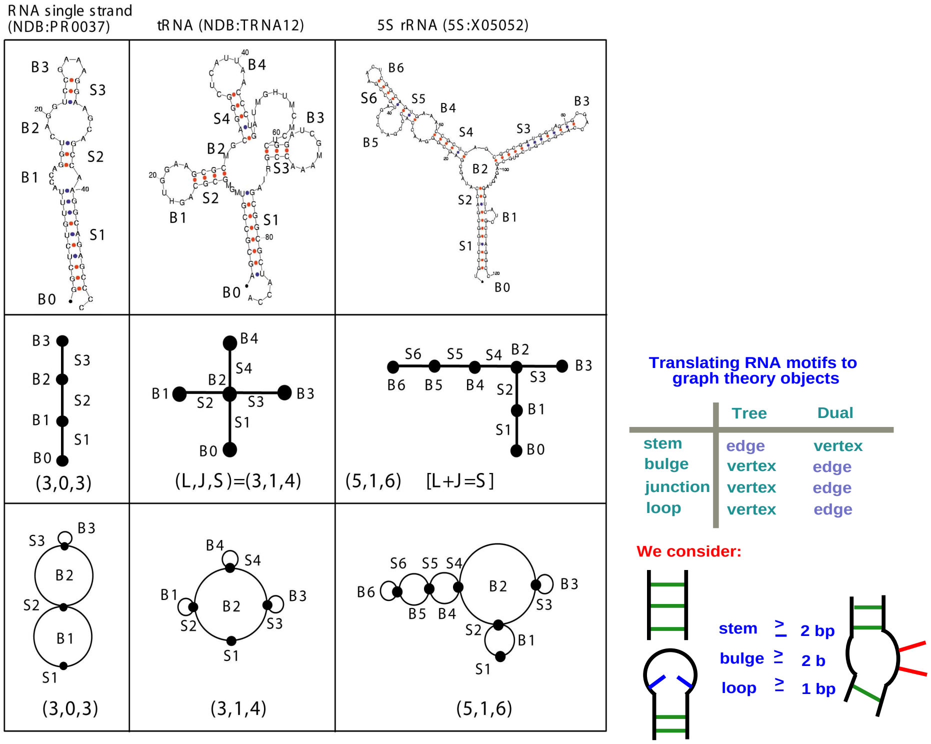
At least five key challenges concerning RNA naturally arise: finding novel RNA genes, identifying the biological roles of these RNA genes, determining the structural repertoire of RNA, determining RNA tertiary folds from sequence, and designing novel RNAs.

5.1 RNA Gene Location

Identifying locations of RNA genes in genome sequences is much more difficult than protein genes, since the start and stop codons for protein transcripts do not apply. Thus, searching for RNA genes in intron and intergenic regions, which comprise over 90% of the genomes, presents a challenge (see Ref.[49] for example). Current programs like tRNAscan-SE, FAStRNA, and Snoscan for identifying RNA genes are based on existing sequences of functional RNA, conservation of RNA secondary structure in identified sequences, and comparative genome analysis[50,51,52]. However, these programs often lead to many false identifications and are not successful for non-conserved ncRNA sequences and as high-throughput discoveries. Most methods instead rely on biochemical and molecular biology techniques. In the case of mRNAs, their associated functions are inferred from complementary target mRNAs where possible[53].

5.2 RNA Gene Function

Once potential sequences that correspond to RNA genes are identified, experiments are required to verify the expression of these tran-



	Tree	Dual
stem	edge	vertex
bulge	vertex	edge
junction	vertex	edge
loop	vertex	edge

Figure 1.6. Graphical representations of RNA secondary structures as trees (middle row) and dual graphs (bottom row).

scripts in cells. Typically, genes are assayed by Northern blotting and microarray techniques[54], and real-time PCR is also used to verify RNA transcripts. Comparative genome analysis to indicate conservation of secondary structures also aids in the verification process of RNA gene candidates. True functional characterization is a laborious and time-consuming process, and thus predictions by modeling and computation need to be as reliable or as discriminating as possible.

5.3 RNA's Structural Repertoire

Defining the structural repertoire of RNA involves delineating all possible folded motifs of functional RNAs. For proteins, catalogues of known folds are collected in many databases, such as SCOP, CATH, PROSITE, or PFAM, and organized by classes, folds, superfamilies, families, and domains (see illustrations in Ref.[55], Chapter 4). The enormous interest in the *protein folding problem* has led to structural genomics initiatives that seek to define and design new protein sequences that will fold into novel motifs. Funding opportunities from the National Institutes of Health have already led to important discoveries in this area[56,57].

However, experience has also taught structural genomicists that Nature is tricky. While some pairs of disparate sequences lead to similar folds, despite expectations to be different, other sequence pairs thought to lead to similar architectures produce new unanticipated folds. These cause-and-effect patterns are likely explained by various scenarios of structural evolution of protein active sites. Besides protein folding and protein structural genomics, scientists in the protein field are also speculating on the total number of existing protein folds; this number is likely in the range of several thousands. Discussions on to how best to employ experimental technology and computation to find all Nature's folds have led to concentrated initiatives by large teams of structural genomicists.

In this regard, the RNA field lags behind considerably. The number of known RNA folds is an order-of-magnitude less than that available for proteins (e.g., 825 vs. $\sim$ 25,000 as of Summer 2005), and the best way to organize RNA motifs is not known. Possibly, the number of functional RNA motifs may grow with sequence size rather than approach a limit.

The application of graph theory to describe secondary structural motifs of RNA (e.g., Refs.[58,59,60]) holds great promise in this area of RNA structure analysis. Together with many reports of newly discovered RNA motifs, our group's RNA-As-Graphs database (RAG) (http://monod.biomath.nyu.edu/rna/rna.html)[61,62,63] suggests that the currently-known RNAs represent only a *small fraction* of the possible stable and functional RNAs.

5.4 The RNA Folding Problem

Our graph theory application involves defining RNA secondary structures as two types of graphical objects: trees and dual graphs (see Figure 1.6)[64]. These classic representations, coupled with linear algebra tools, such as the Laplacian second eigenvalue corresponding to a graph (see Figure 1.7), and graph enumeration theorems allow us to enumerate the possible RNA secondary motifs (see Figure 1.8 and also Figure 1.7, bottom). The color coding (red/blue/black) in the library segment distinguishes motifs corresponding to existing RNAs (red) from those that are hypothetical and RNA-like (blue) and hypothetical but non-RNA-like (black). The latter classification of the hypothetical motifs has been accomplished by statistical clustering methods[63]. In fact, this clustering suggested ten novel RNA motifs (Figure 1.9a and 1.9b) which contain sub-components corresponding to existing RNA motifs. This allowed us to use build-up procedures in combination with existing 2D folding

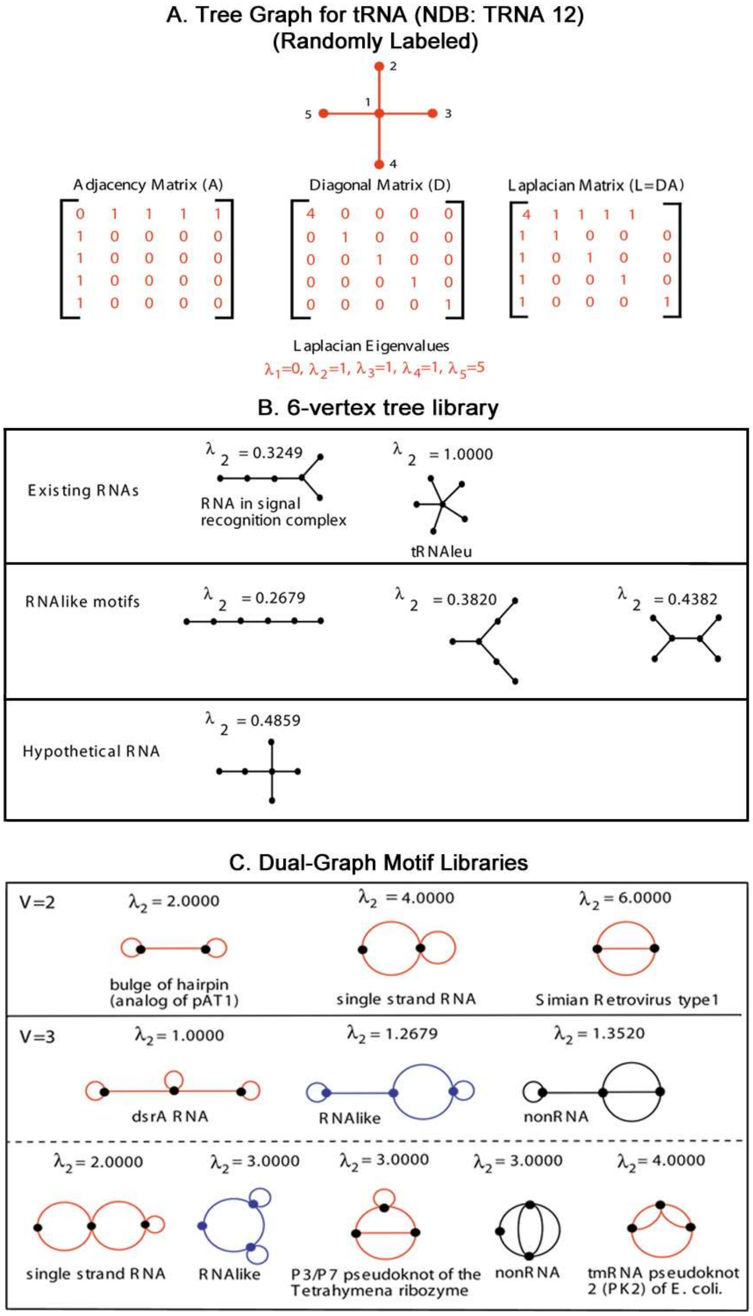


Figure 1.7. Top: Quantifying the tRNA graph using Laplacian matrix and eigenvalues. Middle and bottom: Tree (B) and dual graph (C) motif library segments from the RAG database. They are categorized into functional RNA, RNA-like, and hypothetical RNA (black graphs in C) using motif clustering. RNA-like tree and dual graphs can be used to search for novel ncRNAs.

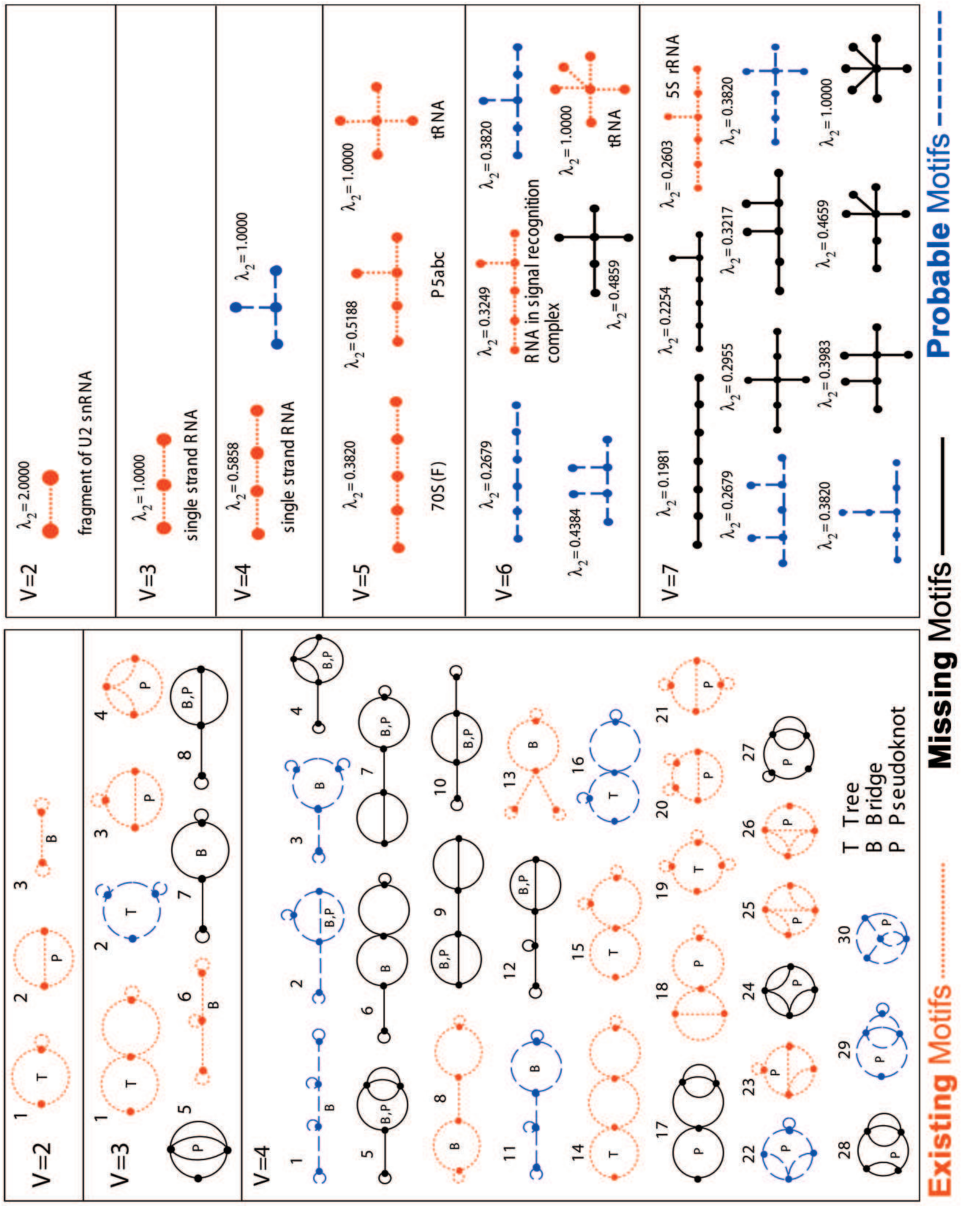


Figure 1.8. Segments of the RAG (http://monod.biomath.nyu.edu/rna) library for dual graphs (bottom) and tree graphs (top).

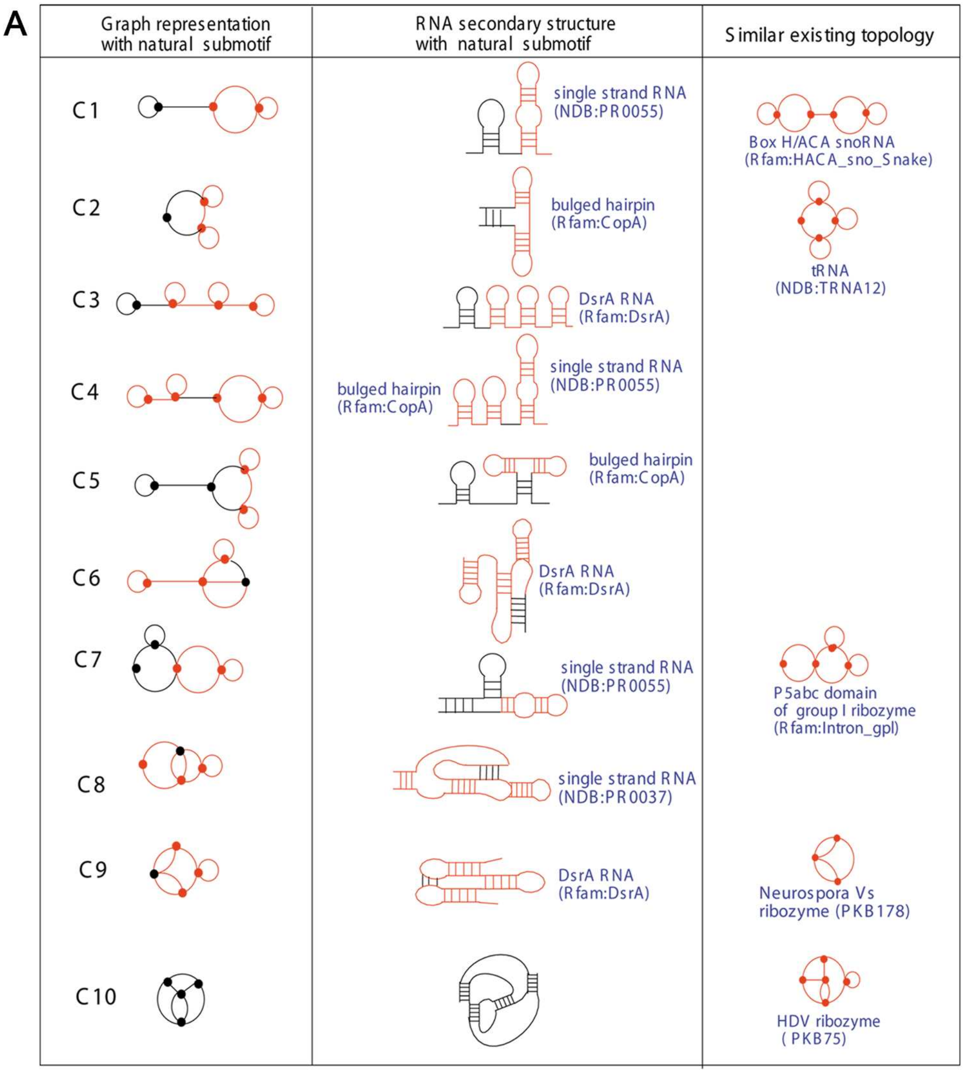

A
Graph representation with natural submotif
RNA secondary structure with natural submotif
Similar existing topology
C1
single strand RNA (NDB:PR0055)
Box H/ACA snoRNA (Rfam:HACA_sno_Snake)
C2
bulged hairpin (Rfam:CopA)
tRNA (NDB:TRNA12)
C3
DsrA RNA (Rfam:DsrA)
C4
bulged hairpin (Rfam:CopA)
single strand RNA (NDB:PR0055)
C5
bulged hairpin (Rfam:CopA)
C6
DsrA RNA (Rfam:DsrA)
C7
single strand RNA (NDB:PR0055)
P5abc domain of group I ribozyme (Rfam:Intron_gpI)
C8
single strand RNA (NDB:PR0037)
C9
DsrA RNA (Rfam:DsrA)
Neurospora Vs ribozyme (PKB178)
C10
HDV ribozyme (PKB75)

Figure 1.9a. Ten examples of predicted novel RNA-like (dual graph) topologies (1st column, labeled C1, ...,C10) shown with their secondary structures (2nd column) and natural submotifs (red lines) that occur in known RNAs. The sequences of those submotifs are used in a build-up procedure to generate candidate sequences for motifs C1 through C10, as shown in Figure 1.9b[63].

B

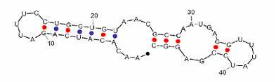
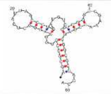
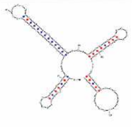
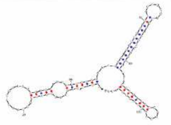
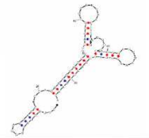
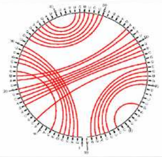
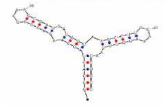
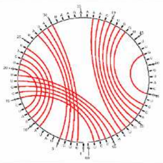
ID	Designed sequence	Novel RNA structure
C1	AACACAUCAGAUUUCCUGGUGUAA CGCCAAUGAGGUUUAUCCGAGGC	
C2	AGCGCCGUGGCAGGGCUCAUAACC CUGAUGUCCUCGGAUCGAAACCGA GCGGCGCUACCA	
C3	AACACUCAGAUUUCCUGGUGUAAC GAAUUUUUUAAGUGCUUCUUGCUU AAGCAAGUUUCUACCCGACCCCCU CAGGGUCGGGAUUUUGGACCUCCA UGACGUUAUGGUCC	
C4	AACACUCAGAUUGGACCUCAUGAC GUUAUGGUCCUUCCUGGUGUAACG AAUUUUUUAAGUGCUUCUUGCUUA AGCAAGUUUCUACCCGACCCCCUC AGGGUCGGGAUUU	
C5	CCUGGUAUUGCAGUACCUCCAGGU AGCGCCGUGGCAGGGCUCAU AACC CUGAUGUCCUCGGAUCGAAACCG AGCGGCGCUACCA	
C6	AGACCGUCAAACACAGACUAAAUGU CGGUCGGGGAAGAUGUAUUCUUCU CAUAAGAUAUAGUCGGCCUGGUAU UGCAGUACCUCCAGGU	
C7	GGCAGUACCAAGUCGCGAAAGCGA UGAUGGUAAGCCUUGCAAAGGGUU AAGCUGCC	
C8	not yet found	
C9	CUUCUUAUAUGAUUAGGUUGUCAU UUAGAAUAAGAAAACCUGGUAUUG CAGUACCUCCAGGUUAACCUG	
C10	not yet found	

Figure 1.9b. Designed candidate sequences that "fold" into the target RNA-like motifs (see Figure 1.9a) using a modular assembly approach where fragments from existing RNAs are assembled and folded[63]. Note that motifs C6 and C9 are pseudoknots.

packages to propose candidate sequences that will generate these target RNA motifs (Figure 1.7)[63]. Such cataloging of RNA structures, both existing and hypothetical, and their applications are important for the development of the RNA field on a large scale, *ribonomics.*

While the protein folding problem (of course to *us*, not to Nature...) has received much attention — due to the enormous intellectual challenge, not to speak of practical applications to drug design — there is an analogous problem in RNA. In fact, because RNA folding is hierarchical, with secondary-structural elements forming first and then forming tertiary interactions independently, deducing RNA tertiary folds from the primary sequence might be simpler[65,66]. Unfortunately, only a tiny fraction of the number of protein folding aficionados are addressing the analogous challenge for RNA.

In this context, a challenge in RNA folding is to understand how strong electrostatic repulsions between closely packed phosphates in RNA are alleviated. Indeed, the stability of compact RNA forms is strongly maintained through interactions with both monovalent and divalent cations and by pseudoknotting.

Predicting the secondary and tertiary folding of RNA is a difficult and ongoing enterprise[67,68,69]. Various 2D algorithms like MFOLD[70] predict the patterns of base pairing, the presence of base pair mismatches, and regions of unpaired bases (loops, bulges, junctions, etc.). Other programs have been developed[71,72,73,74]. Secondary structural elements are easier to identify through modeling combined with evolutionary and database relationships[75,76]. Though imperfect, especially for long RNAs, these predictions provide opportunities for learning what works, as well as what fails, in structure prediction for RNA.

Emerging themes in RNA structure include the importance of metal ions and loops for structural stability, various groove binding motifs, architectural motifs tailored for intermolecular interactions[77], hierarchical folding, fast establishment of secondary structural elements, and extreme flexibility of the molecule as a whole[78,69,6,66].

At present, relatively successful algorithms are available to predict secondary structure of RNA molecules up around 200nt by calculating the most energetically favorable base-pairing schemes. However, discriminating among the possible tertiary interactions to obtain the final folded state remains a challenge[66].

A direct measurement of the complete folding pathway of the *Tetrahymena* ribozyme suggests that thetertiary structure of the P4–P6 domain forms cooperatively within three seconds, but several minutes are required for complete folding of the catalytic center of the enzyme[79]. The folding pathways of large functional RNAs may also take minutes or longer.

Still, findings concerning the folding kinetics of *Tetrahymena* ribozyme [79] have suggested that, as thermodynamic data on tertiary structure interactions become available[66], the RNA folding problem might be easier to solve than protein folding[65,66]. Therefore, with advances in RNA synthesis and structure determination[80] and in the availability of thermodynamic data on tertiary interactions[66], it is likely that our understanding of RNA structure, RNA folding, and RNA's role in enzyme evolution will dramatically increase in the coming decade. These advances will undoubtedly propel *ribonomics*[81] and RNA design (e.g., Refs.[20,18,46,48]).

5.5 Designing Novel RNAs

Because of the many potential applications of RNA in biomedicine and technology[82], designing novel functional RNAs is a promising enterprise.

Typically, *in vitro* selection experiments are used to explore systematically the ability of large sequence pools of nucleic-acid building blocks to form RNAs with desired function[47,45,46,48]. Essentially, huge pools (of order 10^{15}) of substrate nucleotides are mixed in special apparati and amplified by PCR (polymerase chain reaction) to enhance the success of producing functional molecules with respect to desired function (catalysis) or binding activity. The process is iterated many times, thereby mimicking an evolutionary process by which the "fittest molecules" (i.e., those with high binding affinities for ligands) survive. While such *in vitro* technology has produced RNA *aptamers* — synthetic RNAs that bind to target biomolecules, antibiotics, or viruses — the success rate is quite low. This is because the sequence space 4^N, where N is the number of nucleotides in the RNA sequence, has a very low function to information ratio[83]. In other words, only a small fraction of the theoretically possible RNA sequences leads to actual functional RNAs. Thus, random sequence pools tend to produce a biased pattern of resulting products, and these may not encourage structural diversity[84,85].

Thus, enhanced technology using targeted sequence pools can potentially improve this yield. Novel RNAs could also be designed by novel computational techniques under development[63,74,50,54,51]. Such techniques may involve structured RNA libraries[86,87]. Recent work has also shown that ribozyme engineering[88] and molecular design of RNAs by combining self-folding building blocks and optimizing connecting regions has great promise[89,90,63]. New mathematical tricks to analyze *in vitro* pools (e.g., Ref.[85]) and additional engineering tools to describe targeted libraries may prove fruitful as complementary techniques.

6. Invitation to Computational Biologists

Given the exciting new discoveries concerning RNA's starring role in gene regulation and the endless possibilities of engineering and utilizing RNA structure for medicine and technology, it is not difficult to understand the sharp rise of experimental efforts to discover and exploit RNA's diversity. Indeed, numerous companies focusing on RNA technology and its biomedical applications, for example for antibiotic and antiviral agents.

At the same time, numerous new opportunities become available to computational scientists who can apply and develop tools in dynamic programming, molecular modeling, cluster analysis, statistics, and network theory to pressing problems in RNA structure and function.

For example, work in the RAG group has shown that graph theory can be used to catalogue and enumerate RNA motifs[62,61]; identify RNA submotifs within larger RNAs and find new structural and functional similarity between pairs of RNAs[62,91] (see Figure 1.10); predict and design novel RNAs by a build-up procedure that targets the most RNA-like motifs among all the candidate, hypothetical RNAs[63]; enhance *in vitro* selection by a targeted design approach[85]; and search for small functional RNAs in genomes[92,93]. These applications rely on the critical advantages of graph theory: the much smaller motif space (enumerated via graphical objects) compared to the sequence space available for RNA and the automation that such compact mathematical descriptions allow. These critical differences allow exhaustive analysis of the RNA pattern space, subject to the usual modeling limitations.

For now, predicting the *tertiary* folds of RNA, even given the secondary structure, lurks at a distance, but there is no doubt that the dedicated RNA community will make great advances in this challenge too in the coming decade.

Acknowledgments

I thank Carse Ramos, Melanie Hill-Cantey, Yanli Wang, and Karunesh Arora for their excellent technical assistance and Hin Hark Gan for his critical reading of this chapter and contribution to Table 1. This work is supported by Human Frontier Science Program (HFSP), by the National Institute of Health (NIGMS award R01-GM055164-05), and by NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160).

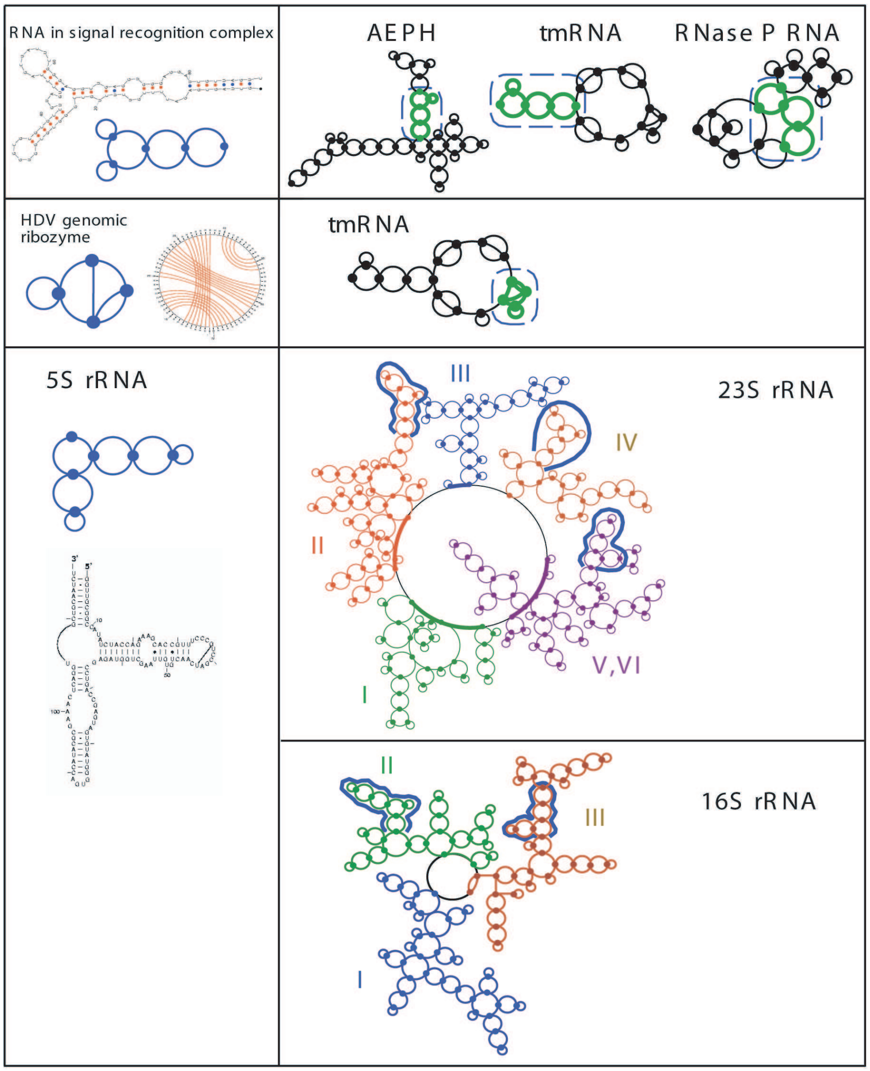


Figure 1.10. RNA topologies within larger RNAs. The search for submotifs can be performed automatically using the concept of graph isomorphism[91].

Bibliography

[1] Eddy, S. R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2:919–929.

[2] Mattick, J. S. 2004. The hidden genetic program of complex organisms. *Sci. Amer.* 291:61–67.

[3] Schlick, T. 2005. The critical collaboration between art and science: *Applying an Experiment on a Bird in an Air Pump to the Ramifications of Genomics on Society. Leonardo* 38:323–329.

[4] Moore, P. B. 1999. Structural motifs in RNA. *Ann. Rev. Biochem.* 68:287–300.

[5] Scott, W. G., J. T. Finch, and A. Klug. 1995. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell* 81:991–1002.

[6] Ferré-D'Amaré, A. R., and J. A. Doudna. 1999. RNA folds: Insights from recent crystal structures. *Ann. Rev. Biophys. Biomol. Struc.* 28:57–73.

[7] Shi, H., and P. Moore. 2000. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* 6:1091–1105.

[8] Ferré-D'Amaré, A. R., K. Zhou, and J. A. Doudna. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature* 395:567–574.

[9] Cate, J. H., M. M. Yusupov, C. Z. Yusupova, T. N. Earnest, and H. F. Noller. 1999. X-ray crystal structure of 70S ribosome functional complexes. *Science* 285:2095–2104.

[10] Cech, T. R. 2000. The ribosome is a ribozyme. *Science* 289:878–879.

[11] Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289:905–920.

[12] Nissen, P., J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science* 289:920–930.

[13] Williamson, J. R. 2000. Small subunit, big science. *Nature* 407:306–307.

[14] Wimberly, B. T., D. E. Brodersen, W. M. Clemons Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. 2000. Structure of the 30S ribosomal subunit. *Nature* 407:327–339.

[15] Carter, A. P., W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. 2000. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* 407:340–348.

[16] Schluenzen, F., A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F.Franceschi, and A. Yonath. 2000. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* 102:615–623.

[17] Yusupov, M. M., G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. D. Cate, and H. F. Noller. 2001. Crystal structure of the ribosome at 5.5 Å resolution. *Science* 292:883–896.

[18] Schultes, E. A., and D. B. Bartel. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* 289:448–452.

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

[20] Tang, J., and R. R. Breaker. 2001. Structural diversity of self-cleaving ribozymes. *Proc. Natl. Acad. Sci. USA* 97:5784–5789.

[21] Read, J. S., and G. F. Joyce. 2002. A ribozyme composed of only two different nucleotides. *Nature* 420:841–844.

[22] Lau, N. C., and D. P. Bartel. 2003. Censors of the genome. *Sci. Amer.* 289:34–41.

[23] Gregory, R. I., and R. Shiekhattar. 2005. MicroRNA biogenesis and cancer. *Cancer Res.* 65:3509–3512.

[24] Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferando,

J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. 2005. MicroRNA expression profiles classify human cancers. *Nature* 435:834–838.

[25] McManus, M. T. 2003. MicroRNAs and cancer. *Semin. Cancer Biol.* 13:253–258.

[26] Jacob, F., and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.

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

[28] Carrington, J. C., and V. Ambros. 2003. Role of microRNAs in plant and animal development. *Science* 301:336–338.

[29] Plasterk, R. H. A. 2002. RNA silencing: The genome's immune system. *Science* 296:1263–1265.

[30] Zamore, P. D. 2002. Ancient pathways programmed by small RNAs. *Science* 296:1265–1269.

[31] Felsenfeld, G., and M. Groudine. 2003. Controlling the double helix. *Nature* 421:448–453.

[32] Ahlquist, P. 2002. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* 296:1270–1273.

[33] Li, H., W. X. Li, and S. W. Ding. 2002. Induction and suppression of RNA silencing by an animal virus. *Science* 296:1319–1321.

[34] Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. L. Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen, and J. Ahringer. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421:231–237.

[35] Ashrafi, K., F. Y. Chang, J. L. Watts, A. G. Fraser, R. S. Kamath, J. Ahringer, and G. Ruvkun. 2003. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* 421:268–272.

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

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

[38] Ke, A., K. Zhou, F. Ding, J. Cate, and J. Doudna. 2004. A conformational swithc controls hepatitis delta virus ribozyme catalysis. *Nature* 429:201–205.

[39] Szostak, J. W. 2002. RNA gets a grip on translation. *Nature* 419:890–891.

[40] Rose, G. 1997. Protein folding and the Paracelsus challenge. *Nature Struc. Biol.* 4:512–514.

[41] Couzin, J. 2002. Small RNAs make big splash. *Science* 298:2296–2297.

[42] Pearson, N. D., and C. D. Prescott. 1997. RNA as a drug target. *Chem. & Biol.* 4:409–414.

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

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

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

[46] Ellington, A. D., and J. W. Szostak. 1990. *In Vitro* selection of RNA molecules that bind specific ligands. *Nature* 346:818–822.

[47] Jäschke, A. 2001. Artificial ribozymes and deoxyribozymes. *Curr. Opin. Struct. Biol.* 11:321–326.

[48] Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–570.

[49] Hershberg, R., S. Altuvia, and H. Margalit. 2003. A survey of small RNA-encoding genes in Escherichia Coli. *Nucl. Acids Res.* 31:1813–1820.

[50] Rivas, E., R. Klein, T. Jones, and S. Eddy. 2001. Computational identification of noncoding RNAs in E-coli by comparative genomics. *Curr. Biol.* 11:1369–1373.

[51] Carter, R., I. Dubchak, and S. Holbrook. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.* 29:3928–3938.

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

[53] Rhoades, M., B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel. 2002. Prediction of plant microRNA targets. *Cell* 110:513–520.

[54] Wassarman, K., F. Repoila, C. Rosenow, G. Storz, and S. Gottesman. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15:1637–1651.

[55] Schlick, T. 2002. Molecular Modeling and Simulation: An Interdisciplinary Guide. Springer-Verlag, New York, NY.

[56] Vitkup, D., E. Melamud, J. Moult, and C. Sander. 2001. Completeness in structural genomics. *Nat. Struc. Biol.* 8:559–565.

[57] Burley, S., and J. Bonanno. 2004. Structural genomics. *Methods Biochem. Anal.* 44:591–612.

[58] Waterman, M., and T. Smith. 1978. RNA secondary structure: A complete mathematical analysis. *Math. Biosci.* 42:257–266.

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

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

[61] Fera, D., N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, N. Kim, and T. Schlick. 2004. RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* 5:88.

[62] Gan, H., D. Fera, J. Zorn, M. Tang, N. Shiffieldrim, U. Laserson, N. Kim, and T. Schlick. 2004. RAG: RNA-As-Graphics database – concepts, analysis, and features. *Bioinformatics* 20:1285–1291.

[63] Kim, N., N. Shiffeldrim, H. Gan, and T. Schlick. 2004. Candidates for novel RNA topologies. *J. Mol. Biol.* 341:1129–1144.

[64] Gan, H. H., S. Pasquali, and T. Schlick. 2003. A survey of existing RNAs using graph theory with implications to RNA analysis and design. *Nucl. Acids Res.* 31:2926–2943.

[65] Batey, R. T., and J. A. Doudna. 1998. The parallel universe of RNA folding. *Nature Struc. Biol.* 5:337–340.

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

[67] Pyle, A. M., and J. B. Green. 1995. RNA folding. *Curr. Opin. Struct. Biol.* 5:303–310.

[68] Schuster, P., P. F. Stadler, and A. Renner. 1997. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol.* 7:229–235.

[69] Brion, P., and E. Westhof. 1997. Hierarchy and dynamics of RNA folding. *Ann. Rev. Biophys. Biomol. Struc.* 26:113–137.

[70] Zuker, M. M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.* 31:3406. http://www.bioinfo.rpi.edu/∼zukerm.

[71] McCaskill, J. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–1119.

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

[73] Xayaphoummine, A., T. Bucher, F. Thalmann, and H. Isambert. 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA* 100:15310–15315.

[74] Andronescu, M., A. Fejas, F. Hutter, H. Hoos, and A. Condon. 2004. A new algorithm for RNA secondary structure design. *J. Mol. Biol.* 336:607–624.

[75] Zuker, M., and P. Stiegler. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* 9:133–148.

[76] Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–940.

[77] Hermann, T., and D. J. Patel. 2000. Adaptive recognition by nucleic acid aptamers. *Science* 287:820–825.

[78] Hagerman, P. J. 1997. Flexibility of RNA. *Ann. Rev. Biophys. Biomol. Struc.* 26:139–156.

[79] Sclavi, B., M. Sullivan, M. R. Chance, M. Brenowitz, and S. A. Woodson. 1998. RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* 279:1940–1943.

[80] Holbrook, S. R., and S.-H. Kim. 1997. RNA crystallography. *Biopolymers* 44:3–21.

[81] Doudna, J. A. 2000. Structural genomics of RNA. *Nature Struc. Biol.* 7:954–956. (Structural Genomics Supplement).

[82] Puerta-Fernández, E., C. Romero-López, A. Barroso-delJesus, and A. Berzal-Herranz. 2003. Ribozymes: Recent advances in the development of RNA tools. *FEMS Microbiol. Rev.* 27:75–97.

[83] Szostak, J. 2003. Molecular messages. *Nature* 423:689.

[84] Carothers, J., S. Oestreich, J. Davis, and J. Szostak. 2004. Informational complexity and functional activity of RNA structures. *J. Amer. Chem. Soc.* 126:5130–5137.

[85] Gevertz, J., H. Gan, and T. Schlick. 2005. *In vitro* random rools are not structurally diverse: a computational analysis. *RNA* 11:853–863.

[86] Davis, J. H., and J. W. Szostak. 2002. Isolation of high-affinity gtp aptamers from partially structured RNA libraries. *Proc. Natl. Acad. Sci.* 99:11616–11621.

[87] Lau, M. W., K. E. Cadieux, and P. J. Unrau. 2004. Isolation of fast purine nucleotide synthase ribozymes. *J. Amer. Chem. Soc.* 126:15686–15693.

[88] Cech, T. 1992. Ribozyme engineering. *Curr. Opin. Struct. Biol.* 2:605–609.

[89] Ikawa, Y., K. Fukada, S. Watanabe, S. Shiraishi, and T. Inoue. 2002. Design, construction, and analysis of a novel class of self-folding RNA. *Structure* 10:527–534.

[90] Chworos, A., I. Severcan, A. Koyfman, P. Weinkam, E. Oroudjev, H. Hansma, and L. Jaeger. 2004. Building programmable jigsaw puzzles with RNA. *Science* 306:2068–2072.

[91] Pasquali, S., H. Gan, and T. Schlick. 2005. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucl. Acids Res.* 33:1384–1398.

[92] Laserson, U., H. Gan, and T. Schlick. 2004. Searching for 2D RNA geometries in bacterial genomes. *In* Proceedings of the Twentieth Annual ACM Symposium on Computational Geometry, June 9–11. ACM Press, 373–377.

[93] Laserson, U., H. Gan, and T. Schlick. 2005. Exploring the connection between synthetic and natural RNAs in genomes via a novel computational approach. *In New Algorithms for Macromolecular Simulation*, Proceedings of the Fourth International Workshop on Algorithms for Macromolecular Modelling, Leicester, UK, August 2004, C. Chipot, R. Elber, A. Laaksonen, B. Leimkuhler, A. Mark, T. Schlick, R. D. Skeel, and C. Schuette, editors, volume 49 of *Lecture Notes in Computational Science and Engineering.* Springer-Verlag, New York.