30
Comparative Analysis of RNA Secondary Structure: 6S RNA

James W. Brown and J. Christopher Ellis

30.1 Introduction

The purpose of this chapter is to provide a "primer" on the comparative analysis of RNA secondary structure. The emphasis here is on the initial stages of the analysis; in other words, how one goes about creating a working model of the secondary structure de novo using the comparative approach. This is a common scenario; you, a student or coworker have discovered that an RNA is involved in a biological system under investigation. The sequence of the RNA is determined, usually either from cDNA or from the gene. Or perhaps it is discovered that a region of a messenger RNA or viral RNA is important in some process and it is suspected that the structure of this region is critical for that function. You are interested, then, it obtaining information about the structure of this RNA in order to help guide experiments and to organize data about the RNA. The determination of the three-dimensional (3-D) structure of the RNA is unlikely to be cost-effective or feasible (certainly not as a first step), but you correctly realize that the single most thermodynamically favorable predicted secondary structure is not going to suffice. How, then, to proceed? Usually, the answer is by creating a secondary structure model based on comparative sequence analysis. The detailed analysis of very-high-resolution secondary structure, the identification and evaluation of tertiary interactions, and the construction of 3-D models based on comparative analysis will not be considered here; these aspects of comparative analysis of RNA structure require specialized experience. Comparative analysis, like X-ray crystallography, is as much art as science, but the creation of a basic secondary structure is well within the range of a newcomer to the "RNA World", the target audience for this chapter.

The approach taken here is to follow the construction of a basic secondary structure of an example RNA: 6S RNA. The 6S RNA was discovered in *Escherichia coli* in 1971 [1], but its function remained unknown until very recently [2]. 6S RNA is not essential for viability [3], but accumulates during the stationary phase, binds directly and specifically to RNA polymerase, and regulates RNA polymerase function in a growth stage-specific manner [2]. The secondary structure of the 6S RNA has not been examined in any detail; the existing secondary structure proposed for this
RNA was based on a comparison of only the *E. coli* and *Pseudomonas aeruginosa* sequences [4, 5].

30.1.1

**RNA Secondary Structure**

What is an RNA secondary structure? Although most researchers would agree on simple definitions of primary structure (sequence) and tertiary structure (3-D coordinates), there is a surprising extent of disagreement about *exactly* what RNA secondary structure is, even (perhaps especially!) among established RNA researchers that work with secondary structures on a daily basis [6]. At its most basic, however, a secondary structure is a list of adjacent, antiparallel Watson–Crick (or G•U) base pairs in an RNA chain; these are the pairings for which the rules are clear and that are readily predicted by comparative sequence analysis. Uncertainty about what exactly is “secondary structure” deal primarily with the distinction between secondary and tertiary interactions. For example, are non-Watson–Crick base pairs other than G•U included? Are isolated base pairs included? What about helical stacks? In the case of a pseudo-knot, which helices are considered secondary and which, if any, are tertiary? All of these are subject to some level of disagreement. It is also worth remembering that “secondary” does not mean 2-D; secondary structures contain a plethora of 3-D information, beginning with the presumption that the helices are generally A-form in structure. However, in the comparative analysis of secondary structure of an RNA, the basic definition of secondary structure is generally most useful.

Secondary structure can be represented in a variety of ways, but is most often presented as a string of letters, the sequence, twisted around on a page (i.e. in two dimensions) such that these antiparallel adjacent interactions can be shown as dashes between each pair of bases. By formal convention, G•U pairs are shown with a hollow dot instead of a dash and non-Watson–Crick pairings with a closed dot, such as G•A [7]. Typically (tRNA is the exception, here) structures are drawn to flow generally clockwise 5' to 3'. A convenient way to specify whether or not there is specific evidence for a base pairing is to only put in the dash (or dot or circle) if there is such evidence.

RNA secondary structure is very specific and highly defined; secondary structure is the central organizing principle in RNA structure. This is a fundamental difference between protein and RNA (and DNA secondary structure, of course). Pragmatically, experiments are almost always developed and results represented in the context of the secondary structure of an RNA.

30.1.2

**Comparative Sequence Analysis**

Comparative sequence analysis is the process of extracting information about a macromolecule (in this case RNA) from the similarities and differences between
different, but homologous, sequences (for review, see [8–11]). The underlying assumption is that the higher order structure of the molecule is more highly conserved than is the sequence; in other words, the sequence is free to change during evolution as long as the 3-D structure is generally maintained. In terms of secondary structure, this means that changes in the identity of a base involved in a pairing should generally be allowed by a compensatory change in its pairing partner so that the ability of the two to form isosteric base pairs is retained. The two bases that pair then vary together, or covary. The work involved in the construction of a secondary structure of an RNA by comparative analysis is primarily the search for these sequence covariations. If sufficient numbers of sequences are available, these covariations can be identified statistically directly from a sequence alignment [12, 13]. Comparative analysis, then, is an iterative process in which improvements in the alignment result in additional structural information, which can be used in turn to improve the alignment. Although attempts have been made to automate this process (see, e.g. [14–16]), with varying levels of success, in practice this is generally still a manual process.

30.1.3

Strengths and Weakness of Comparative Analysis

Comparative analysis is the “Gold Standard” method for determining secondary structure of RNAs; computational methods for predicting secondary structure are typically validated by comparison with “true” secondary structures as determined by comparative analysis (see, e.g. [17, 18]). However, other methods for determining secondary structure can be very useful supplements to comparative analysis or serve as last resort alternatives if comparative analysis is not feasible, e.g. if few or only one sequence is available for analysis.

A particularly useful supplement to comparative analysis is the genetic analysis of mutation and second-site compensatory mutation; in fact, these methods are formally equivalent, the difference being whether you create the variations or observe them in nature. This method is typically laborious and so has not been used generally as an alternative to comparative analysis, but can be especially useful either to confirm the presence of a particular feature of secondary structure (see, e.g. [19]) or to probe secondary structure than cannot be assessed by the comparative method, such as pairings involving invariant sequences. For instance, the 6S RNA secondary structure used as an example of comparative analysis in this chapter contains a stem–loop in which none of the base pairing are specifically supported by sequence covariations; the paired sequences are invariant among the sequences available. An alternative to obtaining additional 6S RNA sequences in hope of finding covariations in this potential stem–loop would be to make point mutations in this region of the RNA in E. coli that affect the function of the molecule and then make the compensatory change. If the RNA with two substitutions, such that the potential base pairing is maintained, functions better than the RNA with either single substitution that disrupts the potential pairing, then the pairing is presumed
to be legitimate. Genetic analysis has also been used in the absence of comparative data in cases where only a single instance of the functional RNA is known, such as the delta virus ribozyme [20].

Another useful supplement to comparative analysis, as we will see, is the prediction of structures thermodynamically. This is, in reality, where secondary structure modeling usually begins. These predictions are steadily improving, especially with the ability to predict a variety of structures near the minimum free energy and assess the frequency that particular base pairings are predicted in these collections of structures [21–23]. Thermodynamic predictions are routinely used to predict the structures of idiosyncratic elements of structure that appear as insertions in specific instances of an RNA. The danger of thermodynamic prediction is the tendency to consider these structures endings rather than beginnings. A measure of the success of thermodynamic prediction is that the predicted lowest free energy structures contain, on average, about 73% of base pairs that would exist in a “true” secondary structure determined by comparative analysis [17].

The last commonly used method for assessing secondary structure in RNA is chemical and enzymatic probing. Although these methods have been used extensively in attempt to determine structure, their utility is mostly in the examination of changes or differences in structure that result from mutation, binding to other molecules, and the like. Chemical and enzymatic probing data are notoriously difficult to judge directly in terms of the secondary structure of the RNA.

30.1.4
Comparison with Other Methods

Comparative analysis is similar to, but more sensitive than, genetic experiments because natural selective pressure is more sensitive than our biochemical or genetic methods. Comparative analysis past the initial stages is objective, quantitative and conceptually automatable. Given sufficient numbers of variable sequences, a secondary structure can be very high resolution, in which every base pair is assessed individually. Only biologically relevant base pairings are identified by comparative analysis. Nevertheless, there are limitations to comparative analysis. The most important of these is that no structure can be assessed in the absence of sequence variation; as a result, the most important aspects of structure, those comprised of the most highly conserved sequences, are the most difficult to prove by comparative analysis. The initial stage of a comparative analysis, the subject of this chapter, is basically a manual process. No specific information is provided about unique sequences that cannot be meaningfully aligned. Although tertiary interactions can also be detected by comparative analysis (although this typically required large collections of sequences), only base-base interactions in which more than one isosteric possibility is structurally acceptable will be detected. Nevertheless, comparative analysis is certainly the method of choice whenever possible. The list of structures determined definitively by comparative analysis is nearly as long as the list of known RNA types: large and small subunit ribosomal RNAs [24], transfer RNAs [25], RNase P [26] and MRP [27] RNAs, SRP RNA [28],
tmRNA [29], group I [30] and II [31] introns, nuclear splicing RNAs (e.g. [32]), H/ACA [33] and box C/D [34] snoRNAs, telomerase RNA [35], etc.

30.2 Description

30.2.1 Collecting Sequence Data

The raw material needed to determine the secondary structure of an RNA by comparative analysis is sequence data; more specifically, what is needed is a collection of different, but functional and homologous sequences. There are two ready sources for sequences: nature and GenBank [36, 37]. The first step, then, is to mine the available databases for homologous sequences. Very often there are sufficient numbers of suitable sequences available for the generation of at least an initial secondary structure. If this is not the case or if a higher-resolution secondary structure is desired, it will be necessary to obtain additional sequences experimentally.

A variety of approaches are needed to identify as many homologous sequences in GenBank as possible. A good starting point is to search the GenBank using BLAST [38] with your sequence of interest. In our example, the E. coli 6S RNA sequence (X01238) returned a number of other sequence records containing the E. coli 6S RNA:

AE016766.1 Escherichia coli CFT073 section 12 of 18 of the complete genome
X01238.1 E. coli 6S ribosomal RNA
AE005521 Escherichia coli O157:H7 EDL933 genome, contig 3 of 3, section 140 of
AE000374 Escherichia coli K12 MG1655 section 264 of 400 of the complete genome
U28377.1 Escherichia coli K-12 genome; approximately 65 to 68 minutes
M12965.1 E.coli ssr gene encoding 6S RNA
AP002563.1 Escherichia coli O157:H7 DNA, complete genome, section 14/20

These are all identical to the original sequence and so of no use to us. Note that this need not be the case; for some RNAs there may be useful variants in different strains of the same species. Other sequences obtained in this search were:

AE016988.1 Shigella flexneri 2a str. 2457T section 11 of 16 of the complete genome
AE015303.1 Shigella flexneri 2a str. 301 section 266 of 412
of the complete
AE016844.1 Salmonella enterica subsp. enterica serovar
Typhi Ty2, section 11 of
AL627277.1 Salmonella enterica serovar Typhi (Salmonella
typhi) strain CT18
AE008840.1 Salmonella typhimurium LT2, section 144 of 220
of the complete genome
AE008841.1 Salmonella typhimurium LT2, section 145 of 220
of the complete genome
AE013931.1 Yersinia pestis KIM section 331 of 415 of the
complete genome
AJ414145.1 Yersinia pestis strain CO92 complete genome;
segment 5/20

The sequences from the two Shigella flexneri strains are identical, as are those of
the two strains of Salmonella enterica and Yersinia pestis. The Salmonella typhimu-
rium sequences represent the same sequence from the genome sequence, split in
two by the separation of sections 144 and 145 of the genome record. This is fre-
quently the case for RNA-encoding genes because genome sequences are divided
into sections with an eye toward larger “intergenic” regions (spaces between open
reading frames) that often turn out to be RNA-encoding genes. It is simply a mat-
ter of extracting the two fragments of sequence and merging them. For most of the
other sequences, the entire sequence can be extracted simply by cutting and past-
ing from the BLAST results page. Sometimes, however, it is necessary to go to the
original sequence record. For example, in the case of the Y. pestis sequence, the 3’
end of the sequence is different enough from that of E. coli that it was not returned
in the BLAST alignment and had to be retrieved from the original.

Additional sequences can often be identified by repeating the BLAST searches
with the sequences identified in the initial search. In the case of our example, how-
ever, a search using the most disparate sequence identified so far, that of Yersinia
pestis, yielded the same list of sequences.

Another obvious approach is to search using the name of the RNA, but unlike
protein-encoding genes, RNA-encoding genes (except those of rRNAs and tRNAs)
are often not annotated even in genome sequences. Using “6S RNA” as the search
term for our example locates the sequence from E. coli (X01238), all of the se-
quencies listed above and that of P. aeruginosa (Y00334). However, already we see
one of the weaknesses of relying on sequence annotations; the E. coli 6S RNA
is misannotated as the “E. coli 6S ribosomal RNA”! In addition, a number of
“6S RNA” sequence annotations are typographical errors where “16S RNA” was
meant. A number of other matches are spurious because of the presence of the
term “6S” in strain or clone names, enzyme name (e.g. the “6Fe–6S prismane
cluster-containing protein”) or other RNAs with the same names (it seems there
are different “6S RNAs” in vertebrates and in λ). Annotations must always be scru-
tinized critically.
Nevertheless, the identification of the annotated 6S RNA sequence from *P. aeruginosa* provides a fresh avenue for the search; a BLAST search using this sequence identified a homolog in the *Pseudomonas syringae* genome (AE016875) (as well as several instances of the *P. aeruginosa* sequence, of course). In addition, a weak match to the *P. aeruginosa* 6S RNA sequence was found in the *Pasteurella multocida* genome (AE006208); this region of the genome sequence was extracted and used, in turn, in a BLAST search that identified a homologous sequence is *Haemophilus influenzae* Rd (U32767). In the cases of both of these sequences, the ends of the RNA are not obvious from sequence similarity and so a generous amount for sequence was taken from either end.

The 6S RNA is encoded by the *ssrS* gene and has in the past been referred to as the "ssr RNA" [3]; a search of the GenBank using these terms did not identify any additional sequences.

A number of complete genome sequences are available for organisms that are related to those for which 6S RNA sequences had been identified, but in which homologous sequences had not been found in general BLAST searches of the GenBank. The genomes of all of the *γ* proteobacteria in which the 6S RNA had not yet been found were then searched individually from the NCBI genome-specific web pages in the hope of extracting additional sequences. (Phylogenetic information about these organisms can be found on the Taxonomy Browser at http://www.ncbi.nlm.nih.gov/Taxonomy/ [36]). The *Pseudomonas putida* KT2440 (NC_002947) 6S RNA sequence was identified in a search using the *P. aeruginosa* sequence as the query; it is perhaps surprising that this sequence failed to be identified in the original search of the entire GenBank, but this is not unusual. More surprising still is that no 6S-like sequences could be identified in the complete genomes of *Vibrio cholerae*, *Vibrio parahaemolyticus* or *Vibrio vulnificus* using the sequence from the closely related *E. coli* as the query. Nor could a 6S-like sequence be identified in the *Haemophilus ducreyi* complete genome sequence using the *H. influenzae* sequence as query.

Another source of sequences are secondary databases, such as, in this case, the Small RNA Database (http://mbcr.bcm.tmc.edu/smallRNA/) [39], the Noncoding RNAs Database (http://biobases.ibch.poznan.pl/ncRNA/) [40] or the Washington University Rfam Database (http://rfam.wustl.edu/index.html) [41]. The first two of these include only the *E. coli* and *P. aeruginosa* sequences, but the Washington University Rfam site contains an alignment of 6S-like sequences from a number of bacterial genomes, including three that were not found in our previous searches: *Shewanella oneidensis* (AE015522), *V. vulnificus* (AE016802) and *V. cholerae* (AE004317). Using these sequences in turn to search the global GenBank and individual genome sequences using BLAST yielded a sequence from the *Vibrio parahaemolyticus* genome using the *V. vulnificus* (but surprisingly not the *V. cholerae*) sequence as the query.

At this point, 14 presumptive 6S RNA sequences have been identified and extracted from the GenBank. These sequences range from nearly identical (those of *E. coli* and *S. flexneri* differ by only 1 nt) to less than 50% identical; a reasonable collection to begin a comparative analysis. It is important to have a wide range
of sequence variation. The closely related sequences are useful because they are readily aligned and allow the initial identification of structure in the most variable parts of the RNA, but provide no useful information in the conservative regions of the sequence. The distantly related sequences are needed (often at later stages of the analysis) as a source of sequence variation for the analysis of the conservative (and therefore most important) regions of the RNA.

If additional sequences are needed, either because homologous sequences cannot be found by mining sequence databases or to increase the resolution of a secondary structure based on available sequences, they will have to be obtained experimentally. PCR amplification is typically used to obtain these sequences, but because the sequences flanking the gene are unlikely to be conserved, primers for amplification most often are within the gene itself and so only partial sequences are obtained. Although partial sequences have been very useful in comparative analyses of RNA structure, the entire sequence can usually be obtained using a variety of technologies available in “kit” form. It is important to note that the primer target sequences at either end of a PCR product should not be used in a comparative analysis; these sequences are derived from the primers, not the target.

A particularly useful approach to collecting large numbers of sequences quickly has been the use of PCR amplification from DNA extracted from complex microbial natural populations, rather than pure cultures [42]. The amplification products are populations of sequences, and so must be separated by cloning, but hundreds of sequences can be obtained in a single experiment. The species from which any particular sequence originates is unknown, but this information is unnecessary for the purposes of comparative analysis; all that matters is that the sequence is a valid sequence. In any case, the phylosite of the sequence itself can be determined after the fact by the construction of phylogenetic trees based on the final sequence alignment.

30.2.2

Thermodynamic Predictions

It is useful, early in the process, to have the thermodynamic predictions of the structures of all of the RNA sequences in the collection. These are generated using mfold, most conveniently using the Mfold Web Server [43] at http://www.bioinfo. rpi.edu/applications/mfold/old/rna/form1.cgi. For the purposes of initial comparative analysis, the default settings should suffice for most RNAs of reasonable

Fig. 30.1. Potential structures of the E. coli 6S RNA predicted thermodynamically. These are all of the structures predicted by mfold using the default parameters; in particular, only structures within 5% of the minimum free energy were allowed and a window parameter (which defines how dissimilar two structures must be to be considered distinct) of 10 was used. Structures from left to right are from most to least favorable, respectively. Any base pairings in the helices identified in the comparative analysis (see Fig. 30.4) are boxed. Structures were downloaded from the mfold server [43] as connect (.ct) files and displayed using LoopDloop [55].
length. If an unmanageable number of structures are predicted, the window parameter can be increased. If only one or two structures are generated, increase the percent suboptimality parameter to 10. The predicted structures can be downloaded as images for printing, but also download and print the energy table; this represents all of the predicted suboptimal foldings. Consistencies in these folding predictions among the different RNA sequences provide a starting point for comparative analysis.

In the case of the 6S RNAs, mfold consistently predicts pairing of the middle regions (roughly bases 60–130) of the RNAs in a stem–loop, and the two ends (the first and last around 20 nt) of the RNA as a terminal helix (see Figs 30.1 and 30.2 for the predicted E. coli 6S RNA structures). The interior of this extended stem–loop structure is less consistently predicted. The most common alternatives for the central region of the RNA (between the consistently predicted terminal stem and the medial stem–loop) are base-pairing across this internal region such that the entire RNA would form an extended irregular hairpin or the presence of local stem–loops on either side of the “conserved” central stem–loop. A stem–loop on the 3′ side (position around 130–150) is predicted more frequently and the placement of this predicted stem–loop is more consistent than predictions on the 5′ side.

30.2.3
Initial Alignment

A comparative analysis requires that the homologous sequences be aligned; in fact, it is the continuous building and refinement of the alignment that drives the structure analysis. Comparative analysis is an iterative process; additions to or improvements in the alignment result in additional structural information that, in turn, allows the alignment to be refined and provides insight required to add increasingly distantly related sequences to the analysis.

The first step, of course, is to collect all of these sequences into a sequence alignment editor. A variety of alignment editors are available for various computer platforms and many of them are freely available from the authors. For Windows/PC computers, a particularly useful alignment editor and analysis program, available at no cost, is BioEdit (http://www.mbio.ncsu.edu/BioEdit/). Most commercial DNA manipulation and analysis software packages include an alignment editor. Because you will most often be adding sequences by extracting them from larger (often much larger) sequence records, it is usually most convenient to move them to the alignment editor by simply cutting and pasting. Retyping sequences manually, although it might seem to be a small task, is a last resort; any hand-typed sequences will need to be painstakingly checked and rechecked for errors.

Once the sequences are all added to the alignment editor, they will need to be aligned preliminarily. If the sequences are all fairly closely related, this might be easily done “by eye”, but generally one would use an automated method, CLUSTAL [44] being the most common method incorporated into most alignment editors. Note that this is your initial alignment, not your final alignment! Much of
Fold of Eco 6S RNA at 37 C.

deltaG in Plot File = 3.8 kcal/mole

Lower Triangle Shows Optimal Energy

Upper Triangle Basepairs Plotted: 183

Optimal energy:
-81.5 < energy <= -80.2
-80.2 < energy <= -79.0
-79.0 < energy <= -77.7

Fig. 30.2. Energy table of the E. coli 6S RNA from mfold. This represents the same structures shown in Fig. 30.1. The x- and y-axis each represent the sequence in the 5' to 3' direction. Each "dot" indicates a predicted base pairing; the single best predicted structure (the left-hand-most structure in Fig. 30.1) is shown below the diagonal, all of the predicted structures are shown above the diagonal. Helices identified in the comparative analysis (see Fig. 30.4) are boxed above the diagonal. This energy table was generated by the mfold server [22, 43].
the work of a comparative analysis is the iterative improvement of the alignment. Even a novice can usually scan through a preliminary CLUSTAL alignment and find room for improvement. There is a fundamental difference between protein sequence alignments, which are generally based only on some maximizing measure of similarity between all pairs of sequences, and RNA alignments, that are based on the higher-order structure of the molecules. Ultimately, of course, the goal of any sequence alignment is to have homologous residues in alignment, but protein alignments attempt to achieve this by maximizing sequence similarity, because the richness of amino acid variation provides substantially more information on which to base an alignment than do the 4 bases in nucleic acid alignments. On the other hand, protein secondary structure is less informative than the highly organized secondary structures of RNA, which are based on one-to-one interactions between bases, and so RNA alignments are more easily based directly on higher-order structure.

Before proceeding further, it is important to arrange the sequences phylogenetically within the alignment (see Fig. 30.3). The NCBI Taxonomy Browser website (http://www.ncbi.nlm.nih.gov/Taxonomy/) is a useful guide to general phylogenetic relationships. In our example, the 6S RNAs of *E. coli* and *S. flexneri* are nearly identical; they should therefore be adjacent in the alignment, as should the two sequences from *Salmonella* species (*S. typhimurium* and *S. enterica*), *Vibrio* species, and *Pseudomonas* species. *E. coli*, *S. flexneri* and *Salmonella* species form a larger cluster, and so should be brought together; likewise, all of the sequences from the enteric bacteria (the species just mentioned and *Vibrio* species) should be clustered. *Haemophilus* and *Pasteurella* are relatives, and so these two sequences belong together as well. It is convenient to have our prototype sequence, that of *E. coli*, at the top of the alignment, with increasingly distant sequences arranged downwards.

### 30.2.4 Terminal Helix (P1a)

Pairing of the sequences near the 5′ and 3′ ends to form a terminal helix is a common element of RNA structure, and is a good starting point in the construction of the secondary structure of an RNA. Assuming that the ends of at least one example of the RNA of interest has been determined experimentally, the identification of a terminal helix allows the prediction of the location of the ends of the remaining RNAs in the alignment. In the case of our example, the 6S RNA, a terminal helix is also consistently predicted thermodynamically (Figs 30.1 and 30.2). In fact, all of

---

**Fig. 30.3.** Alignment of 6S RNA sequences following comparative analysis. Sequences are ordered phylogenetically (see text) and the alignment (not any particular sequence) is numbered at the top. The base pairing identified by comparative analysis are defined using parentheses in the last line. The structure is also shown diagrammatically at the bottom; the upstream (5′) and downstream (3′) nucleotides in each helix are shown with arrows. Four regions of absolutely conserved sequence longer than 5 nt are labeled “Conserved Region” (CR) I–IV, as used for the RNase P RNA [56].
the sequences in the collection are complementary near the ends, but the length of that complementarity varies somewhat and two sequences contain a bulged "A" interrupting this helix. Alignment of the nucleotides on either strand of this helix is straightforward, however, on the basis of sequence conservation; only minor alteration of the CLUSTAL alignment is required to bring the bases in each position of the helix into the same columns (Fig. 30.3). When aligned on the basis of sequence similarity, it becomes clear that the variation in helix length results from the addition of 2 bases to the distal ends of the helix (i.e. then end of the helix that contains the 5' and 3' tails) in *Pseudomonas* and *S. onedensis*. If the alignment is correct, there is a one to one correspondence in pairing partners in columns of the alignment; notice how the helix is opened by one column to accommodate the bulged “A” in two of the *Pseudomonas* sequences in the 3’ strand of this helix. To solidify the specific base pairs and their homology among the sequences, a new line in the sequence alignment is added to hold right- and left-facing parenthesis to specify pairing partners (see Fig. 30.3). Additional lines can be added to the alignment for annotations to make it easier to visualize the helices. Once the alignment of this terminal helix (if present) is finalized, the alignment can be trimmed at the ends to match the native ends of any RNAs in which the ends have been determined experimentally. In our example, the ends of the *P. multocida* and *H. influenzae* sequences could not be clearly defined on the basis of sequence conservation relative to either the *E. coli* or *P. aeruginosa* sequences (in which the native ends are known [1, 4]), but these ends can now be predicted on the basis of the structure, and the alignment trimmed to match. The predicted terminal helix in these organisms is 1 bp shorter (at the distal end) than the other sequences other than those of *Pseudomonas* species and *S. onedensis*. Following the nomenclature used for group I introns and RNase P [45, 46], we will call this helix “P1a”; P for “pairing”, “1” because it is the first helix counting from the 5’ end and “a” because, as we will see, P1 continues after an interruption.

Before moving on, it is important to evaluate in detail the evidence supporting the existence of this helix. The basic bits of evidence upon which secondary structures are built are sequence *covications*. Two positions in an alignment are said to covary (for the purposes of secondary structure analysis) if both positions vary while maintaining the ability to form A-U, G-C or G-U base pairs. Covariation of 2 bp in a potential helix is generally accepted as proof that the helix exists. In our example, the presence of the terminal helix P1a is supported by sequence covariation in most of the base pairs of the helix with only a few discrepancies. The ends of a helix can be harder to define; what is needed at each end of the helix is sequence covariations supporting pairing on the terminal base pair and clear failure of the adjacent 2 bases to covary arguing against their pairing. Ultimately, one would like to have evidence supporting the pairing of every base pair shown in a secondary structure; one useful way to denote how close you are to this is to only draw the line (or open dot, in the case of G-U pairs, by convention) connecting base pairs in a secondary structure if these positions in the alignment covary, i.e. if that individual base pair is supported by sequence covariation (see Fig. 30.4). In our example, we have covariation supporting the terminal base pair of the con-
Fig. 30.4. Secondary structure of the *E. coli* 6S RNA. Helices are labeled as described in the text. Base pairings supported individually by sequence covariance are indicated by the connecting lines or dots; unsupported pairings lack these markers. The sequence is numbered 5' to 3' every 20 nt, with a tick mark every 10 nt. Four regions of absolutely conserved sequence longer than 5 nt are boxed and labeled “Conserved Region” (CR) I–IV as in Fig. 30.3. This structure was generated in connect (.ct) format directly from the alignment in Fig. 30.3 using a Hypertalk script and displayed using LoopDloop.
served structure and the consistent inability of the three 5' and four 3' nucleotides to pair, so this end of the helix is well defined, except in *Pseudomonas* and *S. onedensis*, in which this helix is potentially lengthened at this end by an additional base pairs and two flanking unpaired nucleotides (Fig. 30.3). Whether or not this extra potential base pair is really paired is not clear, because the 5' base is a C and the 3' base always G in these sequences; in the absence of sequence variation, comparative analysis provides no evidence for or against the pairing. At the proximal end of this helix, the last base pair is likewise uncertain; the 5' base is U or C, the 3' base is an invariant G. This is consistent with their pairing, but does not constitute specific evidence for it. The penultimate base pair, on the other hand, is supported by covariation; this is typically G=C, but is A–U in *P. putida*. The G=G at this position in *P. syringae* constitutes a covariation neither with G=C nor with A–U and so is not evidence for or against this pairing. The flanking two bases are often U and G, and so might be thought to pair, but these fail to covary and so should not be included as part of this helix; the 3' G is invariant, and the 5' base, although U in most sequences, is a G in *P. putida* and an A in *S. onedensis*. There are three adjacent unsupported base pairs in the interior of this helix, but they are given provisional acceptance given that they have the potential to base pair and are flanked on both sides by well-supported pairings.

30.2.5

**Subterminal Helix (P1b)**

The sequence immediately interior to the 11–13 base pairings that make up the terminal helix P1a of the 6S RNA cannot pair, but complementarity resumes after only a few bases on either side. The pairing of these sequences is predicted in most of the mfold structures from all of the sequences, although there are usually some idiosyncratic alternatives (see Figs 30.1 and 30.2). Adjustment of the alignment of the 5' region of this potential helix is needed to accommodate an extra nucleotide present only in *V. vulnificus* and *V. parahaemolyticus*; assignment of homology is straightforward if you keep conserved purines (G or A) or pyrimidines (U or C) aligned (Fig. 30.3). There is reasonable covariation of all but one of the positions in the potential six base pair helix, which we will call “P1b”, with an occasional mismatch and a bulged nucleotide representing the extra nucleotide in two *Vibrio* sequences.

30.2.6

**Apical Helix (P2a)**

In addition to the consistently predicted terminal helix, all of the 6S RNA sequences are predicted by mfold to have a stem–loop in the middle of the RNA containing some conservative sequence elements: CUCGG on the 5' side, and CCGAG on the 3' side (Figs 30.1 and 30.2). Attention is also draw to the potential pairing of these sequences because of the presence in most of the 6S RNAs of the conserved tetraloop sequences UNCG or CUYG (GNRA is the other conserved tet-
raloop motif [47]), although these would be unusual stem–loops in that the tetraloop sequences are in this case followed (except in Y. pestis) by 1 or 2 extraneous nucleotides before the 3’ strand of the stem (Fig. 30.3). Nevertheless, the alignment of these sequences is straightforward based on sequence conservation. The minor exception is the RNA of *H. influenzae*, which contains extra nucleotides between the tetraloop sequence (CUAC) and the conserved complementary sequences on both sides; these extra nucleotides are generally complementary and so would presumably create a terminal extension to the stem–loop. There are sequence covariations confirming all of the base pairs of the helix, which we will call “P2a”, with the exception of the terminal pairing, which is always G=C except in the *H. influenza* sequence in which this is an A-C mismatch in the middle of the extension of this stem.

30.2.7 Subapical Helices (P2b and P2c)

Flanking the apical helix P1a are highly conserved sequences that are not complementary and what sequence variation that does occur does not support the specific interaction of these sequences. However, flanking these conserved sequences in turn are variable sequences that are generally complementary (Fig. 30.3). Variation in these regions makes them difficult to assign definitive homologies solely on the basis of sequence, but they are readily aligned, by default, simply based on their conserved distance from the flanking conserved sequences; the only adjustment necessary is the addition of a gap downstream of the apical helix corresponding to an obviously absent U in the *Vibrio* sequences. A similar gap is required in the *S. onedens* sequence, although the location of this gap is questionable. This helix is 1 bp shorter in the *Pseudomonas* sequences than in the others; this deletion seems to be from the proximal end (furthest from the apex). There are covariations supporting all of the base pairings in this helix, which will be designated “P2b”. One RNA, that of *S. onedens*, has two non-Watson–Crick mispairs in P2b. Although these might seem to argue against the pairing of these sequences, the remaining sequences covary cleanly. More importantly, the mispairs in *S. onedens* are adjacent G-A/A-G pairs, a 3-D motif that is sometimes is seen as an alternative to Watson–Crick pairs in helices and is known not to interrupt the flanking A-form helical structure [48, 49]. Otherwise, there is only a single instance of a mispair between these two sequences (a C-C pairing in *P. putida*). The internal loop between P2b and P2a is nearly symmetrical, and is comprised of highly conserved sequences, suggesting an important functional role.

Closely flanking P2b are additional complementary sequences. In this case, these sequences are so divergent from one group of sequences to another that alignment based on sequence is impossible (Fig. 30.3). Furthermore, there is significant variation in the spacing between these complementary sequences and the surrounding conserved sequences and helix. Nevertheless, because of the conservation of general complementarity, the presence of this helix in the favorable structures predicted by mfold, and that fact that there is no apparent covariation
between either of these sequence regions and anything else in the RNA, we will include this potential helix "P2c" in the secondary structure on a preliminary basis. After aligning these sequences based on this complementarity, there is good covariation of all of the potential base pairings except that a G-A mispairing is present in the penultimate position of the helix in *E. coli* and its closest relatives (*S. flexneri* and *Salmonella* species).

30.2.8

**Potential Interior Stem–loop (P3)**

The secondary structure of the 6S RNA, as we understand it at this point, is an extended terminal helix P1 and an extended apical helix P2 flanking a central loop of as-yet undefined structure. Very highly conserved sequences on either side of this loop have the potential to pair to form a 9-bp helix with a single bulged A. *mfold* includes this helix in many of the structures it produced. However, the predictions in this region are not consistent; a number of equivalently favorable structures are predicted for each sequence, and the pairings predicted are idiosyncratic for each sequence (Figs 30.1 and 30.2). Only three sequence variants exist in this region; all three of these changes (an A to U in *S. onedensis* and a G to A in *H. influenzae* and *P. multocida*) disrupt the complementarity between these sequences, arguing against their pairing (Fig. 30.3).

A search for frequently occurring tetraloop sequences (UNCG, CUYG and GNRA) [47] flanked by complementary sequences, however, reveals an alternative; a potential stem–loop on the 3′ strand of this interior loop. This 9-bp stem would be composed of very highly conserved sequences; the only sequence variation among the potentially paired nucleotides is the change of a conserved C to U in the terminal base pair in the *Pseudomonas* sequences (Figs 30.3 and 30.4). This is consistent with the pairing of this nucleotide with the conserved G opposite, but does not constitute specific evidence for that pairing. In the absence of evidence for any of the pairings in this helix, the conservative approach would be not to propose this helix until additional sequences with variation in this region can be obtained or a genetic experiment performed. However, two aspects of the potential loop sequences argue for the provisional acceptance of this stem–loop, "P3". First, in the *Pseudomonas* sequence, this loop is a UUCG tetraloop motif, implying a stem–loop structure for at least these RNAs and, by extension, the others as well given the conservation of this region. Second is the observation that all of the sequence and length variation in this region is very specifically located in this potential loop (and the closing base pair in *Pseudomonas*), and that variation is consistent with loop structure.

30.2.9

**Is There Anything Else?**

At some point in the initial analysis of the secondary structure of your RNA, you will reach a point where no additional structure is obvious. What else can you do in
attempt to find structure? Perhaps the most useful approach is to draw all of the RNAs in the secondary structure as it is at this point and compare them with an eye towards common potential helices; how useful this is likely to depend on how much structure you have already gleaned; the more you already know the better. Some will find it convenient to draw a single "reference" structure and then annotate this with sequence variants. Another fruitful approach is to go back to mfold and generate another round of structures, using the structure you already know as constraints, i.e. force the pairing of all of the helices you are sure of and see what structures are predicted in this context. These structures would then be scrutinized from sequence to sequence in search of commonalities, as before. If you have already identified a large part of the secondary structure, there are likely to be only a small number of favorable structures generated. Another useful approach is to generate a sequence logo from the alignment (a web server for this can be found at http://weblogo.berkeley.edu/logo.cgi) [50]. This allows you to consider potential pairing in the context of sequence variation; sequences are expected to pair with other sequences with similar extents of variation. In the case of the 6S RNA, these methods failed to provide any additional insight into the structure, perhaps because so much structure has already been identified.

30.2.10
Where To Go From Here

Once you are satisfied that you have extracted all of the secondary structure information you can out of your sequences, you have a useful "working model" for the structure of the RNA. Have you identified all of the base pairing in the RNA? Not likely. Do you have extraneous base pairs in the structure? Probably. What direction do you take from here? In the words of the Cheshire Cat, "That depends a good deal on where you want to get to" [51]. If this working model is sufficient for your needs, then you are finished. If you wish to learn more about the secondary structure of the RNA or identify tertiary interactions, then you will need to continue the comparative analysis after additional sequences are obtained.

The choice of where to get additional sequences depends on what you want to learn about the RNA and how well the initial secondary structure analysis went. Most likely, you will want to know more about both the variable and most highly conserved regions of the RNA, and so you will want to obtain sequences that are closely and distantly related to those already in hand. Sequences similar to those already in hand are typically easy to obtain experimentally, and can often be obtained in large numbers. Distantly related sequence are much harder to obtain, but are needed to provide details about the regions of the RNA that are most important for function and so are very highly conserved. Thermophiles are a good source of useful sequences; these RNAs typically contain the fewer irregularities than those of mesophiles, and are much better fodder for thermodynamic prediction [21, 52, 53]. With new sequences in hand, of course, you have the opportunity to mine the sequence databases again and potentially identify sequences that were there all along but remained unrecognized.
30.3 Troubleshooting

Comparative analysis is a straightforward process, but as with any approach, it is possible to run into trouble. Below are listed some of the common problems that arise and how you might try to get around them.

- **But I only have one sequence!** This is a major problem; you cannot do a comparative analysis with only one sequence. However, this is the usual starting point; you are interested in the structure of a specific RNA from a specific organism and that is the only one you have in hand. Usually you will be able to get at least one or two additional sequences from the genomes of related species. If data mining fails to yield the sequences you need, you have no choice but to get the sequences experimentally.

- **I just do not get it — how do I get started once I have some sequences?** If you are having trouble getting started and you cannot seem to get a handle on the alignment, then reduce the problem by starting with a smaller collection of very similar sequences that you can align easily by eye. Look at every difference in the sequences — can you find one change that corresponds to another change that means sequences could remain complementary? Again, start out, as we did in our example, by looking to see if the two ends of the RNA might form a helix. Also be on the lookout for common tetraloop sequences (UUCG, UCCG, GAGA, GAAA, GUGA, GUAA, GCCA, GCAA and CUUG [9]) flanked by complementary sequences — these are very likely to form stem-loops. Another approach is to start with the single best structure predicted by mfold and then pick through each helix by comparative analysis to prove or disprove each one. You could then move on to the unique helices predicted in the less favorable structures or the structures of the other sequences.

- **Some of my sequences have the most highly conserved sequences, but otherwise cannot be aligned.** It is common to have sequences that you can only align to others in the conservative regions of the molecules, at least at first. In our example of the 6S RNAs, the *H. influenzae* and *P. multocida* sequences do not align well to the others at first. This is especially a problem for sequences that are quite different in length that the others; sometimes localizing the sites of the insertions or deletions can be difficult. This is usually best dealt with from both directions: aligning those regions you can to identify structure in common and dividing the alignment into smaller groups of similar sequences to identify structure in the regions unique to each group. Once some insight on both is obtained, the alignments can be merged on the basis of structure rather than just sequence.

- **PCR or sequencing artifacts.** It is critically important that the sequences that go into a comparative analysis be valid. If you must enter sequences manually, check them very carefully for errors. The qualities of sequences are only as good as the abilities of the person or machine that did the sequence determination; there is always a chance that the sequence is incorrect. Genome sequences are usually reliable, but even here errors occasionally arise. In some cases, it might even be worth your effort to confirm an unusual sequence experimentally. As
was seen above, sequence annotations are imperfect; if a sequence does not look
like what you expect, it probably is not what you want, and even if it is you will
not (yet) be able to use it. Generally speaking, the more recent the sequence, the
less likely it is to contain errors. A common source of problematic sequences is
PCR amplification and there are two commonly seen types of these errors: point
"mutations" and chimeras [54]. Point mutations are a problem, but a limited
one. These changes will most often appear in the analysis as the occasional mis-
match or idiosyncrasy. Chimeric sequences are sequences that have been artifi-
cially spliced together during the amplification process; this is a common prob-
lem when amplifying genes from DNA extracted from microbial populations
rather than pure cultures. Two aspects of chimeric RNAs usually reveal their na-
ture; their failure to conform to long-range structure that is well-maintained
among the remaining sequences or their similarity to one sequence at one end
of the RNA, but a very different sequence at the other end. Suspected chimeric
sequences should, of course, be removed from the analysis.

Acknowledgments

Research in the authors laboratory is supported by NIH grant GM52894 to J. W. B.
and is dedicated to the memory of Dr Elizabeth Suzanne Haas (1957–2002).

References

14 Parsch, J., Braverman, J. M., Stephan, W., Genetics 2000, 154, 909–921.