Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families

Lars Barquist,[1,2] Sarah W. Burge,[2] and Paul P. Gardner[3,4]

[1]Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany
[2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom
[3]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand
[4]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand

Emerging high-throughput technologies have led to a deluge of putative non-coding RNA (ncRNA) sequences identified in a wide variety of organisms. Systematic characterization of these transcripts will be a tremendous challenge. Homology detection is critical to making maximal use of functional information gathered about ncRNAs: identifying homologous sequence allows us to transfer information gathered in one organism to another quickly and with a high degree of confidence. ncRNA presents a challenge for homology detection, as the primary sequence is often poorly conserved and de novo secondary structure prediction and search remain difficult. This unit introduces methods developed by the Rfam database for identifying "families" of homologous ncRNAs starting from single "seed" sequences, using manually curated sequence alignments to build powerful statistical models of sequence and structure conservation known as covariance models (CMs), implemented in the Infernal software package. We provide a step-by-step iterative protocol for identifying ncRNA homologs and then constructing an alignment and corresponding CM. We also work through an example for the bacterial small RNA MicA, discovering a previously unreported family of divergent MicA homologs in genus *Xenorhabdus* in the process. © 2016 by John Wiley & Sons, Inc.

Keywords: covariance model • homology • RNA • Rfam • alignment • ncRNA • conservation

How to cite this article:
Barquist, L., Burge, S.W. and Gardner, P.P. 2016. Studying RNA homology and conservation with infernal: from single sequences to RNA families *Curr. Protoc. Bioinform.* 54:12.13.1-12.13.25.
doi: 10.1002/cpbi.4

INTRODUCTION

Over the past twenty years, the development and commodification of high-throughput technologies for reading DNA has led to dramatic changes in the way biology is done. The development of reliable and inexpensive transcriptomic technologies, in particular, has lead to major changes to our understanding of the role RNA plays in both bacterial and eukaryotic systems (Rinn and Chang, 2012; Barquist and Vogel, 2015). While exceptions to the central dogma have been known since the discovery of tRNA in the late 1950's, it was only with the development of genome-scale sequencing technologies that the pervasive nature of regulation mediated by non-coding RNA (ncRNA) has become clear. This has led to a diverse menagerie of RNA classes, ranging from the microRNAs (miRNAs) and bacterial small RNAs (sRNAs), which operate to modulate gene expression through RNA:RNA antisense interactions, to small nucleolar RNAs (snoRNAs), which guide RNA modification enzymes, to the riboswitches capable of

Current Protocols in Bioinformatics 12.13.1-12.13.25, June 2016
Published online June 2016 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.4

sensing metabolites with so-called RNA aptamers, to the diverse long non-coding RNAs (lncRNAs) suspected to blanket eukaryotic genomes. Characterizing a single ncRNA is a challenge in itself; to maximize the utility of this work, it is often desirable to transfer functional annotations between organisms. For instance, one might identify an ncRNA associated with disease in humans, but want to study it in a system where technical and (more importantly) ethical considerations allow for genetic manipulation, such as mice. Similarly, in bacteria, molecular tools are often best developed in model strains such as *E. coli*, making them attractive platforms for characterizing ncRNAs, which can then be inferred to operate in a similar function in related bacteria. However, this inference depends crucially on computational tools capable of transferring these hard-won functional annotations through homology prediction.

This problem of homology prediction is usually approached as the problem of finding sequences that align well to our source sequence, on the reasonable assumption that similar sequences are more likely to be evolutionarily and functionally related. A wide range of critical applications in genomics rely on our ability to produce "good" alignments. Single-sequence homology search, as implemented in tools such as BLAST (Altschul et al., 1990), is an (often heuristic) application of alignment. The sensitivity and specificity of homology search can be improved by the use of evolutionary information in the form of accurate substitution and insertion-deletion (indel) rates derived from multiple sequence alignments (MSAs), captured in the statistical models used by HMMER (Finn et al., 2011, 2015; Eddy, 2011) and Infernal (Nawrocki et al., 2009; Nawrocki and Eddy, 2013), for protein and RNA alignments respectively. These models can be thought of as defining "families" of homologous sequences, as in the Pfam and Rfam databases (Finn et al., 2014; Nawrocki et al., 2015). By using these models to classify sequences, we can infer functional and structural properties of uncharacterized sequences.

Unfortunately, producing the high-quality "seed" alignments of RNA that these methods require remains difficult. While proteins can be aligned accurately using only primary sequence information, with pairwise sequence identities as low as 20% for an average-length sequence (Rost, 1999; Thompson et al., 1999), it appears that the "twilight zone" where blatantly erroneous alignments occur between RNA sequences may begin at above 60% identity (Gardner et al., 2005; Lindgreen et al., 2014). The inclusion of secondary structure information can improve alignment accuracy (Freyhult et al., 2007), but predicting secondary structure is not trivial (Gardner and Giegerich, 2004; Puton et al., 2014). An instructive example of the difficulties this can lead to is the case of the 6S gene, a bacterial ncRNA that modulates σ^{70} activity during the shift from exponential to stationary growth. The *Escherichia coli* 6S sequence was determined in 1971 (Brownlee, 1971) and its function determined in 2000 (Wassarman and Storz, 2000). However, the extent of this gene's phylogenic distribution was not realized until 2005 when Barrick and colleagues carefully constructed an alignment from a number of deeply diverged putative 6S sequences, and through successive secondary-structure-aware homology searches demonstrated its presence across large swaths of the bacterial phylogeny (Barrick et al., 2005). Even now, new homologs are discovered on a regular basis (Sharma et al., 2010; Weinberg et al., 2010; Wehner et al., 2014), and 6S appears to be an ancient and important component of the bacterial regulatory machinery. Similar examples of the power of enhanced homology search considering both sequence and structural information can be found for other classes of ncRNAs, such as riboswitches (Barrick and Breaker, 2007), ribosomal leader sequences (Fu et al., 2013), ribozymes (Weinberg et al., 2015), and snoRNAs (Gardner et al., 2010).

In this unit, it is our hope to make these techniques accessible to sequence analysis novices. We introduce the techniques necessary to construct a high-quality RNA alignment from a single seed sequence, and then use the information contained in this alignment to identify additional more distant homologs, expanding the alignment in an iterative fashion. These methods, while time-consuming, can be far more sensitive than a BLAST search (Menzel et al., 2009). We present a brief protocol that starts with a single sequence, and then use a collection of Web and command-line based tools for alignment, structure prediction, and search to construct an Infernal covariance model (CM), a probabilistic model that captures many important features of structured RNA sequence variation (Nawrocki and Eddy, 2013). These models may then be used in the iterative expansion of alignments or for homology search and genome annotation. CMs are also are used by the Rfam database in defining RNA sequence families, and are the subject of a dedicated RNA families track at the journal *RNA Biology* (Gardner and Bateman, 2009). We include as an instructive example the construction of an RNA family for the enterobacterial small RNA MicA, used as the basis for the current Rfam model, discovering a convincing divergent clade of homologs in the process.

STRATEGIC PLANNING

A large variety of tools exist for RNA sequence analysis. Given the diversity of ncRNA sequence, structure, and function, it is likely that no one set of tools will work optimally for every ncRNA. Here we review a number of alternative methods and tools that can be easily substituted at various stages of our procedure.

Single Sequence Search

We rely on NCBI BLAST (Altschul et al., 1990) to quickly identify close homologs of RNA sequences in this unit, although other methods can be substituted (Table 12.13.1). NCBI and EMBL-EBI both maintain servers (Boratyn et al., 2013; McWilliam et al., 2013) with slightly different interfaces, though there are no substantive differences in the implementations. We use the NCBI server here. EBI also maintains servers for a number of BLAST and FASTA derivatives, which may be helpful. Both sites also allow users to BLAST against databases of expressed sequences including GEO at NCBI and high-throughput cDNA and transcriptome shotgun assembly databases at EMBL-EBI (Barrett et al., 2013; Silvester et al., 2015). Such searches can be helpful for gathering comparative expression data for your ncRNA.

A nucleotide version of the HMMER3 package (Eddy, 2011) for profile-based sequence search provides both increased sensitivity and specificity over BLAST at little additional computational cost (Wheeler and Eddy, 2013). We hope that a Web server similar to the one currently available for protein sequences (Finn et al., 2015) will be forthcoming. In the meantime, RNAcentral (Bateman et al., 2011; RNAcentral Consortium, 2015) offers an nhmmer-based search facility; however, it is limited to searching known ncRNA sequences. If it is possible that homologous sequences are spliced—e.g., introns in the

Table 12.13.1 Resources for Single Sequence Homology Search

Resource	Reference	URL
NCBI-BLAST	(Johnson et al., 2008)	*http://blast.ncbi.nlm.nih.gov/Blast.cgi*
EMBL-EBI NCBI-BLAST	(McWilliam et al., 2013)	*http://www.ebi.ac.uk/Tools/sss/ncbiblast/*
EMBL-EBI Sequence Search	(McWilliam et al., 2013)	*http://www.ebi.ac.uk/Tools/sss/*
HMMER3	(Finn et al., 2015, 2011)	*http://www.ebi.ac.uk/Tools/hmmer/*
RNAcentral nhmmer search	(RNAcentral Consortium, 2015; Bateman et al., 2011)	*http://rnacentral.org/sequence-search/*

Table 12.13.2 Resources for RNA Sequence Alignment

Resource	Reference	URL
Web server for Aligning structural RNAs (WAR)	(Torarinsson and Lindgreen, 2008)	*http://genome.ku.dk/resources/war/*
Vienna RNA	(Gruber et al., 2008b)	*http://rna.tbi.univie.ac.at/*
Freiburg RNA Tools	(Smith et al., 2010)	*http://rna.informatik.uni-freiburg.de*
CBRC Functional RNA Project	(Asai et al., 2008)	*http://software.ncRNA.org*
RTH Resources	N/A	*http://rth.dk/pages/resources.php*
EMBL-EBI Alignment Tools	(McWilliam et al., 2013)	*http://www.ebi.ac.uk/Tools/msa/*

U3 snoRNA (Myslinski et al., 1990)—then a splice-site-aware search method may be useful, such as BLAT (Kent, 2002) or GenomeWise (Birney et al., 2004), but we are not aware of any publicly available Web servers.

Alignment and Secondary Structure Prediction Tools

We find it best to run a variety of alignment and secondary structure prediction tools simultaneously (see Table 12.13.2). Each has its own peculiarities, and our hope is that by looking for shared homology and secondary structure predictions we can mitigate some of the problems discussed in the introduction. In this unit, we use the WAR Web server (Torarinsson and Lindgreen, 2008), which allows the user to run 14 different methods simultaneously. These include Sankoff-type methods: FoldalignM (Torarinsson et al., 2007), LocARNA (Will et al., 2007), MXSCARNA (Tabei et al., 2008), Murlet (Kiryu et al., 2007), and StrAL (Dalli et al., 2006) with PETcofold (Seemann et al., 2011); align-then-fold methods, which use a traditional alignment tool [ClustalW (Larkin et al., 2007) or MAFFT (Katoh and Standley, 2013)] followed by structure prediction [RNAalifold (Bernhart et al., 2008) or Pfold (Knudsen and Hein, 2003)]; fold-then-align methods, which predict structures in all the input sequences and attempt to align these structures [RNAcast (Reeder and Giegerich, 2005) plus RNAforester (Höchsmann et al., 2003)]; sampling methods that attempt to iteratively refine alignment and structure [MASTR (Lindgreen et al., 2007) and RNASampler (Xu et al., 2007)]; and other methods that do not fit into the above traditional categories—CMfinder (Yao et al., 2006) and LaRA (Bauer et al., 2007). Finally, WAR also computes a maximum consistency alignment using all the alignment predictions with T-Coffee (Notredame et al., 2000).

However, WAR is by no means exhaustive, and the methods may not be the most recent versions available. A number of groups maintain their own servers for RNA sequence analysis. Notable servers include the Vienna RNA WebServers (Gruber et al., 2008b), the Freiburg RNA Tools (Smith et al., 2010), the CBRC Functional RNA Project (Asai et al., 2008; Mituyama et al., 2009), and the Center for Non-Coding RNA in Technology and Health (RTH) Resources page. In addition, EMBL-EBI maintains a number of Web servers for popular multiple sequence alignment tools (McWilliam et al., 2013). Ultimately, as you become more comfortable with RNA sequence analysis, you may want to begin installing and running new tools on a local *NIX machine; however, this is beyond the scope of the current unit.

Genome Browsers

Genome browsers are essential for checking the context of putative homologs, and several are available online that offer access to genome annotations (see Table 12.13.3). The ENA

Table 12.13.3 Web-Based Resources for Genome Annotations

Resource	Reference	URL
European Nucleotide Archive	(Silvester et al., 2015)	*http://www.ebi.ac.uk/ena*
UCSC Genome Browser	(Karolchik et al., 2014)	*http://genome.ucsc.edu/*
Ensembl	(Flicek et al., 2014)	*http://www.ensembl.org*
UCSC Microbial Genome Browser	(Chan et al., 2012; Schneider et al., 2006)	*http://microbes.ucsc.edu/*
ENCODE Project	(ENCODE Project Consortium, 2011)	*https://www.encodeproject.org/*

Table 12.13.4 RNA Alignment Editors

Resource	Reference	URL
BoulderALE	(Stombaugh et al., 2011)	*http://boulderale.sourceforge.net/*
JalView	(Waterhouse et al., 2009)	*http://www.jalview.org*
MultiSeq	(Roberts et al., 2006)	*http://www.scs.illinois.edu/schulten/multiseq/*
RALEE	(Griffiths-Jones, 2005)	*http://sgjlab.org/ralee/*
Ribo	(Waldispühl et al., 2015)	*http://ribo.cs.mcgill.ca/*
S2S	(Jossinet and Westhof, 2005)	*http://bioinformatics.org/S2S/*
SARSE	(Andersen et al., 2007)	*http://sarse.ku.dk/*

(Silvester et al., 2015) provides a no-frills sequence browser perfect for quickly checking annotations. For deeper annotations, the UCSC genome browser (Karolchik et al., 2014) and Ensembl (Flicek et al., 2014) both contain a wide range of information for the organisms they cover. For bacterial and archaeal genomes, the Lowe lab maintains a modified version of the UCSC genome browser (Schneider et al., 2006; Chan et al., 2012) that provides a number of tracks of particular interest to those working with ncRNA. Finally, the ENCODE Project (ENCODE Project Consortium, 2011) provides access to a large number of functional assays in a number of eukaryotic organisms, many of which have relevance to ncRNA. Offline genome browsers are also available for use on your own computer, for example IGV and Artemis (Carver et al., 2012; Thorvaldsdóttir et al., 2013). There are also methods available for comparing synteny (gene order) information between genomes (Carver et al., 2005; Alikhan et al., 2011; Sullivan et al., 2011).

Alignment Editors

It is possible to edit alignments in any text editor; however, we highly recommend using a secondary structure–aware editor such as Emacs with the RALEE major mode (Griffiths-Jones, 2005). RALEE allows you to color bases according to base identity, secondary structure, or base conservation. It also allows the easy manipulation of sequences involved in structural interactions but which are not close in sequence space, through the use of split screens. A number of other specialized RNA editors are available (Table 12.13.4). BoulderALE (Stombaugh et al., 2011) and S2S (Jossinet and Westhof, 2005) both allow the end user to visualize and manipulate tertiary structure in addition to secondary structure, and may be particularly useful if crystallographic information is available for your RNA. Other alternatives for editing RNA secondary structure are SARSE (Andersen et al., 2007) and MultiSeq (Roberts et al., 2006). Recent versions of JalView (Waterhouse et al., 2009) have begun to support RNA secondary structure as well. Finally, a recent attempt has been made to "gamify" RNA alignment editing. This approach, called Ribo, samples poorly resolved regions of alignments, and feeds these regions to gamers who manually refine the alignments. These re-aligned regions can then be reinserted into the original alignment (Waldispühl et al., 2015).

Infernal

The centerpiece of our unit is the Infernal package for constructing covariance models (CMs) from RNA multiple alignments (Nawrocki and Eddy, 2013). We will use this to construct models of our RNA family. CMs model the conservation of positions in an alignment in a similar way to a hidden Markov model (HMM), while also capturing *co-variation* in structured regions (Eddy and Durbin, 1994; Sakakibara et al., 1994; Durbin et al., 1998). Covariation is the process whereby a mutation of a single base in a hairpin structure will lead to selection in subsequent generations for compensatory mutations of its structural partner in order to preserve canonical base-pairing, i.e., Watson-Crick plus G-U pairs, and a functional structure. This combination of structural-evolutionary information has been shown to provide the most sensitive and specific homology search for RNA of any tools currently available (Freyhult et al., 2007; Gardner, 2009). Unfortunately, this sensitivity and specificity come at a high computational cost, and Infernal searches can be time consuming, with genome-scale searches often taking hours on desktop computers. The development of heuristics to reduce this computational cost is an area of active research for the Infernal team, and has already been mitigated to some extent by the use of HMM filters and query-dependent banding of alignment matrices (Nawrocki and Eddy, 2007, 2013). We refer the reader to Eric Nawrocki's primer on annotating functional RNAs in genomic sequence for a friendly introduction to the mechanics of the Infernal package (Nawrocki, 2014; also see Nawrocki et al., 2009; Nawrocki and Eddy, 2013; *http://eddylab.org/software/infernal/*).

CHOOSING THE RIGHT PROTOCOLS

We assume for the sake of this unit that you are starting with a single sequence of interest. We will illustrate our method using the example of MicA, an Hfq-dependent bacterial trans-acting antisense small RNA (sRNA). Many bacterial sRNAs are similar in function to eukaryotic microRNAs, pairing to target mRNA transcripts through a short antisense-binding region, generally targeting the transcript for degradation (Barquist and Vogel, 2015). MicA is known to target a wide range of outer membrane protein mRNAs using a 5′ binding-region in both *E. coli* (Gogol et al., 2011) and *S. enterica* (Vogel, 2009) in response to membrane stress. The previous covariance model for MicA (accession RF00078) in Rfam (release 10.1) was largely restricted to *E. coli*, *S. enterica*, and *Y. pestis*. Here, as an example, we produce an improved model, which has served as the core for the current Rfam model (release 12.0). In the process, we discover previously unreported MicA homologs in the nematode symbionts of the γ-proteobacterial genus *Xenorhabdus*.

If you already have a set of putative homologs, you may wish to further diversify your collection of sequences using the methods described in Basic Protocol 1, or you may skip directly to Basic Protocol 2 or 3 if a secondary structure is known. No matter how many sequences you are starting with, it is always a good idea to run the sequence search tools available on the Rfam Web site (*http://rfam.xfam.org/*) on them. This will verify that there is no CM already available that covers your sequences. There are a number of other specialist databases that may also be worth searching if you have reason to believe your RNA sequence is a member of a well-defined class of RNAs, i.e., microRNAs, snoRNAs, rRNAs, tRNAs, etc. We have recently reviewed these specialist RNA databases (Hoeppner et al., 2014). A centralized RNA sequence database aiming to capture all known RNA sequences, RNAcentral (Bateman et al., 2011; RNAcentral

Consortium, 2015), has recently launched that provides an integrated resource for easily identifying similar sequences with some evidence of transcription via an integrated nhmmer search.

GATHERING AN INITIAL SET OF HOMOLOGOUS SEQUENCES

BASIC PROTOCOL 1

Now that you have confirmed that your sequence is novel using Rfam, RNAcentral, or appropriate specialist data bases, we will use NCBI-BLAST to identify additional homologous sequences. Once you have navigated to the nucleotide BLAST server (*http://blast.ncbi.nlm.nih.gov/*) there are a number of important options to set.

Necessary Resources

Computer with an up-to-date Web browser (e.g., Firefox, Chrome, Internet Explorer)
Text editor

Setting NCBI-BLAST parameters

1. First, it is important to choose a search set appropriate to your sequence. At this initial phase, we want to limit our exposure to sequences that are very distant from ours to limit the number of obviously spurious alignments we will need to examine, thereby increasing the power of our search. So, if your initial sequence is of human origin, you may want to limit your search to Mammalia, Tetrapoda, or Vertebrata depending on sequence conservation. Similarly, if you are working with an *Escherichia coli* sequence, you may want to limit your initial searches to Enterobacteriaceae or the Gammaproteobacteria. NCBI-BLAST searches are relatively fast, so try several search sets to get a feel for how conserved your sequence is.

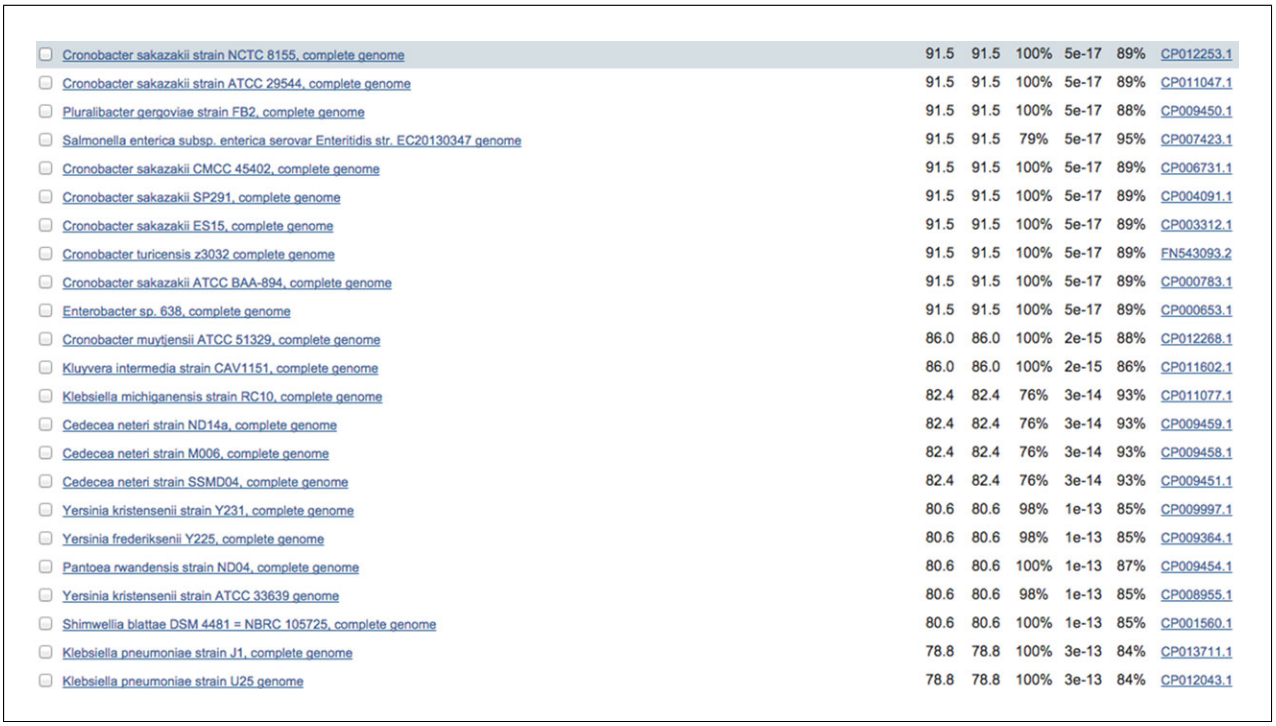
2. The second set of options to set are the Program Selection and Algorithm Parameters. We recommend **blastn**, as it allows for smaller word sizes. The word size describes the minimum length of an initial perfect match needed to trigger an alignment between our query sequence and a target. Smaller word sizes provide greater sensitivity, and seem to perform better for non-coding RNAs. We recommend a word size of 7, the smallest the NCBI-BLAST server allows.

3. Finally, you should set "Max Target Sequences" parameter to at least 1000. NCBI-BLAST returns hits in a ranked list from best match to worst by E-value (or the number of matches with the same quality expected to be found in a search over a database of this size), and will only display as many as Max Target Sequences is set to. We are primarily interested in matches on the edge of what NCBI-BLAST is capable of detecting reliably, and these will naturally fall toward the end of this list.

4. Our example sequence, MicA, is from *E. coli*, so we will limit our initial searches to Enterobacteriaceae. This sequence, obtained from Gisela Storz's collection of known *E. coli* sRNAs (*http://cbmp.nichd.nih.gov/segr/ecoli_rnas.html*) is:

```
GAAAGACGCGCATTTGTTATCATCATCCCTGAATTCAGAGATGAAATTTTGGCCACTCACGAGTGGCCTTTTT
```

   Paste this sequence into the query sequence box at the top of the page. Limit the space of your search by entering `Enterobacteriaceae` (or, equivalently `taxid:91347`) in the Organism box. Under Program Selection, click the radio button to select "blastn" search. Access additional options by clicking the plus sign next to "Algorithm parameters." Under General Parameters set "Max target sequences" to 1000 and "Word size" to 7. Click the BLAST button to run the search.



☐ Cronobacter sakazakii strain NCTC 8155, complete genome	91.5	91.5	100%	5e-17	89%	CP012253.1
☐ Cronobacter sakazakii strain ATCC 29544, complete genome	91.5	91.5	100%	5e-17	89%	CP011047.1
☐ Pluralibacter gergoviae strain FB2, complete genome	91.5	91.5	100%	5e-17	88%	CP009450.1
☐ Salmonella enterica subsp. enterica serovar Enteritidis str. EC20130347 genome	91.5	91.5	79%	5e-17	95%	CP007423.1
☐ Cronobacter sakazakii CMCC 45402, complete genome	91.5	91.5	100%	5e-17	89%	CP006731.1
☐ Cronobacter sakazakii SP291, complete genome	91.5	91.5	100%	5e-17	89%	CP004091.1
☐ Cronobacter sakazakii ES15, complete genome	91.5	91.5	100%	5e-17	89%	CP003312.1
☐ Cronobacter turicensis z3032 complete genome	91.5	91.5	100%	5e-17	89%	FN543093.2
☐ Cronobacter sakazakii ATCC BAA-894, complete genome	91.5	91.5	100%	5e-17	89%	CP000783.1
☐ Enterobacter sp. 638, complete genome	91.5	91.5	100%	5e-17	89%	CP000653.1
☐ Cronobacter muytjensii ATCC 51329, complete genome	86.0	86.0	100%	2e-15	88%	CP012268.1
☐ Kluyvera intermedia strain CAV1151, complete genome	86.0	86.0	100%	2e-15	86%	CP011602.1
☐ Klebsiella michiganensis strain RC10, complete genome	82.4	82.4	76%	3e-14	93%	CP011077.1
☐ Cedecea neteri strain ND14a, complete genome	82.4	82.4	76%	3e-14	93%	CP009459.1
☐ Cedecea neteri strain M006, complete genome	82.4	82.4	76%	3e-14	93%	CP009458.1
☐ Cedecea neteri strain SSMD04, complete genome	82.4	82.4	76%	3e-14	93%	CP009451.1
☐ Yersinia kristensenii strain Y231, complete genome	80.6	80.6	98%	1e-13	85%	CP009997.1
☐ Yersinia frederiksenii Y225, complete genome	80.6	80.6	98%	1e-13	85%	CP009364.1
☐ Pantoea rwandensis strain ND04, complete genome	80.6	80.6	100%	1e-13	87%	CP009454.1
☐ Yersinia kristensenii strain ATCC 33639 genome	80.6	80.6	98%	1e-13	85%	CP008955.1
☐ Shimwellia blattae DSM 4481 = NBRC 105725, complete genome	80.6	80.6	100%	1e-13	85%	CP001560.1
☐ Klebsiella pneumoniae strain J1, complete genome	78.8	78.8	100%	3e-13	84%	CP013711.1
☐ Klebsiella pneumoniae strain U25 genome	78.8	78.8	100%	3e-13	84%	CP012043.1

Figure 12.13.1 Partial results of a BLAST search using the *E. coli* MicA sequence from the "nr" sequence database. The tabular view provides important information that can be used to pick putative homolog sequences including species and strain information (column 1), query sequence coverage (column 4), E-value (column 5), and percent identity (column 6). Also note the accession number in column 7; this will be useful looking up sequences in other databases (e.g., ENA).

Selecting sequences

Our goal at this stage is to pick a representative set of homologous sequences with which to "seed" our alignment. As discussed in the introduction, single sequence alignment for nucleotides is generally only reliable to around 60 percent pairwise sequence identity. At the same time, picking a large number of sequences with high percent identity can lead to *overfitting* of the secondary structure—that is, if our sequences are too similar, we can end up predicting alignments and secondary structures that capture accidental features of a narrow clade, rather than conserved structure and sequence variation.

5. There are three major criteria we pick additional sequences based on, in rough order of importance: percent sequence identity, taxonomy, and sequence coverage. Handily, the NCBI-BLAST output displays measures of all of these (Fig. 12.13.1). Our first selection criterion, percent identity, should fall between 65% and 95%—if much lower, the sequence will be difficult to align, if higher it will be too similar to have any meaningful variation.

6. The second selection criterion, taxonomy, will depend somewhat on the organisms your sequence is associated with, but we generally want to limit the inclusion to a single (orthologous) instance per species. The exception to this rule is for diverged paralogous sequences within the species; if paralogs exist, you will need to decide how broadly you wish to define your family, although it can be difficult to construct families that capture the full phylogenetic range of an element while retaining the ability to discriminate between paralogs within single genomes. Additionally, it may be useful to further limit the maximum percent identity to, say, 90% within a densely sequenced genus to further limit the number of highly similar sequences in your initial alignment.

7. Finally, assuming that you are sure of your sequence boundaries, we want to select sequences that cover the entire starting sequence. If you see many matches covering

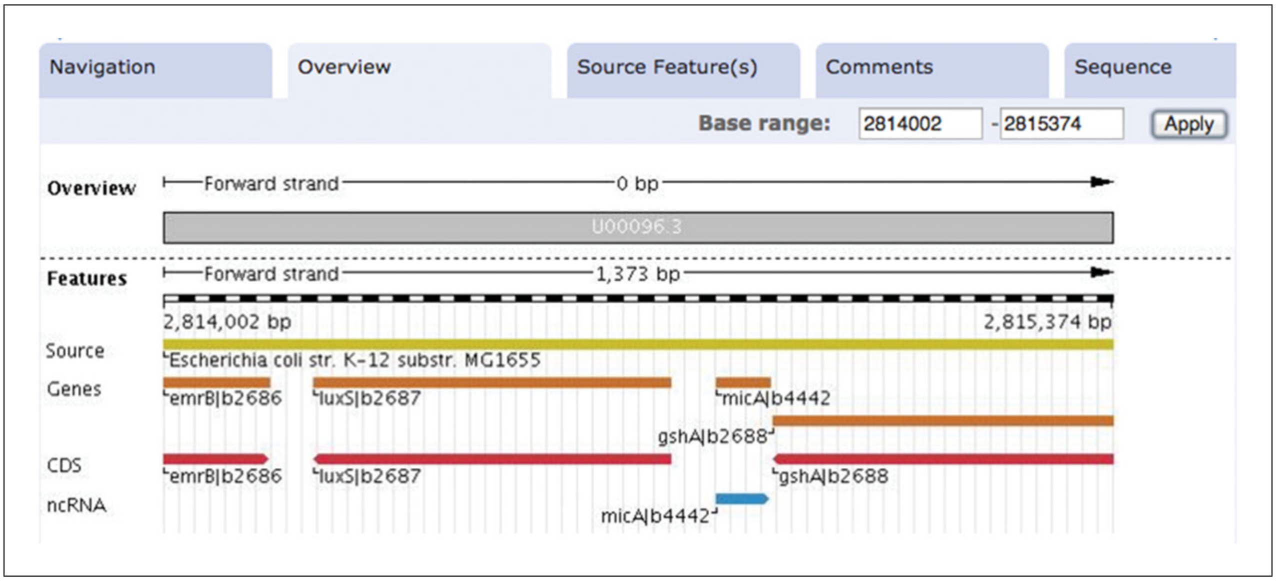
only a short section of your sequence, this may be due to the matching of a short convergent motif. This most commonly happens with the relatively long, highly-constrained bacterial rho-independent terminators (Gardner et al., 2011), but may occur with other motifs (Gardner and Eldai, 2015). Alternatively, if you do not have well-defined sequence boundaries, you will need to determine these from the conservation you see in your BLAST hits—look for taxonomically diverse hits covering the same segment of your query sequence. In some cases, such as the long non-coding RNAs, conserved domains may be much shorter than the complete transcribed sequence (Burge et al., 2013), but stay aware of the potential motif issue. A taxonomic distribution of sequences that makes biological sense given your knowledge of the molecule's function and that can be explained by direct inheritance of the sequence will be your best guide.

8. Continuing our MicA example, we want to select a group of sequences with a reasonably diverse taxonomic range and as much sequence diversity as possible, while being reasonably confident that they are true homologs. In this case we will choose, based on maximizing genus diversity, a percent identity between 75% and 90%, and 100% sequence coverage, as we are fairly confident in the MicA gene boundaries. You can retrieve individual sequences by clicking Download on the alignment, then selecting the "FASTA (aligned sequence)" option. Note that you may have to reverse-complement sequences (e.g., using the Web server *http://www.bioinformatics.org/sms/rev_comp.html*), as by default NCBI-BLAST returns sequences from the forward strand of the genomic sequence. For our example alignment, in addition to our original sequence from *E. coli* (EMBL-Bank accession: U00096), we have chosen sequences from *Salmonella typhimurium* (FQ312003), *Klebsiella pneumoniae* (CP002910), *Enterobacter cloaca* (CP002272), *Yersinia pestis* (AM286415), *Pantoea* sp. At-9b (CP002433), and *Erwinia pyrifoliae* (FP236842). Collect these in FASTA format:

```
>U00096.3
GAAAGACGCGCATTTGTTATCATCATCCCTGAATTCAGAGATGAAATTTTGGCCACTCACGAGTGGCCTTTTT
>FQ312003
GAAAGACGCGCATTTGTTATCATCATCCCTGTTTTCAGCGATGAAATTTTGGCCACTCCGTGAGTGGCCTTTTT
>CP002272
GAAAGACGCGCATTTGTTATCATCATCCCTGACTTCAGAGATGAAATGTTTGGCCACAGTGATGTGGCCTTTTT
>CP002910
GAAAGACGCGCATTTATTATCATCATCATCCCTGAATCAGAGATGAAAGTTTGGCCACAGTGATGTGGCCTTTTT
>AM286415
GAAAGACGCGCATTTGTTATCATCATCCCTGTTATCAGAGATGTTAATTTGGCCACAGCAATGTGGCCTTTT
>CP002433
GAAAGACGCGCATTTGTTATCATCATCCCTGACAACAGAGATGTTAATTCGGCCACAGTGATGTGGCCTTTT
>FP236842
GAAAGACGCGTATTTGTTATCATCATCTCATCCCTGACAACAGAGATGTTAATTTAGGCCACAGTGACGTGGCCTTTTT
```

Examining your initial homolog set

9. Once you have assembled a set of sequences fitting the criteria described above, it is worth taking a closer look at them. Remember that these sequences will form the core of your alignment and CM, and errors at this stage can dramatically bias your results. A good first test is to examine the taxonomy of your sequences, and make sure that it makes sense. Can you identify a clear pattern of inheritance that might explain the taxonomic distribution you see at this stage? Another good check is to examine your sequences in the ENA browser, or a domain-specific browser if one exists for your organisms. For many independently transcribed RNAs, genomic context is better conserved than sequence, and ncRNA genes will often fall in homologous intergenic



Base range: 2814002 - 2815374

Figure 12.13.2 Genomic context of *micA*. ENA browser view of the region surrounding *micA* in *E. coli*. All of our selected homologs show conserved synteny with the *luxS* and *gshA* genes, providing additional evidence for their evolutionary relationship. Note the Rfam annotation in this region, matching our sequence. This view can be generated by navigating to *http://www.ebi.ac.uk/ena/data/view/U00096.3* and entering the genome coordinates in the "Base range" boxes.

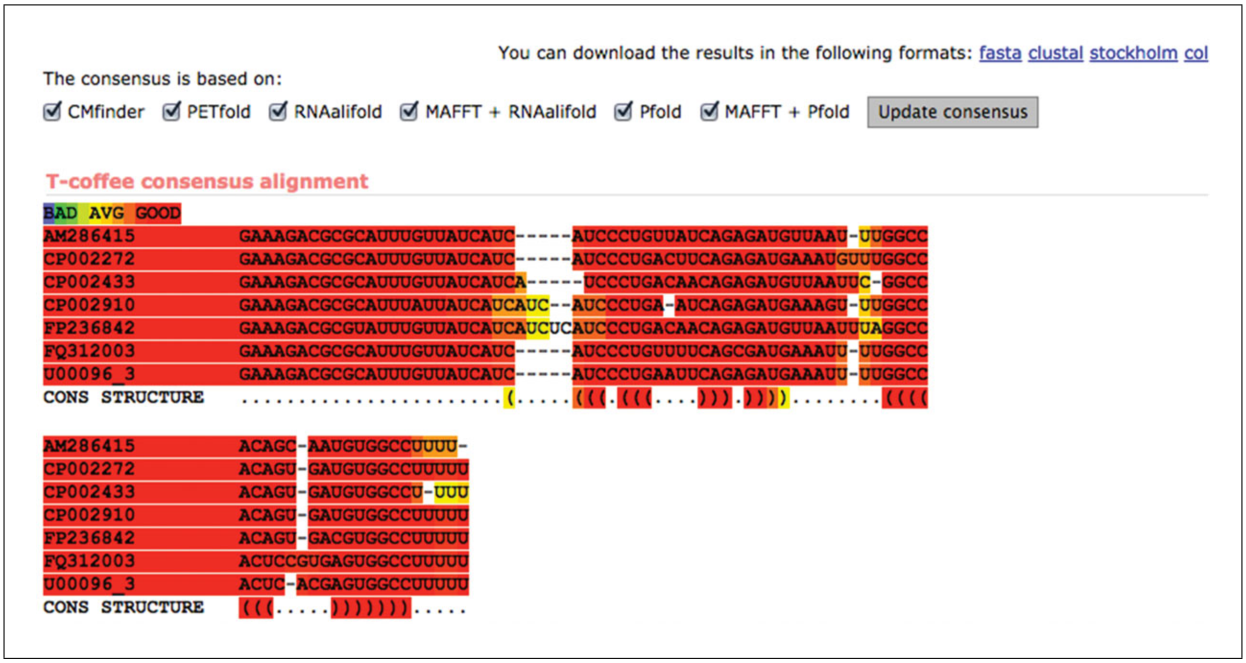
or intronic regions even over large evolutionary distances. If you are particularly ambitious, and the tools are available for your organisms of interest, you may wish to try to identify promoter sequences upstream of your candidate or terminator sequences downstream. If your sequence is a putative cis-regulatory element, such as a riboswitch, thermosensor, or attenuator, you may want to check that it occurs upstream of genes with similar functions or in similar pathways (Weinberg et al., 2015). Finally, it is always worth searching your putative homologs through the Rfam Web site even if your initial sequence had no matches . Rfam's models are not perfect, and may miss distant homologs of known families.

10. You can quickly examine your chosen sequences with the ENA browser, accessed by appending the accession number to *http://www.ebi.ac.uk/ena/data/view/* (Fig. 12.13.2). It appears that all of these sequences fall in an intergenic region between a *luxS* protein homolog and a *gshA* protein homolog, further increasing our confidence that these are true homologs. From our results, we can also see a few promising hits that do not quite meet our criteria, such as *Dickeya*, *Xenorhabdus*, *Photorhabdus*, and *Wigglesworthia*. We will keep these in mind later to expand our coverage.

BASIC PROTOCOL 2

ALIGNING AND PREDICTING SECONDARY STRUCTURE

We will use the WAR server to construct an initial alignment. Because of the criteria that we have set for sequence similarity in our gathering step, all of the sequences in our initial homolog set should have at least 60% pairwise sequence identity with at least one other sequence in the set. Under these conditions, alignment methods using primary sequence information only can perform adequately, as discussed in Background Information. These methods, combined with alignment folding tools that identify conserved structural signals and covariation, can produce reasonable secondary structures predictions (Gardner and Giegerich, 2004). However it is still often useful to observe the behavior of as many alignment tools as possible. Using WAR, for a fairly fast alignment we recommend running CMfinder, StrAL+PETfold, ClustalW, and MAFFT with RNAalifold and Pfold. WAR will also produce a consensus alignment using T-Coffee, which will attempt to



You can download the results in the following formats: fasta clustal stockholm col

The consensus is based on:

☑ CMfinder ☑ PETfold ☑ RNAalifold ☑ MAFFT + RNAalifold ☑ Pfold ☑ MAFFT + Pfold [Update consensus]

T-coffee consensus alignment

```
BAD AVG GOOD
AM286415         GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUAUCAGAGAUGUUAAU-UUGGCC
CP002272         GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGACUUCAGAGAUGAAAUGUUUGGCC
CP002433         GAAAGACGCGCAUUUGUUAUCAUCA-----UCCCUGACAACAGAGAUGUUAAUUC-GGCC
CP002910         GAAAGACGCGCAUUUAUUAUCAUCAUC--AUCCCUGA-AUCAGAGAUGAAAGU-UUGGCC
FP236842         GAAAGACGCGUAUUUGUUAUCAUCAUCUCAUCCCUGACAACAGAGAUGUUAAUUUAGGCC
FQ312003         GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUUUCAGCGAUGAAAUU-UUGGCC
U00096_3         GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGAAUUCAGAGAUGAAAUU-UUGGCC
CONS STRUCTURE   .......................(.....(((.(((....))).))))........((((

AM286415         ACAGC-AAUGUGGCCUUUU-
CP002272         ACAGU-GAUGUGGCCUUUUU
CP002433         ACAGU-GAUGUGGCCU-UUU
CP002910         ACAGU-GAUGUGGCCUUUUU
FP236842         ACAGU-GACGUGGCCUUUUU
FQ312003         ACUCCGUGAGUGGCCUUUUU
U00096_3         ACUC-ACGAGUGGCCUUUUU
CONS STRUCTURE   (((.....))))))).....
```

Figure 12.13.3 Consensus alignment of MicA sequences. Visualization of a T-coffee consensus alignment incorporating information from six different alignment and structure prediction methods on the WAR Web server.

find an alignment consistent with all of the individual alignments produced by other methods.

Necessary Resources

Computer with a modern Web browser (e.g., Firefox, Chrome, Internet Explorer)
Text editor

Submit your sequences to the WAR Web server

1. You will need to format your input sequence in FASTA format, an example can be found in step 8 of Basic Protocol 1. Navigate to *http://genome.ku.dk/resources/war/*. Paste this into the sequence box. Uncheck the tick boxes next to all methods besides CMfinder, Stral+PETfold, ClustalW+Pfold, ClustalW+RNAalifold, MAFFT+Pfold, and MAFFT+RNAalifold; this will limit WAR to running relatively fast methods. You can include additional methods in later runs to see how this affects the resulting alignments. Submit this together with your e-mail address. Using fast alignment methods you should get results within 10 min, assuming that the server is not busy.

Understanding WAR results

2. Once WAR returns your alignment results, there are a number of things you should take note of that will assist you in picking an alignment and in manual refinement. First, the consensus alignment page will display a graphical representation of the consistency of the alignments, which will allow you to quickly tell which areas of the alignment may require attention during manual refinement, or areas that may harbor structure not captured by the majority consensus. For instance, in our example MicA alignment (Fig. 12.13.3), there is some alignment and structural disagreement around the first alignment gap, suggesting that this area might require some manual curation to correct. The consensus can be recomputed based on differing subsets of alignment methods, if you believe one method (or set of methods) may be unduly influencing the consensus. Once you have carefully looked over the consensus alignment, examine each alignment produced by WAR in turn: What structures are shared? Where do

the alignments differ from each other? Can you identify any sequence or structural motifs which may help to guide your alignment? At this level of sequence identity, you should hope to see fairly consistent alignments in functional regions of the sequence, interspersed with more difficult-to-align regions, presumably under weaker selective pressure. Often the consensus alignment is a good choice to move forward with. However, there are cases where certain classes of tools will obviously misalign regions of the sequence and bias the consensus. Keep in mind what you have seen in the alternative alignments as well; this information may be useful in manual refinement. You will want to save the stockholm file for the alignment that you have chosen to your local computer at this point. This can be done by clicking the Stockholm link at the top right of any result page.

Aligning more distantly related sequences

3. Later in the family-building process when you have identified more distant homologs, the average pairwise identity of the sequences in your data set may have dropped below 60%. At this point, you may want to begin including some of the Sankoff-type alignment methods available in WAR. Using these methods can dramatically increase the runtime for your sequence alignment jobs, however, particularly for sequences over a couple of hundred of bases long. We will discuss alternatives to re-aligning sequences during the iterative expansion of the alignment in Basic Protocol 5.

BASIC PROTOCOL 3

GUIDANCE FOR MANUALLY REFINING ALIGNMENTS

Our goal in manual refinement is to attempt to correct errors made by automatic alignment tools. We generally use RALEE (Griffiths-Jones, 2005), an RNA editing mode for Emacs, for editing alignments. However, any editor you are comfortable with in which you can easily visualize sequence and structural conservation will work; a number of alternative editors are listed in the Strategic Planning section, above. RALEE offers coloring based on structure, sequence conservation, and compensatory mutations. These can be accessed from the Structure menu in Emacs (Fig. 12.13.4).

Necessary Resources

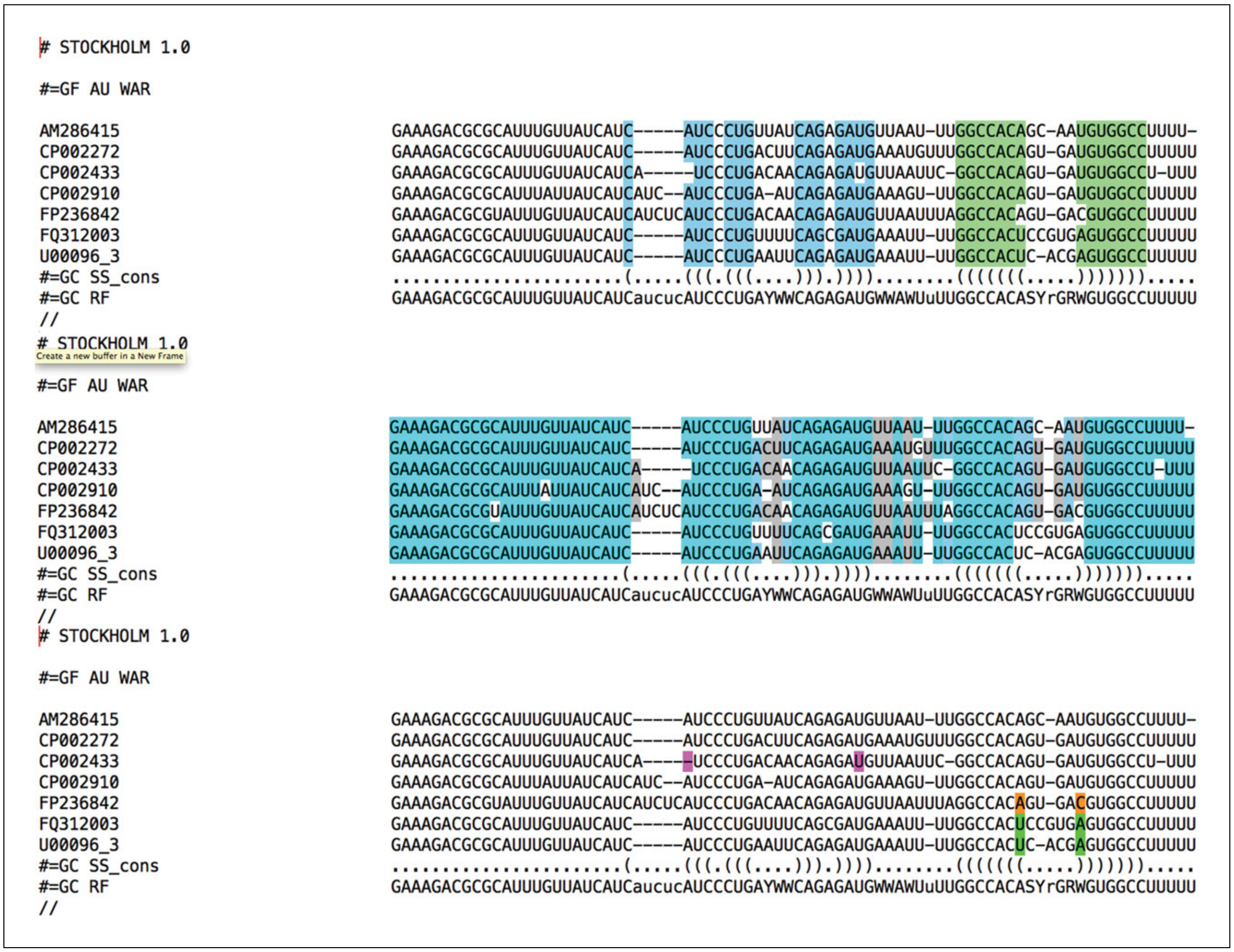
Computer, preferably running a *NIX-based operating system (e.g., Linux, MacOS X)
Emacs with RALEE mode installed (see *http://sgjlab.org/ralee/*)

1. A good place to start editing is around the edges of predicted hairpin structures. Are there base pairs that appear to be misaligned? Can you add base pairs to the structure? Are there predicted base pairs that do not appear to be well conserved that should be trimmed? Can individual bases be moved in the alignment to create more convincing support for the predicted structure?

2. Once you are satisfied with your manual refinement of predicted secondary structure elements, next, you should turn your attention to areas identified as uncertain in the WAR/T-Coffee consensus alignment. Were there alternative structures predicted in these regions? Do you see support for these structures in the sequences? If these regions are unstructured, can you identify any conserved sequence motifs in the region?

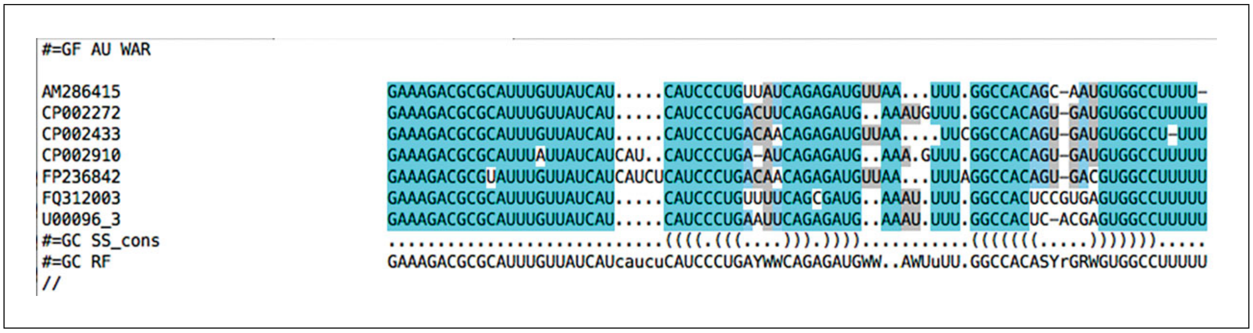
   If you will be regularly working with a particular class of ncRNA, it can be useful to familiarize yourself with predicted binding motifs of associated RNA-binding proteins, as these are likely to be conserved but can have many variable positions (Gardner and Eldai, 2015).

3. Our MicA consensus alignment provides a few opportunities for improvement (Fig. 12.13.5). The first base-pair in the first stem in CP002433 can be rescued



```
# STOCKHOLM 1.0

#=GF AU WAR

AM286415             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUAUCAGAGAUGUUAAU-UUGGCCACAGC-AAUGUGGCCUUUU-
CP002272             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGACUUCAGAGAUGAAAUGUUUGGCCACAGU-GAUGUGGCCUUUUU
CP002433             GAAAGACGCGCAUUUGUUAUCAUCA-----UCCCUGACAACAGAGAUGUUAAUUC-GGCCACAGU-GAUGUGGCCU-UUU
CP002910             GAAAGACGCGCAUUUAUUAUCAUCAUC--AUCCCUGA-AUCAGAGAUGAAAGU-UUGGCCACAGU-GAUGUGGCCUUUUU
FP236842             GAAAGACGCGUAUUUGUUAUCAUCAUCUCAUCCCUGACAACAGAGAUGUUAAUUUAGGCCACAGU-GACGUGGCCUUUUU
FQ312003             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUUUCAGCGAUGAAAUU-UUGGCCACUCCGUGAGUGGCCUUUUU
U00096_3             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGAAUUCAGAGAUGAAAUU-UUGGCCACUC-ACGAGUGGCCUUUUU
#=GC SS_cons         .......................(.....(((.(((....))).))))........(((((((.....))))))).....
#=GC RF              GAAAGACGCGCAUUUGUUAUCAUCaucucAUCCCUGAYWWCAGAGAUGWWAWUuUUGGCCACASYrGRWGUGGCCUUUUU
//
# STOCKHOLM 1.0

#=GF AU WAR

AM286415             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUAUCAGAGAUGUUAAU-UUGGCCACAGC-AAUGUGGCCUUUU-
CP002272             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGACUUCAGAGAUGAAAUGUUUGGCCACAGU-GAUGUGGCCUUUUU
CP002433             GAAAGACGCGCAUUUGUUAUCAUCA-----UCCCUGACAACAGAGAUGUUAAUUC-GGCCACAGU-GAUGUGGCCU-UUU
CP002910             GAAAGACGCGCAUUUAUUAUCAUCAUC--AUCCCUGA-AUCAGAGAUGAAAGU-UUGGCCACAGU-GAUGUGGCCUUUUU
FP236842             GAAAGACGCGUAUUUGUUAUCAUCAUCUCAUCCCUGACAACAGAGAUGUUAAUUUAGGCCACAGU-GACGUGGCCUUUUU
FQ312003             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUUUCAGCGAUGAAAUU-UUGGCCACUCCGUGAGUGGCCUUUUU
U00096_3             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGAAUUCAGAGAUGAAAUU-UUGGCCACUC-ACGAGUGGCCUUUUU
#=GC SS_cons         .......................(.....(((.(((....))).))))........(((((((.....))))))).....
#=GC RF              GAAAGACGCGCAUUUGUUAUCAUCaucucAUCCCUGAYWWCAGAGAUGWWAWUuUUGGCCACASYrGRWGUGGCCUUUUU
//
# STOCKHOLM 1.0

#=GF AU WAR

AM286415             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUAUCAGAGAUGUUAAU-UUGGCCACAGC-AAUGUGGCCUUUU-
CP002272             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGACUUCAGAGAUGAAAUGUUUGGCCACAGU-GAUGUGGCCUUUUU
CP002433             GAAAGACGCGCAUUUGUUAUCAUCA-----UCCCUGACAACAGAGAUGUUAAUUC-GGCCACAGU-GAUGUGGCCU-UUU
CP002910             GAAAGACGCGCAUUUAUUAUCAUCAUC--AUCCCUGA-AUCAGAGAUGAAAGU-UUGGCCACAGU-GAUGUGGCCUUUUU
FP236842             GAAAGACGCGUAUUUGUUAUCAUCAUCUCAUCCCUGACAACAGAGAUGUUAAUUUAGGCCACAGU-GACGUGGCCUUUUU
FQ312003             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGUUUUCAGCGAUGAAAUU-UUGGCCACUCCGUGAGUGGCCUUUUU
U00096_3             GAAAGACGCGCAUUUGUUAUCAUC-----AUCCCUGAAUUCAGAGAUGAAAUU-UUGGCCACUC-ACGAGUGGCCUUUUU
#=GC SS_cons         .......................(.....(((.(((....))).))))........(((((((.....))))))).....
#=GC RF              GAAAGACGCGCAUUUGUUAUCAUCaucucAUCCCUGAYWWCAGAGAUGWWAWUuUUGGCCACASYrGRWGUGGCCUUUUU
//
```

Figure 12.13.4 Editing alignments in Emacs RALEE mode. Screenshots showing three different color markups of the MicA consensus alignment highlighting secondary structure (top), sequence conservation (middle), and compensatory mutations (bottom).



```
#=GF AU WAR

AM286415             GAAAGACGCGCAUUUGUUAUCAU.....CAUCCCUGUUAUCAGAGAUGUUAA...UUU.GGCCACAGC-AAUGUGGCCUUUU-
CP002272             GAAAGACGCGCAUUUGUUAUCAU.....CAUCCCUGACUUCAGAGAUG..AAAUGUUU.GGCCACAGU-GAUGUGGCCUUUUU
CP002433             GAAAGACGCGCAUUUGUUAUCAU.....CAUCCCUGACAACAGAGAUGUUAA....UUCGGCCACAGU-GAUGUGGCCU-UUU
CP002910             GAAAGACGCGCAUUUAUUAUCAUCAU..CAUCCCUGA-AUCAGAGAUG..AAA.GUUU.GGCCACAGU-GAUGUGGCCUUUUU
FP236842             GAAAGACGCGUAUUUGUUAUCAUCAUCUCAUCCCUGACAACAGAGAUGUUAA...UUUAGGCCACAGU-GACGUGGCCUUUUU
FQ312003             GAAAGACGCGCAUUUGUUAUCAU.....CAUCCCUGUUUUCAGCGAUG..AAAU.UUU.GGCCACUCCGUGAGUGGCCUUUUU
U00096_3             GAAAGACGCGCAUUUGUUAUCAU.....CAUCCCUGAAUUCAGAGAUG..AAAU.UUU.GGCCACUC-ACGAGUGGCCUUUUU
#=GC SS_cons         ............................((((.(((....))).))))...........(((((((.....))))))).....
#=GC RF              GAAAGACGCGCAUUUGUUAUCAUcaucuCAUCCCUGAYWWCAGAGAUGWW..AWUuUU.GGCCACASYrGRWGUGGCCUUUUU
//
```

Figure 12.13.5 MicA alignment following manual refinement. Manual refinement using RALEE restores full conservation to the first stem-loop and removes an unlikely insertion within the hairpin structure. Additionally, careful manipulation of the sequence between the two hairpins reveals strong conservation of a short A/U-rich sequence motif that was previously obscure.

by shifting a few nucleotides, and by pulling apart the alignment between the first and second stem we reveal what appears to be a well-conserved AAUUU sequence motif that was previously hidden (Fig. 12.13.5). The bacterial RNA chaperone Hfq is known to bind to A/U rich sequences, so this motif may have some functional significance (Schu et al., 2015).

4. At this stage, it is also possible to include information from experimental data. Crystal structure information from a single sequence in the seed alignment can be used to validate and improve a predicted secondary structure. Tertiary structure–aware editors such as BoulderAle (Stombaugh et al., 2011) can help in applying this information to the alignment. Other experimental evidence, such as chemical footprinting, particularly when coupled with high-throughput sequencing (Spitale et al., 2014; Kwok et al., 2015), can also provide valuable information. Knowing whether even a single base is involved in a pairing interaction can drastically reduce the space of possible structures the sequence can fold into, simplifying the problem of predicting secondary structure. Both the RNAfold and RNAalifold Web servers available through the Vienna RNA Web site (Gruber et al., 2008b) are capable of taking advantage of this information in the form of folding constraints. Structural mapping datasets are beginning to be archived in consistent formats (Rocca-Serra et al., 2011; Cordero et al., 2012), and will be increasingly available for a variety of non-coding RNA molecules in the future.

BASIC PROTOCOL 4

BUILDING A COVARIANCE MODEL WITH INFERNAL

For those comfortable with the *NIX command line, building an Infernal CM is fairly straightforward. We refer the reader to the User's Guide available from the Infernal Web site (*http://eddylab.org/software/infernal/*) for installation instructions and a detailed tutorial.

Necessary Resources

Computer running a *NIX-based operating system (e.g., Linux, OS X)
Infernal (see *http://eddylab.org/infernal/*)

Run cmbuild

1. Assuming you have successfully installed Infernal somewhere in your path and you have saved your edited alignment in a stockholm file named `my.sto` (see Fig. 12.13.5 for an example of stockholm format), run:

```
> cmbuild my.cm my.sto
```

This will construct a CM and save the results in a file named `my.cm`. This command also returns some useful statistics on your model, such as the effective number of sequences and the relative entropy of your CM compared to an HMM with the same sequences, i.e., how much information is gained by including structural information in your model.

Run **`cmcalibrate`**

2. To generate E-values in search results, the CM must now be calibrated. This calibration depends on a sampling procedure, and so can be time-consuming. For our MicA CM, it will take about 5 min. Run:

```
> cmcalibrate my.cm
```

Note that calibration can take a long time—hours for longer models. You can get a quick estimate of the time calibration will take using the command:

```
> cmcalibrate --forecast 1 my.cm
```

Congratulations! You should now have a working CM for your RNA family. This is a fully capable model, and can be used as is for homology search and genome annotation. However, as it stands, your CM will only capture the sequence diversity which was able to be detected by our initial BLAST search. In order to fully take advantage of the power of CMs, you may want to expand the diversity of the sequence it is trained on through iterative expansion of our initial set of sequence homologs.

STRATEGIES FOR EXPANDING MODEL COVERAGE

BASIC PROTOCOL 5

Now that you have constructed your CM from BLAST results, you will want to try to expand your model coverage to fully capture the phylogenetic diversity of your sequence. There are several strategies for doing this that we discuss below, not all of which will work for every ncRNA.

Plan A: Iterative search of sequence databases

The method Rfam used to identify more divergent homologs to seed sequences prior to recent HMM-based acceleration of the Infernal pipeline (Nawrocki and Eddy, 2013; Nawrocki et al., 2015) was to pre-filter CM-based searches with WU-BLAST. This allows us to cover a large sequence space with a (comparatively) modest investment of computational time. Any of the single sequence search tools mentioned in the Strategic Planning section would make an effective pre-filter.

The easiest way to perform filtering yourself is to use the NCBI BLAST Web server to search each sequence in your seed alignment following the methods outlined for collecting your initial set of homologs in Basic Protocol 1. You may wish to relax the criteria slightly, then use the CM to perform a more sensitive search on this set of filtered sequences. This will enable you to detect more distantly related sequences, though you should always examine sequence context and the phylogenetic relationship between sequences as a sanity check before including them in your seed. These methods can be automated with basic scripting and bioinformatics modules such as BioPerl (Stajich et al., 2002) or Biopython (Cock et al., 2009), though this is beyond the scope of this chapter.

Once you have identified a new set of homologs, you can align them to your previous CM using Infernal's cmalign:

```
> cmalign --mapali my.sto my.cm newsequences.fasta >
  mynewalignment.sto
```

This `--mapali` option includes your original training alignment in the `mynewalignment.sto` output. This alignment can then be used to build a new CM, which will capture the additional sequence variation you have discovered in your BLAST searches.

The disadvantage of this method is that each search uses only the information available in a single sequence, meaning that valuable information about variation is lost, and, as a result, the power of the search suffers. Fast profile-based methods implemented in HMMER3 (Wheeler and Eddy, 2013) can be used to remedy this (Lindgreen et al., 2014), but require significant local computational resources to be used, in a manner similar to NCBI-BLAST.

Plan B: Directed search of chosen sequences

Another approach is to run the unfiltered CM over selected genomes or genomic regions. While the greater sensitivity and specificity of this method can help identify more distant homologs than is possible with BLAST, it has the disadvantage that it requires a much larger investment of computational resources to provide an equivalent phylogenetic coverage. This method can be particularly powerful in bacterial and archaeal genomes, where small genome size allows us to search a phylogenetically representative sample of genomes in less than a day on a desktop computer. In the case of larger eukaryotic genomes, it may be necessary to search a few genomes to determine if homologs of your

RNA are likely to exist in certain lineages, then extract homologous intergenic regions to continue searching. Our rationale here is much the same as in limiting the database for our initial BLAST search: by only looking in genomes where we have some prior belief that they may contain homologous sequence we reduce the noise in our low-scoring hits, meaning that we have to manually examine less hits to establish a score threshold for likely homologs.

Once you have examined candidates following the principles outlined earlier, it is easy to incorporate your new sequences using the easel package included with Infernal. First, search the genome, generating a tabfile:

```
>cmsearch --tblout searchfile.tab my.cm genome.fasta
```

Then use easel to index the genome and extract the hits:

```
> esl-sfetch --index genome.fasta
> esl-sfetch -Cf searchfile.tab genome.fasta >
  hits.fasta
```

These sequences can then be aligned and merged as with BLAST hits. Alternatively, if you discover a divergent lineage, it may be easiest to construct a separate alignment for these sequences, then use shared structural and sequence motifs to manually combine the two alignments. A Sankoff-type alignment method may also be useful for aligning divergent clades.

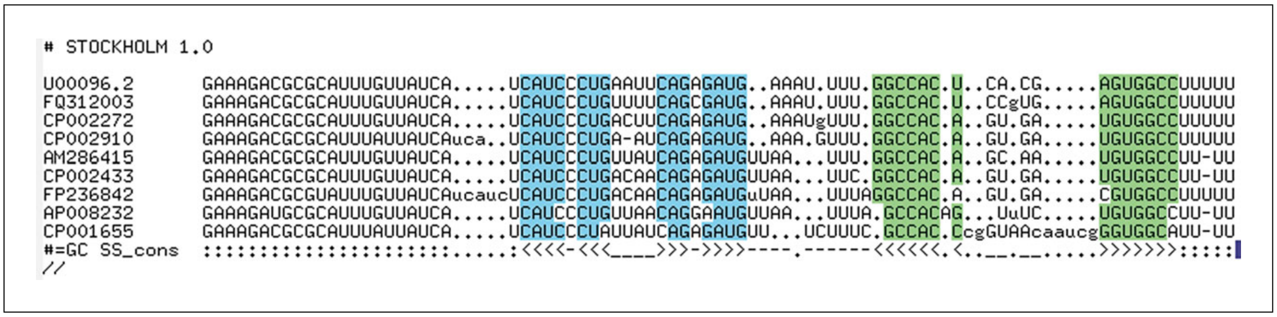
Plan C: When A and B fail...

In some cases, it will be very difficult to identify homologs of a candidate RNA across its full phylogenetic range. This can be because of high sequence variability, as in the Vault RNAs (Stadler et al., 2009). Alternatively, some longer RNAs, such as the RNA component of the telomerase ribonucleoprotein, consist of structurally conserved segments interspersed with long variable regions that cannot be easily discovered by standard search with naive covariance models (Xie et al., 2008; Marz et al., 2012).

A number of computational techniques exist for approaching these difficult cases, reviewed by Mosig and colleagues (Mosig et al., 2009). These methods include fragrep2 (Mosig et al., 2007), which allows the user to search fragmented conserved regions [including pol III promotor and terminator motifs (Gruber et al., 2008a; Stadler et al., 2009; Marz et al., 2009)], fragrep3, which allows the user to incorporate custom structural motifs with fragmented search, and GotohScan (Hertel et al., 2009), which implements a *semi-global* alignment algorithm that will align a query sequence to a (potentially) extended genomic region.

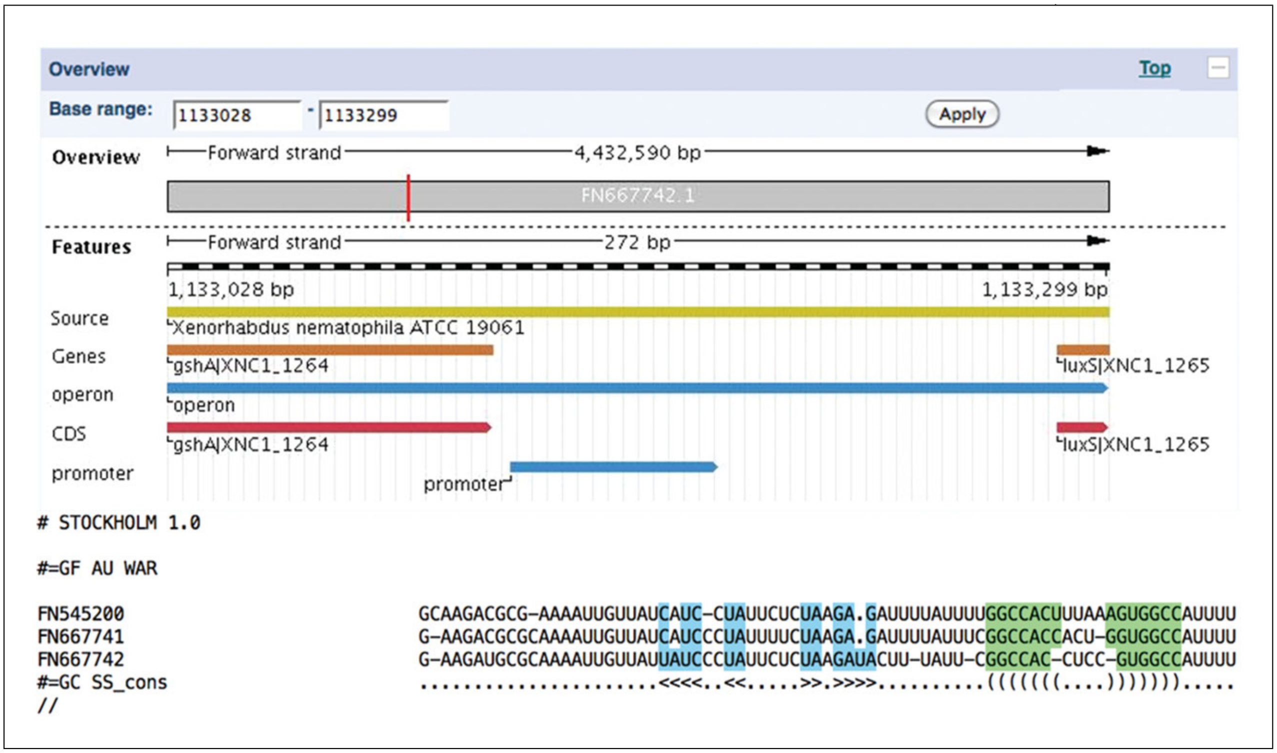
Applying plans A & B to MicA

Now we will follow Plan B to add sequences to our alignment using the genomes for the low-scoring BLAST hits we had previously made a note of while collecting our initial set of sequences, though you could also choose these sequences based on your knowledge of your organisms phylogeny or the suspected function of your RNA. The genomes we have chosen here are *Dickeya zeae* (CP001655), *Sodalis glossinidius* (AP008232), *Xenorhabdus nematophila* (FN667742), and *Wigglesworthia glosinidia* (BA000021); these can all be downloaded from the EMBL bacterial genomes pages (*http://www.ebi.ac.uk/genomes/bacteria.html*). Searching these genomes allows us to identify strong hits in *D. zeae* and *S. glossinidius* with E-values of 10^{-12} and 10^{-10}, respectively, which we can merge into our alignment using the methods described in Plan B above. You should then manually refine the resulting merged alignment with an eye towards maintaining conserved sequence motifs and structure. Already, at this



```
# STOCKHOLM 1.0

U00096.2       GAAAGACGCGCAUUUGUUAUCA.....UCAUCCCUGAAUUCAGAGAUG..AAAU.UUU.GGCCAC.U..CA.CG.....AGUGGCCUUUUU
FQ312003       GAAAGACGCGCAUUUGUUAUCA.....UCAUCCCUGUUUUCAGCGAUG..AAAU.UUU.GGCCAC.U..CCgUG.....AGUGGCCUUUUU
CP002272       GAAAGACGCGCAUUUGUUAUCA.....UCAUCCCUGACUUCAGAGAUG..AAAUgUUU.GGCCAC.A..GU.GA.....UGUGGCCUUUUU
CP002910       GAAAGACGCGCAUUUAUUAUCAuca..UCAUCCCUGA-AUCAGAGAUG..AAA.GUUU.GGCCAC.A..GU.GA.....UGUGGCCUUUUU
AM286415       GAAAGACGCGCAUUUGUUAUCA.....UCAUCCCUGUUAUCAGAGAUGUUAA...UUU.GGCCAC.A..GC.AA.....UGUGGCCUU-UU
CP002433       GAAAGACGCGCAUUUGUUAUCA.....UCAUCCCUGACAACAGAGAUGUUAA...UUC.GGCCAC.A..GU.GA.....UGUGGCCUU-UU
FP236842       GAAAGACGCGUAUUUGUUAUCAucaucUCAUCCCUGACAACAGAGAUGuUAA...UUUAGGCCAC.A..GU.GA.....CGUGGCCUUUUU
AP008232       GAAAGAUGCGCAUUUGUUAUCA.....UCAUCCCUGUUAACAGGAAUGUUAA...UUUA.GCCACAG...UuUC.....UGUGGCCUU-UU
CP001655       GAAAGACGCGCAUUUAUUAUCA.....UCAUCCCUAUUAUCAGAGAUGUU...UCUUUC.GCCAC.CcgGUAAcaaucgGGUGGCAUU-UU
#=GC SS_cons   ::::::::::::::::::::::.....:<<<<-<<<____>>>->>>>----.------<<<<<<.<..__.__.....>>>>>>>:::::
//
```

Figure 12.13.6 MicA alignment containing additional homologs found using Plan B. Two new sequences have been added to the alignment (*D. zeae*, CP001655; *S. glossinidius,* AP008232), adding structural and sequence diversity.



```
# STOCKHOLM 1.0

#=GF AU WAR

FN545200                      GCAAGACGCG-AAAAUUGUUAUCAUC-CUAUUCUCUAAGA.GAUUUUAUUUUGGCCACUUUAAAGUGGCCAUUUU
FN667741                      G-AAGACGCGCAAAAUUGUUAUCAUCCCUAUUUUCUAAGA.GAUUUUAUUUCGGCCACCACU-GGUGGCCAUUUU
FN667742                      G-AAGAUGCGCAAAAUUGUUAUUAUCCCUAUUCUCUAAGAUACUU-UAUU-CGGCCAC-CUCC-GUGGCCAUUUU
#=GC SS_cons                  .......................<<<<..<<.....>>.>>>>..........((((((((....)))))))).....
//
```

Figure 12.13.7 Xenorhabdus MicA homologs found using Plan A. The top panel shows the location of a putative *micA* homolog with a marginal Infernal *E*-value in *X. nematophila*, sharing synteny with established *micA* sequences in *E. coli*. The bottom panel shows a manually refined alignment of three putative *Xenorhabdus* MicA sequences, displaying structural and sequence similarity to our previously constructed alignment of enterobacterial MicA sequences.

evolutionary distance, there has been some apparent small decay in secondary structure, as well as an expansion of the sequence contained in the loop region of the second stem in *D. zeae* (Fig. 12.13.6).

We observe a number of hits in *X. nematophila* with E-values in the range of 10^{-2}, and we can apply Plan A to investigate these. By checking each of these individually in the ENA browser, we can identify one that falls in the same genomic context as our previous MicA homologs (Fig. 12.13.7). By using this sequence as the starting point for a BLAST search, we can identify a number of other divergent *Xenorhabdus* homologs. As these are quite diverged from the *E. coli* sequence, we first construct an alignment for them using WAR (Fig. 12.13.7), then attempt to merge our alignments manually using shared structural features as our guide. Interestingly, the target-binding region of MicA appears to have suffered a poly(A) insertion along this lineage, suggesting that there may be changes in the regulon it targets, in line with recent findings that RNA:RNA interactions are frequently poorly conserved (Richter and Backofen, 2012; Lai and Meyer, 2015).

Using this model to search all of the bacterial genomes in EMBL-Bank (approximately 6 GB of sequence, taking ~30 hr on a 2.26-GHz Intel Core 2 Duo processor) shows that our CM now has high-scoring hits exclusively in Enterobacteriales, while covering a broader range than our initial BLAST searches. This search also reveals a number of possible sources of additional diversity. *Photorhabdus asymbiotica* and *Edwardsiella ictaluri* both have strong hits below the average score for other enterobacterial genomes —incorporating them may further increase the sensitivity of our model, and is left as an exercise to the reader.

GUIDELINES FOR UNDERSTANDING RESULTS

Given the diversity of ncRNAs in sequence, structure, function, and conservation, it is difficult to provide general guidelines for interpreting results, though we will attempt to provide some guidance here. The best guard against spurious results is maintaining a skeptical attitude during family building. We have suggested some questions to ask yourself during family building above, but they bear repeating: does the phylogenetic distribution make sense? Can it be explained by vertical inheritance? The appearance of a sequence from a distantly related phylum in Infernal search results is far more likely to be a false positive than horizontal transfer in most cases. Do the expanded alignments make sense? Are important structural and sequence features (e.g., intermolecular interaction sites, protein binding sites, stabilizing secondary structures) conserved? If additional sequences no longer contain these important features, you can assume you have hit the limit of your family's conservation.

Knowing a "good" alignment is more art than science, and often requires experience and domain-specific knowledge of the molecules being aligned. Articles in the RNA Families track at *RNA Biology* provide good, reviewed examples of RNA alignments for a range of RNA classes, and may serve as instructive examples. Covariation is often taken as gold-standard evidence for the conservation of secondary structure, and can be visualized in RALEE. However, it should be noted that many secondary structure–aware RNA alignment tools explicitly attempt to maximize covariation; this combined with the tendency of many aligners to produce alignments for any given set of sequences, even if they are not homologous (see discussion in the Commentary below), can lead to an artificial prediction of high levels of covariation in secondary structure. This can generally be avoided by not including low-E-value sequences in your alignment without strong additional evidence for homology (e.g., experimental results, conserved synteny, conservation of defining structural and sequence features). Even in the case of good alignments, knowing what degree of covariation to expect in bona fide secondary structure is often difficult. Ultimately, an RNA family only provides a hypothesis concerning the evolutionary relationship between a set of sequences. How good that hypothesis is depends on many factors discussed in this article, but an attitude of healthy skepticism and understanding of the biology of the family on the part of the builder are perhaps the most important.

COMMENTARY

Background Information

RNA sequence alignment remains a challenge despite at least 30 years of work on the problem (Woese et al., 1983; Gutell et al., 1985; Lane et al., 1985; Pace et al., 1989). As discussed in the introduction, alignments based on primary sequence become untrustworthy below ~60% pairwise sequence identity, likely due to the lower information content of individual nucleic acids as compared to amino acids in protein alignments. This can be intuitively understood by recalling the fact that 3 nucleotides are required to encode an individual amino acid; so, an amino acid carries 3 times as much information as a nucleic acid (a bit less, actually, due to the redundancy of the genetic code). In addition, the larger alphabet size of protein sequences allows for the easy deployment of more complex substitution models, and a glut of protein sequence data allows for highly effective parameterization of these models.

The incorporation of secondary structure—i.e., base-pairing—information has been proposed as a means to make up for these difficulties in RNA alignment methods. The first proposal for such a method is now known as the Sankoff algorithm (Sankoff, 1985). The Sankoff algorithm uses dynamic programming, an optimization technique central to sequence analysis. A full explanation of dynamic programming is beyond the scope of this unit, but for a brief introduction see two excellent primers by Sean Eddy covering applications to alignment (Eddy, 2004b) and secondary structure prediction (Eddy, 2004a); for those seeking a deeper understanding, Durbin et al. (1998) provides detailed coverage of dynamic programming as well as covariance models. Dynamic programming had previously been applied to the problems of sequence alignment (Needleman and Wunsch, 1970) and RNA folding (Nussinov et al., 1978). Sankoff proposed a union of these two methods. Unfortunately, the resulting algorithm has a time requirement of $O(L^{3N})$ and space requirements of $O(L^{2N})$, where is the sequence length and is the number of sequences aligned. This is impractical, even for small numbers of short sequences. A number of faster heuristic algorithms have been developed to approximate Sankoff alignment. Recent examples include CentroidAlign (Hamada et al., 2009b), mLocARNA (Will et al., 2007), and FoldalignM (Torarinsson et al., 2007). These methods can push the RNA alignment twilight zone as low as 40% identity (Gardner et al., 2005).

However, for the purpose of family building, we are often starting with a single sequence of unknown secondary structure and have to gather additional homologs using a fast alignment tool, such as BLAST. This is a common starting point in analyzing noncoding coding regions from RNA sequence experiments, for example (Perkins et al., 2009; Chaudhuri et al., 2011; Lindgreen et al., 2014). Unfortunately, BLAST and similar single-sequence homology search methods are not able to reliably detect homologs below 60% sequence identity. In this range of pairwise sequence identities, the slight increases in accuracy of Sankoff-type algorithms over non-structural alignment is only rarely worth the additional computational costs involved. [For relatively recent benchmarks of alignment tools on ncRNA sequences, see Hamada et al. (2009b) and the supplementary information of Bradley et al. (2008); Hamada includes comparisons of aligner runtimes, while Bradley examines relative performance over a range of pair-wise sequence percent identities.] Alignments generated with standard alignment tools can then be used as a basis for predictions of secondary structure using tools like Pfold (Knudsen and Hein, 2003), RNAalifold (Bernhart et al., 2008), or CentroidFold (Hamada et al., 2009a).

Regardless, all modern alignment tools, Sankoff-type or standard, suffer from a number of known problems. Most alignment tools use *progressive* alignment. This means that the aligner decomposes the alignment problem into a series of pair-wise alignment problems along a guide tree built using some measure of sequence similarity, so that highly similar sequences are aligned first before alignments between more divergent sequences are attempted. This greatly reduces the computational complexity of the alignment problem, but also means that errors in early pairwise alignment steps are propagated through the entire alignment. A number of solutions have been proposed to this problem, such as explicitly modeling insertion-deletion histories (Löytynoja and Goldman, 2008) or using modified or alternative optimization methods such as consistency-guided progressive alignment (Notredame et al., 2000) or sequence annealing (Schwartz and Pachter, 2007). A second issue is that it is not clear which function of the alignment aligners should be optimizing (Iantorno et al., 2014), and many appear to over-predict homology (Schwartz et al., 2005; Bradley et al., 2008, 2009). Finally, many parameters commonly used in alignment, such as gap opening and closing probabilities and substitution matrices, appear to vary across organisms, sequences, and even positions within an alignment. All of this leads to considerable uncertainty in alignment (Wong et al., 2008), which is not easily captured by most current alignment methods, although sampling-based methods do exist that attempt to account for this (Suchard and Redelings, 2006; Westesson et al., 2012). The additional parameters and uncertainty introduced by RNA secondary structure prediction only compound these problems.

An additional problem with alignment is the issue of determining whether two sequences are similar due to *homology* or *analogy*. Homology describes a similarity in features based on common descent; for instance, all bird wings are homologous wings.

Analogy, on the other hand, describes a similarity in features based on common function without common descent; bat and bird wings perform the same function and appear superficially similar; however, their evolutionary histories are quite different. In sequence analysis, we often assume that aligned residues within an alignment share common ancestry, but this assumption can be confounded by analogous sequence. These analogs often take the form of *motifs*, short sequences that perform specific functions within the RNA molecule and that can arise easily through convergent evolution (Gardner and Eldai, 2015). An example of such a motif is the bacterial rho-independent terminator (Gardner et al., 2011), a short hairpin responsible for halting transcription in many species. While such motifs can be a boon for discovering novel ncRNA genes (Argaman et al., 2001; Livny et al., 2005) or aligning homologs that contain them (Nawrocki et al., 2015), they can also be a source of false positives when attempting to build an alignment of homologous sequences.

Acknowledgements

This work was supported by the Wellcome Trust, grant number WT098051. L.B. was supported by a Research Fellowship from the Alexander von Humboldt Stiftung/Foundation. PPG is supported by a Rutherford Discovery Fellowship.

Literature Cited

Alikhan, N.-F., Petty, N.K., Ben Zakour, N.L., and Beatson, S.A. 2011. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410. doi: 10.1016/S0022-2836(05)80360-2.

Andersen, E.S., Lind-Thomsen, A., Knudsen, B., Kristensen, S.E., Havgaard, J.H., Torarinsson, E., Larsen, N., Zwieb, C., Sestoft, P., Kjems, J., and Gorodkin, J. 2007. Semiautomated improvement of RNA alignments. *RNA* 13:1850-1859. doi: 10.1261/rna.215407.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G.H., Margalit, H., and Altuvia, S. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11:941-950. doi: 10.1016/S0960-9822(01)00270-6.

Asai, K., Kiryu, H., Hamada, M., Tabei, Y., Sato, K., Matsui, H., Sakakibara, Y., Terai, G., and Mituyama, T. 2008. Software.ncrna.org: Web servers for analyses of RNA sequences. *Nucleic Acids Res.* 36:W75-W78. doi: 10.1093/nar/gkn222.

Barquist, L. and Vogel, J. 2015. Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.* 49:367-394. doi: 10.1146/annurev-genet-112414-054804.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., and Soboleva, A. 2013. NCBI GEO: Archive for functional genomics data sets-update. *Nucleic Acids Res.* 41:D991-D995. doi: 10.1093/nar/gks1193.

Barrick, J.E. and Breaker, R.R. 2007. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol.* 8:R239. doi: 10.1186/gb-2007-8-11-r239.

Barrick, J.E., Sudarsan, N., Weinberg, Z., Ruzzo, W.L., and Breaker, R.R. 2005. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA* 11:774-784. doi: 10.1261/rna.7286705.

Bateman, A., Agrawal, S., Birney, E., Bruford, E.A., Bujnicki, J.M., Cochrane, G., Cole, J.R., Dinger, M.E., Enright, A.J., Gardner, P.P., Gautheret, D., Griffiths-Jones, S., Harrow, J., Herrero, J., Holmes, I.H., Huang, H.D., Kelly, K.A., Kersey, P., Kozomara, A., Lowe, T.M., Marz, M., Moxon, S., Pruitt, K.D., Samuelsson, T., Stadler, P.F., Vilella, A.J., Vogel, J.H., Williams, K.P., Wright, M.W., and Zwieb, C. 2011. RNAcentral: A vision for an international database of RNA sequences. *RNA* 17:1941-1946. doi: 10.1261/rna.2750811.

Bauer, M., Klau, G.W., and Reinert, K. 2007. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* 8:271. doi: 10.1186/1471-2105-8-271.

Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., and Stadler, P.F. 2008. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474. doi: 10.1186/1471-2105-9-474.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988-995. doi: 10.1101/gr.1865504.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., and Zaretskaya, I. 2013. BLAST: A more efficient report with usability improvements. *Nucleic Acids Res.* 41:W29-W33. doi: 10.1093/nar/gkt282.

Bradley, R.K., Pachter, L., and Holmes, I. 2008. Specific alignment of structured RNA: Stochastic grammars and sequence annealing. *Bioinformatics* 24:2677-2683. doi: 10.1093/bioinformatics/btn495.

Bradley, R.K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. 2009. Fast statistical alignment. *PLoS Comput. Biol.* 5:e1000392. doi: 10.1371/journal.pcbi.1000392.

Brownlee, G.G. 1971. Sequence of 6S RNA of *E. coli. Nature* 229:147-149. doi: 10.1038/229147a0.

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* 41:D226-D232. doi: 10.1093/nar/gks1005.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., and McQuillan, J.A. 2012. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28:464-469. doi: 10.1093/bioinformatics/btr703.

Carver, T.J., Rutherford, KM., Berriman, M., Rajandream, M.-A., Barrell, B.G., and Parkhill, J. 2005. ACT: The artemis comparison tool. *Bioinformatics* 21:3422-3423. doi: 10.1093/bioinformatics/bti553.

Chan, P.P., Holmes, A.D., Smith, A.M., Tran, D., and Lowe, T.M. 2012. The UCSC archaeal genome browser: 2012 update. *Nucleic Acids Res.* 40:D646-D652. doi: 10.1093/nar/gkr990.

Chaudhuri, R.R., Yu, L., Kanji, A., Perkins, T.T., Gardner, P.P., Choudhary, J., Maskell, D.J., and Grant, A.J. 2011. Quantitative RNA-seq analysis of the *Campylobacter jejuni* transcriptome. *Microbiology* 157:2922-2932. doi: 10.1099/mic.0.050278-0.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422-1423. doi: 10.1093/bioinformatics/btp163.

Cordero, P., Lucks, J.B., and Das, R. 2012. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics* 28:3006-3008. Available at: *http://bioinformatics.oxfordjournals.org/content/28/22/3006.short*. doi: 10.1093/bioinformatics/bts554.

Dalli, D., Wilm, A., Mainz, I., and Steger, G. 2006. STRAL: Progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* 22:1593-1599. doi: 10.1093/bioinformatics/btl142.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

Eddy, S.R. 2004a. How do RNA folding algorithms work? *Nat. Biotechnol.* 22:1457-1458. doi: 10.1038/nbt1104-1457.

Eddy, S.R. 2004b. What is dynamic programming? *Nat. Biotechnol.* 22:909-910. doi: 10.1038/nbt0704-909.

Eddy, S.R. 2011. Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Eddy, S.R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* 22:2079-2088. Available at: *http://nar.oxfordjournals.org/content/22/11/2079.short*. doi: 10.1093/nar/22.11.2079.

ENCODE Project Consortium 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9:e1001046. doi: 10.1371/journal.pbio.1001046.

Finn, R.D., Clements, J., and Eddy, S.R. 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29-W37. doi: 10.1093/nar/gkr367.

Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. 2015. HMMER web server: 2015 update. *Nucleic Acids Res.* 43:W30-W38. Available at: *http://dx.doi.org/10.1093/nar/gkv397*. doi: 10.1093/nar/gkv397.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., and Punta, M. 2014. Pfam: The protein families database. *Nucleic Acids Res.* 42:D222-D230. doi: 10.1093/nar/gkt1223.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., and Searle, S.M. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749-D755. doi: 10.1093/nar/gkt1196.

Freyhult, E.K., Bollback, J.P., and Gardner, P.P. 2007. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.* 17:117-125. doi: 10.1101/gr.5890907.

Fu, Y., Deiorio-Haggar, K., Anthony, J., and Meyer, M.M. 2013. Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria. *Nucleic Acids Res.* 41:3491-3503. doi: 10.1093/nar/gkt055.

Gardner, P.P. 2009. The use of covariance models to annotate RNAs in whole genomes. *Brief. Funct. Genomic Proteomic* 8:444-450. doi: 10.1093/bfgp/elp042.

Gardner, P.P. and Bateman, A.G. 2009. A home for RNA families at RNA Biology. *RNA Biol.* 6:2-4. doi: 10.4161/rna.6.1.7635.

Gardner, P.P. and Eldai, H. 2015. Annotating RNA motifs in sequences and alignments. *Nucleic Acids Res.* 43:691-698. doi: 10.1093/nar/gku1327.

Gardner, P.P. and Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5:140. doi: 10.1186/1471-2105-5-140.

Gardner, P.P., Wilm, A., and Washietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* 33:2433-2439. doi: 10.1093/nar/gki541.

Gardner, P.P., Bateman, A., and Poole, A.M. 2010. SnoPatrol: How many snoRNA genes are there? *J. Biol.* 9:4. doi: 10.1186/jbiol211.

Gardner, P.P., Barquist, L., Bateman, A., Nawrocki, E.P., and Weinberg, Z. 2011. RNIE: Genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res.* 39:5845-5852. doi: 10.1093/nar/gkr168.

Gogol, E.B., Rhodius, V.A., Papenfort, K., Vogel, J., and Gross, C.A. 2011. Small RNAs endow a transcriptional activator with essential repressor functions for single-tier control of a global stress regulon. *Proc. Natl. Acad. Sci. U.S.A.* 108:12875-12880. doi: 10.1073/pnas.1109379108.

Griffiths-Jones, S. 2005. RALEE-RNA ALignment editor in Emacs. *Bioinformatics* 21:257-259. doi: 10.1093/bioinformatics/bth489.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. 2008b. The Vienna RNA websuite. *Nucleic Acids Res.* 36:W70-W74. doi: 10.1093/nar/gkn188.

Gruber, A.R., Kilgus, C., Mosig, A., Hofacker, I.L., Hennig, W., and Stadler, P.F. 2008a. Arthropod 7SK RNA. *Mol. Biol. Evol.* 25:1923-1930. doi: 10.1093/molbev/msn140.

Gutell, R.R., Weiser, B., Woese, C.R., and Noller, H.F. 1985. Comparative anatomy of 16-S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* 32:155-216. doi: 10.1016/S0079-6603(08)60348-7.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. 2009a. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* 25:465-473. doi: 10.1093/bioinformatics/btn601.

Hamada, M., Sato, K., Kiryu, H., Mituyama, T., and Asai, K. 2009b. CentroidAlign: Fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics* 25:3236-3243. doi: 10.1093/bioinformatics/btp580.

Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B., and Stadler, P.F. 2009. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res.* 37:1602-1615. doi: 10.1093/nar/gkn1084.

Höchsmann, M., Töller, T., Giegerich, R., and Kurtz, S. 2003. Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 2:159-168.

Hoeppner, M.P., Barquist, L.E., and Gardner, P.P. 2014. An introduction to RNA databases. *In* RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods Methods in Molecular Biology (J. Gorodkin and W.L. Ruzzo, eds.) pp. 107-123. Humana Press, Totowa, N.J.

Iantorno, S., Gori, K., Goldman, N., Gil, M., and Dessimoz, C. 2014. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *In* Multiple Sequence Alignment Methods Methods in Molecular Biology, pp. 59-73. Humana Press, Totowa, N.J.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. 2008. NCBI BLAST: A better web interface. *Nucleic Acids Res.* 36:W5-W9. doi: 10.1093/nar/gkn201.

Jossinet, F. and Westhof, E. 2005. Sequence to Structure (S2S): Display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* 21:3320-3321. doi: 10.1093/bioinformatics/bti504.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., and Kent, W.J. 2014. The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* 42:D764-D770. doi: 10.1093/nar/gkt1168.

Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772-780. doi: 10.1093/molbev/mst010.

Kent, W.J. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* 12:656-664. doi: 10.1101/gr.229202.

Kiryu, H., Tabei, Y., Kin, T., and Asai, K. 2007. Murlet: A practical multiple alignment tool for structural RNA sequences. *Bioinformatics* 23:1588-1598. doi: 10.1093/bioinformatics/btm146.

Knudsen, B. and Hein, J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 31:3423-3428. doi: 10.1093/nar/gkg614.

Kwok, C.K., Tang, Y., Assmann, S.M., and Bevilacqua, P.C. 2015. The RNA structurome: Transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.* 40:221-232. doi: 10.1016/j.tibs.2015.02.005.

Lai, D. and Meyer, I.M. 2015. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Res.* Available at: *http://nar.oxfordjournals.org/content/early/2015/12/15/nar.gkv1477.abstract*. doi: 10.1093/nar/gkv1477.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.* 82:6955-6959. doi: 10.1073/pnas.82.20.6955.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948. doi: 10.1093/bioinformatics/btm404.

Lindgreen, S., Gardner, P.P., and Krogh, A. 2007. MASTR: Multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* 23:3304-3311. doi: 10.1093/bioinformatics/btm525.

Lindgreen, S., Umu, S.U., Lai, A.S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N.E., Biggs, P.J., Thomson, N.R., Barquist, L., Poole, A.M., and Gardner, P.P. 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.* 10:e1003907. doi: 10.1371/journal.pcbi.1003907.

Livny, J., Fogel, M.A., Davis, B.M., and Waldor, M.K. 2005. sRNAPredict: An integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res.* 33:4096-4105. doi: 10.1093/nar/gki715.

Löytynoja, A. and Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320:1632-1635. doi: 10.1126/science.1158395.

Marz, M., Mosig, A., Podlevsky, J.D., and Stadler, P.F. 2012. The common ancestral core of vertebrate and fungal telomerase RNAs. *Nucleic Acids Res.* Available at: *http://nar.oxfordjournals.org/content/early/2012/10/23/nar.gks980.short*. doi: 10.1093/nar/gks980.

Marz, M., Donath, A., Verstraete, N., Nguyen, V.T., Stadler, P.F., and Bensaude, O. 2009. Evolution of 7SK RNA and its protein partners in metazoa. *Mol. Biol. Evol.* 26:2821-2830. doi: 10.1093/molbev/msp198.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. 2013. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* 41:W597-W600. doi: 10.1093/nar/gkt376.

Menzel, P., Gorodkin, J., and Stadler, P.F. 2009. The tedious task of finding homologous noncoding RNA genes. *RNA* 15:2075-2082. doi: 10.1261/rna.1556009.

Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., Yoshizawa, A., Komori, T., and Asai, K. 2009. The Functional RNA Database 3.0: Databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.* 37:D89-D92. doi: 10.1093/nar/gkn805.

Mosig, A., Chen, J.J.-L., and Stadler, P.F. 2007. Homology search with fragmented nucleic acid sequence patterns. *In* Algorithms in Bioinformatics Lecture Notes in Computer Science, pp. 335-345. Springer, Berlin, Heidelberg.

Mosig, A., Zhu, L., and Stadler, P.F. 2009. Customized strategies for discovering distant ncRNA homologs. *Brief Funct. Genomic Proteomic* 8:451-460. doi: 10.1093/bfgp/elp035.

Myslinski, E., Ségault, V., and Branlant, C. 1990. An intron in the genes for U3 small nucleolar RNAs of the yeast *Saccharomyces cerevisiae*. *Science* 247:1213-1216. doi: 10.1126/science.1690452.

Nawrocki, E.P. 2014. Annotating functional RNAs in genomes using Infernal. *Methods Mol. Biol.* 1097:163-197. doi: 10.1007/978-1-62703-709-9_9.

Nawrocki, E.P. and Eddy, S.R. 2007. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.* 3:e56. Available at: *http://dx.plos.org/10.1371/journal.pcbi.0030056*. doi: 10.1371/journal.pcbi.0030056.

Nawrocki, E.P. and Eddy, S.R. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933-2935. doi: 10.1093/bioinformatics/btt509.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* 25:1335-1337. doi: 10.1093/bioinformatics/btp157.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., and Finn, R.D. 2015. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* 43:D130-D137. doi: 10.1093/nar/gku1063.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453. doi: 10.1016/0022-2836(70)90057-4.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205-217. doi: 10.1006/jmbi.2000.4042.

Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* 35:68-82. doi: 10.1137/0135006.

Pace, N.R., Smith, D.K., Olsen, G.J., and James, B.D. 1989. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA—a review. *Gene* 82:65-75. doi: 10.1016/0378-1119(89)90031-0.

Perkins, T.T., Kingsley, R.A., Fookes, M.C., Gardner, P.P., James, K.D., Yu, L., Assefa, S.A., He, M., Croucher, N.J., Pickard, D.J., Maskell, D.J., Parkhill, J., Choudhary, J., Thomson, N.R., and Dougan, G. 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet.* 5:e1000569. doi: 10.1371/journal.pgen.1000569.

Puton, T., Kozlowski, L.P., Rother, K.M., and Bujnicki, J.M. 2014. CompaRNA: A server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.* 42:5403-5406. doi: 10.1093/nar/gku208.

Reeder, J. and Giegerich, R. 2005. Consensus shapes: An alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* 21:3516-3523. doi: 10.1093/bioinformatics/bti577.

Richter, A. and Backofen, R. 2012. Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA Biol.* 9:954-965. doi: 10.4161/rna.20294.

Rinn, J.L. and Chang, H.Y. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81:145-166. doi: 10.1146/annurev-biochem-051410-092902.

RNAcentral Consortium. 2015. RNAcentral: An international database of ncRNA sequences. *Nucleic Acids Res.* 43:D123-D129. doi: 10.1093/nar/gku991.

Roberts, E., Eargle, J., Wright, D., and Luthey-Schulten, Z. 2006. MultiSeq: Unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics* 7:382. doi: 10.1186/1471-2105-7-382.

Rocca-Serra, P., Bellaousov, S., Birmingham, A., Chen, C., Cordero, P., Das, R., Davis-Neulander, L., Duncan, C.D.S., Halvorsen, M., Knight, R., Leontis, N.B., Mathews, D.H., Ritz, J., Stombaugh, J., Weeks, K.M., Zirbel, C.L., and Laederach, A. 2011. Sharing and archiving nucleic acid structure mapping data. *RNA* 17:1204-1212. doi: 10.1261/rna.2753211.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85-94. doi: 10.1093/protein/12.2.85.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjölander, K., Underwood, R.C., and Haussler, D. 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* 22:5112-5120. Available at: *http://nar.oxfordjournals.org/content/22/23/5112.short*. doi: 10.1093/nar/22.23.5112.

Sankoff, D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* 45:810-825. doi: 10.1137/0145048.

Schneider, K.L., Pollard, K.S., Baertsch, R., Pohl, A., and Lowe, T.M. 2006. The UCSC archaeal genome browser. *Nucleic Acids Res.* 34:D407-D410. doi: 10.1093/nar/gkj134.

Schu, D.J., Zhang, A., Gottesman, S., and Storz, G. 2015. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *EMBO J.* 34:2557-2573. doi: 10.15252/embj.201591569.

Schwartz, A.S. and Pachter, L. 2007. Multiple alignment by sequence annealing. *Bioinformatics* 23:e24-e29. doi: 10.1093/bioinformatics/btl311.

Schwartz, A.S., Myers, E.W., and Pachter, L. 2005. Alignment metric accuracy. *arXiv [q-bio.QM]*. Available at: *http://arxiv.org/abs/q-bio/0510052*.

Seemann, S.E., Richter, A.S., Gesell, T., Backofen, R., and Gorodkin, J. 2011. PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* 27:211-219. doi: 10.1093/bioinformatics/btq634.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P.F., and Vogel, J. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464:250-255. doi: 10.1038/nature08756.

Silvester, N., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Ten Hoopen, P., Kay, S., Leinonen, R., Li, W., Liu, X., Lopez, R., Pakseresht, N., Pallreddy, S., Plaister, S., Radhakrishnan, R., Rossello, M., Senf, A., Smirnov, D., Toribio, A.L., Vaughan, D., Zalunin, V., and Cochrane, G. 2015. Content discovery and retrieval services at the european nucleotide archive. *Nucleic Acids Res.* 43:D23-D29. doi: 10.1093/nar/gku1129.

Smith, C., Heyne, S., Richter, A.S., Will, S., and Backofen, R. 2010. Freiburg RNA Tools: A web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Res.* 38:W373-W377. doi: 10.1093/nar/gkq316.

Spitale, R.C., Flynn, R.A., Torre, E.A., Kool, E.T., and Chang, H.Y. 2014. RNA structural analysis by evolving SHAPE chemistry. *Wiley Interdiscip. Rev. RNA* 5:867-881. doi: 10.1002/wrna.1253.

Stadler, P.F., Chen, J.J.-L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A.K., Mosig, A., Prohaska, S.J., Qi, X., Schutt, K., and Ullmann, K. 2009. Evolution of vault RNAs. *Mol. Biol. Evol.* 26:1975-1991. doi: 10.1093/molbev/msp112.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12:1611-1618. doi: 10.1101/gr.361602.

Stombaugh, J., Widmann, J., McDonald, D., and Knight, R. 2011. Boulder ALignment Editor (ALE): A web-based RNA alignment tool. *Bioinformatics* 27:1706-1707. doi: 10.1093/bioinformatics/btr258.

Suchard, M.A. and Redelings, B.D. 2006. BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22:2047-2048. doi: 10.1093/bioinformatics/btl175.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. 2011. Easyfig: A genome comparison visualizer. *Bioinformatics* 27:1009-1010. doi: 10.1093/bioinformatics/btr039.

Tabei, Y., Kiryu, H., Kin, T., and Asai, K. 2008. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* 9:33. doi: 10.1186/1471-2105-9-33.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682-2690. doi: 10.1093/nar/27.13.2682.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinformatics* 14:178-192. doi: 10.1093/bib/bbs017.

Torarinsson, E. and Lindgreen, S. 2008. WAR: Webserver for aligning structural RNAs. *Nucleic Acids Res.* 36:W79-W84. doi: 10.1093/nar/gkn275.

Torarinsson, E., Havgaard, J.H., and Gorodkin, J. 2007. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* 23:926-932. doi: 10.1093/bioinformatics/btm049.

Vogel, J. 2009. A rough guide to the non-coding RNA world of Salmonella. *Mol. Microbiol.* 71:1-11. Available at: *http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2008.06505.x/full*. doi: 10.1111/j.1365-2958.2008.06505.x.

Waldispühl, J., Kam, A., and Gardner, P.P. 2015. Crowdsourcing RNA structural alignments with an online computer game. *Pac. Symp. Biocomput.* 330-341. doi: 10.1142/9789814644730_0032.

Wassarman, K.M. and Storz, G. 2000. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* 101:613-623. doi: 10.1016/S0092-8674(00)80873-9.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-1191. doi: 10.1093/bioinformatics/btp033.

Wehner, S., Damm, K., Hartmann, R.K., and Marz, M. 2014. Dissemination of 6S RNA among bacteria. *RNA Biol.* 11:1467-1478. doi: 10.4161/rna.29894.

Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., and Breaker, R.R. 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* 11:R31. doi: 10.1186/gb-2010-11-3-r31.

Weinberg, Z., Kim, P.B., Chen, T.H., Li, S., Harris, K.A., Lünse, C.E., and Breaker, R.R. 2015. New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat. Chem. Biol.* 11:606-610. doi: 10.1038/nchembio.1846.

Westesson, O., Barquist, L., and Holmes, I. 2012. HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics* 28:1170-1171. doi: 10.1093/bioinformatics/bts058.

Wheeler, T.J. and Eddy, S.R. 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29:2487-2489. doi: 10.1093/bioinformatics/btt403.

Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., and Backofen, R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.* 3:e65. doi: 10.1371/journal.pcbi.0030065.

Woese, C.R., Gutell, R., Gupta, R., and Noller, H.F. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.* 47:621-669.

Wong, K.M., Suchard, M.A., and Huelsenbeck, J.P. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473-476. doi: 10.1126/science.1151532.

Xie, M., Mosig, A., Qi, X., Li, Y., Stadler, P.F., and Chen, J.J.-L. 2008. Size variation and structural conservation of vertebrate telomerase RNA. *J. Biol. Chem.* 283:2049-2059. doi: 10.1074/jbc.M708032200.

Xu, X., Ji, Y., and Stormo, G.D. 2007. RNA Sampler: A new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics* 23:1883-1891. doi: 10.1093/bioinformatics/btm272.

Yao, Z., Weinberg, Z., and Ruzzo, W.L. 2006. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 22:445-452. doi: 10.1093/bioinformatics/btk008.