

Computational identification of functional RNA homologs in metagenomic data

Eric P. Nawrocki and Sean R. Eddy*

HHMI Janelia Farm Research Campus

Ashburn, VA 20147

Phone: 571-209-3112 (EPN)

571-209-4163 (SRE)

Fax: 571-209-4094

E-mail: nawrockie@janelia.hhmi.org

eddys@janelia.hhmi.org

* Corresponding author. Send proofs to:

Sean Eddy

19700 Janelia Farm Blvd

HHMI Janelia Farm Research Campus

Ashburn, VA 20147

eddys@janelia.hhmi.org

An important step in analyzing a metagenomic sequence dataset is identifying functional sequence elements. This is a prerequisite for determining important properties of the environment the sequence data were sampled from, such as the metabolic processes and organismal diversity present there. At least initially, functional sequence element identification is addressed computationally. One class of elements, functional noncoding RNA elements, are especially difficult to identify because they tend to be short, lack open reading frames, and sometimes evolve rapidly at the sequence level even while conserving structure integral to their function (Eddy, 2001, Szymanski, 2003, Backofen, 2007, Hammann, 2007, Jossinet, 2007, Machado-Lima, 2008).

Functional RNA elements include both RNA genes (genes transcribed into functional untranslated RNA) and cis-regulatory mRNA structures. RNA elements play many roles. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are well known and universally present in all cellular life. Bacteria, archaea, and viruses, the organisms predominantly targeted by current metagenomics studies, also use numerous small RNA (sRNA) genes for translational and posttranslational regulation (Gottesman, 2005), as well as many cis-regulatory RNAs such as riboswitches (structural RNAs that respond to binding small molecule metabolites and control expression of nearby genes (Winkler, 2005, Tucker, 2005)). Archaea have numerous small nucleolar RNAs (snoRNAs) homologous to eukaryal snoRNAs that direct site-specific RNA methylation and pseudouridylation (Bachellerie, 2002). Many eukaryotes make extensive use of RNA regulatory mechanisms via pathways related to RNA interference (RNAi) and microRNAs (miRNAs) (Bartel, 2004, Ambros, 2004), and these will become relevant to metagenomic studies as they begin to target eukaryotes. These are only a small list of the most abundant classes of functional RNA elements. There are many other examples: catalytic introns, eukaryotic spliceosomal RNAs, RNA components of ribonucleoprotein complexes including telomerase, ribonuclease P, the signal recognition particle, and more.

It is striking that several of the large classes of RNAs just mentioned were either discovered recently (miRNAs, riboswitches) or have had their numbers greatly expanded by recent analyses (sRNAs, snoRNAs). This highlights the relative difficulty in discovering and analyzing functional RNA sequences, compared to more well-developed methodologies for discovering and analyzing protein coding sequences.

It hints that other RNAs likely remain undiscovered (Eddy, 2002).

In this chapter, we will discuss methods for computationally identifying homologs of known RNA elements, such as riboswitches and sRNA riboregulators. Another problem of great interest is to discover entirely *new* functional RNA elements by computational sequence analysis (Backofen, 2007, Machado-Lima, 2008, Pichon, 2008, Vogel, 2005), but as recent reviews have discussed, *de novo* RNA discovery (“genefinding”) methods (Babek, 2007, Meyer, 2007, Griffiths-Jones, 2007) currently have high false positive rates that are difficult to estimate statistically. Computational methods for *de novo* RNA discovery are unsuited to high-throughput automatic analysis, and instead need to be used as screens that can be followed by experimental confirmation (del Val, 2007). In contrast, RNA homology search programs are sufficiently reliable, and backed by sufficiently well-curated databases of known RNA sequence families, that automated large-scale computational metagenomic analyses are feasible.

Exploiting conserved structure in RNA similarity searches

Protein homology search by amino acid primary sequence comparison is powerful. At the amino acid level, BLASTP has no trouble detecting significant similarity down to about 25-30% amino acid sequence identity. Many protein coding regions conserve this level of similarity even across the deepest divergences in the tree of life amongst archaea, bacteria, and eukaryotes.

In contrast, RNA homology search by nucleotide primary sequence comparison is much less able to detect distant RNA homologies. BLASTN typically requires about 60-65% sequence identity to detect a statistically significant similarity for RNAs of typical length. Although some RNAs are very highly conserved over evolution (notably large and small subunit ribosomal RNAs, which are readily detected by sequence comparison in all species; the so-called human “ultraconserved” regions included regions of rRNA (Bejerano, 2004)), this is not the rule. Many functional RNA homologies are undetectable at the primary sequence level in cross-phylum comparisons (such as nematode/human or fly/human), because weakly or moderately conserved nucleic acid sequences can diverge to the 65% identity level in just a few tens of millions of years.

A striking example of the differing power in detecting protein versus RNA homologs by sequence

analysis comes when searching for homologs of the components of some ribonucleoprotein (RNP) complexes. It is not uncommon to detect homologs of the protein components but not the RNA components of complexes such as the signal recognition particle, ribonuclease P, small nucleolar RNPs, and telomerase. The interpretation upon finding only the protein component is usually (and almost certainly correctly) that the RNP complex is present in the organism, but the RNA component(s) are too difficult to detect. For example, the probable presence of small nucleolar RNAs in archaea could be inferred from the presence of homologs of snoRNP protein components like fibrillarin well before snoRNA homologs were discovered (Amiri, 1994, Omer, 2000). A similar situation can occur when identifying homologous cis-regulatory RNA elements (such as riboswitches) for clearly homologous coding genes.

Table 1 shows some specific anecdotal examples. These data are fairly typical of searching databases with protein versus RNA queries. They demonstrate two key points about the relative difficulty in detecting homologs of functional RNAs. First, notice that for the protein coding genes, the statistical significance of the similarity (the E-value) is always much better (lower and more significant) when comparing their amino acid sequences rather than when comparing their DNA sequences, highlighting the additional statistical power inherent in searches at the amino acid level. This is the reason for the recommended practice of always comparing protein sequences at the amino acid level (Pearson, 1996). Second, notice that RNA components are usually much shorter than the coding sequence of the protein components, further compromising statistical signal and the ability of primary sequence analysis (BLASTN) to resolve homologous relationships from background. (Sequence accessions used in these searches are listed in Table 2.)

What can be done about the weakness of primary sequence based methods for detecting functional RNAs? Some other source of statistical signal needs to be found for functional RNAs. Such a signal exists: many (though not all) functional RNAs conserve a distinctive RNA secondary structure.

Of course, proteins conserve structure more than sequence too. Remote homologies invisible to primary sequence analysis often become apparent when a protein's three-dimensional structure is solved. What makes RNA secondary structure constraints of particular utility for computational sequence analysis

is that they produce a simple and strong statistical signal of pairwise residue correlations in aligned RNA sequences. These correlations may be sufficiently obvious that they are apparent even to the naked eye (analogous to the obviousness of ORFs for coding gene analysis). RNA consensus secondary structures have been accurately inferred by manual “comparative sequence analysis” alone (Michel, 1990, Pace, 1993, Gutell, 2002).

How much extra information does RNA secondary structure conservation contribute in addition to primary sequence conservation? We can ask this question rigorously in the context of homology search applications, across a range of different types of RNAs. Figure 1 shows the average score of search models for about 100 RNA sequence families, comparing models of sequence conservation alone (“profile hidden Markov models”, profile HMMs) to models of sequence plus RNA secondary structure conservation (“covariance models”, CMs). These consensus models are discussed in more detail below, but for the present point, their salient feature is that they are built from an input multiple alignment of homologous sequences, and they represent that alignment using a probabilistic position-specific scoring system.

The unit of score is a “bit” (essentially the same as BLAST “bit scores”), which is a measure of information content (Shannon, 1948, MacKay, 2003). Some intuition can be given for what a bit means, without much mathematics. A single perfectly conserved RNA residue (probability 1.0) contrasted to a uniform expected background (probability 0.25) is $\log_2 \frac{1.0}{0.25} = 2$ bits of information – you need to ask two yes/no questions to narrow four possibilities down to one, thus two “bits” (binary units) of information. A position where each residue occurs with equal probability (same as expected background) has zero bits of information. Imagine two positions that contain a covarying Watson-Crick base pair in which each of the four possible base pairs occurs with equal frequency $\frac{1}{4}$. In a sequence-only model the two positions contribute zero bits of information, but in a structure/sequence model this pair contributes two bits of information from the pairwise correlation (the expected background in these columns is 0.0625 for each of the 16 possible base pairs, but only 4 are observed with probability 0.25 each). In contrast, two

columns that form a Watson-Crick base pair that is perfectly conserved (a GC with probability 1.0 for example) always contribute four bits of information, regardless of whether they are modeled together as a pair ($\log_2 \frac{1.0}{0.0625} = 4$) or independently ($\log_2 \frac{1.0}{0.25} + \log_2 \frac{1.0}{0.25} = 4$). Thus, the best case for extracting useful sequence information from RNA secondary structure that could not be extracted from RNA primary sequence consists of covarying base pairs that are individually not conserved in primary sequence at all. The more highly conserved the aligned RNA sequences are, the more primary sequence information content and less covariation will be seen.

Importantly, for local sequence alignment searches using probabilistic models, there is a direct, and intuitive connection between the bit score and the statistical significance (E-value) of a detected match (Eddy, 2008). Roughly speaking, every 3 or so bits of score improves the E-value by a factor of ten-fold (for high scores, the E-value is an exponential function of the bit score x ; E is proportional to 2^{-x}). So, as a rule of thumb, extracting ten more bits of information for a homology search means shifting E-values by three orders of magnitude. This increase in resolution doesn't matter much if a sequence is already readily detected by primary sequence comparison (improving an already significant E-value of 10^{-30} to 10^{-33} , for example), but it becomes important when lifting a marginally insignificant E-value to significance (0.1 to 10^{-4} , for example).

Figure 1 shows the extra bits of information contributed by including RNA secondary structure in “typical” RNA search models. These models are all position-specific profiles built from alignments in the Rfam RNA families database, described below. There is substantial variation from family to family, but the extra information contributed by secondary structure is usually on the order of 10 to 20 bits or more, depending on the length and conservation of the alignment, which would be expected to improve E-values of homologs by about 3 to 6 orders of magnitude. This improvement can be seen in the results of the anecdotal searches of Table 1 comparing the E-values obtained by primary sequence BLASTN searches to INFERNAL, a sequence+secondary structure RNA homology search, as we will discuss in more detail below. The conclusion here is that while primary sequence is still the dominant source of information (at least for these particular “typical” searches; it is, of course, possible to imagine searching for RNAs with

zero sequence information and only secondary structure information), adding secondary structure contributes enough information content that we can expect a structure+sequence method to resolve some homologs that were not quite resolvable by sequence analysis alone.

Infernal: software for RNA homology search and alignment

Computational methods that combine RNA secondary structure and sequence conservation information in a single consistent statistical model have been developed, based on probabilistic models called “stochastic context-free grammars” (SCFGs) (Eddy, 1994, Sakakibara, 1994, Durbin, 1998, Eddy, 2002). Dynamic programming algorithms exist for optimal alignment of SCFGs to target sequences, analogous to algorithms for sequence alignment except that SCFG algorithms are aligning by base-paired secondary structure in addition to sequence (Kasami, 1965, Younger, 1967, Hopcroft, 1979, Durbin, 1998). A particular formulation of SCFGs, called *covariance models* (CMs), was developed specifically for automatic construction of statistical models from input RNA secondary structures or input multiple alignments annotated with consensus RNA structure. This technology is implemented in a freely available software package called INFERNAL (<http://infernal.janelia.org>).

A variety of other computational tools for RNA homology search exist besides INFERNAL (reviewed in Eddy, 2006, Jossinet, 2007, Backofen, 2007, Machado-Lima, 2008). Some of the most popular tools are ERPIN (Gautheret, 2001), FASTR (Zhang, 2005), RSmatch (Liu, 2005), RNAMotif (Macke, 2001), RNATOPS (Huang, 2008), and PatScan (Dsouza, 1997). INFERNAL is one of the most generally applicable tools, is the basis for a widely used RNA family database (Rfam; described below), and currently appears to be the best overall in performance according to published benchmarks (Freyhult, 2007). Here we will restrict our discussion to INFERNAL.

To demonstrate how scoring structure increases statistical power for RNA homology search, we used INFERNAL to build CMs and perform searches for the single sequence/structure queries in Table 1 (the structures were obtained from the Rfam database, described below). As expected, modeling structure makes the target RNA more distinguishable from background, as evidenced by the decrease in E-values between BLASTN and CM searches of between three and thirteen orders of magnitude.

Figure 2 provides more detail for the cobalamin (B12) riboswitch example from Table 1. It shows the *Escherichia coli* query sequence and secondary structure, and the pattern of conservation in two different homologs found by a CM built from the *E. coli* query. Notice that although many of the residue substitutions between query and target are in the predicted loop regions, those that occur in a position that is basepaired are often accompanied by a compensatory change in the paired position to maintain a Watson-Crick or GU/UG pair. The extra information from the *E. coli* structure allows INFERNAL to find the homologous riboswitch in the *Acinetobacter baumannii* genome as the top scoring hit with a significant E-value of 3.7×10^{-5} , despite it sharing only 49% sequence identity with the *E. coli* riboswitch. The analogous search with BLASTN does not identify the riboswitch homology with a significant E-value (E=2.6).

CMs can be built from single RNAs, but they are most powerful when built from a multiple sequence alignment with consensus secondary structure annotation. CMs implement a position-specific (“profile”) scoring system, where each consensus single-stranded position or base pair is represented by its own set of four or sixteen scores, and insertion/deletion scores are likewise specific to each point where an insertion or deletion can occur. Given enough aligned sequences, a position-specific profile model can learn which residues or base pairs are highly conserved, what substitutions are tolerated by evolution, and where an RNA does and does not frequently tolerate insertion and deletion of sequence residues or structural domains. Given only a single RNA sequence (as in the examples in Table 1 and Figure 2), the CM scoring system reverts to a position-independent parameterization representing the averaged constraints on typical RNAs, essentially analogous to the use of score matrices in pairwise sequence alignment methods like BLAST.

CMs are probabilistic models, meaning that all the scoring parameters are probabilities rather than arbitrary scores and penalties. This helps in managing the complexity of setting a large number of parameters in an objective, automatic, and mathematically justified way; a consensus tRNA CM has about 1500 parameters and a consensus LSU rRNA CM has about 50,000 parameters that need to be determined. Using probabilities as parameters also helps in interpreting the significance of potential matches in a database search, and in calculating confidence values (posterior probabilities) associated with

each residue in a proposed alignment. The use of probabilistic models for RNA structure/sequence analysis follows in the wake of similar techniques in primary sequence analysis, where score profiles (also called position-specific scoring matrices, PSSMs) have been made more powerful and consistent using probabilistic models called profile hidden Markov models (profile HMMs) (Krogh, 1994, Durbin, 1998).

A CM can be used for a variety of alignment and search tasks. For example, very large numbers of RNA sequences can be aligned to a single RNA structure consensus with reasonable accuracy and efficiency: the Ribosomal Database Project (RDP) now uses INFERNAL to produce alignments of hundreds of thousands of small subunit (SSU) ribosomal RNAs (Cole, 2009). For sequence annotation, including metagenomic analysis, the main use of CMs is for homology search.

Because INFERNAL requires that the user provide a consensus RNA secondary structure for the query RNA, and because CMs are most powerful when models are built from multiple sequence alignments, a fair amount of work might be invested in carefully assembling a high-quality multiple sequence alignment annotated with a consensus structure. This investment may be feasible if one is only interested in sequence analysis of a particular RNA family, such as ribosomal RNA. However, if the goal is comprehensive high-throughput annotation of many different functional RNAs, for instance as part of analyzing a new metagenomic sequence dataset, it would be useful to have access to a large number of structure-annotated RNA alignments and prebuilt CMs. Much as protein domain databases like Pfam and SMART have collected on the order of 10,000 protein domain sequence alignments for systematic profile HMM analysis (Letunic, 2008, Finn, 2008), there is a database called Rfam that has systematically collected RNA alignments and CMs (Gardner, 2009).

Rfam: high-throughput RNA homology search and annotation

The Rfam database (Gardner, 2009) is a curated and annotated collection of RNA sequence families, intended for the purpose of systematic, automated, high-throughput annotation of functional RNA elements in genomic and metagenomic sequence data. The current (9.1) version of Rfam contains 1372 families (<http://rfam.sanger.ac.uk>). Each Rfam family consists of three main components: a representative “seed” alignment, a covariance model (CM) built from the seed alignment, and a comprehensive “full”

alignment.

The seed alignment is intended to be a small, stable, and curated alignment of representative members of the sequence family, annotated with a consensus RNA secondary structure. For example, the glycine riboswitch (RF000504; <http://rfam.sanger.ac.uk/family/RF00504>) is represented by an alignment of 53 RNAs.

The full alignment is intended to be comprehensive. It consists of an INFERNAL-generated structural alignment of all homologous RNAs detected by INFERNAL in a search, using the CM built from the seed alignment, of a composite DNA sequence database, RFAMSEQ, which now includes both genomic and metagenomic sequence data (Gardner, 2009).

The alignments are useful for a variety of purposes, such as phylogenetic tree inference, examining the phylogenetic range over which a given RNA family occurs, or as a source of training data for other RNA structure analysis methods. For metagenomic analyses, the main application of INFERNAL and Rfam is homology search, and the main resource is the set of prebuilt Rfam CMs.

The INFERNAL package (<http://infernalia.org>) and Rfam CM files (<http://rfam.sanger.ac.uk>) can be freely downloaded and used to identify homologs of known functional RNAs in a metagenomics dataset. As an example of such an analysis, we performed CM searches of a previously published metagenomics dataset (Tringe, 2005). The dataset includes about 200,000 whole genome shotgun sequencing reads totalling about 230 Mb derived from samples of agricultural soil (~140 Mb, accession AAFX01000000) and three “whale fall” carcasses (~90 Mb, accessions AAFZ00000000, AAFY01000000, AAGA00000000). To simplify the analysis for our illustrative purposes here, we searched only for riboswitches, using the 15 Rfam 9.1 CMs of type ‘cis reg; riboswitch’ (Gardner, 2009). For comparison, we repeated the search with BLAST, using each individual sequence in the Rfam seed alignment as a BLAST query and combining the results to identify any significant matches (Grundy, 1998). Additionally, we performed searches with INFERNAL v1.0 run in a profile HMM mode, which ignores secondary structure and scores only primary sequence conservation. Comparison of the BLAST, profile HMM and CM search results illustrates the relative contribution of the two main differences between BLAST and CMs: the use of probabilistic profiles instead of pairwise comparisons (by comparing BLAST

and HMM results), and scoring both sequence and RNA structure (by comparing CM and HMM results).

Table 3 includes the number of putative riboswitches (hits) with E values less than 10^{-5} found for each family using each method. Also displayed are the number of hits detected by one method but not another for all six possible pairwise combinations of the three methods. Using the strict 10^{-5} E-value cutoff, INFERNAL CM searches found 135 total putative riboswitches in the soil and whale falls dataset; HMM searches found 102 (a subset of the 135); and BLAST found 50. Profile HMM searches detected 61 hits that BLAST did not, and CM searches detected 33 hits that profile HMMs did not, indicating that using profiles and additional scoring of structure both contribute significantly to an increased sensitivity of CMs over BLAST.

We can compare these results to the recently published results of a similar analysis of riboswitch occurrence in the same metagenomic dataset using different search methodology (Kazanov, 2007). Kazanov *et al.* used the pattern based search program RNA-PATTERN to identify candidates of 11 riboswitch families (8 of which we used in our analysis) in the same soil and whale falls data we analyzed. For the 8 families in common, their pattern based search detected 103 candidate riboswitches, compared to 125 identified by CM searches at a stringent threshold. RNA-PATTERN detected 14 candidates that CMs did not, and CMs detected 36 candidates that RNA-PATTERN did not. The largest differences were for the cobalamin family, for which CMs found 18 candidates undetected by RNA-PATTERN, and the glycine family, for which RNA-PATTERN found 10 candidates undetected by CMs using a CM E-value threshold of 10^{-5} . Six of these 10 are found by the glycine riboswitch CM, but with E-values just below the strict threshold, ranging between 10^{-3} and 10^{-5} . For the remaining four, we cannot distinguish whether these are missed by the CM, or whether they are false positive predictions by RNA-PATTERN and Kazanov *et al.*; one disadvantage of pattern search programs is that they do not generally report any objective measure of the statistical significance of a match.

Can we trust that the statistically significant matches to the CM are really homologs, and that increased numbers of predictions really reflect increased detection sensitivity? That is, based solely on the results of the demonstration experiment here, where we are just counting the number of hits detected

below some E-value threshold and asserting that these are all probable homologs, it is possible that INFERNAL is instead merely assigning incorrectly low E-values to nonhomologous sequences. One way to test the accuracy of any program's E-values is to search randomized nonhomologous sequence; one expects the top-scoring random match to have an E-value on the order of 1 (by definition of expectation value: the number of hits you expect to see in this database search with a score this high just by chance). This sort of test is a useful control experiment to run whenever thinking of adopting any new search method. In one recent experiment of ours (Nawrocki, 2009), involving a benchmark of 51 CMs being searched against a 10 megabase synthetically generated target sequence, the highest nonhomologous hit had an E-value of 0.009, about what you'd expect from doing 51 independent searches ($1/51 = 0.019$) if E-values were accurate. In our experience, an E-value threshold of 10^{-5} is conservative. Most importantly, an independent benchmark of a variety of RNA similarity search methods has been published (Freyhult, 2007), which generally found that CM based methods are the most sensitive and specific methods available.

Limitations of CMs

Now the fine print. Users applying INFERNAL and Rfam for metagenomics analysis should be aware of five important limitations of CM similarity search:

1. The principal drawback of CM methods is that they are slow. In the riboswitch example above, the fifteen CM searches took about 71 CPU hours (18 minutes on 250 processors), about 100 times longer than BLAST searches (45 minutes on one processor). Repeating this search using all 1372 Rfam 9.1 models would take roughly 3 CPU years (about 4 days on a 250 CPU cluster). Significant compute power (such as a moderate sized cluster) is required to do large scale analyses with CMs. INFERNAL is parallelized for use on clusters using the Message Passing Interface (MPI) (Gropp, 1996).

Though still slow compared to BLAST, INFERNAL is much faster than it was just a few years ago. The current version (v1.0) (Nawrocki, 2009) is about 100 times faster than version 0.55 (Eddy, 2002). The speedup is due to heuristics, including filtering (Weinberg, 2006, Nawrocki, in preparation) and banded dynamic programming (Nawrocki, 2007), which sacrifice a small amount of sensitivity for the increased

speed. This sensitivity sacrifice, though small, disproportionately impacts remote homology detection (Nawrocki, 2007, Nawrocki, 2009). It may be worthwhile to switch off the heuristic speedups for smaller scale analyses if the requisite compute power is at hand. Conversely, if compute power is limiting, the heuristic speedup parameters can be tuned for greater acceleration at a greater cost in sensitivity (see the Infernal user's guide, <http://infernal.janelia.org>). Further acceleration remains a major goal of INFERNAL development.

Another computationally expensive step of CM similarity search is “calibrating” models in order to obtain E-values for search results, and to determine the appropriate filtering scheme for maximum speed without significant sensitivity loss. INFERNAL's `cmcalibrate` program must run several large computational simulations, and this takes several CPU hours for a typical sized CM. The CMs from the Rfam database come pre-calibrated, so Rfam users do not have to pay this cost, but any custom built models need to be calibrated.

2. A CM models only a single user-provided RNA consensus structure. Many RNA structures are inferred, rather than being determined by crystallographic or NMR methods, so secondary structure annotation may well be at least partially incorrect – especially in large collections like Rfam, where curation of a set of over 1300 consensus structures is challenging. Additionally, a single consensus structure is unable to properly capture the evolutionary variation observed amongst individual homologous secondary structures, except in a crude way (as structural deletions and insertions relative to the consensus). And finally, an assumption that an RNA adopts only a single secondary structure is only an approximation, as RNAs (like proteins) are sure to exist in an ensemble of different structures (perhaps bound and unbound to a protein or substrate). Riboswitches, for example, are a dramatic example of the function of an RNA depending on at least two distinct structural conformations.

3. Using a CM for non-structured RNAs is pointless. Many RNAs may not require a conserved structure for their function. For example, antisense regulatory RNAs that control gene expression simply by basepairing to target mRNAs are acting as primary sequences, and they do not necessarily conserve any intramolecular secondary structure. Though CMs can model RNAs with no consensus base pairs, it is more practical and appropriate to use profile HMMs rather than CMs, avoiding the CMs' computational

costs.

4. CMs ignore some aspects of RNA structure. By their nature, CMs are only able to model a canonical secondary structure consisting of exclusively nested base pairing relationships, meaning a set of base pairs for which no two pairs “overlap“ in sequence position (no two pairs between positions $i:j$ and $k:l$ exist such that $i < k < j < l$). This means CMs do not model RNA pseudoknots, base triples, nor most other contacts found in RNA tertiary structure. The goal of a CM is not to model RNA structure completely, but rather to harness as much additional structural information as possible for more accurate RNA search and alignment, while still allowing for reasonably efficient algorithms. Capturing yet more higher-order RNA structural information is possible, but it violates the constraints of SCFG-type probabilistic models and comes at a disproportionate cost in computational efficiency (Rivas, 1999). Other methods exist for RNA homology search that can model RNA pseudoknots, including ERPIN (Gautheret, 2001) and RNATOPS (Huang, 2008).

5. Truncated sequences present special issues. Metagenomic shotgun sequencing surveys produce sequence fragments from essentially random positions on a host genome. When a shotgun read overlaps a structural RNA element, residues involved in conserved base pairs may be missing. An RNA structural alignment algorithm needs to anticipate and allow this sort of sequence truncation to be useful for shotgun sequence analysis. Until recently, the local alignment algorithms used for CMs and related SCFG-based RNA alignment methods looked for structural deletions (deletions that remove a stem or structural domain, respecting evolutionary base pairing constraints), but did not look for sequence truncations. INFERNAL now includes a new local alignment method, developed by Diana Kolbe, which more effectively deals with missing sequence data in shotgun reads (Kolbe, 2009). As we write this, this feature is not yet incorporated in cmsearch for homology searches (it soon will be), but the new algorithm can be used for sequence alignment with INFERNAL’s prototype trecyk program.

Conclusion

Compensatory base pair changes in RNA sequence alignments are strikingly apparent even to the eye. The deeper the alignment (the more sequences known to conserve roughly the same structure), the more

the RNA structure becomes obvious by sequence analysis alone. Robin Gutell and co-workers were able to predict the secondary structure of ribosomal RNA to greater than 98% accuracy per base pair by essentially manual comparative analysis of careful rRNA alignments (Gutell, 2002), and Francois Michel and Eric Westhof essentially predicted the structure of group I intron catalytic introns in much the same way (Michel, 1990). The automation of comparative RNA structure/sequence analysis is essentially the basis of algorithms that combine RNA secondary structure and sequence analysis to enable identification of more remote RNA homologs than primary sequence methods alone can achieve. These methods can be used to search metagenomics datasets for known families of RNAs using a combination of the INFERNAL software (<http://infernal.janelia.org>) and CMs from the Rfam database (Gardner, 2009).

Acknowledgements

We are grateful to Seolkyoung Jung, Fred Davis, Elena Rivas and Tom Jones for critical comments on the manuscript. We thank Howard Hughes Medical Institute for their financial support.

References

- Ambros, V.** 2004. The functions of animal microRNAs. *Nature*, **431**:350–355.
- Amiri, K. A.** 1994. Fibrillar-like proteins occur in the domain archaea. *J. Bacteriol.*, **176**:2124–2127.
- Babak T., B. J. Blencowe, and T. R. Hughes.** 2007. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*. **8**:33.
- Bachellerie J. P., J. Cavaille, and A. Hüttenhofer.** 2002. The expanding snoRNA world. *Biochimie*. **84**:775–790.
- Backofen R., S. H. Bernhart, C. Flamm, C. Fried, G. Fritsch, J. Hackermüller, J. Hertel, I. L. Hofacker, K. Missal, A. Mosig, S. J. Prohaska, D. Rose, P. F. Stadler, A. Tanzer, S. Washietl, and S. Will.** 2007. RNAs everywhere: Genome-wide annotation of structured RNAs. *J. Exp. Zoolog. B Mol. Dev. Evol.* **308**:1–25.
- Bartel, D. P.** 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**:281–297.
- Bejerano G. , M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler.** 2004. Ultraconserved elements in the human genome. *Science*. **304**:1321–1325.
- Cole J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje.** 2009. The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucl. Acids Res.* **37**:D141–D145.
- del Val C., E. Rivas, O. Torres-Quesada, N. Toro, and J. I. Jiménez-Zurdo.** 2007. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. *Mol. Microbiol.* **66**:1080–1091.
- Dsouza M., N. Larsen, and R. Overbeek.** 1997. Searching for patterns in genomic data. *Trends Genet.* **13**:497–498.
- Durbin, R., S. R. Eddy, A. Krogh, and G. J. Mitchison.** 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK.
- Eddy, S. R.** 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**:919–929.
- Eddy, S. R.** 2002a. Computational genomics of noncoding RNA genes. *Cell*. **109**:137–140.
- Eddy, S. R.** 2002b. A memory-efficient dynamic programming algorithm for optimal alignment of a

- sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**:18.
- Eddy, S. R.** 2006. Computational analysis of RNAs. *Cold Spring Harbor Symp. Quant. Biol.* **71**:117–128.
- Eddy, S. R.** 2008. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **4**:e1000069.
- Eddy S. R., and R. Durbin.** 1994. RNA sequence analysis using covariance models. *Nucl. Acids Res.* **22**:2079–2088.
- Finn, R. D., J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman.** 2008. The Pfam protein families database. *Nucl. Acids Res.* **36**:D281–D288.
- Freyhult, E. K., J. P. Bollback, and P. P. Gardner.** 2007. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.* **17**:117–125.
- Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman.** 2009. Rfam: Updates to the RNA families database. *Nucl. Acids Res.* **37**:D136–D140.
- Gautheret D., and A. Lambert.** 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **313**:1003–1011.
- Gottesman, S.** 2005. Micros for microbes: Non-coding regulatory RNAs in bacteria. *Trends Genet.* **7**:399–404.
- Griffiths-Jones, S.** 2007. Annotating noncoding RNA genes. *Annu. Rev. Genom. Hum. Genet.* **8**:279–298.
- Gropp W., E. Lusk, N. Doss, and A. Skjellum.** 1996. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing.* **22**:789–828.
- Grundy, W. N.** 1998. Homology detection via family pairwise search. *J. Comput. Biol.*, **5**:479–491.
- Gutell, R. R., J. C. Lee, and J. J. Cannone.** 2002. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* **12**:301–310.
- Hammann C., and E. Westhof.** 2007. Searching genomes for ribozymes and riboswitches. *Genome Biol.* **8**:210.

- Hopcroft J. E. and J. D. Ullman.** 1970. *Introduction to Automata Theory, Languages, and Computation*. 1st ed. Addison-Wesley, Reading, MA.
- Huang, Z., Y. Wu, J. Robertson, L. Feng, R. Malmberg, and L. Cai.** 2008. Fast and accurate search for non-coding rna pseudoknot structures in genomes. *Bioinformatics*. 24:2281–2287.
- Jossinet, F., T. E. Ludwig, and E. Westhof.** 2007. RNA structure: Bioinformatic analysis. *Curr. Opin. Microbiol.* 10:279–285.
- Kasami, T.** 1965. An efficient recognition and syntax algorithm for context-free algorithms. *Technical Report AFCRL-65-758*, Air Force Cambridge Research Lab, Bedford, MA.
- Kolbe, D. L. and S. R. Eddy.** Local RNA structure alignment with incomplete sequence. *Bioinformatics*, in press. 2009.
- Kazanov, M. D., A. G. Vitreschak, and M. S. Gelfand.** 2007. Abundance and functional diversity of riboswitches in microbial communities. *BMC Genomics*. 8:347.
- Krogh, A., M. Brown, I. S. Mian, K. Sjölander, and D. Haussler.** 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235:1501–1531.
- Letunic, I., T. Doerks, and P. Bork.** 2009. SMART 6: Recent updates and new developments. *Nucl. Acids Res.* 37:D229–D232.
- Liu, J., J. T. Wang, J. Hu, and B. Tian.** 2005. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics*. 6:89.
- Machado-Lima, A., H. A. del Portillo, and A. M. Durham.** 2008. Computational methods in noncoding RNA research. *J. Math. Biol.* 56:15–49.
- MacKay, D. J. C.** 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK.
- Macke, T. J., D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath.** 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucl. Acids. Res.* 29:4724–4735.
- Meyer, I. M.** 2007. A practical guide to the art of RNA gene prediction. *Brief. Bioinform.* 8:396–414.
- Michel F., and E. Westhof.** 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216:585–610.

- Nawrocki, E. P. and S. R. Eddy.** 2007. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.* **3**:e56.
- Nawrocki, E. P., D. L. Kolbe, and S. R. Eddy.** 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics.* in press.
- Omer, A. D., T. M. Lowe, A. G. Russell, H. Ebhardt, S. R. Eddy, and P. P. Dennis.** 2000. Homologs of small nucleolar RNAs in Archaea. *Science.* **288**:517–522.
- Pace, N. R., B. C. Thomas, and C. R. Woese.** 1993. Probing RNA structure, function and history by comparative analysis. p. 113-142. In R. F. Gesteland and J. F. Atkins (ed.), *The RNA World*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Pearson, W. R.** 2008. Effective protein sequence comparison. *Meth. Enzymol.*, **266**:227–258.
- Pichon C., and B. Felden.** 2008. Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics.* **24**:2807–2813.
- Rivas, E. and S. R. Eddy.** 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285**:2053–2068.
- Sakakibara, Y., M. Brown, R. C. Underwood, I. S. Mian, and D. Haussler.** 1994. Stochastic context-free grammars for modeling RNA. p. 284-293. In Lawrence Hunter (ed.), *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences: Biotechnology Computing*, vol. 5. IEEE Computer Society Press, Los Alamitos, CA.
- Shannon, C. E.** 1948. A mathematical theory of communication. *Bell System Technical Journal.* **27**:379–423,623–656.
- Szymanski, M., M. Z. Barciszewska, M. Zywicki, and J. Barciszewski.** 2003. Noncoding RNA transcripts. *J. Appl. Genet.* **44**:1–19.
- Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. *Science.* **308**:554–557.
- Tucker, B. J., and R. R. Breaker.** 2005. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.* **15**:342–348.

- Vogel, J., and C. M. Sharma.** 2005. How to find small non-coding RNAs in bacteria. *Biol. Chem.* **386**:1219–1238.
- Weinberg, Z. and W. L. Ruzzo.** 2006. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics.* **22**:35–39.
- Winkler, W. C.** 2005. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.* **9**:594–602.
- Younger, D. H..** 1967. Recognition and parsing of context-free languages in time . *Information and Control.* **10**:189–208.
- Zhang, S., B. Haas, E. Eskin, and V. Bafna.** 2005. Searching genomes for noncoding RNA using FastR. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2**:366–379.

Table Footnotes

Table 1: Examples of identifying coding region homologies by amino acid sequence versus nucleic acid sequence comparison (BLASTP vs. BLASTN), compared to identifying RNA homologies by primary sequence versus structure/sequence comparison (BLASTN vs. INFERNAL). For each query/target pair, the query sequence was searched against the target genome (for coding sequence and RNA searches) or predicted proteome (for amino acid sequence searches) using the indicated search programs. “Length” indicates unaligned length of the query. “% id” indicates percent identity of local alignments of hits returned by each method. (BLASTN RNA alignments are usually higher percentage identity than INFERNAL alignments, but are also usually significantly shorter.) Query RNAs were selected from candidates found by INFERNAL in each listed query’s genome sequence using the the Rfam 9.1 CM for the appropriate family (listed below). Each query RNA was used to build a CM using the INFERNAL imposed Rfam structure, and each CM was calibrated and used to search the target genomes. Rfam family IDs for each family listed in “RNA name”, in row order, are: RF00373, RF00168, RF00001, RF00234, RF00169, RF00174. For riboswitches, the protein components are always immediately downstream of the RNA components. Versions used: WU-BLASTN-2.0MP, WU-BLASTP-2.0MP and cmsearch from INFERNAL version 1.0. For BLASTN, the -w=5 option was always used, and the -kap option was used only if it resulted in a more significant (lower) E-value for the target sequence.

Table 2: GenBank genome and protein accessions and RNA genomic coordinates for examples from Table 1.

Table 3: Riboswitch search results. “# seqs found”: the number of hits receiving E-values better than (less than) 10^{-5} for each search method in the 230 Mb soil and whale falls dataset (Tringe, 2005) for each Rfam 9.1 riboswitch family. “# different seqs found”: number of hits detected by one method and not detected by another for each pairwise combination of methods, for example: “BLAST-HMM” for RF00162 is 3 because 3 hits detected by BLAST were undetected by the HMM search. A blank space in a cell indicates 0.

Figure Legends

Figure 1: Additional information (in bits) gained by structure/sequence profiles versus sequence-only profiles for various RNA families. Structure/sequence profiles are most advantageous for families with less primary sequence information (towards left) and more secondary structure information (towards top), so Rfam families that gain the most from including secondary structure terms in a homology search are those toward the upper left quadrant. Data shown for the 95 Rfam release 9.1 (Gardner, 2009) families with 50 or more sequences in the “seed” alignment. For each family, the seed alignment was used to build two profile models, one with structure (sequence/structure profile CM model) and one without (sequence profile HMM model). From each model, 10,000 sequences were generated and scored, and the average score per sampled sequence was calculated. Several of the outlying points are labeled by the name of RNA family as given by Rfam. Note that the x-axis is drawn on a log scale. Models were built and sequences were generated and scored using INFERNAL version 1.0 programs cmbuild, cmemit and cmalign.

Figure 2: Secondary structure of three cobalamin riboswitches. Using the *E. coli* sequence as a query against their respective genomes, BLASTN detects the *Y. enterocolitica* cobalamin riboswitch with a significant E-value, but not the *A. baumannii* riboswitch. INFERNAL searches with a CM constructed from the *E. coli* sequence and structure (from the Rfam seed alignment for family RF00174 (Gardner, 2009)) find both riboswitches with increased significance values. These example searches are listed in Table 1. Structures of the targets and percent identity figures were derived from the highest scoring CM alignment of each target to the query (*E. coli*). Sequence substitutions and insertions in the targets with respect to the query are shown in grey. Inserted residues with respect to the query are shown in lowercase. Basepairs in the Rfam annotated structure are connected by solid lines, except those that are not Watson-Crick, GU, or UG, which are connected by dotted lines. All riboswitches are immediately upstream (5'; within 100 residues) of *btuB* vitamin B12 transporter protein coding genes in their respective genomes.

organism 1 query	organism 2 target	protein name	amino acid sequence			coding sequence			RNA name	RNA				
			length	%id	E-value	length	%id	E-value		length	%id	E-value	%id	E-value
Methanocaldococcus jannaschii	Pyrococcus horikoshii	Rpp29	95	50%	2.3e-19	288	62%	4.2e-09	RNase P RNA	258	72%	9.2e-06	51%	6.7e-09
Bacillus cereus	Bacillus subtilis	lysC	409	67%	1.0e-135	1230	66%	5.4e-94	Lysine riboswitch	187	63%	2.0e-05	59%	2.9e-18
Sulfolobus solfataricus	Thermococcus kodakarensis	rpl30p	158	42%	4.9e-29	477	61%	0.38	5S rRNA	115	73%	4.3e-05	62%	7.6e-10
Bacillus subtilis	Clostridium acetobutylicum	glmS	600	46%	3.2e-138	1803	57%	2.5e-66	glmS riboswitch	168	63%	3.5e-04	56%	1.6e-14
Escherichia coli	Bacillus subtilis	ffh	453	54%	9.8e-124	1362	63%	2.3e-85	SRP RNA	100	66%	2.7e-03	62%	2.2e-13
	Candidatus P. amoebophila			44%	2.5e-102		57%	8.6e-36			68%	0.78	47%	6.7e-06
Escherichia coli	Klebsiella pneumoniae	btuB	614	57%	2.6e-192	1845	65%	7.0e-113	Cobalamin riboswitch	191	78%	5.7e-18	77%	2.5e-33
	Yersinia enterocolitico			52%	1.5e-173		60%	6.4e-83			74%	3.2e-09	67%	9.3e-21
	Vibrio cholerae			38%	8.1e-107		57%	6.4e-34			72%	0.043	57%	4.5e-05
	Acinetobacter baumannii			26%	5.0e-46		61%	2.4e-06			65%	2.6	49%	3.7e-05

Table 1.

organism name	genome	protein	protein	RNA	RNA
	accession	name	accession	name	genomic coordinates
Methanocaldococcus jannaschii	NC_000909.1	Rpp29	NP_247439.1	RNase P RNA	643504-643761
Pyrococcus horikoshii	NC_000961.1	Rpp29	NP_143607.1	RNase P RNA	168208-168414
Bacillus cereus	NC_003909.8	lysC	NP_0978199.1	Lysine riboswitch	1818638-1818452
Bacillus subtilis	NC_000964.2	lysC	NP_390725.1	Lysine riboswitch	2910116-2909946
Sulfolobus solfataricus	NC_002754.1	rpl30p	NP_342208.1	5S rRNA	78064-77946
Thermococcus kodakarensis	NC_006624.1	rpl30p	YP_183933.1	5S rRNA	1769482-1769599
Bacillus subtilis	NC_000964.2	glmS	NP_388059.1	glmS riboswitch	200006-200173
Clostridium acetobutylicum	AE001437.1	glmS	AAK78142.1	glmS riboswitch	179915-180074
Escherichia coli	NC_000913.2	ffh	NP_417101.1	SRP RNA	475679-475778
Bacillus subtilis	NC_000964.2	ffh	NP_389480.1	SRP RNA	26531-26633
Candidatus P. amoebophila	NC_005861.1	ffh	YP_007653.1	SRP RNA	871975-872074
Escherichia coli	NC_000913.2	btuB	NP_418401.1	Cobalamin riboswitch	4161407-4161597
Klebsiella pneumoniae	CP000647.1	btuB	ABR78634.1	Cobalamin riboswitch	4660061-4660248
Yersinia enterocolitica	NC_08800.1	btuB	YP_001004531.1	Cobalamin riboswitch	157101-157301
Vibrio cholerae	NC_009457.1	btuB	YP_001218242.1	Cobalamin riboswitch	2498535-2498369
Acinetobacter baumannii	NC_011586.1	btuB	YP_002320687.1	Cobalamin riboswitch	3485342-3485537

Table 2.

Family	Rfam ID	avg % id	# different seqs found									
			# seqs found			BLAST	BLAST	HMM	HMM	CM	CM	
			BLAST	HMM	CM	-HMM	-CM	-BLAST	-CM	-BLAST	-HMM	
FMN	RF00050	72	8	9	9			1		1		
TPP	RF00059	56	8	23	35			15		27	12	
SAM	RF00162	69	4	8	14	3		7		10	6	
Purine	RF00167	56										
Lysine	RF00168	51		1	1			1		1		
Cobalamin	RF00174	54	20	43	47	2	2	25		29	4	
glmS	RF00234	60		2	2			2		2		
Glycine	RF00504	54	7	8	17	3	1	4		11	9	
SAM_alpha	RF00521	71	1	7	7			6		6		
PreQ1	RF00522	67										
SAM-IV	RF00634	73										
preQ1-II	RF01054	68										
MOCO_RNA	RF01055	59			1					1	1	
Mg_sensor	RF01056	78										
SAH	RF01057	63	2	1	2	1						1
Total	-	-	50	102	135	9	3	61	0	88	33	

Table 3.

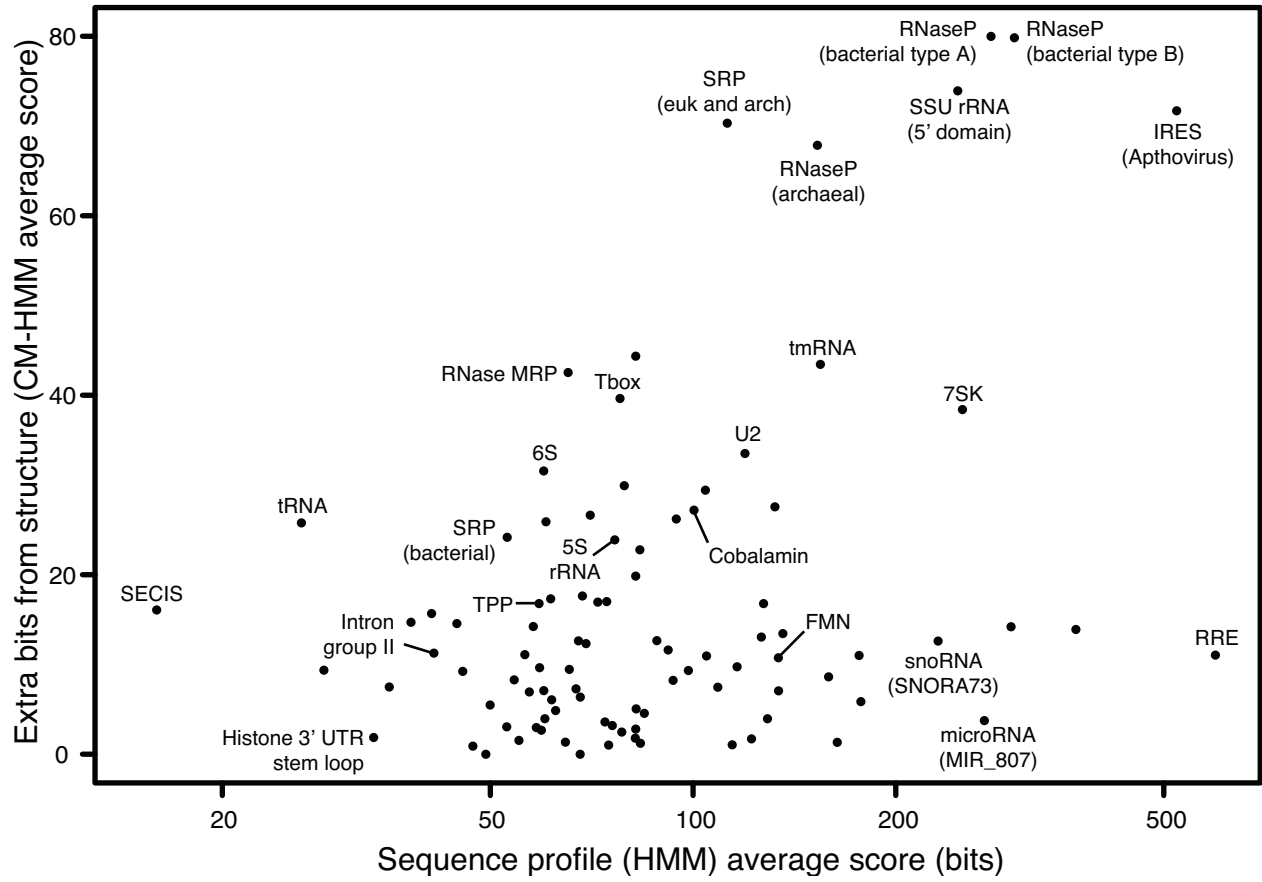


Figure 1.

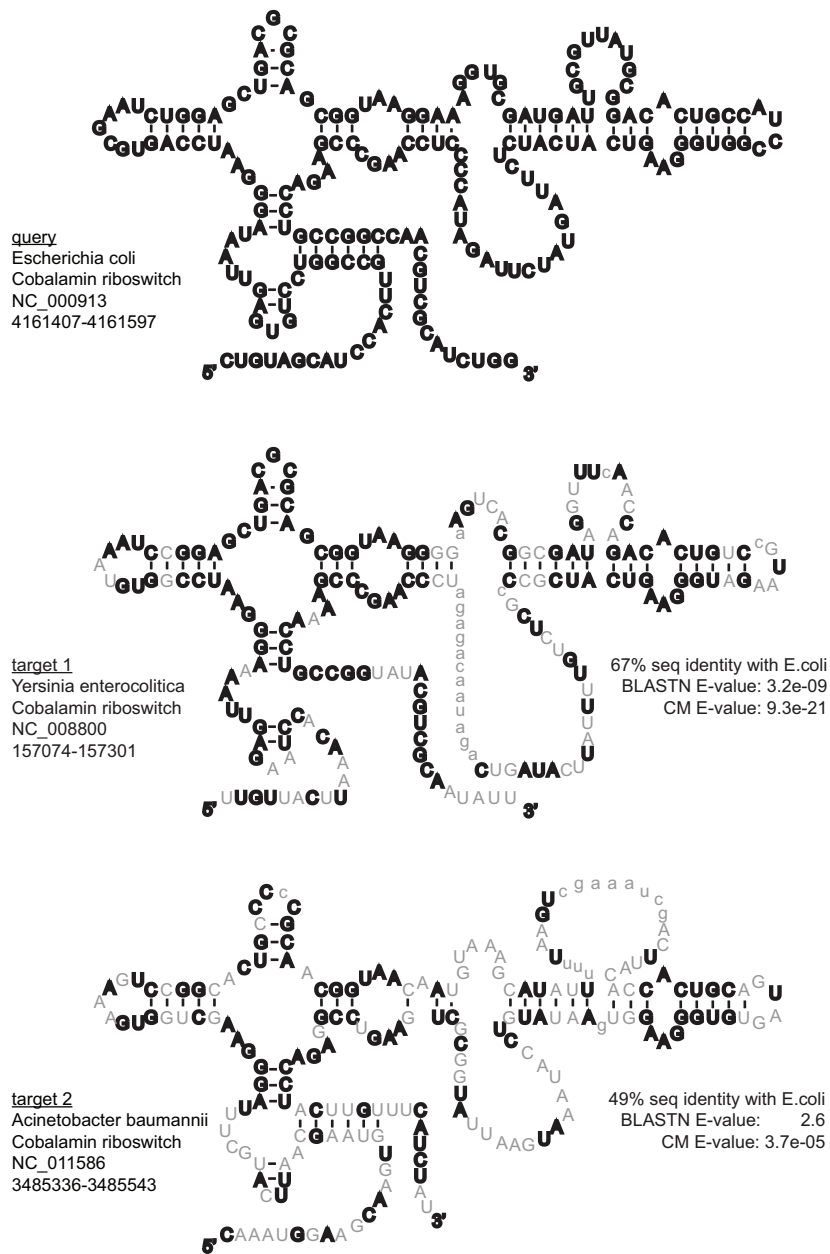


Figure 2.