RNA Biology 10:7, 1170–1179; July 2013; © 2013 Landes Bioscience

Computational identification of functional RNA homologs in metagenomic data

Eric P. Nawrocki* and Sean R. Eddy

Janelia Farm Research Campus; Ashburn, VA USA

Keywords: metagenomics, structural RNA, noncoding RNA, homology search

A key step toward understanding a metagenomics data set is the identification of functional sequence elements within it, such as protein coding genes and structural RNAs. Relative to protein coding genes, structural RNAs are more difficult to identify because of their reduced alphabet size, lack of open reading frames and short length. Infernal is a software package that implements "covariance models" (CMs) for RNA homology search, which harness both sequence and structural conservation when searching for RNA homologs. Thanks to the added statistical signal inherent in the secondary structure conservation of many RNA families, Infernal is more powerful than sequence-only based methods such as BLAST and profile HMMs. Together with the Rfam database of CMs, Infernal is a useful tool for identifying RNAs in metagenomics data sets.

An important step in analyzing a metagenomic sequence data set is identifying functional sequence elements. This is a prerequisite for determining important properties of the biological 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.[1-6]

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 transer 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 post-translational regulation,[7] 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[8,9]). Archaea have numerous small nucleolar RNAs (snoRNAs) homologous to eukaryal snoRNAs that direct site-specific RNA methylation and pseudouridylation.[10] Many eukaryotes make extensive use of RNA regulatory mechanisms via pathways related to RNA interference (RNAi) and micro-RNAs (miRNAs),[11,12] and these will become relevant to metagenomic studies that 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 with more well-developed methodologies for discovering and analyzing protein coding sequences. It hints that other RNAs likely remain undiscovered.[13] 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,[4,6,14,15] but as other reviews have discussed, de novo RNA discovery (gene-finding) methods[16-18] 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.[19] 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[20] 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 among 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 rRNAs, which are readily detected by sequence comparison in all species; the so-called human "ultraconserved" regions included regions of rRNA[21]), this

*Correspondence to: Eric P. Nawrocki; Email: nawrockie@janelia.hhmi.org
Submitted: 02/14/13; Revised: 05/13/13; Accepted: 05/14/13
http://dx.doi.org/10.4161/rna.25038

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 vs. 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 was inferred from the presence of homologs of snoRNP protein components like fibrillarin well before snoRNA homologs were discovered.[22,23] A similar situation can occur when identifying homologous cis-regulatory RNA elements (such as riboswitches) for clearly homologous coding genes.

Figure 1 shows some specific anecdotal examples. These data are fairly typical of searching databases with protein vs. 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.[24] Second, notice that RNA components are usually much shorter than the coding sequence of the protein components, compromising statistical signal and the ability of primary sequence analysis (BLASTN) to resolve homologous relationships from background. (Sequence accessions used in these examples are listed in **Table 1**.)

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 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.[25-27]

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 2** shows the average score of search models for about 150 RNA sequence families, comparing models of sequence conservation alone ("profile hidden Markov models," profile HMMs[28,29]) to models of sequence plus RNA secondary structure conservation ("covariance models," CMs[28,30]). 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 scoring system that is position-specific, with different scores at each position derived from the observed frequencies of the aligned residues at that position.

The unit of score is a "bit" (essentially the same as BLAST "bit scores"), which is a measure of information content.[31,32] 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 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 1/16 for each of the 16 possible base pairs, but only four are observed with probability 1/4 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.[33] Roughly speaking, every three or so bits of score improves the E-value by a factor of 10-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 10 more bits of information for a

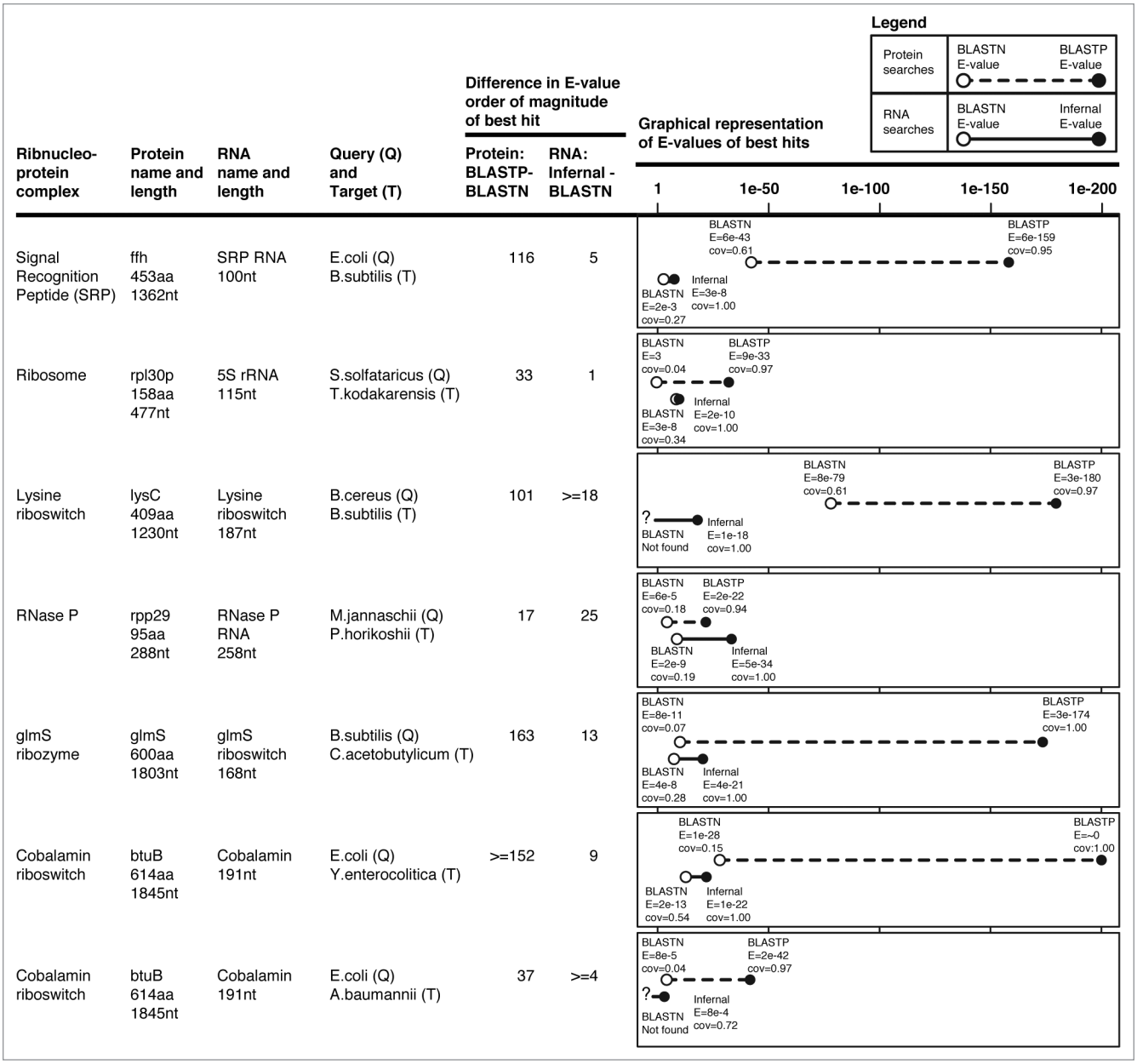

Figure 1. Homology search improvement achieved by utilizing additional information for proteins and structured noncoding RNAs. Examples of identifying coding region homologies by amino acid sequence vs. nucleic acid sequence comparison (BLASTP vs. BLASTN, dashed lines), compared with identifying RNA homologies by primary sequence vs. structure/sequence comparison (BLASTN vs. Infernal, solid lines) for several ribonucleoprotein complexes. Filled circles correspond to BLASTP protein searches and Infernal RNA searches. Open circles correspond to BLASTN coding region searches and BLASTN RNA searches. Question marks indicate targets that were not found by the indicated search method. Each point is labeled with its E-value ("E") and fractional coverage ("cov"), calculated as the fraction of query positions included in the hit alignment. 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. For example, in the leftmost column, when we use the SRP protein *ffh* in *E. coli* as a query against the *B. subtilis* proteome with BLASTP, the top scoring hit is to the *ffh* protein with an E-value of 6×10^{-159} and spans 95% of the full protein sequence. Using the coding sequence of *E. coli*'s *ffh* protein as a BLASTN target against the *B. subtilis* genome returns a top hit comprising 61% of the *ffh* coding sequence with an E-value of 6×10^{-43}. Thus, using the protein sequence instead of the coding sequence increases the statistical significance of the *ffh* homology match by 116 orders of magnitude, indicated by the length of the dashed line in the leftmost box of the figure. Reported E-values are for these single genome/proteome searches, and so would be higher for searches of larger databases. Query RNAs were selected from candidates found by Infernal in each listed query genome's sequence using the Rfam 11.0 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, in row order are: RF00169, RF00001, RF00168, RF00373, RF00234, RF00174. For riboswitches, the protein components are always immediately downstream of the RNA components.

Table 1. GenBank genome and protein accessions and RNA genomic coordinates for examples from **Figure 1**

Organism name	Genome accession	Protein name	Protein accession	RNA name	RNA genomic coordinates
Methanocaldooccus jannaschii	NC_00909.1	Rpp29	NP_247439.1	RNase P RNA	643504–643761
Pyrococcus horikoshii	NC_00961.1	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_000964.3	lysC	NP_390725.1	Lysine riboswitch	2910872–2911051
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_000964.3	glmS	NP_388059.1	Glms riboswitch	200006–200173
Bacillus subtilis	NC_000964.3	ffh	NP_389480.1	SRP RNA	26531–26633
Escherichia coli	NC_000913.2	ffh	NP_417101.1	SRP RNA	475679–475778
Escherichia coli	NC_000913.2	btuB	NP_418401.1	Cobalamin riboswitch	416407–4161597
Klebsiella pneumonia	CP000647.1	btuB	ABR78634.1	Cobalamin riboswitch	4660061–4660248
Yersinia enterocolitica	NC_08800.1	btuB	YP_01004531.1	Cobalamin riboswitch	157101–157301
Vibrio cholera	NC_009457.1	btuB	YP_001218242.1	Cobalamin riboswitch	2498535–2498369
Acinetobacter baumannii	NC_011586.1	btuB	YP_002320687.1	Cobalamin riboswitch	3485342–3485537

homology search means shifting E-values favorably 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 2 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 often 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 three to six orders of magnitude. This improvement can be seen in the results of the anecdotal searches of **Figure 1** comparing the E-values obtained by primary sequence BLASTN searches to those obtained by Infernal,[34] a sequence+secondary structure RNA homology search tool, as we will discuss in more detail below.

The conclusion here is that while primary sequence is still the dominant source of information for these searches, 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).[28,30,35,36] 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.[28,37-39] 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 (www.infernal.janelia.org).

A variety of other computational tools for RNA homology search exist besides Infernal.[3,4,6,40] Some of the most popular tools are ERPIN,[41] FASTR,[42] RSmatch,[43] RNAMotif,[44] RNATOPS,[45] and PatScan.[46] Infernal is one of the most generally applicable and powerful[47] tools and is the basis for a widely used RNA family database (Rfam; described below). 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 **Figure 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 one and 25 orders of magnitude.

Figure 3 provides more detail for the cobalamin (B12) riboswitch example from **Figure 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 base-paired 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 8×10^{-4}, despite it sharing only 55% sequence identity with the *E. coli* riboswitch. The analogous search with BLASTN does not identify the riboswitch homology (no reported hits).

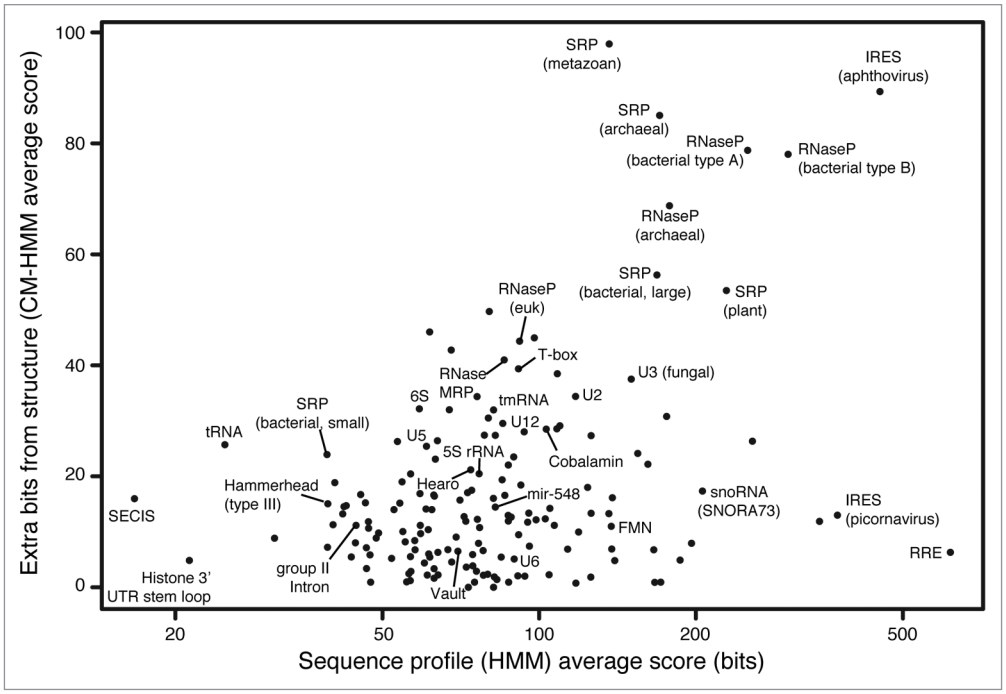

Figure 2. Additional information (in bits) gained by structure/sequence profiles vs. sequence-only profiles for various RNA families. Structure/sequence profiles are most advantageous for families with less primary sequence information (toward left) and more secondary structure information (toward top), 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 164 Rfam release 11.0 families[52] with 50 or more sequences in the "seed" alignment (with the exception of SSU rRNA bacteria and SSU rRNA eukarya which would have been outliers on the plot with x-axis values above 1,900 bits and y-axis values above 150 bits). 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, 10000 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 1.1 programs cmbuild, cmemit and cmalign. A slightly modified version of this figure will appear in a book to be published by Springer Humana Press entitled "RNA sequence, structure and function: computational and bioinformatic methods," edited by Jan Gorodkin and Walter Ruzzo, in chapter 9, entitled "Annotating functional RNAs in genomes using Infernal" as **Figure 2**. This figure is included here with kind permission from Springer Science+Business Media B.V.

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 16 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 **Figs. 1 and 3**), 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.[48]

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 1,500 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).[28,29]

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) uses Infernal to produce alignments of hundreds of thousands of small subunit (SSU) rRNAs.[49] 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

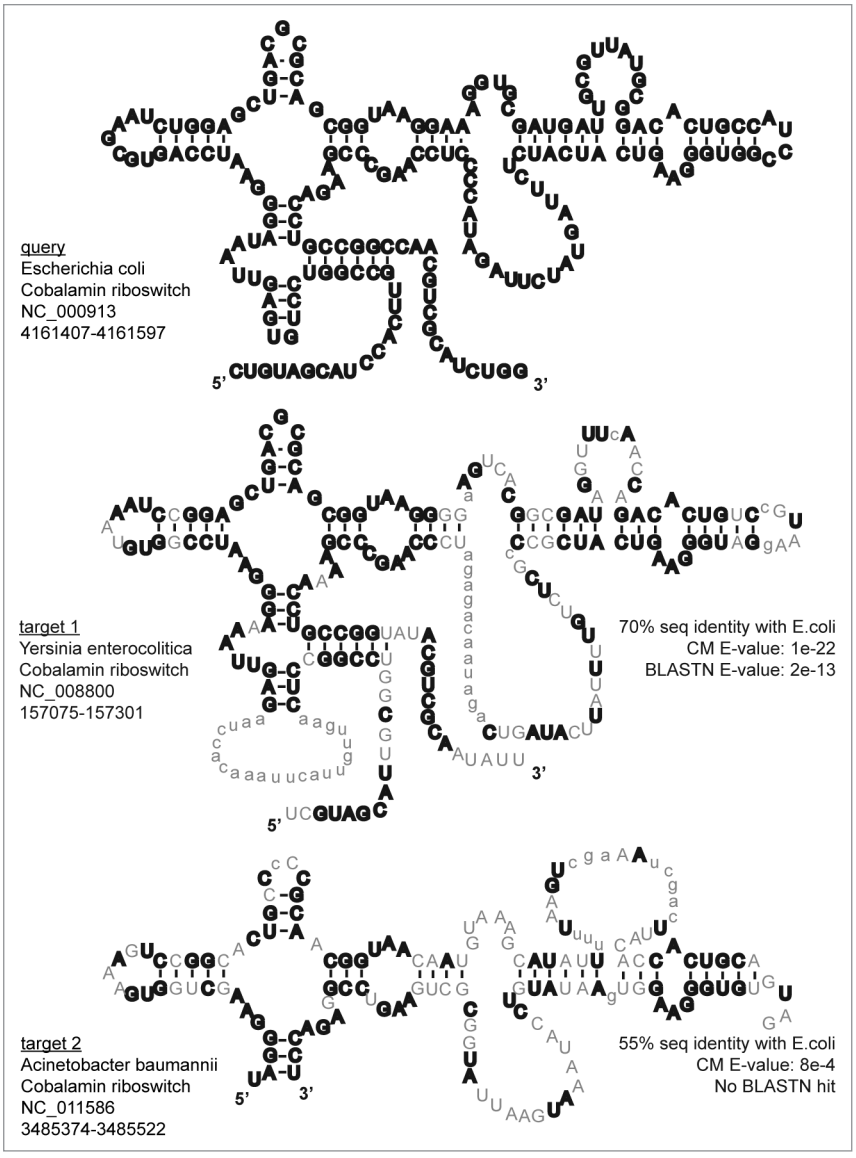

Figure 3. Secondary structure of three cobalamin riboswitches. Using the *E. coli* sequence as a query against their respective genomes, BLASTN detects the *Y. enterocolitica* cobalamin riboswitch with a significant E-value, but not the *A. baumanii* riboswitch. Infernal searches with a CM constructed from the *E. coli* sequence and structure (from the Rfam seed alignment for family RF00174[52]) find both riboswitches with increased significance values. These example searches are also used in **Figure 1**. Note that the *A. baumanii* riboswitch prediction by Infernal is not full length, and excludes the 5' and 3' ends. Presumably, the *A. baumanii* riboswitch extends past the boundaries of the Infernal prediction, but is sufficiently diverged from the *E. coli* sequence and structure to not be included in the optimal hit alignment. 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 gray. Inserted residues with respect to the query are shown in lowercase. Basepairs in the Rfam annotated structure are connected by solid lines. All riboswitches are immediately upstream (5'; within 100 residues) of btuB vitamin B12 transporter protein coding genes in their respective genomes.

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 rRNA. However, if the goal is comprehensive high-throughput annotation of many different functional RNAs, for instance as part of analyzing a new metagenomic sequence data set, 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 10000 protein domain sequence alignments for systematic profile HMM analysis,[50,51] there is a database called Rfam that has systematically collected RNA alignments and CMs.[52]

Rfam: High-throughput RNA homology search and annotation. The Rfam database[52] 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 (11.0) version of Rfam contains 2208 families (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; rfam.sanger.ac.uk/family/RF00504) is represented by an alignment of 44 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.[52]

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 (infernal.janelia.org) and Rfam CM files (rfam.sanger.ac.uk) can be freely downloaded and used to identify homologs of known functional RNAs in a metagenomics data set. As an example of such an analysis, we performed CM searches of a previously published metagenomics data set.[53] The data set

includes about 200000 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 26 Rfam 11.0 CMs of type "cis reg; riboswitch."[52] For comparison, we repeated the search with BLASTN, using each individual sequence in the Rfam seed alignment as a BLASTN query and combining the results to identify any significant matches.[54] Additionally, we performed searches with Infernal v1.1 using non-structured HMM-like models from alignments without secondary structure that ignore secondary structure and score only primary sequence conservation. (These models were created using the --noss option to Infernal's cmbuild program and are essentially equivalent to profile HMMs, so we refer to them as HMMs below). Comparison of the BLAST, 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 2 includes the number of putative riboswitches (hits) with E values less than 10^{-5} found for each family using each method. For the BLAST searches, E-values were multiplied by the number of queries per family. For example, an E-value reported by BLAST of 10^{-5} for the FMN family would be corrected to 1.44×10^{-3} because there were 144 BLAST queries. Also displayed in **Table 2** 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 145 total putative riboswitches in the soil and whale falls data set; Infernal "no structure" HMM-like searches found 133; and BLAST found 96. The HMM-like searches detected 45 hits that BLAST did not, and CM searches detected 16 hits that HMMs did not, indicating that using profiles and additional scoring of structure both contribute significantly to an increased sensitivity of CMs over BLAST. Also note the significant difference in average coverage (fraction of the query sequence covered in the hit alignment) between BLAST (0.66) and the HMM and CM searches (0.98 and 0.97, respectively). BLAST tends to find shorter hits of high identity, while HMMs and CMs often return full-length hits that are more informative for annotation.

We can compare these results to the published results of a similar analysis of riboswitch occurrence in the same metagenomic data set using different search methodology.[55] Kazanov et al. used the pattern based search program RNA-PATTERN to identify candidates of 11 riboswitch families (eight of which we used in our analysis) in the same soil and whale falls data we analyzed. For the eight families in common, their pattern-based search detected 103 candidate riboswitches, compared with 129 identified by CM searches at a stringent threshold. RNA-PATTERN detected 11 candidates that CMs did not, and CMs detected 39 candidates that RNA-PATTERN did not. The largest differences were for the cobalamin family, for which CMs found 20 candidates undetected by RNA-PATTERN, and the glycine family, for which RNA-PATTERN found 11 candidates undetected by CMs using a CM E-value threshold of 10^{-5}. Three of these 11 are found by the glycine riboswitch CM, but with E-values just below the strict threshold, ranging between 10^{-3} and 10^{-5}. The remaining eight are all immediately adjacent to glycine hits that Infernal does find. This suggests they are likely functional riboswitches because glycine riboswitches usually occur in tandem with two similar structure right next to each other. We found when we looked into this that due to an implementation detail, Infernal sometimes misses one of two RNAs if they appear very close together. This is a limitation of the software we hope to remedy in a future version. Currently, the best way to search for families that occur in tandem is to build and search with a single model of both structures simultaneously. Repeating this Infernal search with this kind of model (built from a new "seed" alignment created by simply concatenating two copies of the original Rfam "seed") finds highly significant ($E < 10^{-10}$) tandem glycine structures that completely cover all 11 RNA-PATTERN hits missed by the original Rfam model.

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, in 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 non-homologous sequences. One way to test the accuracy of any program's E-values is to search randomized non-homologous 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 experiment of ours,[34] involving a benchmark of 51 CMs being searched against a 10 megabase synthetically generated target sequence, the highest non-homologous 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,[47] 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 four important limitations of CM similarity search:

(1) Infernal is slower than BLAST. In the riboswitch example above, the 26 CM searches took about 45 min on one processor, about 15 times longer than BLAST searches (3 min on one processor for all 2,555 pairwise searches). Repeating the Infernal search using all 2,208 Rfam 11.0 models against the 230 Mb data set would take roughly 100 h. Significant compute power (such as a moderate sized cluster) is required to do large scale analyses with CMs. Infernal is parallelized to use multiple threads on multicore computers and for use on clusters using the Message Passing Interface (MPI),[56] although neither method was enabled for the searches reported here.

Table 2. Riboswitch search results

			# Seqs found			# Different seqs found						Avg fractional coverage		
						BLAST	BLAST	HMM	HMM	CM	CM			
Family	Rfam ID	#seed seqs	BLAST	HMM	CM	-HMM	-CM	-BLAST	-CM	-BLAST	-HMM	BLAST	HMM	CM
FMN	RF00050	144	9	9	9							0.90	1.00	1.00
TPP	RF00059	115	26	38	37		1	12	1	12		053	0.97	0.99
SAM	RF00162	433	14	14	14	1	1	1		1		0.68	0.99	0.99
Purine	RF00167	133												
Lysine	RF00168	47		1	1			1		1			1.00	1.00
Cobalamin	RF00174	430	19	38	49	3		23		31	11	0.48	0.95	0.92
glmS	RF00234	18	1	2	3			1		2	1	0.36	1.00	1.00
Glycine	RF00504	44	11	14	16	1		4	1	5	3	0.75	1.00	0.99
SAM α	RF00521	40	5	8	7			3	1	2		0.74	1.00	1.00
PreQ1	RF00522	41												
SAM-IV	RF00634	40												
preQ1-II	RF1064	14												
MOCO RNA motif	RF01055	160												
Mg sensor	RF01056	4												
SAH riboswitch	RF01057	52	2	2	2							1.00	1.00	1.00
AdoCbl riboswitch	RF01482	7	1			1	1					0.25		
MFR	RF01510	2												
AdoCbl-variant	RF01689	144												
SAM-I-IV-variant	RF01725	439	3	2	2	1	1		1		1	1.00	1.00	1.00
SAM-SAH	RF01727	53	2	2	2							1.00	1.00	1.00
SMK box riboswitch	RF01767	25												
c-di-GMP-II	RF01786	54	3	3	3							1.00	1.00	1.00
Drz-agam-1	RF01787	7												
Drz-agam-2–2	RF01788	5												
SAM V	RF01826	6												
THF	RF01831	98												
Total	-	2555	96	133	145	7	4	45	4	54	16	0.66	0.98	0.97

"# seed seqs," the number of sequences in the Rfam seed alignments used to build each model for "CM" and "HMM" columns, and number of query sequences for "BLAST" columns. "# 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 data set[53] for each Rfam 11.0 riboswitch family. For BLAST columns, reported E-values were multiplied by the number of query searches as a correction for multiple tests. Only corrected E-values of less than 10^{-5} were counted. "# 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 1 because one hit detected by BLAST was undetected by the HMM search. A blank space in a cell indicates 0. "HMM" refers to the "non-structured" HMM-like Infernal models. "avg fractional coverage" is the average fractional coverage of the alignment of the hit, with respect to the query. For example, a BLAST hit that involves positions 26 to 75 of a length 100 query sequence has a coverage of 0.5. The "total" row for these columns gives the average coverage over all hits to all families.

Though still slow compared with BLAST, Infernal is much faster than it was just a few years ago. The current version (v1.1) is about 10,000 times faster than version 0.55.[13] The speedup is due to heuristics, including profile HMM filtering[57,58] and banded dynamic programming,[59] which sacrifice a small amount of sensitivity for the increased speed. This sensitivity sacrifice, though small, disproportionately impacts remote homology detection.[34,59] 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.[60] Further acceleration remains a 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 2000 consensus structures is challenging. Additionally, a single consensus structure is unable to properly capture the evolutionary variation observed among 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) 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.[61] Other methods exist for RNA homology search that can model RNA pseudoknots, including ERPIN[41] and RNATOPS.[45]

(4) 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 (Eddy 2003), it is more practical and appropriate to use profile HMMs rather than CMs, avoiding the CMs computational costs. For this reason, Infernal's cmsearch program automatically detects when a query model with zero basepairs is being used, and employs the more efficient profile HMM search algorithms instead of the slower CM methods.

Materials and Methods

Software used: NCBI-BLAST 2.2.27+ for BLASTN and BLASTP and Infernal version 1.1. Default settings were used for all searches, with the following exceptions: a single CPU was used, instead of allowing multithreading, using the num threads 1 option for BLAST and the --cpu 0 option for Infernal; and for BLASTN of protein coding sequences in **Figure 1**, word size was set at 8 (word_size 8) because it resulted in lower E-values than the default word size of 11 in some cases. The GenBank genome and protein accessions and RNA genomic coordinates for searches in **Figure 1** are listed in **Table 1**.

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 coworkers were able to predict the secondary structure of rRNA to greater than 98% accuracy per base pair by essentially manual comparative analysis of careful rRNA alignments,[25] and Francois Michel and Eric Westhof essentially predicted the structure of group I intron catalytic introns in much the same way.[26] 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 data sets for known families of RNAs using a combination of the Infernal software (infernal.janelia.org) and CMs from the Rfam database.[52]

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

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

References

1. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2001; 2:919-29; PMID:11733745; http://dx.doi.org/10.1038/35103511
2. Hammann C, Westhof E. Searching genomes for ribozymes and riboswitches. Genome Biol 2007; 8:210; PMID:17472738; http://dx.doi.org/10.1186/gb-2007-8-4-210
3. Jossinet F, Ludwig TE, Westhof E. RNA structure: bioinformatic analysis. Curr Opin Microbiol 2007; 10:279-85; PMID:17548241; http://dx.doi.org/10.1016/j.mib.2007.05.010
4. Machado-Lima A, del Portillo HA, Durham AM. Computational methods in noncoding RNA research. J Math Biol 2008; 56:15-49; PMID:17786447; http://dx.doi.org/10.1007/s00285-007-0122-6
5. Szymański M, Barciszewska MZ, Zywicki M, Barciszewski J. Noncoding RNA transcripts. J Appl Genet 2003; 44:1-19; PMID:12590177
6. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 2007; 3:e65; PMID:17432929; http://dx.doi.org/10.1371/journal.pcbi.0030065
7. Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet 2005; 21:399-404; PMID:15913835; http://dx.doi.org/10.1016/j.tig.2005.05.008
8. Tucker BJ, Breaker RR. Riboswitches as versatile gene control elements. Curr Opin Struct Biol 2005; 15:342-8; PMID:15919195; http://dx.doi.org/10.1016/j.sbi.2005.05.003
9. Winkler WC. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. Curr Opin Chem Biol 2005; 9:594-602; PMID:16226486; http://dx.doi.org/10.1016/j.cbpa.2005.09.016

10. Bachellerie JP, Cavaillé J, Hüttenhofer A. The expanding snoRNA world. Biochimie 2002; 84:775-90; PMID:12457565; http://dx.doi.org/10.1016/S0300-9084(02)01402-5
11. Ambros V. The functions of animal microRNAs. Nature 2004; 431:350-5; PMID:15372042; http://dx.doi.org/10.1038/nature02871
12. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004; 116:281-97; PMID:14744438; http://dx.doi.org/10.1016/S0092-8674(04)00045-5
13. Eddy SR. Computational genomics of noncoding RNA genes. Cell 2002; 109:137-40; PMID:12007398; http://dx.doi.org/10.1016/S0092-8674(02)00727-4
14. Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. Bioinformatics 2008; 24:2807-13; PMID:18974076; http://dx.doi.org/10.1093/bioinformatics/btn560
15. Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria. Biol Chem 2005; 386:1219-38; PMID:16336117; http://dx.doi.org/10.1515/BC.2005.140
16. Babak T, Blencowe BJ, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics 2007; 8:33; PMID:17263882; http://dx.doi.org/10.1186/1471-2105-8-33
17. Meyer IM. A practical guide to the art of RNA gene prediction. Brief Bioinform 2007; 8:396-414; PMID:17483123; http://dx.doi.org/10.1093/bib/bbm011
18. Griffiths-Jones S. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet 2007; 8:279-98; PMID:17506659; http://dx.doi.org/10.1146/annurev.genom.8.080706.092419
19. del Val C, Rivas E, Torres-Quesada O, Toro N, Jiménez-Zurdo JI. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. Mol Microbiol 2007; 66:1080-91; PMID:17971083; http://dx.doi.org/10.1111/j.1365-2958.2007.05978.x
20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-402; PMID:9254694; http://dx.doi.org/10.1093/nar/25.17.3389
21. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science 2004; 304:1321-5; PMID:15131266; http://dx.doi.org/10.1126/science.1098119
22. Amiri KA. Fibrillarin-like proteins occur in the domain Archaea. J Bacteriol 1994; 176:2124-7; PMID:8144483
23. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. Homologs of small nucleolar RNAs in Archaea. Science 2000; 288:517-22; PMID:10775111; http://dx.doi.org/10.1126/science.288.5465.517
24. Pearson WR. Effective protein sequence comparison. Methods Enzymol 1996; 266:227-58; PMID:8743688; http://dx.doi.org/10.1016/S0076-6879(96)66017-0
25. Gutell RR, Lee JC, Cannone JJ. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol 2002; 12:301-10; PMID:12127448; http://dx.doi.org/10.1016/S0959-440X(02)00339-1
26. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol 1990; 216:585-610; PMID:2258934; http://dx.doi.org/10.1016/0022-2836(90)90386-Z
27. Pace NR, Thomas BC, Woese CR. Probing RNA structure, function and history by com-parative analysis. In Gesteland RF, Atkins JF, editors, The RNA World, pages 113-142. Cold Spring Harbor Laboratory Press, New York, 1993.
28. Durbin R, Eddy SR, Krogh A, Mitchison GJ. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge UK, 1998.
29. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994; 235:1501-31; PMID:8107089; http://dx.doi.org/10.1006/jmbi.1994.1104
30. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res 1994; 22:2079-88; PMID:8029015; http://dx.doi.org/10.1093/nar/22.11.2079
31. MacKay DJC. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, UK, 2003.
32. Shannon CE. A mathematical theory of communication. Bell System Technical Journal 1948; 27:379-423
33. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 2008; 4:e1000069; PMID:18516236; http://dx.doi.org/10.1371/journal.pcbi.1000069
34. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics 2009; 25:1335-7; PMID:19307242; http://dx.doi.org/10.1093/bioinformatics/btp157
35. Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 2002; 3:18; PMID:12095421; http://dx.doi.org/10.1186/1471-2105-3-18
36. Sakakibara Y, Brown M, Underwood RC, Mian IS, Haussler D. Stochastic context-free grammars for modeling RNA. In Lawrence Hunter, editor, Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences: Biotechnology Computing, volume V, pages 284-293, Los Alamitos, CA, 1994. IEEE Computer Society Press.
37. Hopcroft JE, Ullman JD. Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, Reading, Massachusetts, 1979.
38. Kasami T. An efficient recognition and syntax algorithm for context-free algorithms. Technical Report AFCRL-65-758, Air Force Cambridge Research Lab, Bedford, Mass., 1965.
39. Younger DH. Recognition and parsing of context-free languages in time n^3. Inf Control 1967; 10:189-208; http://dx.doi.org/10.1016/S0019-9958(67)80007-X
40. Eddy SR. Computational analysis of RNAs. Cold Spring Harb Symp Quant Biol 2006; 71:117-28; PMID:17381287; http://dx.doi.org/10.1101/sqb.2006.71.003
41. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 2001; 313:1003-11; PMID:11700055; http://dx.doi.org/10.1006/jmbi.2001.5102
42. Zhang S, Haas B, Eskin E, Bafna V. Searching genomes for noncoding RNA using FastR. IEEE/ACM Trans. Comput. Biol. Bioinform. 2005; 2:366-79; PMID:17044173 http://dx.doi.org/10.1109/TCBB.2005.57
43. Liu J, Wang JT, Hu J, Tian B. A method for aligning RNA secondary structures and its application to RNA motif detection. BMC Bioinformatics 2005; 6:89; PMID:15817128; http://dx.doi.org/10.1186/1471-2105-6-89
44. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 2001; 29:4724-35; PMID:11713323; http://dx.doi.org/10.1093/nar/29.22.4724
45. Huang Z, Wu Y, Robertson J, Feng L, Malmberg RL, Cai L. Fast and accurate search for non-coding RNA pseudoknot structures in genomes. Bioinformatics 2008; 24:2281-7; PMID:18687694; http://dx.doi.org/10.1093/bioinformatics/btn393
46. Dsouza M, Larsen N, Overbeek R. Searching for patterns in genomic data. Trends Genet 1997; 13:497-8; PMID:9433140; http://dx.doi.org/10.1016/S0168-9525(97)01347-4
47. Freyhult EK, Bollback JP, Gardner PP. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 2007; 17:117-25; PMID:17151342; http://dx.doi.org/10.1101/gr.5890907
48. Klein RJ, Eddy SR. RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 2003; 4:44; PMID:14499004; http://dx.doi.org/10.1186/1471-2105-4-44
49. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 2009; 37(Database issue):D141-5; PMID:19004872; http://dx.doi.org/10.1093/nar/gkn879
50. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. Nucleic Acids Res 2008; 36(Database issue):D281-8; PMID:18039703; http://dx.doi.org/10.1093/nar/gkm960
51. Letunic I, Doerks T, Bork P. SMART 6: recent updates and new developments. Nucleic Acids Res 2009; 37(Database issue):D229-32; PMID:18978020; http://dx.doi.org/10.1093/nar/gkn808
52. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 2013; 41(Database issue):D226-32; PMID:23125362; http://dx.doi.org/10.1093/nar/gks1005
53. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. Science 2005; 308:554-7; PMID:15845853; http://dx.doi.org/10.1126/science.1107851
54. Grundy WN. Homology detection via family pairwise search. J Comput Biol 1998; 5:479-91; PMID:9773344; http://dx.doi.org/10.1089/cmb.1998.5.479
55. Kazanov MD, Vitreschak AG, Gelfand MS. Abundance and functional diversity of riboswitches in microbial communities. BMC Genomics 2007; 8:347; PMID:17908319; http://dx.doi.org/10.1186/1471-2164-8-347
56. Gropp W, Lusk E, Doss N, Skjellum A. A high-performance, portable implementation of the mpi message passing interface standard. Parallel Comput 1996; 22:789-828; http://dx.doi.org/10.1016/0167-8191(96)00024-5
57. Eddy SR. Accelerated profile HMM searches. PLoS Comp. Biol. 2011; 7:e1002195; PMID:22039361; http://dx.doi.org/10.1371/journal.pcbi.1002195
58. Weinberg Z, Ruzzo WL. Sequence-based heuristics for faster annotation of non-coding RNA families. Bioinformatics 2006; 22:35-9; PMID:16267089; http://dx.doi.org/10.1093/bioinformatics/bti743
59. Nawrocki EP, Eddy SR. Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput Biol 2007; 3:e56; PMID:17397253; http://dx.doi.org/10.1371/journal.pcbi.0030056
60. Nawrocki EP, Kolbe DL, Eddy SR. The Infernal user's guide. [http://infernal.janelia.org/], 2009.
61. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol 1999; 285:2053-68; PMID:9925784; http://dx.doi.org/10.1006/jmbi.1998.2436