Computational Identification of Structural RNAs Using Infernal and Rfam

Eric P. Nawrocki(1), Ioanna Kalvari(2), Joanna Argasinska(2), Anton I. Petrov(2) and Sean R. Eddy(3)

1: National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA. 2: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. 3: Howard Hughes Medical Institute, FAS Center for Systems Biology, John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

nawrocke@ncbi.nlm.nih.gov

RNA homology searches based on sequence and structure

Functional RNAs do not encode proteins, but rather function directly as RNAs. Many of these RNAs form stable, evolutionarily conserved three-dimensional structures that are crucial to their functions in various fundamental cellular processes including protein synthesis, gene expression, splicing, protein transport, and more.

Finding homologs of structural RNAs is challenging because the sequences are often short (100-200 nt), lack ORFs, and have regions of high sequence variability even while conserving their three-dimensional structure. The most successful approaches for RNA homology search take advantage of both sequence and secondary structure conservation [1]. The example below from [2] shows how searching for both sequence and secondary structure using a covariance model (CM) can identify a Cobalamin riboswitch which BLAST, a sequence-only based method, fails to identify.

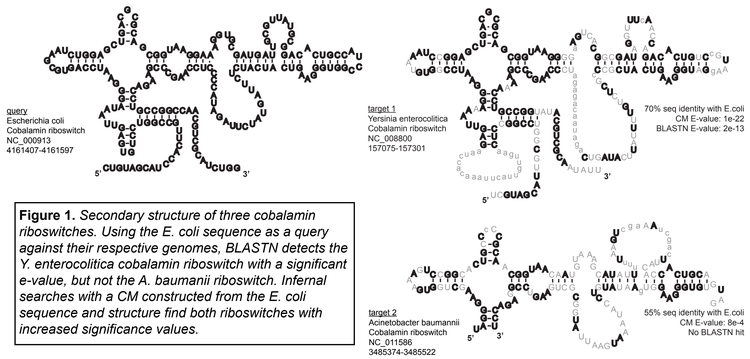


Figure 1. *Secondary structure of three cobalamin riboswitches. Using the E. coli sequence as a query against their respective genomes, BLASTN detects the Y. enterocolitica cobalamin riboswitch with a significant e-value, but not the A. baumanii riboswitch. Infernal searches with a CM constructed from the E. coli sequence and structure find both riboswitches with increased significance values.*

When searching for protein coding genes, the amino acid sequence should be used instead of the nucleotide sequence because the larger amino acid alphabet makes protein searches much more powerful [3]. Incorporating secondary structure into RNA searches offers a similar boost to the statistical power of a nucleotide-only based search for RNAs, albeit not as dramatic. Figure 2 below compares the increase in statistical significance between BLASTN (nucleotide-based search) and BLASTP (protein-based search) for protein-coding genes and between BLASTN and Infernal (CM-based sequence and structure search) for RNA genes, where both the protein-coding gene and RNA gene being searched for are members of the same ribonucleoprotein complex.

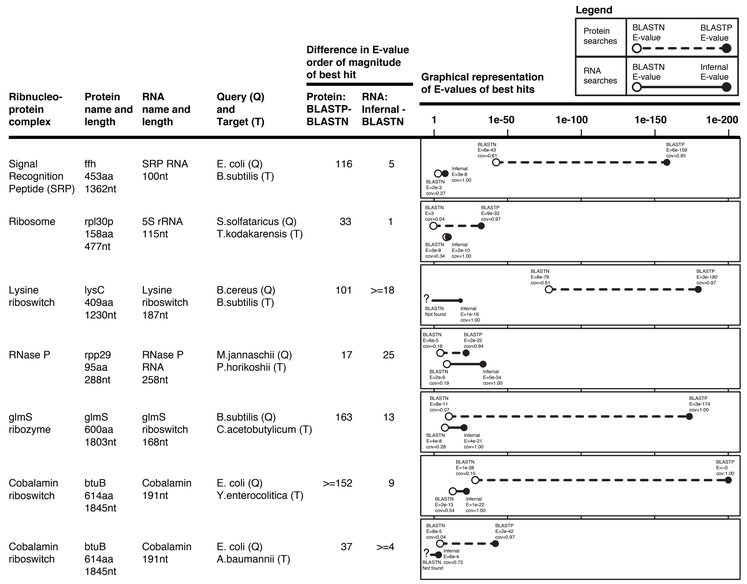


Figure 2: *Homology search improvement achieved by utilizing additional information for proteins and structured noncoding RNAs. This figure is from [2].*

Conserved sequence and structure as statistical signals

The conserved sequence and secondary structure of RNAs offers two statistical signals that can be harnessed when searching databases for homologs using CMs. In Figure 3 below, the amount of information, measured in *bits*, inherent in a sequence-only profile (14 bits) and a sequence-and-structure profile (17 bits) is shown for a toy example of an RNA family. We expect a match to a sequence-only profile for this family once in every $2^{14} = 16,384$ random nucleotides. Additionally modeling structure with a sequence-and-structure based profile (like a CM) reduces this probability 8-fold, to once every every $2^{17} = 131,072$ random nucleotides.

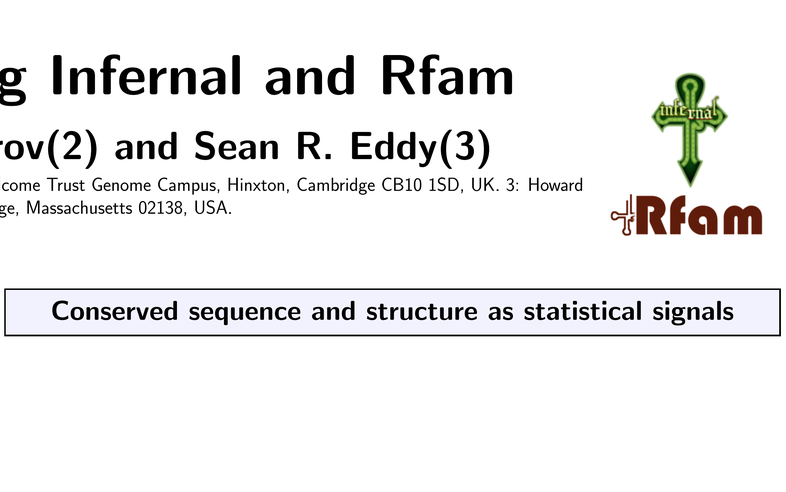


Figure 3: *Information in a sequence-only versus a sequence and structure profile. Boxes with internal numbers at top and bottom of the alignment indicate the number of bits per position from the sequence (black), or per basepair from the structure (white). This figure is from [4].*

The amount of additional information gained from structure varies widely for real RNA families, as shown for about 160 families in Figure 4 below. Note that for most families, modeling structure contributes at least 10 additional bits of information, which corresponds to lowering the expected chance of a false positive in a random database (i.e. the E-value of a database hit) by three orders of magnitude ($2^{10} = 1024$).

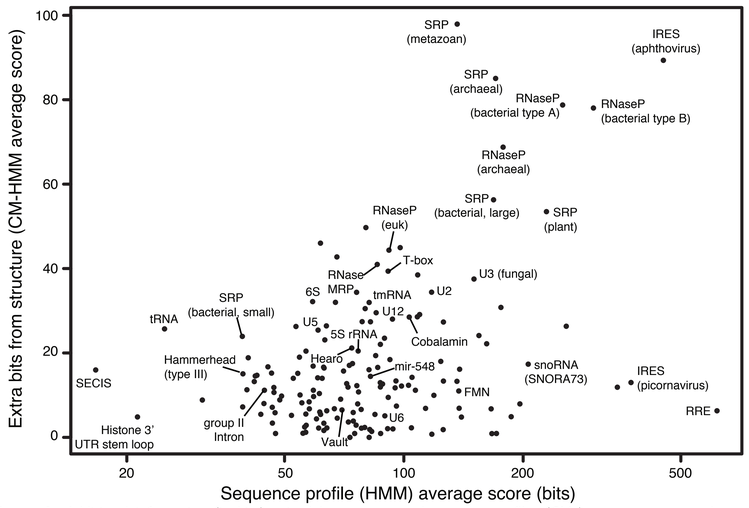


Figure 4: *Additional information (in bits) gained by sequence and structure profiles (CMs) versus sequence-only profiles (HMMs) for various RNA families. Data shown for the 164 Rfam release 11.0 families with 50 or more sequences in the* seed *alignment. For each family, the seed alignment was used to build two profile models, a CM and a profile HMM. From each model, 10,000 sequences were generated and scored, and the average score per sampled sequence was calculated. Infernal version 1.1 was used for all steps. This figure is from [2].*

Internal benchmark shows benefit of modeling sequence and structure conservation

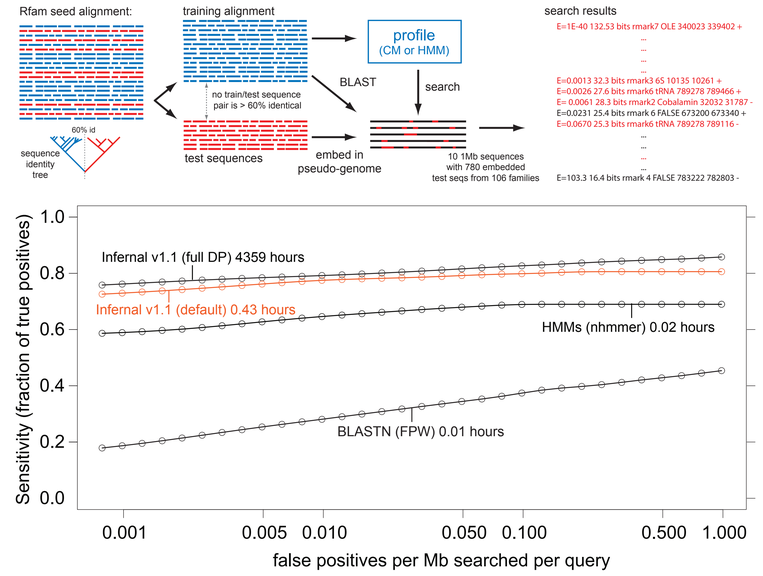


Figure 5: *RMARK benchmark. Top: schematic for benchmark construction. Bottom: Results of benchmark. Plots are shown for the new Infernal 1.1 with and without filters, for profile HMM searches with nhmmer [5] (from the HMMER package included in Infernal 1.1, default parameters) and for family-pairwise-searches with BLASTN (ncbi-blast-2.2.28+ default parameters). The Infernal times do not include time required for model calibration. This figure is from [6]*

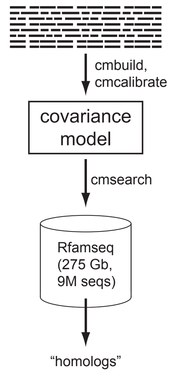
Rfam: the RNA families database [7]

Rfam contains 2588 families, each represented by a
- SEED alignment
- covariance model (CM)
- list of annotated hits in Rfamseq database (8,725,484 total hits)

If you submit a family to Rfam, these tools become available to the community, allowing it to be annotated in genomes and other datasets.

Rfam also includes community annotation of families (wikipedia), RNA motifs, secondary structure diagrams, alignment statistics, taxonomic information, and sequence search capability.

rfam.xfam.org



Types of RNAs in Rfam

count	*type*
239	Cis-reg;
28	Cis-reg; frameshift-element;
32	Cis-reg; IRES;
30	Cis-reg; leader;
31	Cis-reg; riboswitch;
10	Cis-reg; thermoregulator;
76	Gene;
29	Gene; antisense;
11	Gene; antitoxin;
64	Gene; CRISPR;
217	Gene; lncRNA;
530	Gene; miRNA;
23	Gene; ribozyme;
15	Gene; rRNA;
3	Gene; snRNA;
463	Gene; snRNA; snoRNA; CD-box;
266	Gene; snRNA; snoRNA; HACA-box;
24	Gene; snRNA; snoRNA; scaRNA;
15	Gene; snRNA; splicing;
471	Gene; sRNA;
2	Gene; tRNA;
9	Intron;

Funding

This work is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (EPN), by Howard Hughes Medical Institute (EPN (previously), SRE), by the European Bioinformatics institute of the European Molecular Biology Laboratory (IK, JA, AIP), and by the Biotechnology and Biological Sciences Research Council (IK, JA, AIP).

References

[1] E. K. Freyhult, J. P. Bollback, and P. P. Gardner. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, 17:117–125, 2007.

[2] E. P. Nawrocki and S. R. Eddy. Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.*, 10:1170–1179, 2013.

[3] W. R. Pearson. Effective protein sequence comparison. *Methods Enzymol*, 266:227–258, 1996.

[4] E. P. Nawrocki. Annotating functional RNAs in genomes using Infernal. In J. Gorodkin and W. L. Ruzzo, editors, *RNA Sequence, Structure, and Function: Computational and Bioinformatic methods*, pages 163–197. Springer Humana Press, 2014.

[5] T. J. Wheeler and S. R. Eddy. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29:2487–2489, 2013.

[6] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, 2013.

[7] E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, 43:D130–D137, 2015.