Discovery of Archaeal Group I Introns Using Infernal

Eric P. Nawrocki(1), Thomas A. Jones(2), and Sean R. Eddy(2)

1: National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA.
2: 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

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 conserved sequence and secondary structure of RNAs offers two statistical signals that can be harnessed when searching databases for homologs using CMs.

In Figure 1 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.

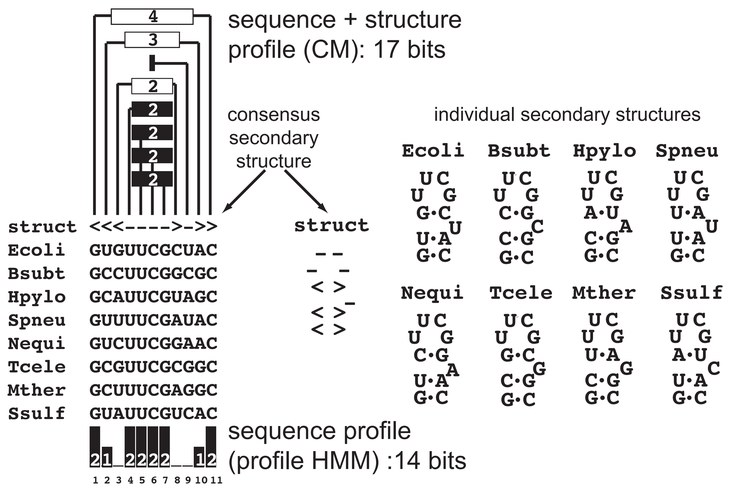


Figure 1: *Information in a sequence-only versus a sequence and structure profile. Boxes with internal numbers at top and bottom of the alignment indicate the number of bits per position from the sequence (black), or per basepair from the structure (white). This figure is from [1].*

The amount of additional information gained from structure varies widely for real RNA families, as shown for about 160 families in Figure 2 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 ($2^{10} = 1024$).

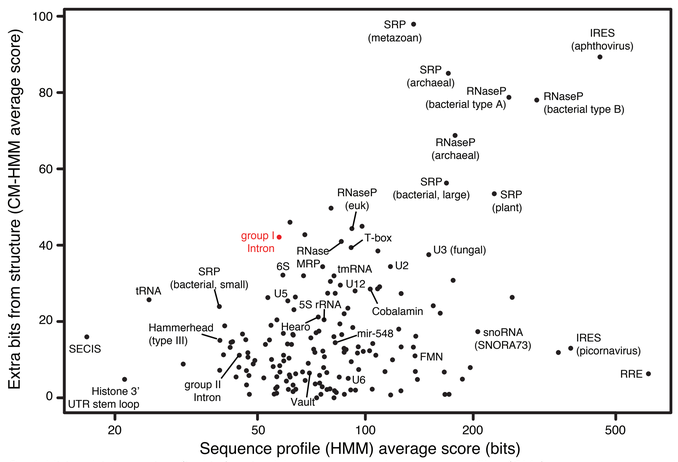


Figure 2: *Additional information (in bits) gained by sequence and structure profiles (CMs) versus sequence-only profiles (HMMs) for various RNA families. Data shown for the 164 Rfam release 11.0 families with 50 or more sequences in the* seed *alignment plus the group I intron model (RF00028, 12 sequences). 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 very similar to one from [2].*

Internal benchmark shows benefit of modeling sequence and structure conservation

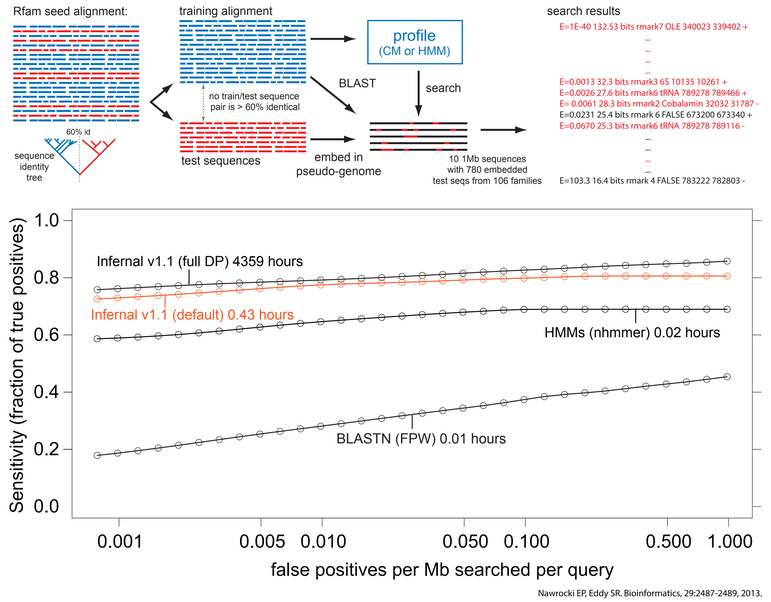


Figure 3: *RMARK benchmark. Top: schematic for benchmark construction. Bottom: Results of benchmark. Plots are shown for the new Infernal 1.1 with and without filters, for profile HMM searches with nhmmer [3] (from the HMMER package included in Infernal 1.1, default parameters) and for family-pairwise-searches with BLASTN (ncbi-blast-2.2.28+ default parameters). The Infernal times do not include time required for model calibration. This figure is from [4]*

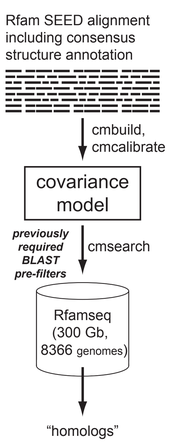
Rfam: the RNA families database [5]

Rfam contains 2791 families, each represented by a
- SEED alignment
- covariance model (CM)
- list of annotated hits in Rfamseq database (more than 2 million 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;

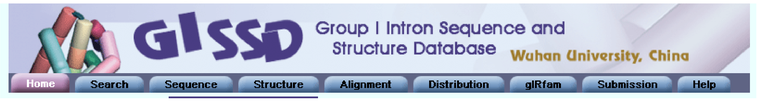
Infernal, without BLAST filters, finds new homologs

Table 1. Comparison of the old Rfam 11.0 BLAST and Infernal 1.0 search strategy versus the new Rfam 12.0 Infernal 1.1 search strategy for 15 of 200 randomly chosen families

Accession	Family ID	Length (nt)	#of seed seqs	Time new (h)	Time old (h)	Time (old/new)	New total hits	Old total hits	New unique hits	Old unique hits
Top five families										
RF00028	Intron_gpI	251	12	125.0	357.2	2.8	71 433	60 264	11 175	1
RF00026	U6	104	188	31.2	181.1	5.8	66 517	62 174	4367	14
RF00003	U1	166	100	11.6	64.0	5.5	15 770	14 867	904	1
RF00162	SAM	108	433	8.3	590.0	70.8	4905	4797	108	0
RF00050	FMN	140	144	17.1	169.9	23.9	4381	4306	76	1

Group I introns

Group I introns are self-splicing ribozymes found in lower eukaryotes, higher plants, bacteria and viruses. Group Is exhibit a sporadic phylogenetic distribution that is consisten with horizontal gene transfer and they often encode homing endonuclease genes which are responsible for their mobility. They are found within highly conserved genes, such as ribosomal RNAs. They are quickly evolving at the sequence level, while maintaining a core secondary structure that is crucial to their catalytic ability. Previous to this work[6] Group I introns had not been discovered in archaea.



The Group I Intron Sequence and Structure Database [7] includes alignments and consensus structures of 15 subtypes of group I introns.

name	# seq	avglen	%id	# bps	# sig	HMM	CM	name	# seq	avglen	%id	# bps	# sig	HMM	CM
RF00028	12	364.8	34	61	8	62.6	41.7	IC1	837	436.0	39	103	84	130.3	40.3
IA1	76	583.6	45	82	51	166.5	38.0	IC2	32	320.2	66	86	27	298.1	51.7
IA2	15	276.8	38	67	24	58.0	47.5	IC3	328	255.8	67	58	16	244.1	13.2
IA3	56	282.3	46	81	50	122.4	39.1	ID	17	242.5	53	66	16	122.8	45.5
IB1	42	298.0	72	87	9	320.4	51.8	IE1	38	362.2	60	95	19	268.0	44.8
IB2	18	242.2	39	65	27	57.1	39.3	IE2	56	399.9	55	112	38	250.9	47.4
IB3	7	277.7	52	72	10	98.5	60.3	IE3	110	405.9	57	119	51	293.4	45.7
IB4	89	282.3	44	72	43	108.3	33.2								

Table 1: *Attributes of the alignments of the 14 GISSD group I subtypes and models, plus the Rfam RF00028 model.*

Searches for archaeal group I introns using Infernal

- downloaded all archaeal sequences in GenBank (6.7Gb as of Sept 2017)
- searched archaeal sequences with all GISSD models + RF00028 with default cm-search parameters and with `-anytrunc`
- 95 non-overlapping hits with $E < 0.01$ corresponding to 39 group I intron candidates (12 IA3 and 27 IB4)
- 30/39 introns have at least one hit with $E < 10^{-10}$
- 36 within LSU rRNA, 3 within SSU rRNA

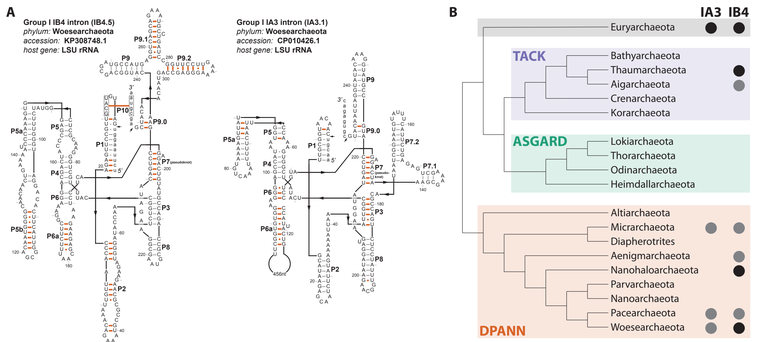


Figure 4: *A) Predicted secondary structure of example archaeal group I IB4 and IB3 introns. B) Archaeal phyla in which candidate group I introns were found.*

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), TAJ and SRE). Rfam is supported by the European Bioinformatics Institute of the European Molecular Biology Laboratory and by the Biotechnology and Biological Sciences Research Council.

References

[1] 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.

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

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

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

[5] 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.

[6] E. P. Nawrocki, T. A. Jones, and S. R. Eddy. Group I introns are widespread in archaea. 46:7970–7976, 2018.

[7] Y. Zhou, C. Lu, Q. J. Wu, Y. Wang, Z. T. Sun, J. C. Deng, and Y. Zhang. GISSD: Group I intron sequence and structure database. 36:D31–D37, 2008.