

Infernal 1.0: inference of RNA alignments

Eric P. Nawrocki, Diana L. Kolbe, and Sean R. Eddy

HHMI Janelia Farm Research Campus

19700 Helix Drive

Ashburn VA 20147

<http://selab.janelia.org/>

December 18, 2008

Summary: INFERNAL is a software package for building consensus RNA secondary structure profiles called covariance models (CMs), and using them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence and structure-based multiple sequence alignments.

Availability: Source code, documentation, and benchmark downloadable from <http://infernal.janelia.org>. Freely licensed under the GNU General Public License version 3 (GPLv3).

Contact: {nawrockie,kolbed,eddys}@janelia.hhmi.org

Introduction

When searching for homologous structural RNAs in sequence databases, it is desirable to score both primary sequence and RNA secondary structure conservation. Many tools for integrating and scoring RNA sequence and secondary structure have been developed. Some implement specialized rules for a specific RNA family [1–7], and others use pattern matching methods and expertly designed query patterns [8]. The most general approaches take as input any RNA (or RNA multiple alignment), and construct an appropriate statistical scoring system that allows quantitative ranking of putative homologs in a target sequence database [9–11]. Stochastic context-free grammars (SCFGs) provide a natural statistical framework for combining sequence and (non-pseudoknotted) secondary structure conservation information in a single consistent scoring system

[12–15].

Here, we announce the 1.0 release of *INFERNAL*, an implementation of a general SCFG-based approach for RNA database searches and multiple alignment. *INFERNAL* builds consensus RNA profiles called *covariance models* (CMs), a special case of SCFGs designed for modeling RNA consensus sequence and structure. It uses CMs to search nucleic acid sequence databases for homologous RNAs, or to create new sequence and structure-based multiple sequence alignments. One use of *INFERNAL* is to annotate RNAs in genomes in conjunction with the *RFAM* database [16], which contains hundreds of RNA families. *RFAM* follows a seed profile strategy, in which a well-annotated “seed” alignment of each family is curated, and a CM built from that seed alignment is used to identify and align additional members of the family. *INFERNAL* has been in use since 2002, but 1.0 is the first version that we consider to be a reasonably complete production tool. It now includes E-value estimates for the statistical significance of database hits, and heuristic acceleration algorithms for both database searches and multiple RNA sequence alignment that allow *INFERNAL* to be deployed in a variety of real RNA analysis tasks with manageable (albeit high) computational requirements.

Usage

A CM is built from a multiple sequence alignment (or single RNA sequence) with consensus secondary structure annotation marking which positions of the alignment are single stranded and which are base paired. CMs assign position specific scores for the four possible residues at single stranded positions and the sixteen possible base pairs at paired positions, as well as position specific scores for insertions and deletions. These scores are log-odds scores derived from the observed counts of residues, base pairs, insertions and deletions in the input alignment, combined with prior information derived from structural ribosomal RNA alignments. Construction and parameterization of CMs have been described in more detail elsewhere [13, 17–19].

INFERNAL is composed of several programs that are used in combination to build models, search databases, and align putative homologs, following four basic steps:

1. Build a CM from an input alignment with *cmbuild*.

cmbuild takes as input a structural multiple RNA alignment in Stockholm format [17] and creates a CM file that is used by other *INFERNAL* programs.

2. Calibrate a CM for similarity search with *cmcalibrate*.

This step is optional and computationally expensive (Table 1), but is required to obtain E-values that estimate the statistical significance of each hit in a database search. *cmcalibrate* will also determine appropriate HMM filter thresholds for accelerating searches without an appreciable loss of sensitivity. Each model only needs to be calibrated once.

3. Search databases for putative homologs with *cmsearch*.

Given a CM file and a target database as input, *cmsearch* searches the target database for high scoring hits to the model and outputs alignments of each hit in a BLAST-like format augmented with structure annotation.

4. Align putative homologs to a CM with *cmalign*.

cmalign takes a CM file and a file of putative homologs, and aligns the full length sequences to the model, creating a structurally annotated multiple alignment in Stockholm format.

Performance

A published benchmark (independent of our lab) [20] and our own internal benchmark used for INFERNAL development [19] both find that INFERNAL and other CM based methods are the most sensitive and specific tools for structural RNA homology search among the several that were tested. Figure 1 shows updated results of our internal benchmark comparing INFERNAL 1.0 to the previous version (0.72) that was benchmarked in Freyhult et al. [20], and also to family-pairwise-search with BLASTN [21, 22]. The sensitivity and specificity of INFERNAL 1.0 have greatly improved relative to 0.72. There have been three relevant improvements in the implementation: a biased composition correction to the raw log-odds scores, the use of the full Inside log-likelihood scores (summed over all alignments) in place of CYK maximum likelihood alignment scores, and the introduction of approximate E-value estimates for the scores.

The benchmark dataset used in Figure 1 was constructed as follows. The sequences of the seed alignments of 503 Rfam (release 7) families were single linkage clustered by pairwise sequence identity, and separated into two clusters such that no sequence in one cluster is more than 60% identical to any sequence in the other. The larger of the two clusters was assigned as the query (preserving their original

Rfam alignment and structure annotation), and the sequences in the smaller cluster were assigned as true positives in a test set. We required a minimum of five sequences in the query alignment. 51 Rfam families met these criteria, yielding 450 test sequences which were embedded at random positions in a 10 Mb “pseudogenome”. Previously we generated the pseudogenome sequence from a uniform residue frequency distribution [19]. Here, we generated a more realistic pseudogenome sequence using a 15-state fully connected hidden Markov model (HMM) trained by Baum-Welch expectation maximization [15] on genome sequence data from a wide variety of species. Each of the 51 query alignments was used to build a CM and search the pseudogenome, a single list of all hits for all families were collected and ranked, and true and false hits were defined (as described in Nawrocki and Eddy [19]), producing the ROC curves in Figure 1.

INFERNAL searches require a large amount of compute time (Table 1). To alleviate this, INFERNAL 1.0 implements two rounds of filtering. When appropriate, the HMM filtering technique described by Weinberg and Ruzzo [23] is applied first with filter thresholds configured by *cmcalibrate* (occasionally a model with little primary sequence conservation cannot be usefully accelerated by a primary sequence based filter). The query-dependent banded (QDB) CYK search algorithm is used as a second filter with relatively tight bands ($\beta = 10^{-7}$) [19]. Any sequence fragments that survive the filters are searched a final time with the Inside algorithm (again using QDB, but with looser bands ($\beta = 10^{-15}$)). In our benchmark, the default filters accelerate similarity search by about 30-fold overall, while sacrificing a small amount of sensitivity (Figure 1). This makes version 1.0 substantially faster than 0.72. BLAST is still orders of magnitude faster, but significantly less sensitive than INFERNAL. Further acceleration remains a major goal of INFERNAL development.

The computational cost of CM alignment with the *cmalign* program has been a limitation of previous versions of INFERNAL. Version 1.0 now uses a constrained dynamic programming approach first developed by Brown [14] that uses sequence specific bands derived from a first-pass HMM alignment. This technique offers a dramatic speedup relative to unconstrained alignment, especially for large RNAs such as small and large subunit (SSU and LSU) ribosomal RNAs, which can now be aligned in roughly 1 and 3 seconds per sequence, respectively (Table 1), as opposed to 12 minutes and 3 hours in previous versions. We expect this to be particularly useful in applications where many large RNA sequences need to be aligned. One of the main ribosomal RNA databases, RDP, has recently adopted INFERNAL in its pipeline [24].

Discussion

INFERNAL is now a faster and more sensitive tool for RNA sequence analysis. Version 1.0's heuristic acceleration techniques make some important applications possible on a single desktop computer in less than an hour, such as searching a prokaryotic genome for a particular RNA family, or aligning a few thousand SSU rRNA sequences. Nonetheless, INFERNAL remains computationally expensive, and many problems of interest require the use of a cluster. The most expensive programs (*cmcalibrate*, *cmsearch*, and *cmalign*) are implemented in coarse-grained parallel MPI versions.

The complete INFERNAL version 1.0 software package, including documentation and ANSI C source code, may be downloaded from <http://infernal.janelia.org>. INFERNAL uses a GNU configure system and should be portable to any POSIX-compliant operating system, including Linux and Mac OS/X.

Acknowledgements

We thank Goran Ceric for his peerless skill in managing Janelia Farm's high performance computing resources.

Funding

INFERNAL development is supported by Howard Hughes Medical Institute. It has also been supported in the past by an NIH NHGRI Institutional Training Grant in Genomic Science (T32-HG000045) to EPN, an NSF Graduate Fellowship to DLK, and by NIH R01-HG01363 and a generous endowment from Alvin Goldfarb.

References

- [1] T. M. Lowe and S. R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25:955–964, 1997.
- [2] D. Laslett and B. Canback. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucl. Acids Res.*, 32:11–16, 2004.
- [3] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283:1168–1171, 1999.
- [4] P. Schattner, S. Barberan-Soler, and T. M. Lowe. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, 12:15–25, 2006.

- [5] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin. Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, 4:R42, 2003.
- [6] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel. Vertebrate microRNA genes. *Science.*, 299:1540, 2003.
- [7] M. Regalia, M. A. Rosenblad, and T. Samuelsson. Prediction of signal recognition particle RNA genes. *Nucl. Acids Res.*, 30:3368–3377, 2002.
- [8] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *NAR*, 29:4724–4735, 2001.
- [9] D. Gautheret and A. Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, 313:1003–1011, 2001.
- [10] S. Zhang, B. Haas, E. Eskin, and V. Bafna. Searching genomes for noncoding RNA using FastR. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2:366–379, 2005.
- [11] Z. Huang, Y. Wu, J. Robertson, L. Feng, R. Malmberg, and L. Cai. Fast and accurate search for non-coding rna pseudoknot structures in genomes. *Bioinformatics*, 24:2281–2287, 2008.
- [12] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, 22:5112–5120, 1994.
- [13] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22:2079–2088, 1994.
- [14] M. P. Brown. Small subunit ribosomal RNA modeling using stochastic context-free grammars. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:57–66, 2000.
- [15] R. Durbin, S. R. Eddy, A. Krogh, and G. J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK, 1998. ISBN 0521629713.
- [16] P. P. Gardner, 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. Rfam: Updates to the RNA families database. *NAR*, in press, 2009.
- [17] S. R. Eddy. The Infernal user's guide. [<http://infernal.janelia.org/>], 2003.
- [18] S. R. Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18, 2002.
- [19] E. P. Nawrocki and S. R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, 3:e56, 2007.
- [20] 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.
- [21] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acids Res.*, 25:3389–3402, 1997.
- [22] W. N. Grundy. Homology detection via family pairwise search. *J. Comput. Biol.*, 5:479–491, 1998.
- [23] Z. Weinberg and W. L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, 22:35–39, 2006.
- [24] J. R. Cole, 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. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. in press, 2009.
- [25] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Miller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics.*, 3:2, 2002.

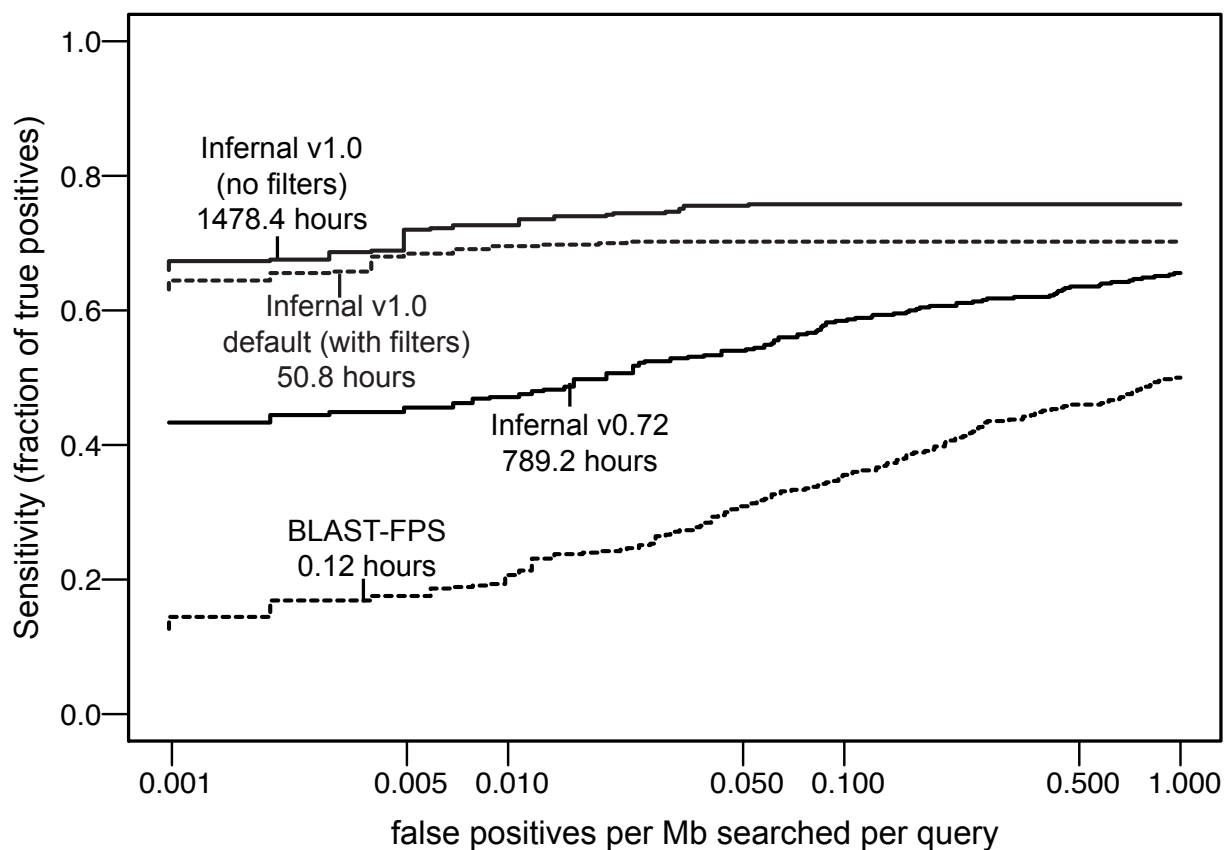


Figure 1: **ROC curves for the benchmark.** Plots are shown for the new INFERNAL 1.0 with and without filters, for the old INFERNAL 0.72, and for family-pairwise-searches (FPS) with BLASTN.

family	length	calibration (hours)	search (min/Mb)		alignment (sec/seq)
			no filters	w/filters	
tRNA	71	3.2h	23.5m	4.4m	0.01s
5S rRNA	119	4.4h	29.3m	1.1m	0.03s
Lysine riboswitch	183	8.9h	100.5m	1.3m	0.06s
SRP RNA	304	13.5h	166.0m	3.0m	0.18s
RNaseP	365	16.8h	205.6m	0.9m	0.19s
SSU rRNA	1466	84.5h	1265.5m	17.6m	1.10s
LSU rRNA	2909	169.7h	3907.6m	740.4m	3.34s

Table 1: Calibration, search, and alignment running times for seven known structural RNAs of various sizes. CPU times are measured on 3.0 GHz Intel Xeon processors with 8 GB RAM, running Red Hat AS4 Linux operating systems. All times were single execution threads except for SSU and LSU calibrations and searches which were run in parallel using MPI (OpenMPI) on 12 CPUs (times reported are actual times multiplied by 12). “Length” is the number of consensus positions (positions that contain gaps in fewer than 50% of the aligned sequences) in the input alignment. Randomly generated sequence of length 20 Mb (for filtered) and 2 Mb (for non-filtered) was used for the searches. Query alignments are all Rfam 9.0 seed alignments (RF00005, RF00001, RF00168, RF00017, RF00011) [16] except for SSU and LSU rRNA which were subsets of alignments at the Comparative RNA Website [25].