Sequence analysis

Infernal 1.0: inference of RNA alignments

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

HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA

Received on January 13, 2009; revised on March 11, 2009; accepted on March 14, 2009
Advance Access publication March 23, 2009

1 INTRODUCTION

When searching for homologous structural RNAs in sequence databases, it is desirable to score both primary sequence and secondary structure conservation. The most generally useful tools that integrate sequence and structure together are used on primary sequence or RNA multiple alignments. INFERNAL builds consensus RNA profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments. 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 includes hundreds of RNA families. INFERNAL builds consensus RNA profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments. INFERNAL builds consensus RNA profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments. INFERNAL builds consensus RNA profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments.

ABSTRACT

Summary: INFERNAL builds consensus RNA secondary structure profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments. INFERNAL is composed of several programs that are used in combination by following four basic steps:

1. Build a CM from a structural alignment with cmbuild.
2. Calibrate a CM for homology search with cmcalibrate.
3. Search databases for putative homologs with cmsearch.
4. Align putative homologs to a CM with cmalign.

The calibration step is optional and computationally expensive (4 h on a 3.0 GHz Intel Xeon for a CM of a typical RNA family of length 100 nt), but is required to obtain E-values that estimate the statistical significance of hits in a database search. cmcalibrate will also determine appropriate hidden Markov model (HMM) filter thresholds for accelerating searches without an appreciable loss of sensitivity. Each model only needs to be calibrated once.

2 USAGE

A CM is built from a Stockholm format 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 (Eddy, 2009). CMs assign position-specific scores for the four possible residues at single-stranded positions, the 16 possible base pairs at paired positions and 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. CM parameterization has been described in more detail elsewhere (Eddy, 2002, 2009; Eddy and Durbin, 1994; Klein and Eddy, 2003; Nawrocki and Eddy, 2007).

INFERNAL is composed of several programs that are used in combination by following four basic steps:

1. Build a CM from a structural alignment with cmbuild.
2. Calibrate a CM for homology search with cmcalibrate.
3. Search databases for putative homologs with cmsearch.
4. Align putative homologs to a CM with cmalign.

The calibration step is optional and computationally expensive (4 h on a 3.0 GHz Intel Xeon for a CM of a typical RNA family of length 100 nt), but is required to obtain E-values that estimate the statistical significance of hits in a database search. cmcalibrate will also determine appropriate hidden Markov model (HMM) filter thresholds for accelerating searches without an appreciable loss of sensitivity. Each model only needs to be calibrated once.

3 PERFORMANCE

A published benchmark (independent of our lab) (Freyhult et al., 2007) and our own internal benchmark used during development (Nawrocki and Eddy, 2007) both find that INFERNAL and other CM-based methods are the most sensitive and specific tools for structural RNA homology search among those tested. Figure 1 shows updated results of our internal benchmark comparing INFERNAL 1.0 with the previous version (0.72) that was benchmarked in Freyhult et al. (2007), and also to family-pairwise search with BLASTN (Altschul et al., 1997; Grundy, 1998). INFERNAL’s sensitivity and specificity have greatly improved, due to mainly three relevant
To alleviate this, INFERNAL 1.0 implements two all 51 family searches measured for single execution threads on 3.0 GHz Intel Xeon processors. The INFERNAL 1.0 times do not include time required for model calibration.

Improvements in the implementation (Eddy, 2009): a biased composition correction to the raw log-odds scores, the use of inside log likelihood scores (the summed score of all possible alignments of the target sequence) in place of CYK scores (the single maximum likelihood alignment score) and the introduction of approximate E-value estimates for the scores.

The benchmark dataset used in Figure 1 includes query alignments and test sequences from 51 Rfam (release 7) families [details in (Nawrocki and Eddy, 2007)]. No query sequence is longer than 6 km (2000 nucleotides). The test set consists of all possible alignments of the target sequence to a test sequence. The 450 total test sequences were embedded in a single list of all hits for all families.

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 on a separate processor. When appropriate, the HMM filtering technique was 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 as explained in (Eddy, 2009)]. The query-dependent banded (QDB) CYK maximum likelihood search algorithm is used as a second filter with relatively tight bands [\( \beta = 10^{-5} \)], the \( \beta \) parameter is the subtree length probability mass excluded by imposing the bands as explained in Nawrocki and Eddy (2007)]. 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 (Fig. 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 cmaligne has been a limitation of previous versions of INFERNAL. Version 1.0 now uses a constrained dynamic programming approach first developed by Brown (2000) 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, respectively) ribosomal RNAs, which can now be aligned in roughly 1 and 3 s per sequence, respectively, as opposed to 12 min and 3 h in previous versions. This acceleration has facilitated the adoption of INFERNAL by RDP, one of the main ribosomal RNA databases (Cole et al., 2009).

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 which divide the workload into independent units, each of which is run on a separate processor.

ACKNOWLEDGEMENTS

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

Funding: INFERNAL development is supported by the Howard Hughes Medical Institute. It has been supported in the past by an NIH NHGRI training grant (T32-HG000045) to E.P.N., an NSF Graduate Fellowship to D.L.K.; NIH R01-HG01363 and a generous endowment from Alvin Goldfarb.

Conflict of Interest: none declared.

REFERENCES


E.P.Nawrocki et al.
Infernal 1.0


Corrigendum

Infernal 1.0: inference of RNA alignments

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

We regret that the above paper was published with the wrong copyright line. The correct copyright line for this Open Access paper should have been:

© The Author 2009. Published by Oxford University Press. All rights reserved.