Rfam: updates to the RNA families database

Paul P. Gardner1,*, Jennifer Daub1, John G. Tate1, Eric P. Nawrocki2, Diana L. Kolbe2, Stinus Lindgreen3, Adam C. Wilkinson1, Robert D. Finn1, Sam Griffiths-Jones4, Sean R. Eddy2 and Alex Bateman1

1Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK, 2Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, USA, 3Center for Bioinformatics, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark and 4Faculty of Life Sciences, The University of Manchester, Manchester M13 9PL, UK

Received October 1, 2008; Revised October 5, 2008; Accepted October 6, 2008

ABSTRACT

Rfam is a collection of RNA sequence families, represented by multiple sequence alignments and covariance models (CMs). The primary aim of Rfam is to annotate new members of known RNA families on nucleotide sequences, particularly complete genomes, using sensitive BLAST filters in combination with CMs. A minority of families with a very broad taxonomic range (e.g. tRNA and rRNA) provide the majority of the sequence annotations, whilst the majority of Rfam families (e.g. snoRNAs and miRNAs) have a limited taxonomic range and provide a limited number of annotations. Recent improvements to the website, methodologies and data used by Rfam are discussed. Rfam is freely available on the Web at http://rfam.sanger.ac.uk/ and http://rfam.janelia.org/.

DATA AND METHODOLOGICAL IMPROVEMENTS

RFAMSEQ

All Rfam models are searched against an underlying nucleotide sequence database, known as RFAMSEQ, which is derived from the EMBL nucleotide sequence database (4). Prior to release 9.0, RFAMSEQ represented only the various species sections of EMBL. These sections contained only sequences that were considered to be of finished quality and excluded sequences from many important genomes. With release 9.0, RFAMSEQ has been expanded to include the whole genome shotgun (WGS) and environmental sequence (ENV) divisions. These changes have increased the number of sequences in RFAMSEQ by more than an order of magnitude (2,225,030 sequences in Rfam 8.0 versus 29,574,458 sequences in Rfam 9.0).

Sequence filters

In order to make it feasible to search more than 120 gigabases of sequence with hundreds of covariance models in a reasonable time, we use sequence-based filters to prune the search space prior to applying the more accurate and more computationally expensive CMs. One of the primary limitations of the Rfam annotation pipe-line has been the use of BLAST-based sequence filters, which are likely to compromise search sensitivity. In order to address this issue at least partially, NCBI-BLAST has been replaced with a WU-BLAST search, which has been tuned for high sensitivity and low sequence similarity. A benchmark of several homology search tools has shown WU-BLAST to be more accurate than the two methods on nucleotide data (5). Additionally, in order to make the BLAST filters more

*To whom correspondence should be addressed. Tel: +44 1223 494 983; Fax: +44 1223 494 919; Email: pg5@sanger.ac.uk

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
similar to profile HMMs, a sequence mask has been applied to each sequence in the alignment. Any nucleotide in an alignment column that has either a low frequency or is an insert relative to the majority of the rest of the sequences is ‘soft masked’ and not used for the BLAST word matches. These masked nucleotides do, however, still contribute to alignments that were seeded in the flanking regions. This approach has resulted in many fewer spurious hits with no detectable cost to sensitivity (data not shown), thus allowing \( E \)-value thresholds to be further relaxed. These observations together mean that the BLAST filters have been improved in terms of specificity and sensitivity.

**Iteration of families**

In order to improve the species and sequence depth of individual Rfam families, more than 370 families have been expanded by an ‘iteration’ process, in which some sequences in the FULL alignment are chosen for promotion to the SEED alignment. The sequences selected from FULL alignments for inclusion in the SEED must pass a series of stringent quality control requirements and be manually approved by a curator. The quality control steps include: ensuring that the sequence and secondary structure are consistent with the existing SEED sequences; the sequence identity with existing SEED sequences falls within 60–95%; the sequence is not truncated with respect to the SEED alignment. The curator also ensures that the new sequences make phylogenetic sense before allowing them to be incorporated into an updated SEED. An example of the utility of iteration is the snoRNA U103 SEED from Rfam 8.1 (accession: RF00188), which contained just three sequences and spanned two eutherian mammals (human and mouse). The SEED in Rfam 9.0 after iteration contains 42 sequences and spans Eutheria, Teleost (ray-finned fishes), Iguanidae, Monotremes, Marsupials, Placentals, Aves and Chondrichthyes (cartilaginous fishes).

Phylogenetic trees have been estimated for both the SEED and FULL alignments. For the majority of the alignments we produced the trees using an accurate maximum-likelihood approach, which included models of indels (6). However, the computational complexity of tree estimation meant that maximum-likelihood was not always possible and hence, where the number of sequences in the alignment was greater than 64, a neighbour-joining method was used instead (7). Large alignments and trees are problematic, both in terms of the computational cost of generation and the challenges of displaying them. Therefore, where the number of sequences in the alignment was greater than 1024, the highly similar sequences were filtered by sequence similarity, resulting in relatively sparse and easily presented trees that required comparatively little computing power to generate.

**PRESENTATION IMPROVEMENTS**

**Website redesign**

We are currently developing a new Rfam website, with the aim of improving the presentation of Rfam data and providing more and better tools for searching the data. The new site is now available from http://rfam.sanger.ac.uk/ and can be used to access Rfam 9.0 data. The new site lacks some features of the old site, but we aim to add all existing features and add many new ones over the coming months. Note that, at time of writing, the new website was available only at the Wellcome Trust Sanger Institute (http://rfam.sanger.ac.uk/). The two mirror sites will be updated to run the same website to coincide with the release of Rfam 9.1. The new site provides detailed overviews of Rfam families, including: a snapshot of the latest community-contributed annotation from Wikipedia (see below); tools for viewing and downloading the sequence alignments in various formats; representations of predicted secondary structure (see below); the taxonomic tree for the family; and phylogenetic trees for the SEED and FULL alignments.

Additionally, we provide several search tools in the new site. We currently support interactive searches, allowing a single RNA sequence to be searched against the whole Rfam database, and a batch search tool for searching multiple sequences against Rfam, the results of which are returned to the user via email. A new taxonomy search tool allows the user to find Rfam families that are specific to a given taxonomic level, or those found in a set of taxonomic levels that are specified by a complex, boolean query. For example, the query ‘Drosophila AND Caenorhabditis NOT Mammalia’ returns the two Rfam families (RF00047 and RF00533) that contain sequences from both drosophila and caenorhabditis but no sequences from any mammalian species.

**Structure graphics**

New graphical representations of secondary structures have been added to the Rfam website, based on software
from the Vienna RNA package (8). We now annotate several statistics directly on secondary structure diagrams, including sequence conservation, covariation, base-pair conservation and the maximum CM scores (Figure 2).

The sequence conservation metric uses a metric computed for each column in the alignment; this is the frequency of the most common nucleotide in each column (Figure 2A). The covariation metric is based upon that used by RNAalifold (9). For each base pair in the consensus structure and for each pair of sequences in the alignment, the difference in structurally consistent and inconsistent mutations is taken. Each mutation is weighted using a tree-weighting scheme (10) and this value is then normalized by the number of possible mutations (Figure 2B). The base-pair conservation metric is the fraction of canonical base pairs (Watson–Crick and G:U) in any two columns that correspond to a base pair in the consensus structure (Figure 2C). The maximum covariance model score and corresponding nucleotide/base pair is computed for each node in the CM. The resulting sequence, structure and bit-scores are used to produce a marked up secondary structure (Figure 2D).

**Wikipedia**

The Rfam website now draws textual annotation of RNA families directly from the scientific community, through the online encyclopedia Wikipedia. Any updates to relevant Wikipedia articles are downloaded on a nightly basis using the MediaWiki API, verified by members of the consortium and presented on the Rfam site (11). We consider the resulting articles to be a great improvement on the original static text because they are frequently updated, provide cross links to related articles and are generally considerably more comprehensive and informative than the original Rfam annotations that they replace.

**FUTURE CHALLENGES**

The rate of discovery of new RNA families is accelerating rapidly, facilitated by advancements in new sequencing technologies (12,13) and targeted computational screens (14–17). Keeping abreast of these updates whilst still ensuring the quality of alignments and secondary structures is an ongoing challenge for Rfam. We continue to evaluate new technologies and techniques as they emerge and will adopt new procedures for building and checking Rfam families as necessary.

We have been actively updating Rfam families and database crosslinks using more specialized RNA databases such as miRBase (18), IRESite (19), Pseudobase (20), snoRNABase (21), the plant snoRNA database (22), TransTerm (23) and the Yeast snoRNA database (24). As a result of these efforts, the next release of Rfam (version 9.1) will contain more than 700 entirely new families, bringing the total number of Rfam families to over 1300.

A new version of Infernal (v1.0) is now available (http://infernal.janelia.org) and we plan to use this
A further area of active research at Rfam is how best to distribute genome annotations. We plan to make annotations available in a variety of formats including the distributed annotation service (DAS) (27), General Feature Format (GFF) (http://song.sourceforge.net/gff3.shtml) and EMBL format, together with links to relevant genome browsers, e.g. ENSEMBL, UCSC and Genome Reviews.

ACKNOWLEDGEMENTS

We would like to thank Ivo Hofacker and Andreas Gruber for secondary structure graphics code, and Zasha Weinberg, Jeffrey Barrick, Chris Brown, Tom Jones for contributions to the database.

FUNDING

Wellcome Trust (to P.P.G., J.D., J.T., R.D.F. and A.B.); Howard Hughes Medical Institute (to E.P.N., D.L.K. and S.R.E); University of Manchester (to S.G.J.); University of Copenhagen (to S.L.). Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES


