

Rfam: updates to the RNA families database

Paul P. Gardner^{1,*}, 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/>.

INTRODUCTION

Rfam is a database of sequence families of structural RNAs, including non-coding RNA genes as well as *cis*-regulatory RNA elements. Rfam release 9.0 contains 603 families, each represented by a multiple sequence alignment of known and predicted representative members of the family, annotated with a consensus base-paired secondary structure. This so-called SEED alignment is used to build a covariance model (CM) with the Infernal software (1). Each Rfam covariance model is searched against a nucleotide sequence database, producing a list of putative hits. Matches that score above a curated threshold are then aligned to the CM to produce a so-called FULL alignment. This process is outlined diagrammatically in Figure 1. The Rfam database was developed as a generic approach to the annotation of structured RNA families on genomic sequences (2,3), but it has been widely used as

a source of reliable alignments and structures for the purposes of training and benchmarking RNA sequence and secondary structure analysis software.

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 the more accurate of 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

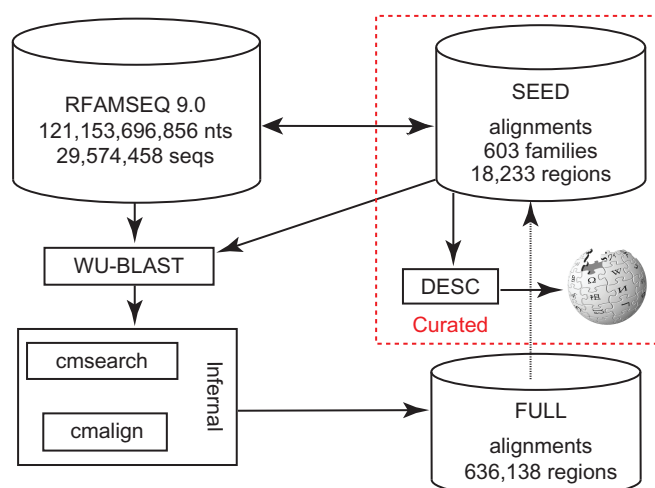


Figure 1. An outline of the Rfam 9.0 databases and methods. RFAMSEQ is drawn from EMBL excluding only the EST, synthetic and patented divisions. There are 603 Rfam families in release 9.0, which are used to scan RFAMSEQ for homologues using first WU-BLAST filters followed by the more accurate CM-based methods cmsearch and cmalign. This results in 603 FULL alignments annotating 636 138 regions.

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

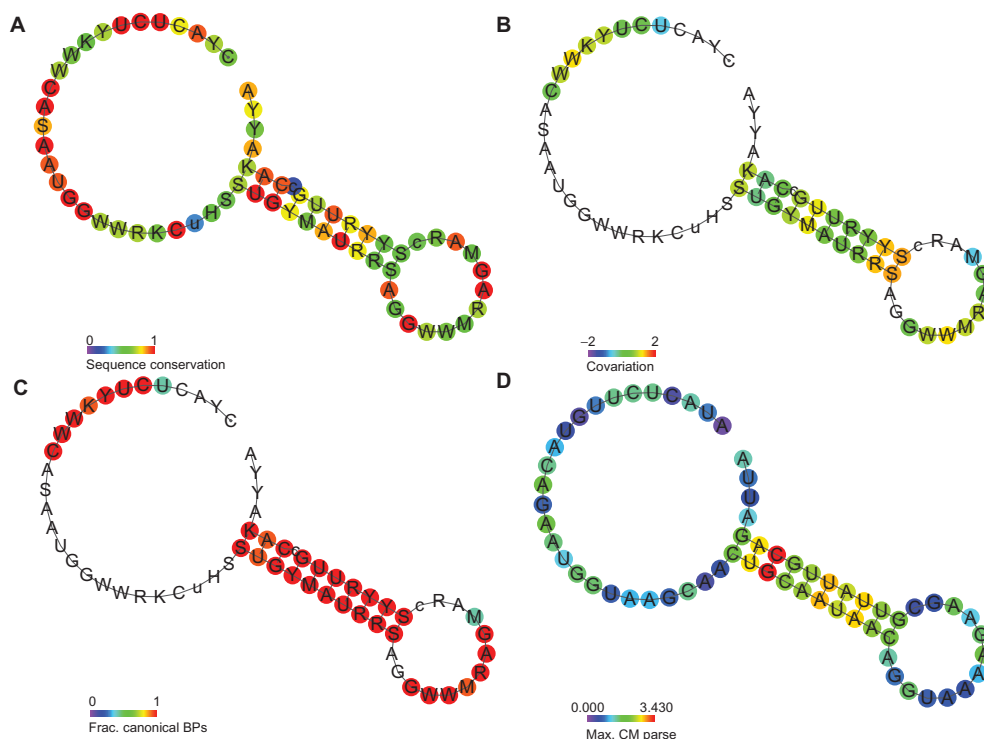


Figure 2. An example of the new secondary markups used by Rfam. The coronavirus 3'-UTR pseudoknot is shown (Rfam Accession RF00165). We display coloured markups of sequence conservation (A), covariation (B), base-pair conservation also known as the fraction of canonical base pairs (C) and CM scores (D).

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



Figure 3. An example of how PDB structures are displayed in Rfam. In this case, the structure 1l ng, containing the SRP19-7S.S RNA Complex from *M. jannaschii*, is rendered as cartoons using Jmol. Protein regions are coloured using the following scheme: beta-sheets (yellow), helices (magenta) and unstructured regions (white). RNA bases that match the Rfam model are coloured according to the key given in the web page (not shown here). In this structure, green represents a match to the eukaryotic SRP model, whereas those unmatched bases are coloured orange.

latest version to prepare the next major release of Rfam. Testing suggests that, compared with the version used for Rfam 9 (v0.72), v1.0 is faster and slightly more sensitive, whilst introducing for the first time *E*-values for hits returned from database searches. Although the speed increase will not be sufficient to obviate the need for BLAST filters in the Rfam production pipeline, this remains a major goal for Infernal development. Importantly, Infernal v1.0 is not compatible with the Rfam 9 CM files. Rfam/Infernal users may wish to generate new CMs from Rfam 9 SEED or FULL alignment files.

We have mapped a subset of three-dimensional RNA structures found in the Protein DataBank (PDB) (25) (primarily SRP and ribosomal RNAs) to corresponding sequences in Rfam. In an initial feasibility study, we have demonstrated that RNA sequences can be retrieved from PDB files and mapped to Rfam sequences reliably. The mapping is currently performed using BLAT (26) to detect local regions of high similarity with high specificity. The positions of matches to Rfam entries are transferred to the PDB sequences, allowing us to colour three-dimensional structures as in Figure 3. We intend to roll-out this mapping across all Rfam families and PDB entries using both local similarities and global matches to Rfam models. This sequence-to-structure mapping will allow us to use determined tertiary structures to calculate secondary structure as a quality control for existing families, and catalogue interactions between RNA–RNA and RNA–protein families.

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

- Nawrocki, E.P. and Eddy, S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, **3**, e56.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
- Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., Bhattacharyya, S., Bonfield, J., Bower, L., Browne, P. *et al.* (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **36**, D5–D12.
- Freyhult, E.K., Bollback, J.P. and Gardner, P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.
- Rivas, E. and Eddy, S.R. (2008) Probabilistic phylogenetic inference with insertions and deletions. *PLoS Comput. Biol.*, **4**, e1000172.
- Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Gruber, A.R., Neuböck, R., Hofacker, I.L. and Washietl, S. (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.*, **35**, W335–W338.
- Hofacker, I.L., Fekete, M. and Stadler, P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
- Gerstein, M., Sonnhammer, E.L. and Chothia, C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.
- Daub, J., Gardner, P.P., Tate, J., Ramsdell, D., Manske, M., Scott, W.G., Weinberg, Z., Griffiths-Jones, S. and Bateman, A. (2008) The RNA WikiProject: community annotation of RNA families. *RNA*, **14**, 12.
- Sittka, A., Lucchini, S., Papenfuss, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C. and Vogel, J. (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.*, **4**, e1000163.

13. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
14. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform.*, **2**, 8.
15. Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–30.
16. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
17. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
18. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
19. Mokrejs,M., Vopalensky,V., Kolenaty,O., Masek,T., Feketova,Z., Sekyrova,P., Skaloudova,B., Kriz,V. and Pospisek,M. (2006) IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Res.*, **34**, D125–D130.
20. van Batenburg,F.H., Gulyaev,A.P. and Pleij,C.W. (2001) PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res.*, **29**, 194–195.
21. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
22. Brown,J.W., Echeverria,M., Qu,L.H., Lowe,T.M., Bachellerie,J.P., Hüttenhofer,A., Kastenmayer,J.P., Green,P.J., Shaw,P. and Marshall,D.F. (2003) Plant snoRNA database. *Nucleic Acids Res.*, **31**, 432–435.
23. Jacobs,G.H., Stockwell,P.A., Tate,W.P. and Brown,C.M. (2006) Transterm-extended search facilities and improved integration with other databases. *Nucleic Acids Res.*, **34**, D37–D40.
24. Piekna-Przybylska,D., Decatur,W.A. and Fournier,M.J. (2007) New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA*, **13**, 305–312.
25. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
26. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
27. Jenkinson,A.M., Albrecht,M., Birney,E., Blankenburg,H., Down,T., Finn,R.D., Hermjakob,H., Hubbard,T.J., Jimenez,R.C., Jones,P. *et al.* (2008) Integrating biological data – the Distributed Annotation System. *BMC Bioinform.*, **9** (Suppl 8), S3.