

Rfam: annotating non-coding RNAs in complete genomes

Sam Griffiths-Jones*, Simon Moxon, Mhairi Marshall, Ajay Khanna¹, Sean R. Eddy¹
and Alex Bateman

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton,
CAMBS, CB10 1SA, UK

¹Howard Hughes Medical Institute and Department of Genetics, Washington
University School of Medicine, Saint Louis, Missouri 63108, USA.

* Corresponding author

Email: sgj@sanger.ac.uk

Tel: +44 1223 834244

Fax: +44 1223 494919

Abstract

Rfam is a comprehensive collection of non-coding RNA families, represented by multiple sequence alignments and profile stochastic context free grammars. Rfam aims to facilitate the identification and classification of new members of known sequence families, and distributes annotation of ncRNAs in over 200 complete genome sequences. The data provide the first glimpses of conservation of multiple ncRNA families across a wide taxonomic range. A small number of large families are essential in all three kingdoms of life, with large numbers of smaller families specific to certain taxa. Recent improvements in the database are discussed, together with challenges for the future. Rfam is available on the web at <http://www.sanger.ac.uk/Software/Rfam/> and <http://rfam.wustl.edu/>.

Introduction

Non-coding RNA (ncRNA) genes produce a functional RNA product instead of a translated protein. These products are components of some of the most important cellular machines, such as the ribosome (ribosomal RNAs), the spliceosome (U1, U2, U4, U5 and U6 RNAs) and the telomerase (telomerase RNA). The known repertoire of ncRNA cellular functions is expanding rapidly. Small nucleolar RNAs (snoRNAs) guide essential modifications of ribosomal and spliceosomal RNAs (reviewed in [1]). Ribozymes catalyse a range of reactions, such as self-cleavage of hepatitis delta virus transcripts, and 5' maturation of tRNAs by the ubiquitous RNase P. A class of small RNAs almost unknown before 2000, the microRNAs (miRNAs), are found to be involved in regulation of ever more processes in higher eukaryotes – including development, cell death, and fat metabolism – by repressing the translation of mRNA targets (reviewed in [2]). Similar mRNA-binding regulatory roles in bacteria are fulfilled by distinct families of small RNAs (reviewed in [3]).

Like protein-coding genes, ncRNA sequences can be grouped into families, and much can be learnt about structure and function from multiple sequence alignments of such families. Unlike proteins, ncRNAs often conserve a base-paired secondary structure with low primary sequence similarity. The combined secondary structure and primary sequence profile of a multiple sequence alignment of ncRNAs can be captured by statistical models, called profile stochastic context free grammars (SCFGs), analogous to profile hidden Markov models (HMMs) of protein alignments.

Rfam is a database of ncRNA families represented by multiple sequence alignments and profile SCFGs, available via the web at <http://www.sanger.ac.uk/Software/Rfam/> and <http://rfam.wustl.edu/>. All data are also available for download, local installation and sequence searching using the INFERNAL software package (<http://infernal.wustl.edu/>) [4]. The Rfam/INFERNAL model is much like the Pfam/HMMER system [5], extended to deal with RNA secondary structure consensus, and has been discussed previously [6]. Here we concentrate on recent

improvements, and discuss challenges that we expect to address through future development.

Recent developments

The database has grown dramatically over the past two years: from 25 families annotating around 55000 regions in the nucleotide sequences databases in release 1.0, to 379 families annotating over 280000 regions in release 6.1. This growth is partly due to a significant increase in scope. The evolution of some large gene families such as miRNAs and snoRNAs are constrained partially by inter-molecular base-pairing, and thus do not conserve significant sequence or primary structure. Whilst we cannot therefore represent all C/D box snoRNAs, or all miRNAs, with a single alignment and model, subfamilies are conserved and are now well represented in the database. Rfam also now includes not only *bona fide* ncRNA genes, but also structured regions of mRNA transcripts. These fall into two broad classes: self-splicing introns, and cis-regulatory elements in untranslated regions (UTRs). The latter can be used as detectors for a wide range of environmental conditions (for example, bacterial riboswitches bind a range of metabolites (reviewed in [7,8]), and the 5' UTR of the PrfA acts as a temperature-dependent switch [9]) to regulate message stability or translational efficiency.

This increased scope has led to the introduction of a limited type ontology, with the top-level types representing the three classes of structured RNA discussed above – ‘Gene’, ‘Intron’ and ‘Cis-reg’. The database currently contains 308 gene families, 69 cis-regulatory elements, and 2 self-splicing introns. The type field provides one of the primary entry points for family browsing and searching, enabling the user to quickly identify all snoRNA gene families for example, or to find all riboswitches in the database.

One of the primary uses of the Rfam database is to search for homologues of known RNAs in a query sequence, including a complete genome. Indeed, the profile SCFG library has been used to annotate a number of newly sequenced genomes (e.g. *C.*

briggsae [10], chicken [11] and *Erwinia caratova* [12]). In addition, we calculate hits in over 200 complete genomes and chromosomes. These data are available through the web interface and are discussed briefly in the following section.

Non-coding RNAs in complete genomes

Rfam makes available annotation of over 13400 candidate ncRNA genes (plus 172 self-splicing introns and 1285 cis-regulatory RNA elements) belonging to 172 families in 224 completed chromosomes and genomes. The average bacterial genome contains over 80 hits, dominated by the number of tRNAs. 170 regions are annotated in *E. coli*, in which most experimental validation of computationally predicted ncRNAs has been carried out. Rfam annotated regions in *Bacillus* genomes (*B. anthracis* shown in **figure 1**) include a number of recently described riboswitches [7,8].

These data provide the first comprehensive view of the distribution of ncRNAs in the three kingdoms of life. There are a small number of very large families representing some of the best-understood RNAs. **Figure 2** shows that these few large families are the only RNAs that are ubiquitous between all three domains of life – only the essential translation components, transfer RNA and ribosomal RNA, together with RNase P (tRNA maturation) and SRP RNA (protein export) are found in eukaryotes, bacteria and archaea. It is tempting to believe that very few families will be added to the catalogue of universally conserved RNAs. However, it is clear that members of some families are so highly divergent so as to be computationally almost unrecognisable. For example, although most eukaryotes would be expected to have a telomerase RNA, current computational techniques are unable to identify homologues in even well-studied model organisms such as *C. elegans*.

Only snoRNAs are found in eukaryotes and archaea and not in bacteria, but no RNA families have yet been identified that are common to bacteria and archaea but not eukaryotes, or eukaryotes and bacteria but not archaea. The vast majority of Rfam families are small, and are often specific to one taxonomic group, and in some cases

to one organism, suggesting relatively recent evolution of function or divergence beyond our ability to recognise homologues. Many novel bacterial ncRNAs have been identified by a number of recent computational screens in *E. coli* (reviewed in [13]), but comparatively few have been experimentally verified. Rfam contains more than 30 ncRNA families based on the verified genes. Few large-scale studies have been conducted in archaea or eukaryotes, and it is clear that such efforts will identify many more small families.

Future challenges

Profile SCFG searches are computationally expensive. Rfam at present uses a BLAST-based heuristic [14] as described previously [6], reducing the search space with an inevitable sensitivity cost. This allows us to search a 5 megabase bacterial genome against the entire Rfam library in around 24 hours. Annotation of large eukaryotic genomes is just feasible using this approach. Recent advances allow the speed of profile SCFGs to be increased by a factor of around 100 for most families, and provably do not reduce the sensitivity of the full SCFG search [15]. Work is ongoing to incorporate such algorithms into the Rfam/INFERNAL approach. We also recognise that the current approach is restricted to RNAs with defined secondary structures, precluding inclusion of important families of essentially unstructured RNAs like XIST (X-Inactive Specific Transcript), RoX (RNA on X), and IPW (Imprinted in Prader-Willi). We plan to evaluate how the use of profile HMMs may allow detection of homologues of such sequences.

Perhaps the biggest challenge for annotation of higher eukaryotic genomes is the problem of ncRNA-derived pseudogenes and repeats. For example, the B2 repeat in mouse is evolutionarily related to an Ala-tRNA, and Alu repeats in human derive from SRP RNA. Over 10% of the draft human genome sequence is made up of 1.1 million Alu sequences [16], and there are over 350000 B2 repeat sequences in mouse [17]. The human genome also contains over 1000 sequences that are closely related to U6 spliceosomal RNA, yet sensible estimates of the U6 gene count suggest that fewer than 50 are functional. Other problem families include the pol III transcribed Y

and 7SK RNAs. Distinguishing the functional copies from the large numbers of pseudogenes is an unsolved problem and presents a significant challenge to RNA computational biologists.

It seems likely that computational and experimental screens will continue to identify numerous novel ncRNAs. Most of these genes are predicted to fall into small families with narrow taxonomic ranges. In contrast, we believe that very few universally conserved RNAs will be found, and the large, well-studied and ubiquitous families will continue to make up the large majority of ncRNAs in a single genome. Rfam will continue to translate novel discoveries of ncRNA genes into alignments and models that are immediately useful for genome annotation and phylogenetic analysis.

Acknowledgements

We thank all those who have contributed data and annotation and developed tools and algorithms for ncRNA detection, alignment and structure prediction. Work at the Sanger Institute is funded by the Wellcome Trust. AK and SRE are supported by the Howard Hughes Medical Institute, the NIH National Human Genome Research Institute, and Alvin Goldfarb.

References

1. Bachellerie JP, Cavaille J and Huttenhofer A (2002). The expanding snoRNA world. *Biochimie*, **84**, 775-790.
2. Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.
3. Storz G, Opdyke JA and Zhang A (2004). Controlling mRNA stability and translation with small, noncoding RNAs. *Curr. Opin. Microbiol.*, **7**, 140-144.

4. Eddy SR (2002). A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
5. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy SR (2003). The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138-D141.
6. Griffiths-Jones S, Bateman A, Marshall M, Khanna A and Eddy SR (2003). Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439-441.
7. Mandal M and Breaker RR (2004). Gene regulation by riboswitches. *Nat. Rev. Mol. Cell. Biol.*, **5**, 451-463.
8. Vitreschak AG, Rodionov DA, Mironov AA and Gelfand MS (2004). Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.*, **20**, 44-50.
9. Johansson J, Mandin P, Renzoni A, Chiaruttini C, Springer M and Cossart P (2002). An RNA thermosensor controls expression of virulence genes in *Listeria monocytogenes*. *Cell*, **110**, 551-561.
10. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R and Waterston RH (2003). The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. *PLoS Biol.*, **1**, E45.
11. International Chicken Genome Sequencing Consortium. Sequencing and comparative analysis of the chicken genome (2004). *Nature*, submitted.

12. Bell KS, Sebahia M, Pritchard L, Holden MT, Hyman LJ, Holeva MC, Thomson NR, Bentley SD, Churcher LJ, Mungall K, Atkin R, Bason N, Brooks K, Chillingworth T, Clark K, Doggett J, Fraser A, Hance Z, Hauser H, Jagels K, Moule S, Norbertczak H, Ormond D, Price C, Quail MA, Sanders M, Walker D, Whitehead S, Salmond GP, Birch PR, Parkhill J and Toth IK (2004). Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc. Natl. Acad. Sci. USA*, **101**, 11105-11110.
13. Hershberg R, Altuvia S and Margalit H (2003). A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.*, **31**, 1813-1820.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.
15. Weinberg Z and Ruzzo WL (2004). Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, **20**, I334-I341.
16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome (2001). *Nature*, **409**, 860-921.
17. Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.

Figures

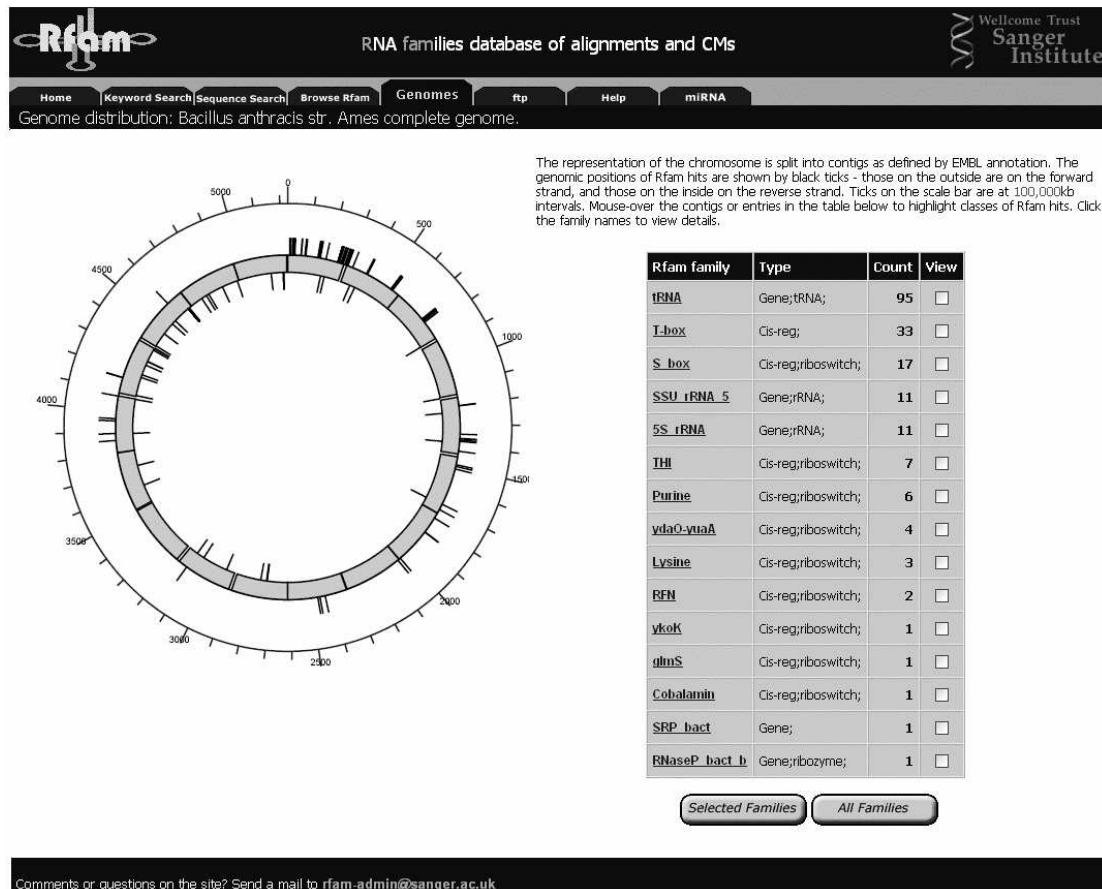


Figure 1: Rfam genome page for *Bacillus anthracis*. The table contains a summary of the number of members of each Rfam family in the genome, with the distribution of hits shown on the map.

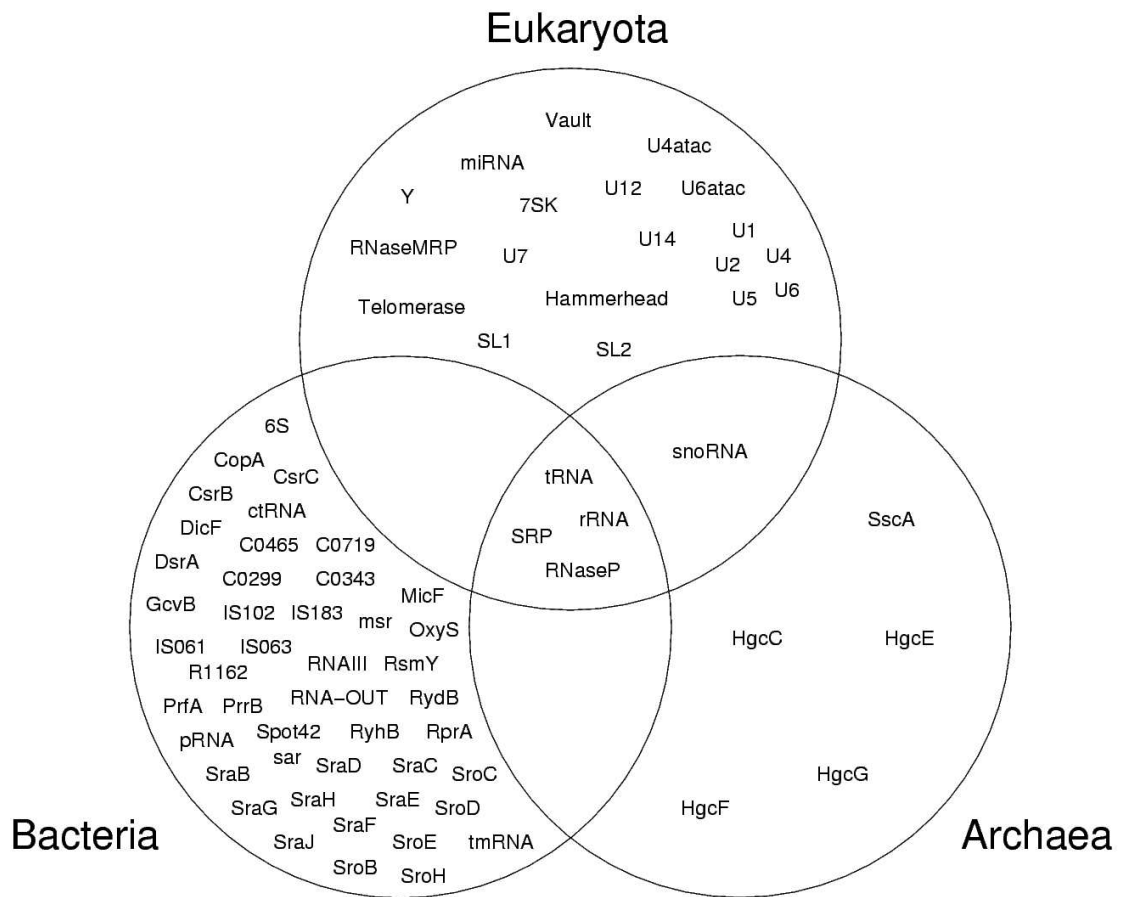


Figure 2: Taxonomic distribution of Rfam family members in the three kingdoms of life.