Published online 17 December 2014

doi: 10.1093/nar/gku1327

Annotating RNA motifs in sequences and alignments

Paul P. Gardner[1,2,*] and Hisham Eldai[1,2]

[1]School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand and [2]Biomolecular Interaction Centre, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

Received March 26, 2014; Revised November 30, 2014; Accepted December 05, 2014

ABSTRACT

RNA performs a diverse array of important functions across all cellular life. These functions include important roles in translation, building translational machinery and maturing messenger RNA. More recent discoveries include the miRNAs and bacterial sRNAs that regulate gene expression, the thermosensors, riboswitches and other cis-regulatory elements that help prokaryotes sense their environment and eukaryotic piRNAs that suppress transposition. However, there can be a long period between the initial discovery of a RNA and determining its function. We present a bioinformatic approach to characterize RNA motifs, which are critical components of many RNA structure–function relationships. These motifs can, in some instances, provide researchers with functional hypotheses for uncharacterized RNAs. Moreover, we introduce a new profile-based database of RNA motifs—RMfam—and illustrate some applications for investigating the evolution and functional characterization of RNA. All the data and scripts associated with this work are available from: https://github.com/ppgardne/RMfam.

INTRODUCTION

Characterizing functional RNAs is an extraordinarily difficult task. Even highly transcribed RNAs from model organisms have remained uncharacterized for decades after their discovery. A specific example is the 6S sRNA, which was discovered in 1971. The 6S sRNA is conserved across Bacteria and is highly expressed in stationary-phase cells (1,2). But the role of 6S as a regulator of RNA polymerase remained an enigma for almost three decades (3). Likewise, Y RNA, which was discovered in 1981, is broadly conserved across metazoans and is highly expressed (4). It took two and a half decades before Y RNAs were shown to be essential for the initiation of DNA replication (5). However, the mechanism for Y RNA function still remains unclear. These and similar examples show that it is remarkably difficult to functionally characterize RNAs, even after decades of work.

A new generation of tools for RNA discovery is now available thanks to powerful new sequencing technologies. Entire transcriptomes from species at different life stages, tissue types and conditions can be studied with RNA-seq (6–8). The total complement of RNA structures encoded in transcriptomes is also accessible with SHAPE-seq (9) and functional regions of entire genomes of bacteria can be probed with techniques like TraDIS and Tn-seq (10,11). The data obtained by these tools are unearthing novel RNAs at an unprecedented rate, many of which are evolutionarily conserved, highly expressed, activated under specific conditions, essential and fold into conserved secondary structures. Annotation efforts such as those by the Rfam consortium (12–14) are useful. However, many RNAs are not found in this database and many that have been curated remain uncharacterized (8). To make sense of the volumes of transcriptome data that is now being generated, annotating this data and functionally characterizing the cohort of RNAs of Unknown Function (RUFs) is critical. A complication for such work is that evolutionary turnover, as well as sequence variation can be high for ncRNAs (15,16). Consequently, homology searches and other sequence-alignment-based analyses can be very challenging.

For the purposes of this work we define a RNA motif as a functional RNA structure that recurs within or across different RNA families. A motif may be characterized by a blend of primary, secondary and tertiary structural features. The motifs that have been characterized to date are involved in a diverse number of functions, including increasing structural stability (e.g. the GNRA tetraloop (17–19)), facilitating interactions with other biomolecules (e.g. the CsrA-binding motif (20–22)), specifying sub-cellular localization (e.g. the SRP S-domain (23)) and coordinating gene regulatory signals (e.g. the HuR mRNA binding motif (24)).

A number of publications detail bioinformatic methods for the *de novo* discovery of RNA secondary structure motifs from RNA primary sequences (25,26). There are also tools that can screen predicted RNA secondary structures (27) and RNA tertiary structures (28) for shared structural features. The knowledge-based approaches for the annotation of RNA motifs include sequence and structure descriptors (Eddy,S., unpublished data,29), primary and secondary structure-based profile methods for specific motifs, e.g. (30,31), and methods that combine primary, secondary

[*]To whom correspondence should be addressed. Tel: +64 3 364 2987; Fax: +64 3 364 2590; Email: paul.gardner@canterbury.ac.nz

and tertiary data (32). We complement these approaches by introducing a resource that identifies a range of previously characterized RNA motifs in RNA sequences and alignments using profile hidden Markov models (HMMs) (33–35) and covariance models (CMs) (35–37).

We present 34 alignments, consensus structures and corresponding probabilistic models of published RNA motifs. We call this resource RMfam, or RNA Motif Families (all associated data and computer code are freely available from our repository hosted on GitHub: http://github.com/ppgardne/RMfam). These have been used to predict ~1900 conserved motifs in the Rfam (v11.0) alignments of RNA families (these are available in Rfam (v12.0) (14)); many of which are confirmed in the published literature. Finally, we show examples of the applicability of our approach for studying RNA function, evolution and alignment curation.

MATERIALS AND METHODS

The distinction between Rfam and RMfam

The Rfam database collects and curates 'seed alignments' of RNA families. These are non-coding RNAs, cis-regulatory elements and self-splicing introns. The alignments are manually constructed and annotated with consensus secondary structures, and used to seed probabilities for CMs for each family. The Rfam CMs are widely used for genome annotation projects to identify RNA loci (e.g. (38)). A requirement before each family can pass Rfam quality-control is that it is specific. In other words, there exists a bit score threshold for each CM that distinguishes between sequence matches that are related to the family and obvious false-positive matches. Consequently, many RNA motifs are not included in Rfam as they lack the required specificity (12–13,39–41). However, the Rfam 12.0 (2014) release of the Rfam database includes RMfam annotations for the first time (14).

What is an RNA motif?

As described in the Introduction, we define RNA motifs as functional RNA structures that recur within or across different RNA families. These are a blend of primary, secondary or tertiary structure. Fortunately, for the purposes of this work, the majority of internal RNA contacts are local (i.e. within 100 nucleotides) (42), therefore, the local probabilistic models (described below) can be used to capture the bulk of the information.

An example RNA motif is the GNRA tetraloop (see Figure 3). This RNA motif is one of the most prevalent hairpins found in a number of RNA families, including rRNA, RNase P, a variety of riboswitches, self-splicing RNAs and many others. It is characterized by a hairpin loop that contains a 4 nucluetide sequence matching the pattern 'GNRA'. The most prevalent of these are GAAA, GUGA, GCAA and GAGA. The terminal 'G' and 'A' are frequently involved in a non-canonical base pair (18), however, the loop may also be involved in some long-range tertiary interactions (17) which can be mimicked by a range of alternative conformations (19). Therefore, the GNRA tetraloop meets our criteria for a RNA motif as it is a functional structure that recurs in multiple families.

Accurate local probabilistic methods for annotating structured RNAs on DNA sequences called HMMs and CMs are now available (33–37,43). From a given alignment, probabilistic models of conserved sequence (HMMs) and conserved sequence plus secondary structure (CMs) can be built and used to filter large numbers of sequences for candidate homologous and/or analogous regions (44). CMs cater to the characteristics of RNA sequence evolution that are imposed by base pairing (i.e. variation tends to preserve base pairing), the result is that the accuracy of CMs is greater than alternative approaches (45). The computational speed of CMs has tended to be poor, however, a lot of effort has been expended on improving the speed of the approach while maintaining the accuracy. The improvements include using HMMs as pre-filters to accelerate CMs, query-dependent banding and Dirichlet mixture priors (43–44,46–48).

RMfam sequences, structures and alignments were collated from a variety of heterogeneous and sometimes overlapping data repositories (12,24,28,31,49–56). Where possible we sourced data from publicly accessible RNA motif resources, these included the FR3D MotifLibrary (54), the models supplied with RMDetect (32), the comparative RNA website (52) and SCOR (51). We also used information from specialized resources, such as the k-turn structural database (49) and SRPDB (57), as well as generating our own alignments for motifs, such as the Shine–Dalgarno and Rho-independent terminators based upon the context of genome annotations (e.g. (31)). RNAFrabase was frequently consulted for RNA secondary structures derived from Protein Data Bank (PDB) structures (58,59). Finally, where necessary, we extracted sequences from publications. This was often a manual effort, involving manually transcribing sequences and structures from figures in published manuscripts. Where possible, these were mapped to nucleotide sequences derived from the PDB (downloaded June 2014) (60–62), the EMBL nucleotide archive (63) and Rfam (v11.0) (12,13). The provenance of each data set is stored in the corresponding Stockholm alignment. Each of these motifs were then passed through quality control steps, where the sensitivity and specificity of the resulting motif is assessed (see Figure 1 and Supplementary Figures S10–S43). If these failed (e.g. the CM cannot identify member sequences or the false-positive rate is extremely high), then the motif was not included in the database. Each motif is also assigned a curated score threshold. This threshold (in bits) provides a reasonable distinction between true and false matches.

A benchmark of motif annotations

In the following we briefly describe the benchmarks we have used to evaluate our motif annotations. The benchmarks are described in further detail and with more elaborate results in the Supplementary Results.

In order to determine the accuracy of our approach we ran a series of three benchmarks. These were evaluated on individual motifs (see Figure 1B and Supplementary Figures S10–S43), as well as on the collective RMfam results (see Figure 1A and Supplementary Figure S9). The first uses 'RMfam sequences' which are taken from the

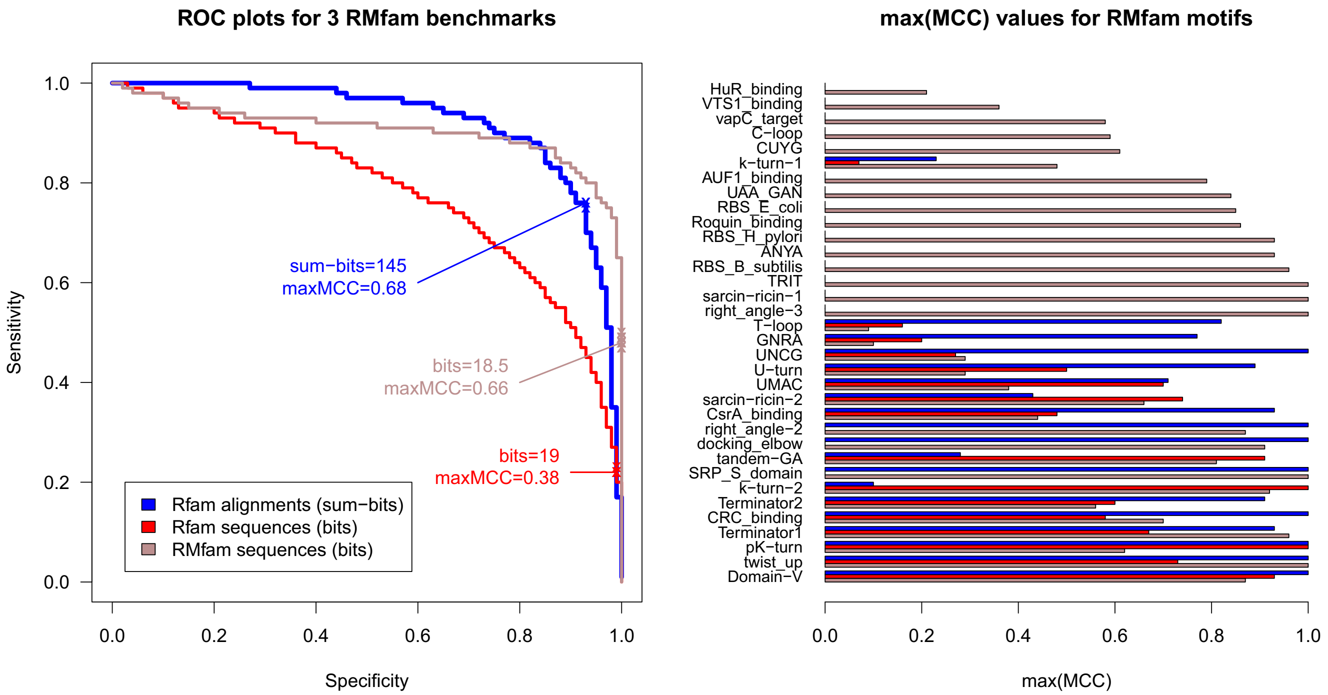

Figure 1. In the above plots we assess the accuracy of motif annotation and test whether annotating alignments instead of sequences improves the prediction accuracy. We have applied three different benchmarks (described in the text). In sub-figure (A) we show a ROC plot for pooled RMfam annotations. This plots the sensitivity versus specificity of all the motif annotations on sequences or alignments at different score thresholds. The '*x*'s illustrate where on the curve the maximum MCC is located, and the corresponding bit scores are indicated. In sub-figure (B) we illustrate the maximum MCC of CM annotation for each motif from the three different benchmarks. See the Supplementary Results for further details regarding these benchmarks.

seed alignments. Ten shuffled sequences, with identical di-nucleotide distributions, were generated for each RMfam seed sequence (64). Together these serve as positive and negative controls for our test.

We constructed two further tests based upon Rfam (v11.0) families. We identified Rfam families where there exists good evidence (primarily based upon reviewing the RNA literature) that a motif is conserved in the family of related sequences (Supplementary Table S1, also available at http://github.com/ppgardne/RMfam/benchmark/true_positives.txt). These serve as positive controls for two further tests. For the 'Rfam sequences' benchmark we randomly selected at least five sequences from each Rfam seed alignment (if fewer than five sequences were available, then all were included). We generated 10 shuffled versions of each sequence; all had an identical di-nucleotide distribution to the native sequence. These sequences were all annotated with RMfam motifs, their CM scores were recorded and used to evaluate the accuracy of the annotations. Finally, for a 'Rfam alignments' benchmark, we evaluated the accuracy of RMfam annotations in an alignment context. Each Rfam alignment was filtered, removing sequences more than 90% identical. The remaining sequences were annotated with RMfam CMs, retaining only those that cover more than 10% of the seed sequences and more than two Rfam seed sequences. The summary statistic we use for this final benchmark is a 'sum-bits' score, this is the sum of the bit scores for each match in all the sequences in a seed.

The accuracy metrics that we report here are the Matthew's correlation coefficient (MCC) (65), sensitivity and specificity. All of our secondary structure diagrams are illustrated with R2R (66).

The CMs built from RNA motifs tend to be short and contain little sequence information. In RMfam the mean sequence length is just 34.3 nucleotides and the mean number of base pairs is 10.9. Therefore, a scan of a large sequence database with these models will result in a number of false-positives. We propose that annotating sequence alignments of ncRNAs has the potential to improve the specificity of our predictions. This assumes that evolutionarily conserved motifs are more likely to be correct. In theory, this approach could be extended to genome alignments of, e.g. transcribed regions.

RESULTS

In this study we present 34 RMfam alignments and probabilistic models of published RNA motifs (all freely available from our repository hosted on GitHub: http://github.com/ppgardne/RMfam). These have been used to predict ∼2500 conserved motifs in the Rfam (v11.0) seed alignments; many of which are confirmed in the published literature. Furthermore, our permutation tests have shown that both the sensitivity and specificity of this approach is remarkably high given the short motifs we use (see Figure 1 and Supplementary Figures S9–S44).

Inference of RNA function with motifs

One of the most labour-intensive stages of RNA research is identifying the function of newly discovered RNAs. In order to illustrate the utility of RMfam for this task we show the matches between a model of the CsrA-binding site and two RNA families of unknown function, TwoAYGGAY and Bacillaceae-1 (Rfam IDs RF01731 and RF01690, see Figure 2). CsrA is a bacterial RNA-binding protein that regulates the translation and stability of mRNAs (20). It binds mRNAs carrying CsrA binding motifs, these physically occlude ribosome-binding sites. This binding can itself be regulated by competition between the mRNAs and highly expressed sRNAs that host numerous CsrA binding sites. However, this class of sRNA (CsrB, CsrC, RsmX, RsmY and RsmZ) has only been identified in Gammaproteobacteria (21,22). The TwoAYGGAY and Bacillaceae-1 families were initially discovered by a large-scale bioinformatic screen (67). Some further analysis identified two tandem-GAs in one of the stems that characterize the structure of TwoAYGGAY (32). Our motif-based analyses have identified strong matches between the TwoAYGGAY family, the reverse-complement of the Bacillaceae-1 family and the CsrA-binding motif. These provide a testable hypothesis for further validation, that there are also CsrA-binding sRNAs in Clostridia (TwoAYGGAY), and Bacillales and Lactobacillales (Bacillaceae-1). The validation of these predictions is a work in progress with our collaborators.

Evolution of RNA motifs

Non-coding RNAs are remarkably tolerant of genetic variation, as evident by the wide degree of sequence variation that can be found between evolutionarily related ncRNAs (16,68–70). However, structure frequently constrains the evolution of RNA sequences. That said, structures can also be dynamic. For example, motifs that confer structural stability can be exchanged over time, resulting in a rich and complex evolutionary history. This illustrates that studying the gain and loss of RNA motifs over evolutionary timescales can help characterize the dynamic evolution of RNA sequences and structures.

A good example of this is the Lysine riboswitch. This is a convenient example, that for illustrative purposes we will describe in further detail. As illustrated in Figure 3 many motifs may be exchanged, e.g. the U-turn motif with a k-turn in the P2 stem or the T-loop and the GNRA tetraloop in stem P4. Interestingly, the motif distributions are relatively clade-like, with closely related riboswitches more likely to share motifs, e.g. the GNRA tetraloop is found in Lysine riboswitches from the Pasteurellales and Vibrionales taxonomic groups. This type of annotation information is valuable for researchers investigating the structure and evolution of RNA families.

RNA motifs for curating RNA alignments

Another use of the results presented in this work is of importance for the curators of RNA alignments and sequences (12,71–72). Until now it has been difficult to analyse the evolutionary conservation of motifs in the context of an alignment, although some progress has been made when crystallographic data is available, e.g. the RNASTAR collection of structural RNA alignments (72). With the help of RMfam, malformed alignments can be detected and corrected where conserved RNA motifs are incorrectly aligned. We illustrate an example of this for the Rfam (v11.0) 5S rRNA alignment that contains a misaligned, yet highly conserved sarcin-ricin motif (see Supplementary Figure S45), and for the Rfam RsmY alignment, which is a CsrA-binding sRNA. The RsmY alignment has a misannotated consensus structure that does not include a further CsrA-binding motif, which are hairpin motifs that host a 'GGA' sequence in the loop (see Supplementary Figure S46). These motifs generally occur in pairs, as CsrA is a homodimeric protein, with each half of the protein binding a RNA motif (73,74).

DISCUSSION AND CONCLUSION

The chief motivation for this work is to functionally characterize novel ncRNAs. Our vision for the RMfam resource is to annotate RUFs (e.g. (8)). These motif annotations will help develop further functional hypotheses and accelerate experimental characterization.

In this work, we have shown that RMfam is surprisingly accurate. Despite the fact that the average RMfam motif consists of just 34.3 nucleotides and 10.9 base pairs, we show that the CMs are specific enough to distinguish between motif-hosting sequences and negative control sequences (see Figure 1 and Supplementary Figures S10–S43). Our approach shows improved performance when evolutionary information encoded in Rfam sequence alignments is incorporated into the predictions. We hypothesize that annotated genome alignments may be a useful source of motifs and we will investigate this idea further in future.

One apparent weakness of employing CMs and HMMs for motif annotation is that the 5′ and 3′ halves of internal-loop motifs (e.g. the k-turns, sarcin-ricin and tandem-GA loops) can in theory be a large distance apart, and therefore outside the QDB window covered by CMs (47). In practice, the distance between nucleotides involved in intramolecular contacts rarely exceeds 100 nucleotides (42), therefore, the majority of these motifs are still captured by CMs and HMMs.

As a discovery tool the RMfam resource has already made some useful predictions. We have predicted the existence of two new CsrA-binding ncRNAs, potentially the first of this class of regulatory molecules to be found outside of the Gammaproteobacteria. However, further work needs to be carried out to validate this claim.

Future work and potential applications

We have identified some future developments and applications for the RMfam resource. We plan to continue developing the accuracy of the motif annotation tools as well as increase the access to RMfam annotations via other databases, such as Rfam (v12.0) (14), and expand the number of motifs included in RMfam. Furthermore, it may be possible to boost the accuracy of RNA secondary structure prediction tools by constraining these with predicted motifs. We elaborate further on these ideas below.

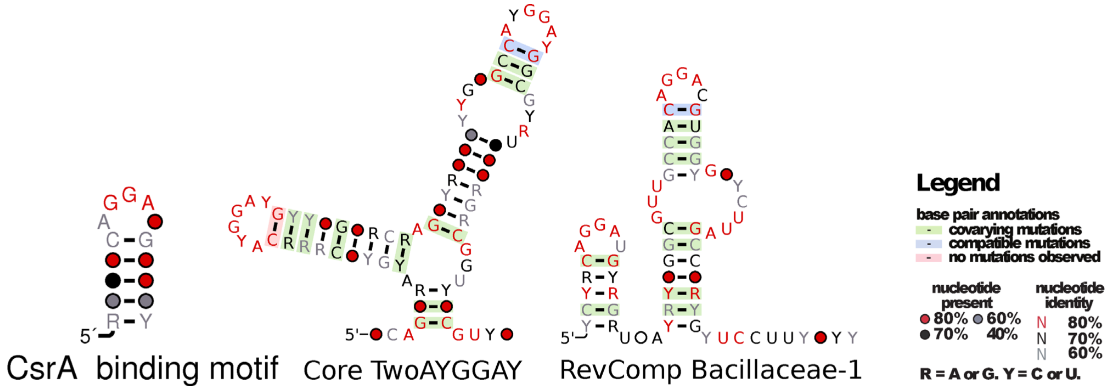

Figure 2. The secondary structures and sequence conservation of CsrA-binding motif and two new candidate CsrA binding sRNAs, TwoAYGGAY and Bacillaceae-1 family illustrated with R2R (66). These families each have two strong matches to the CsrA-binding motif, this new evidence provides a strong case that these RNAs regulate the activity of the regulatory protein, CsrA, by sequestering this nucleotide-binding protein. The 'core' of the TwoAYGGAY structure is shown, the Rfam (v11.0) model contains a further external stem that is not well conserved. Also, the reverse-complement (RevComp) of the Bacillaceae-1 is illustrated, this strand has the matches to the CsrA-binding motif and the original discoverers of this ncRNA are not confident of the strand (personal communication, Zasha Weinberg).

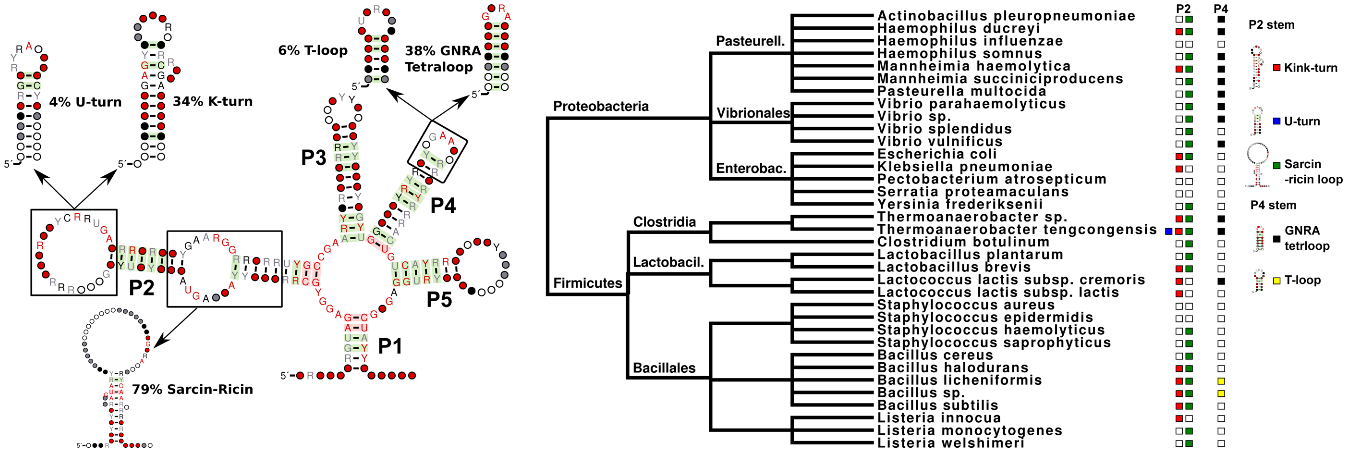

Figure 3. The Lysine riboswitch has substituted different motifs through its evolution. On the left is a representation of the consensus Lysine riboswitch secondary structure (66). This has been annotated with the most frequent motifs the RMfam annotates in the Lysine Rfam (v11.0) seed alignment, the percentage of seed sequences hosting each motif is also indicated. On the right is an annotated species taxonomy that illustrates the phylogenetic nature of the motif distributions. We have also annotated each tip with the motifs hosted in the P2 and P4 stems. The red, blue, green, black and yellow boxes illustrate kink-turn, U-turn, sarcin-ricin loop, GNRA tetraloop and the T-loop, respectively.

The Lysine riboswitch example raises the possibility that certain types of motif are preferentially exchanged during the evolution of ncRNAs. Do stable hairpin motifs, such as the GNRA and T-loops, replace each other more frequently than we expect by chance? This would blur the lines between our understanding of homologous and analogous structures (75). Another possibility is that certain motifs co-occur more frequently than we expect. For example, are k-turns more frequently closed by U-turns than we expect? If correct, these enrichments of favoured exchanges and co-occurances could be used to increase our confidence in motif annotations and can assist with the design of functional RNAs.

Typical RNA structure prediction methods to not incorporate information about RNA motifs. We propose that RMfam predictions can be used as constraints for existing RNA structure prediction software, thus improving the accuracy of structure prediction tools which can often be inaccurate (76). This approach is analogous to the fragment-library approach that is frequently used for tertiary structure prediction (77).

Another application for RMfam CMs is as a pre-filter to accelerate the more complex methods, for example, the Bayesian network approach implemented in RMdetect (32).

Increasing the access of motif annotations is another goal of the authors. We are active in the Rfam consortium which curates non-coding RNAs alignments (12–14). Our results show that curators can benefit greatly from motif annotations (see Supplementary Figures S44 and S45) and it is likely that RMfam annotations will be incorporated into further databases in future releases.

New technologies, such as the sequencing of cross-linked RNA and protein, are a potential source of new RNA-protein motifs. In the future we will mine these data sets (78–80) for new additions to the RMfam database. Furthermore, we will continue to add new motifs to RMfam as they are published.

Finally, as previously mentioned, the specificity of the RMfam annotations is generally low. However, incorporating the genomic and taxonomic context of annotations into the predictions may result in performance gains. For example, Shine–Dalgarno and Rho-independent terminators are generally located in bacterial sequences and at the extremities of annotated genes. A probabilistic incorporation of contextual information will likely result in further performance gains.

In summary, we have developed a resource for annotating diverse sets of RNA motifs in nucleotide sequences and alignments. We have proven the accuracy using benchmarks, and the utility of this resource for alignment curation, evolutionary analyses and shown that it has some promise for the prediction of RNA function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work would not be possible without input from a large community of RNA researchers that openly share their results. It has benefited from many discussions with members of the Xfam Consortium, the RNA Ontology Consortium and attendees of the 2012 Benasque Meeting on RNA. A special thanks to Lars Barquist, Elena Rivas, Rob Knight, Eric Westhof, Zasha Weinberg, Anthony Poole, Peter Fineran and two anonymous reviewers for their valuable contributions.

FUNDING

Rutherford Discovery Fellowship from Government, administered by the Royal Society of New Zealand [to P.P.G.]. Funding for open access charge: Rutherford Discovery Fellowship from NZ Government, administered by the Royal Society of New Zealand. Funding for open access charge: Rutherford Discovery Fellowship.

Conflict of interest statement. None declared.

REFERENCES

1. Brownlee,G.G. (1971) Sequence of 6S RNA of E. coli. *Nat. New Biol.*, **229**, 147–149.
2. Barrick,J.E., Sudarsan,N., Weinberg,Z., Ruzzo,W.L. and Breaker,R.R. (2005) 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, **11**, 774–784.
3. Wassarman,K.M. and Storz,G. (2000) 6S RNA regulates E. coli RNA polymerase activity. *Cell*, **101**, 613–623.
4. Lerner,M.R., Boyle,J.A., Hardin,J.A. and Steitz,J.A. (1981) Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*, **211**, 400–402.
5. Christov,C.P., Gardiner,T.J., Szüts,D. and Krude,T. (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.*, **26**, 6993–7004.
6. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
7. Perkins,T.T., Kingsley,R.A., Fookes,M.C., Gardner,P.P., James,K.D., Yu,L., Assefa,S.A., He,M., Croucher,N.J., Pickard,D.J. *et al.* (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet.*, **5**, e1000569.
8. Lindgreen,S., Umu,S.U., Lai,A.S., Eldai,H., Liu,W., McGimpsey,S., Wheeler,N.E., Biggs,P.J., Thomson,N.R., Barquist,L. *et al.* (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, **10**, e1003907.
9. Westhof,E. and Romby,P. (2010) The RNA structurome: high-throughput probing. *Nat. Methods*, **7**, 965–967.
10. Barquist,L., Langridge,G.C., Turner,D.J., Phan,M.D., Turner,A.K., Bateman,A., Parkhill,J., Wain,J. and Gardner,P.P. (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. *Nucleic Acids Res.***41**, 4549–4564.
11. Barquist,L., Boinett,C.J. and Cain,A.K. (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol.*, **10**, 1161–1169.
12. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the decimal release. *Nucleic Acids Res.*, **39**, D141–D145.
13. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
14. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2014) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, doi:10.1093/nar/gku1063.
15. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
16. Hoeppner,M.P., Gardner,P.P. and Poole,A.M. (2012) Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput. Biol.*, **8**, e1002752.
17. Jaeger,L., Michel,F. and Westhof,E. (1994) Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J. Mol. Biol.*, **236**, 1271–1276.
18. Jucker,F.M., Heus,H.A., Yip,P.F., Moors,E.H. and Pardi,A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, **264**, 968–980.
19. Abramovitz,D.L. and Pyle,A.M. (1997) Remarkable morphological variability of a common RNA folding motif: the GNRA tetraloop-receptor interaction. *J. Mol. Biol.*, **266**, 493–506.
20. Lucchetti-Miganeh,C., Burrowes,E., Baysse,C. and Ermel,G. (2008) The post-transcriptional regulator CsrA plays a central role in the adaptation of bacterial pathogens to different stages of infection in animal hosts. *Microbiology*, **154**, 16–29.
21. Valverde,C., Lindell,M., Wagner,E.G. and Haas,D. (2004) A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of Pseudomonas fluorescens. *J. Biol. Chem.*, **279**, 25066–25074.
22. Toledo-Arana,A., Repoila,F. and Cossart,P. (2007) Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.*, **10**, 182–188.
23. Batey,R.T., Rambo,R.P., Lucast,L., Rha,B. and Doudna,J.A. (2000) Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science*, **287**, 1232–1239.
24. López de Silanes,I., Zhan,M., Lal,A., Yang,X. and Gorospe,M. (2004) Identification of a target RNA motif for RNA-binding protein HuR. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2987–2992.
25. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
26. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
27. Höchsmann,M., Töller,T., Giegerich,R. and Kurtz,S. (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, **2**, 159–168.
28. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.
29. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
30. Naville,M., Ghuillot-Gaudeffroy,A., Marchais,A. and Gautheret,D. (2011) ARNold: a web tool for the prediction of Rho-independent transcription terminators. *RNA Biol.*, **8**, 11–13.

31. Gardner,P.P., Barquist,L., Bateman,A., Nawrocki,E.P. and Weinberg,Z. (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res.*, **14**, 5845–5852.
32. Cruz,J.A. and Westhof,E. (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods*, **8**, 513–521.
33. Haussler,D., Krogh,A., Mian,I. and Sjolander,K. (1993) Protein modeling using hidden Markov models: analysis of globins. In: *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences, 1993*, Vol. i, pp. 792–802.
34. Krogh,A. (1994) Hidden Markov models for labelled sequences. *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Vol. **2**, pp. 140–144.
35. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
36. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
37. Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.
38. Gardner,P.P., Fasold,M., Burge,S.W., Ninova,M., Hertel,J., Kehr,S., Steeves,T.E., Griffiths-Jones,S. and Stadler,P.F. (2014) Conservation and losses non-coding RNA associated loci in avian genomes. Genomics , arXiv:1406.7140.
39. 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.
40. 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.
41. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
42. Doshi,K.J., Cannone,J.J., Cobaugh,C.W. and Gutell,R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105–105.
43. Weinberg,Z. and Ruzzo,W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, **22**, 35–39.
44. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
45. 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.
46. Kolbe,D.L. and Eddy,S.R. (2011) Fast filtering for RNA homology search. *Bioinformatics*, **27**, 3102–3109.
47. Nawrocki,E.P. and Eddy,S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, **3**, e56.
48. Weinberg,Z. and Ruzzo,W.L. (2004) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, **20(Suppl. 1)**, i334–i341.
49. Schroeder,K.T., McPhee,S.A., Ouellet,J. and Lilley,D.M. (2010) A structural database for k-turn motifs in RNA. *RNA*, **16**, 1463–1468.
50. Klosterman,P.S., Hendrix,D.K., Tamura,M., Holbrook,S.R. and Brenner,S.E. (2004) Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.*, **32**, 2342–2352.
51. Tamura,M., Hendrix,D.K., Klosterman,P.S., Schimmelman,N.R., Brenner,S.E. and Holbrook,S.R. (2004) SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.*, **32**, D182–D184.
52. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Müller,K.M. *et al.* (2002) 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.
53. Zhong,C. and Zhang,S. (2012) Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment. *Nucleic Acids Res.*, **40**, 1307–1317.
54. Petrov,A.I., Zirbel,C.L. and Leontis,N.B. (2011) WebFR3D–a server for finding, aligning and analyzing recurrent RNA 3D motifs. *Nucleic Acids Res.*, **39**, W50–W55.
55. Zirbel,C.L., Sponer,J.E., Sponer,J., Stombaugh,J. and Leontis,N.B. (2009) Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.*, **37**, 4898–4918.
56. Grabow,W.W., Zhuang,Z., Swank,Z.N., Shea,J.E. and Jaeger,L. (2012) The right angle (RA) motif: a prevalent ribosomal RNA structural pattern found in group I introns. *J. Mol. Biol.*, **424**, 54–67.
57. Rosenblad,M.A., Larsen,N., Samuelsson,T. and Zwieb,C. (2009) Kinship in the SRP RNA family. *RNA Biol.*, **6**, 508–516.
58. Popenda,M., Blazewicz,M., Szachniuk,M. and Adamiak,R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res.*, **36**, D386–D391.
59. Popenda,M., Szachniuk,M., Blazewicz,M., Wasik,S., Burke,E.K., Blazewicz,J. and Adamiak,R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, **11**, 231.
60. Bluhm,W.F., Beran,B., Bi,C., Dimitropoulos,D., Prlic,A., Quinn,G.B., Rose,P.W., Shah,C., Young,J., Yukich,B. *et al.* (2011) Quality assurance for the query and distribution systems of the RCSB Protein Data Bank. *Database*, **2011**, bar003.
61. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
62. Consortium,PDB (2014) PDB derived data, PDB FTP site. ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt.gz.
63. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.
64. Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**, 4816–4822.
65. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
66. Weinberg,Z. and Breaker,R.R. (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, **12**, 3.
67. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea and their metagenomes. *Genome Biol.*, **11**, R31.
68. Leonardi,J., Box,J.A., Bunch,J.T. and Baumann,P. (2008) TER1, the RNA subunit of fission yeast telomerase. *Nat. Struct. Mol. Biol.*, **15**, 26–33.
69. Webb,C.J. and Zakian,V.A. (2008) Identification and characterization of the Schizosaccharomyces pombe TER1 telomerase RNA. *Nat. Struct. Mol. Biol.*, **15**, 34–42.
70. Gardner,P.P., Bateman,A. and Poole,A.M. (2010) SnoPatrol: how many snoRNA genes are there? *J. Biol.*, **9**, 4.
71. Bateman,A., Agrawal,S., Birney,E., Bruford,E.A., Bujnicki,J.M., Cochrane,G., Cole,J.R., Dinger,M.E., Enright,A.J., Gardner,P.P. *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA*, **17**, 1941–1946.
72. Widmann,J., Stombaugh,J., McDonald,D., Chocholousova,J., Gardner,P., Iyer,M.K., Liu,Z., Lozupone,C.A., Quinn,J., Smit,S. *et al.* (2012) RNASTAR: an RNA STructural Alignment Repository that provides insight into the evolution of natural and artificial RNAs. *RNA*, **18**, 1319–1327.
73. Duss,O., Michel,E., Yulikov,M., Schubert,M., Jeschke,G. and Allain,F.H. (2014) Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature*, **509**, 588–592.
74. Schubert,M., Lapouge,K., Duss,O., Oberstrass,F.C., Jelesarov,I., Haas,D. and Allain,F.H. (2007) Molecular basis of messenger RNA recognition by the specific bacterial repressing clamp RsmA/CsrA. *Nat. Struct. Mol. Biol.*, **14**, 807–813.
75. Barquist,L., Burge,S.W. and Gardner,P.P. (2012) Building non-coding RNA families. Biomolecules , arXiv:1206.4087.

76. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
77. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
78. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9613–9618.
79. Jungkamp,A.C., Stoeckius,M., Mecenas,D., Grn,D., Mastrobuoni,G., Kempa,S. and Rajewsky,N. (2011) In vivo and transcriptome-wide identification of RNA binding protein target sites. *Mol. Cell*, **44**, 828–840.
80. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.