The tedious task of finding homologous noncoding RNA genes

PETER MENZEL,^{1,2} JAN GORODKIN,¹ and PETER F. STADLER^{2,3,4,5,6}

¹Section for Genetics and Bioinformatics, IBHV, and Center for Applied Bioinformatics, University of Copenhagen, DK-1870 Frederiksberg, Denmark

²Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany

³Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany

⁴RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, D-04103 Leipzig, Germany

⁵Santa Fe Institute, Santa Fe, New Mexico 87501, USA

⁶Department of Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria

ABSTRACT

User-driven *in silico* RNA homology search is still a nontrivial task. In part, this is the consequence of a limited precision of the computational tools in spite of recent exciting progress in this area, and to a certain extent, computational costs are still problematic in practice. An important, and as we argue here, dominating issue is the dependence on good curated (secondary) structural alignments of the RNAs. These are often hard to obtain, not so much because of an inherent limitation in the available data, but because they require substantial manual curation, an effort that is rarely acknowledged. Here, we qualitatively describe a realistic scenario for what a “regular user” (i.e., a nonexpert in a particular RNA family) can do in practice, and what kind of results are likely to be achieved. Despite the indisputable advances in computational RNA biology, the conclusion is discouraging: BLAST still works better or equally good as other methods unless extensive expert knowledge on the RNA family is included. However, when good curated data are available the recent development yields further improvements in finding remote homologs. Homology search beyond the reach of BLAST hence is not at all a routine task.

Keywords: RNA secondary structure prediction; noncoding RNA homology search; RNA structural alignments

INTRODUCTION

The derivation of a secondary structure model is an important part of understanding the functional constraints of an RNA. While RNA folding programs can produce plausible predictions, comparative information is required in general to obtain reliable structures and to confirm predictions based on single sequences. The analysis of patterns of sequence and structure conservation over larger evolutionary time scales has been an important source of information, as it provides insights, for example, into the location of binding sites for proteins. For large RNAs, in particular ribosomal RNAs, structures are still most reliably derived using the “phylogenetic method,” that is, by investigating covariations of homologous sequence positions. Covariations beyond the helical regions provide insights into

tertiary interactions and allow the discovery of aggregate motifs (Leontis and Westhof 2003; Leontis et al. 2003), such as K-turns (Klein et al. 2001) or UA-handles (Jaeger et al. 2009), which are functionally important hallmarks of many RNA families.

All this, however, relies on the availability of large sets of homologous representatives. The Rfam database collects such information and provides it in a ready-to-use fashion (Griffiths-Jones et al. 2005). Given this convenient starting point, it should be straightforward to mine the rapidly growing collection of completely sequenced genomes for homologous RNAs—or is it? In fact, most genomes—with the exception of the vertebrates collected in the ENSEMBL system (Hubbard et al. 2009), the 12 *Drosophilids* (*Drosophila* 12 Genomes Consortium 2007; Rose et al. 2007), and *Caenorhabditis elegans* (Stricklin et al. 2005)—come with little or no noncoding RNA (ncRNA) annotation. This is in particular also true for almost all prokaryotes with the notable exception of *Escherichia coli*, although EBI’s *Genome Reviews* (<http://www.ebi.ac.uk/GenomeReviews/>) is now starting to integrate ncRNA annotations for noneukaryotic genomes to the extent that this information is available.

Reprint requests to: Peter F. Stadler, Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; e-mail: peter.stadler@bioinf.uni-leipzig.de; fax: +49-341-97-16709.

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.1556009>.

Finding homologs of ncRNA genes can be a surprisingly hard problem: Many noncoding RNAs (ncRNAs) are very short (often, 100 nucleotides [nt] or less) (Griffiths-Jones et al. 2005); they are very poorly conserved at the sequence level (the telomerase RNAs of *Saccharomyces* and *Kluyveromyces* species cannot even be aligned unambiguously) (Tzfati et al. 2003); and they may vary dramatically in length. Programs that are based on exact seed matches such as BLASTN in addition suffer from frequent small indels since ncRNAs do not have to preserve reading frames.

With structural features playing an important role, a series of software tools has been developed that attempts to utilize the constraints of secondary structures. We can distinguish two types of such approaches: Tools such as RNAMotif, RNABOB, or Palingol that require the user to explicitly specify a search pattern in dedicated descriptor languages; and systems such as ERPIN and Infernal that start from a structure-annotated alignment and infer structural models. Following a very brief review of the most commonly used tools in the next couple of paragraphs, we will focus on the inherent limitations of homology search approaches for ncRNAs, which so far have precluded or at least hampered comprehensive RNA annotation efforts.

One of the earliest implementations of a descriptor-based search algorithm was RNAMOT (Gautheret et al. 1990), whose language allowed the specification of stems and unpaired strands with variable lengths and primary sequence constraints; hits were automatically scored by stem lengths, nucleotide mismatches, and the number of wobble pairs in stems. RNABOB (<http://selab.janelia.org/software.html>) (S Eddy, unpubl.) extended this language and allowed for specifying a certain number of mispairs in a stem and a notation for permitting arbitrary pairing rules at certain positions in a stem. Palingol (Billoud et al. 1996) provides a powerful descriptor language that—inspired by functional programming languages—syntactically differs a lot from its predecessors. Another descriptor syntax was introduced by PatScan (Dsouza et al. 1997), which also allows matching against position weight matrices. One of the most recent and most advanced descriptor-based homology search tools is RNAMotif (Macke et al. 2001), which encompasses the capabilities of the earlier programs and also features a procedural language for evaluating and scoring pattern matches. In practice, a major drawback of descriptor-based approaches is the need to construct the search patterns by hand. The Locomotif tool (Reeder et al. 2007) solves many of the technical issues of specifying a descriptor. Nevertheless, the fundamental issue remains that a human researcher has to know what to search for in the first place. We argue here, that this knowledge is in many cases limited even for experienced experts.

The second class of homology search tools is based on automatic learning of statistical models given a structure-annotated sequence alignment. The most commonly used tool, Infernal (Eddy 2002; Nawrocki and Eddy 2007), is

based on covariance models and stochastic context-free grammars. This approach is extremely time-consuming. RaveNnA (Weinberg and Ruzzo 2006) was thus developed to provide an efficient pre-filter for Infernal by converting the covariance models into profile HMMs. A different approach is taken by ERPIN (Gautheret and Lambert 2001), which transforms a training alignment into a set of weight matrices for each structural element and then matches this matrix set on the sequence database. The advantage of these approaches is also their major disadvantage. The user not only has little effort with generating of the model, but also little chance to modify the search pattern. A recent evaluation of several training set based programs, in terms of specificity and sensitivity, has been presented by Freyhult et al. (2007).

Anecdotal evidence—in part from our own attempts to identify RNA by homology—suggests, however, that neither class of tools provides a ready solution whenever the phylogenetic range of the examples used for training or constructing the descriptor does not cover the genome to be searched. In other words, we have a hard time generalizing ncRNA patterns. In some cases, it is even hard to recognize a particular ncRNA. The Infernal server provided by the Rfam, for example, does not recognize the RNase MRP or the U17 snoRNA of *Trichoplax adhaerens* (even though these sequences are neither particularly derived nor are they outside the phylogenetic range of the training set) (Hertel et al. 2009). Other examples are the 7SK snRNAs of *Ciona intestinalis* and *Drosophila melanogaster*, both of which were detected by RNaz in two different studies (Missal et al. 2005; Rose et al. 2007), but neither one was recognized as a 7SK RNA by any available tool until a recent systematic analysis of this family (Gruber et al. 2008a,b). This is an excellent example demonstrating that ncRNAs that are missed even by extensively curated homology screens can be (re-) discovered in a de novo screen of related species that are far away from the phylogenetic range of the seed sequences.

USER-DRIVEN HOMOLOGY SEARCH

Experienced experts for a particular RNA family can, of course, construct descriptors that pretty much recover all the known examples of a given family. For several RNA families, however, only a (very) small set of examples is known and available in the Rfam seed set. How well do ncRNA gene-finding methods generalize for these families? In order to be able to assess how well one can generalize from a small set of examples, we decided to conduct an experiment starting from a phylogenetically restricted seed set for several RNA families. We chose eight ncRNA families representing the different classes, sizes, and phylogenetic ranges that can be encountered when dealing with ncRNAs. SRP and RNase MRP RNAs are long molecules with big structural variation between clades. SnoRNAs and microRNAs have typical conserved sequence motifs essential for

their function, while Y and vault RNAs are poorly understood and highly variable.

In order to ensure that no knowledge on the RNA families beyond the artificially restricted seed sets is included in the search patterns, we replaced the “expert” with a newly hired Ph.D. student (the first author of this work) with a computer science/bioinformatics background and some education in RNA bioinformatics, but without specific knowledge of the RNA families to be tested. The “expert” was asked to construct RNAMotif descriptors based on a small seed set, to search a broad range of available metazoan genomes, to evaluate the candidate hits, and to modify the descriptors using the newly found putative homologs. Depending on the number of hits produced for the target genomes, descriptors were modified to be less restrictive or to be more restrictive. The specificity of a descriptor can be loosened, for example, by allowing more mismatches in primary sequence constraints, by reducing the minimum length of a stem, by allowing an increased number of nonstandard base pairs, or by extending the length ranges of unpaired sequences in bulges and loop regions. We did not allow the complete loss of entire stems or require the insertion of specific structural motifs. Note that the latter is covered implicitly by the weakening of length constraints. We did allow, however, for the disappearance of small bulge loops. We decided to perform three iterations in each case.

For comparison, the same seed was used as BLASTN queries and to train an ERPIN model. Additionally, we refer to the supplemental material for a more-detailed description of the search procedures. We emphasize that this experiment was not conducted to compare the quality, performance, and usefulness of the software tools. Instead, our aim was to get some insight into the intrinsic difficulties of RNA homology search—which, at least in our experience, makes this seemingly routine task a demanding and technically challenging research topic.

Our interest therefore focuses on the “expert’s” ability to create descriptors that can detect homologous ncRNAs with high sensitivity and specificity, not on the computational efficiency of the search tools. We therefore used the most recent software with the most expressive language, RNAMotif, since descriptors written in other languages can be translated to RNAMotif, but not necessarily vice versa.

The results of our experiment are summarized in Figure 1. Details including all sequence data can be found in the supplemental material. Clearly, the phylogenetic range of detected homologs varies substantially between RNA families. The SRP RNA, U5 snRNA, and U3 snoRNA are quite well conserved at the sequence level already. For these three families, manually constructed descriptors and ERPIN perform comparably, although the descriptors tend to produce a significant number of false positives along with the true hits in the U5 RNA. The quite complex secondary structure of SRP RNA prevents RNAMotif and ERPIN

from capturing the family members in the invertebrates, since the seed set only contained mammalian SRPs. However, BLASTN had no problems in finding the SRPs over the full species range (Supplemental Table 3). Also in the U3 (Supplemental Table 9) and U5 (Supplemental Table 8) families, some of the invertebrate sequences could not be recovered with RNAMotif and ERPIN, but were recovered with BLASTN. For the U3 and SRP, we also screened the invertebrate genomes (except *Hydra magnipapillata*) with RaveNnA using a covariance model derived from the seed alignments. In both families, all known homologs were retrieved and RaveNnA also captured the *C. elegans* U3 snoRNA, which was missed by the other three programs. For most species, the known homologs are among the top three scoring hits both with RaveNnA as well as BLASTN. In the case of RNase MRP RNA (Supplemental Table 4), BLASTN yields a higher recovery rate than RNAMotif and ERPIN. It recovers all the known sequences across diverse invertebrates. Both RNAMotif and ERPIN generalize poorly in this family, which is known to contain structural variation (not to mention pseudoknots). The secondary structure model of the descriptors was not able to capture the structure diversities outside the eutharia. The mediocre performance of ERPIN can be explained in retrospect by the pseudo-knotted structure of RNase MRP RNA, in which exactly the region around the pseudoknot is the best conserved and contains the most informative patterns (Piccinelli et al. 2005; Woodhams et al. 2007). Thus, we also screened the teleostei and invertebrate genomes with RaveNnA using a CM model based on our training set and found that all annotated MRP RNAs were found, except the *Apis mellifera* sequence, which remains undetected by all four methods. For let-7, all methods produce similar results (Supplemental Table 6), also the RNAMotif descriptors recovered almost all family members in most species with high specificity. In the case of vault RNAs (Supplemental Table 7) and Y RNAs (Supplemental Table 5), on the other hand, all methods produced many false positives outside the range of the training data. BLASTN missed the *Fugu* Y RNA, but found more of the known vault RNAs compared to the other two methods, although it did not recover the vault RNA candidates outside the Sarcopterygii, which were predicted in Stadler et al. (2009). We note in this context that the latest release of ENSEMBL (version 52) provides RaveNnA/Infernal-based annotations of Y RNAs in most vertebrates. The vault RNAs, on the other hand, are still limited to Mammalia and *Xenopus*. Thus we screened the genomes of *Ciona intestinalis*, *Branchiostoma floridae*, *Strongylocentrotus purpuratus*, and the two teleostei with RaveNnA, recovering most of the candidate vault RNAs from Stadler et al. (2009).

Comparing the results of the three methods and including RaveNnA scans of the teleostei and invertebrate genomes for some of the families, we find that all of the methods have strengths and weaknesses. With BLASTN we

FIGURE 1. Homology search results. Members of the training set are indicated by boxes: Except for Y and vault RNAs, only mammalian sequences were used to construct the search patterns. For E2 and let-7, Rfam 8.1 provided only multiple human paralogs as seed sequences. For SRP, RNase MRP RNA, U3, and vault RNAs, we also ran RaveNnA on the small teleostei and invertebrate genomes, where ERPIN did not find the already annotated sequences. (Arrow) The range of the RaveNnA screens. (X) False-negative results, i.e., the fact that a homolog is known to exist but was not detected by any method. Complete sequences and detailed result tables are found at the Supplemental website.

find family members in all species with some misses, for example, the *C. elegans* U3 RNA and the vault RNAs outside the Sarcopterygii. RaveNnA has the highest sensitivity even with a very limited training set, but it has the

largest computational efforts of the presented methods, requiring high-end computational equipment for systematic whole-genome screens. The ERPIN results strongly depend on the search parameters derived from the training set, so that those parameters might not be chosen well enough automatically, explaining the moderate recovery rate in some cases. The descriptor-based search with RNAMotif did not generalize well enough to find distant homologs in most of the families. At the very least, several iterations of descriptor modification are required, and even then the descriptors are far from perfect. The effects of loosening the constraints in the descriptor are most prominently visible in the U3 snoRNAs and U5 snRNAs, where we found most homologs with RNAMotif in the third iteration. Note, however, that this also incurred a much higher false discovery rate than the previous two iterations. In the case of larger molecules, $n \gg 100$ nt, a descriptor covering the full structure is bound to fail, for example, the SRP and MRP results. Here, an automatic choice of suitable subpatterns for searching would be helpful. On the other hand, the descriptors for those two families were highly specific and, for example, the noise from all the SRP-derived Alu repeats was filtered out. See the supplemental material for further discussion of the results for each RNA family.

Despite the availability of several computationally efficient specific tools for RNA homology search, this task is thus still an excruciatingly hard one. Both a series of systematic analyses of specific RNA families (U7 snRNA) (Marz et al. 2007), Y RNAs (Mosig et al. 2007), plant enod40 (Gulyaev and Roussis 2007), telomerase RNAs (Chen et al. 2000; Xie et al. 2008), spliceosomal snRNAs (Marz et al. 2008; López et al. 2008), 7SK RNA (Gruber et al. 2008a,b), and nematode Sm Y RNAs (Jones et al. 2009) and our little experiment point at the same main difficulty: neither the “expert user,”

based on the examples at hand, nor the statistical models behind ERPIN managed to capture the nature of sequence/structure variation in sufficient detail to outperform the simple, blind, search for conserved subsequences. Even

when using covariance models, the problem of structural variation is a nontrivial issue. The bottom line is that if the structural variation is not part of the training data, one cannot expect to find it in the candidates produced by genome-wide screens either.

The limiting factor is the generalization of the search pattern beyond the phylogenetic range of the training data. We suggest that this is due to our limited understanding of the structural evolution of ncRNAs—as opposed to a shortcoming of the existing software in incorporating our knowledge. For instance, many RNA families exhibit clade-specific insertions and deletions, and different parts of the molecules can evolve with extremely different rates (Fig. 2). We have not yet learned, however, which rules govern this type of variation.

Similar types of structural variation as that of the telomerase RNA already have been observed a decade ago for other RNA families, such as tRNA (Zwieb et al. 1999).

The situation is similar for RNase P and MRP RNAs, which also present extensive structural variations. An extreme case is the RNase P RNA of *Candida glabrata*, with a length of ~700 nt (Kachouri et al. 2005). Pseudoknots present a serious practical problem in themselves, because the currently used implementations of covariance models do not handle pseudoknots. Rfam, therefore, cannot make full use of the annotations provided by some well-curated structural alignments, such as those stored in tmRDB (Andersen et al. 2006), although pseudoknot annotations are included in some families.

DISCUSSION AND PERSPECTIVES

A deeper understanding of the evolutionary patterns of structured RNAs, however, depends on the availability of diverse and detailed sets of examples. The only practical way to amass the necessary data is to systematically collect

and organize the information collected by the research community—an effort that, of course, is ongoing, as exemplified by the long-standing requirement to submit sequences to GenBank (Benson et al. 2005), and by the curation of dedicated RNA databases such as Rfam (Gardner et al. 2009), MirBase (Griffiths-Jones et al. 2006), and a plethora of smaller endeavors specializing in specific families (many of which are included in the upcoming Database Issue of *Nucleic Acids Research*). Nevertheless, these efforts cover only a fraction of the data that are available in principle: many—in particular, prokaryotic—small RNA families never entered one of the public sequence databases. They remain hidden in supplemental files of research publications, in practice excluding them from global analyses. Even the available structural alignments of the ncRNA families, for example, from Rfam, have been observed to be non-optimal in some cases (Andersen et al. 2007). Despite the continuous updates and improvements, therefore, it is still necessary to critically review the seed data set for the homology search before using it. The systematic annotation of ncRNAs in newly sequenced genomes is still a nontrivial and sometimes frustrating task—at least in part because of a lack of comparative data for homology-based approaches.

While the matching of novel ncRNAs to known families already poses big

FIGURE 2. Vertebrate telomerase structures. (A) Secondary structures of medaka (*Oryzias latipes*, $n = 312$), human ($n = 451$), and dogfish shark (*Squalus acanthias*, $n = 559$). Data adapted from Xie et al. (2008). (B) Sequence conservation. The panel includes data exported from the UCSC Genome Browser (Karolchik et al. 2008), showing the PhastCons (Siepel et al. 2005) conservation track based on the 28 vertebrate MULTIZ alignments (Blanchette et al. 2004), as well as a selection of pairwise alignments with the human locus. Note that outside the mammals only partial alignments are available in the automatic comparative genomics tracks. In particular, the homologs in *Xenopus* and teleost fishes are known in the literature but not identified in the genome-wide alignments.

problems due to structural variation over large phylogenetic distances, many novel structured RNA candidates can be inferred from covarying patterns in structurally conserved RNAs, as demonstrated, for example, on the ENCODE data (Washietl et al. 2007; Torarinsson et al. 2008). Once a novel ncRNA has been identified by one of these approaches, however, we are back to the problems of homology-based methods to identify additional family members. The emerging ability of computational methods to cope with large-scale clustering based on structural features (Havgaard et al. 2007; Torarinsson et al. 2007; Will et al. 2007) may be a step forward to recognizing faint homology signals. Such approaches might supplement or work in conjunction with covariance models. At present, it remains unclear however, whether our preconceptions on the structural variation of distantly related RNA (which necessarily enter the design of these algorithms) are close enough to reality to really solve the problem. In fact, the problem of structural variation (not to be confused with structural inserts) exceeds the *ansatz* in the Sankoff (1985) framework for structural alignment of RNAs. Thus, the structural variation, indeed, poses novel challenges in constructing efficient RNA search tools.

While this contribution was under review, a new version of Infernal became available to the public (Nawrocki et al. 2009). It has an improved support for local alignments that increases the sensitivity and provides a dramatic improvement in computing time. This new version might make pre-filtering, as in RaveNnA, unnecessary. For example, Infernal 1.0 trained on our seed alignment identified a 91-nt subsequence of the RNase MRP in the genome of *A. mellifera*, a homolog that had remained undetected by all other methods in our experiment.

Nevertheless, the homology search problem cannot be solved with present technologies in many cases of practical interest. For instance, none of the experimentally detected telomerase RNA sequences of *Candida* species (Gunisova et al. 2009) is recognizable by any method, including Infernal 1.0, using even the phylogenetically most closely related *Saccharomyces* telomerase RNAs in the training set.

Several approaches toward de novo prediction of structured ncRNAs have been proposed, and they all use different strategies to trade off between speed and accuracy. A range of methods, for example, QRNA (Rivas and Eddy 2001), RNAz (Washietl et al. 2005), and EvoFold (Knudsen and Hein 1999; Knudsen and Hein 2003; Pedersen et al. 2006), employ sliding fixed windows (excised from sequence-based alignments) in which the RNA structure prediction is carried out. Others are more expensive and directly perform local structural alignments (with a range of limitations to lower computational resources), for example, FOLDALIGN (Havgaard et al. 2007) and CMfinder (Yao et al. 2006). Dynalign uses a framework of local structural realignments in sliding windows over the sequence (Harmanci et al. 2007). In principle, these methods

can, of course, also be used in a homology search. At present, their practical application is hampered by the substantial computational costs. For a more-detailed review of the current status of de novo screening, see Gorodkin et al. (2010).

The need for not only a de novo search, but also for a homology search, is becoming apparent also when considering the strong increase of publications about particular noncoding RNAs in recent years (see Supplemental Fig. 1). With a doubling time of 3–4 yr and close to 10,000 publications in 2007, the need for well-curated and well-annotated repositories of such data has become a pressing problem.

Fortunately, the community value of collecting ncRNA sequences, preferably in the form of well-curated alignments, has led to the development of the RNA Family Database Rfam (Griffiths-Jones et al. 2005), which has become a central resource for RNA-related research. Most recently, the RNA community has been encouraged to contribute to this effort directly in a way that acknowledges the complexity of the task and ensures proper credit for individual annotators (Butler 2008; Gardner and Bateman 2009). This should help to facilitate the inclusion into public databases of both experimentally verified and computationally identified RNAs more quickly and more comprehensively; as we hope, this will also stimulate research into the structural evolution of RNA and eventually lead to much improved approaches for RNA gene finding.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at <http://www.rnajournal.org>. Also, a website containing the machine-readable data can be found at <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-025/>.

ACKNOWLEDGMENTS

We thank Ivo L. Hofacker for his critical reading of the manuscript. P.M. is funded by the Danish Research School for Biotechnology through a grant from the Danish Research Council for Technology and Production.

REFERENCES

- Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C. 2006. The tmRDB and SRPDB resources. *Nucleic Acids Res* **33**: D163–D168.
- Andersen E, Lind-Thomsen A, Knudsen B, Kristensen S, Havgaard J, Torarinsson E, Larsen N, Zwieb C, Sestoft P, Kjems J, et al. 2007. Semiautomated improvement of RNA alignments. *RNA* **13**: 1850–1859.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. Genbank. *Nucleic Acids Res* **33**: D34–D38.
- Billoud B, Kontic M, Viari A. 1996. Palingol: A declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res* **24**: 1395–1403.

- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–715.
- Butler D. 2008. Publish in Wikipedia or perish. *Nature News* doi: 10.1038/news.2008.1312.
- Chen JL, Blasco MA, Greider CW. 2000. Secondary structure of vertebrate telomerase RNA. *Cell* **100**: 503–514.
- Drosophila* 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.
- Dsouza M, Larsen N, Overbeek R. 1997. Searching for patterns in genomic data. *Trends Genet* **13**: 497–498.
- Eddy SR. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**: 18. doi: 10.1186/1471-2105-3-18.
- Freyhult EK, Bollback JP, Gardner PP. 2007. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* **17**: 117–125.
- Gardner PG, Bateman AG. 2009. A home for RNA families at RNA Biology. *RNA Biol* **6**: 2–4.
- Gardner J, Daub PP, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, et al. 2009. Rfam: Updates to the RNA families database. *Nucleic Acids Res* **37**: D136–D140.
- Gautheret D, Lambert A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol* **313**: 1003–1011.
- Gautheret D, Major F, Cedergren R. 1990. Pattern searching/alignment with RNA primary and secondary structures: An effective descriptor for tRNA. *Comput Appl Biosci* **6**: 325–331.
- Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL. 2010. Advances in predicting de novo RNA structure from genomic data. *Trends Biotechnol* (in press).
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: Annotating noncoding RNAs in complete genomes. *Nucleic Acids Res* **33**: D121–D124.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**: D140–D144.
- Gruber A, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF. 2008a. Arthropod 7SK RNA. *Mol Biol Evol* **25**: 1923–1930.
- Gruber AR, Koper-Emde D, Marz M, Tafer H, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ. 2008b. Invertebrate 7SK snRNAs. *J. Mol. Evol* **66**: 107–115.
- Gulyaev AP, Roussis A. 2007. Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. *Nucleic Acids Res* **35**: 3144–3152.
- Gunisova S, Elboher E, Nosek J, Gorkovoy V, Brown Y, Lucier J-F, Laterreux N, Wellinger RJ, Tzfati Y, Tomaska L. 2009. Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements. *RNA* **15**: 546–559.
- Harmanci AO, Sharma G, Mathews DH. 2007. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* **8**: 130. doi: 10.1186/1471-2105-8-130.
- Havgaard JH, Torarinsson E, Gorodkin J. 2007. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput Biol* **3**: 1896–1908. <http://foldalign.ku.dk>.
- Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF. 2009. Noncoding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res* **37**: 1602–1615.
- Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al. 2009. Ensembl 2009. *Nucleic Acids Res* **37**: D690–D697.
- Jaeger L, Verzemnieks EJ, Geary C. 2009. The UA _ handle: A versatile submotif in stable RNA architectures. *Nucleic Acids Res* **37**: 215–230.
- Jones TA, Otto W, Marz M, Eddy SR, Stadler PF. 2009. A survey of nematode SmY rnas. *RNA Biol* **6**: 5–8.
- Kachourii R, Stribinskis V, Zhu Y, Ramos KS, Westhof E, Li Y. 2005. A surprisingly large RNase P RNA in *Candida glabrata*. *RNA* **11**: 1064–1072.
- Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**: D773–D779.
- Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: A new RNA secondary structure motif. *EMBO J* **20**: 4214–4221.
- Knudsen B, Hein J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **31**: 3423–3428.
- Knudsen B, Hein J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**: 446–454.
- Leontis NB, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol* **13**: 300–308.
- Leontis NB, Lescoute A, Westhof E. 2003. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol* **13**: 300–308.
- López MD, Alm Rosenblad M, Samuelsson T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res* **36**: 3001–3010.
- Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29**: 4724–4735.
- Marz M, Mosig A, Stadler BMR, Stadler PF. 2007. U7 snRNAs: A computational survey. *Genomics Proteomics Bioinformatics* **5**: 187–195.
- Marz M, Kirsten T, Stadler PF. 2008. Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* **67**: 594–607.
- Missal K, Rose D, Stadler PF. 2005. Noncoding RNAs in *Ciona intestinalis*. *Bioinformatics* **21** Suppl 2: i77–i78.
- Mosig A, Guofeng M, Stadler BMR, Stadler PF. 2007. Evolution of the vertebrate Y RNA cluster. *Theory Biosci* **126**: 9–14.
- Nawrocki EP, Eddy SR. 2007. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* **3**: e56. doi: 10.1371/journal.pcbi.0030056.
- Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.
- Pedersen J, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander E, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**: e33. doi: 10.1371/journal.pcbi.0020033.
- Piccinelli P, Rosenblad MA, Samuelsson T. 2005. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res* **33**: 4485–4495.
- Reeder J, Reeder J, Giegerich R. 2007. Locomotif: From graphical motif description to RNA motif search. *Bioinformatics* **23**: i392–i400.
- Rivas E, Eddy S. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**: 8. doi: 10.1186/1471-2105-2-8.
- Rose DR, Hackermüller J, Washietl S, Findeiß S, Reiche K, Hertel J, Stadler PF, Prohaska SJ. 2007. Computational RNomics of *Drosophilids*. *BMC Genomics* **8**: 406. doi: 10.1186/1471-2164-8-406.
- Sankoff D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* **45**: 810–825.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.
- Stadler PF, Chen JJ-L, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, et al. 2009. Evolution of vault RNAs. *Mol Biol Evol* **26**: 1975–1991.

- Stricklin SL, Griffiths-Jones S, Eddy SR. 2005. *C. elegans* noncoding RNA genes. In *WormBook* (ed. The *C. elegans* Research Community). doi: 10.1895/wormbook.1.7.1, http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html.
- Torarinsson E, Havgaard JH, Gorodkin J. 2007. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* **23**: 926–932. <http://foldalign.ku.dk>.
- Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J. 2008. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. *Genome Res* **18**: 242–251.
- Tzfati Y, Knight Z, Roy J, Blackburn EH. 2003. A novel pseudoknot element is essential for the action of a yeast telomerase. *Genes & Dev* **17**: 1779–1788.
- Washietl S, Hofacker I, Stadler P. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102**: 2454–2459.
- Washietl S, Pedersen JS, Korb J, Gruber A, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Stocsits C, Tanzer A, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**: 852–864.
- Weinberg Z, Ruzzo WL. 2006. Sequence-based heuristics for faster annotation of noncoding RNA families. *Bioinformatics* **22**: 35–39.
- Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* **3**: e65. doi: 10.1371/journal.pcbi.0030065.
- Woodhams MD, Stadler PF, Penny D, Collins LJ. 2007. RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol Biol* **7** Suppl 1: S13. doi: 10.1186/1471-2148-7-S1-S13.
- Xie M, Mosig A, Qi X, Li Y, Stadler PF, Chen JJ-L. 2008. Structure and function of the smallest vertebrate telomerase RNA from teleost fish. *J Biol Chem* **283**: 2049–2059.
- Yao Z, Weinberg Z, Ruzzo W. 2006. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**: 445–452.
- Zwieb C, Wower I, Wower J. 1999. Comparative sequence analysis of tmRNA. *Nucleic Acids Res* **27**: 2063–2071.

RNA

A PUBLICATION OF THE RNA SOCIETY

The tedious task of finding homologous noncoding RNA genes

Peter Menzel, Jan Gorodkin and Peter F. Stadler

RNA 2009 15: 2075-2082 originally published online October 27, 2009

Access the most recent version at doi:[10.1261/rna.1556009](https://doi.org/10.1261/rna.1556009)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2009/10/23/rna.1556009.DC1>

References

This article cites 62 articles, 12 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/15/12/2075.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:

<http://rnajournal.cshlp.org/subscriptions>