The Tedious Task of Finding Homologous Non-coding RNA Genes

Peter Menzel [a,b], Jan Gorodkin [a], Peter F. Stadler [b,c,d,e,f,*]

[a] *Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark*
{ptr,gorodkin}@genome.ku.dk

[b] *Bioinformatics Group, Dept. of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
{ptr,studla}@bioinf.uni-leipzig.de

[c] *Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*

[d] *RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany*

[e] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[f] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

Abstract

User driven *in silico* RNA homology search is still a non-trivial 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 the what a "regular user" (i.e., a non-expert 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. Homology search beyond the reach of `BLAST` hence is not at all a routine task.

Key words: RNA secondary structure prediction; non-coding RNA homology search; RNA structural alignments.

Manuscript *25 January 2010*

1 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 *e.g.*, 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", *i.e.*, by investigating covariations of homologous sequence positions. Covariations beyond the helical regions provide insights into tertiary interactions and allow the discovery of aggregate motifs [35, 34], such as K-turns [31] or UA-handles [27], that 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 [18, 12]. 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 [26], the 12 Drosophilids [8, 48], and *C. elegans* [52] — come with little or no non-coding RNA annotation. This is in particular also true for almost all procaryotes with the notable exception of *E. coli*, although EBI's Genome Reviews (`http://www.ebi.ac.uk/GenomeReviews/`) is now starting to integrate ncRNA annotations for non-eukaryotic genomes to the extent that this information is available.

Finding homologs of ncRNA genes can be a surprisingly hard problem: Many ncRNAs are very short (often 100nt or less) [18], they are very poorly conserved at sequence level (the telomerase RNAs of *Saccharomyces* and *Kluyveromyces* species cannot even be aligned unambiguously [55], 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 have been developed that attempt 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` [15], 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` (Sean Eddy 1996, unpubl., `http://selab.janelia.org/software.html`) extended this language and allowed for specifying a certain number of mis-pairs in a stem and a notation for permitting arbitrary pairing rules at certain positions in a stem. `Palingol` [4] provides a powerful descriptor language which — inspired by functional programming languages — syntactically differs a lot from its predecessors. Another descriptor syntax was introduced by `PatScan` [9], which also allows matching against position weight matrices. One of the most recent and most advanced descriptor-based homology search tools is `RNAMotif` [37], 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 [46] 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` [10, 42], is based on covariance models and stochastic context-free grammars. This approach is extremely time-consuming. `RaveNnA` [58] 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` [14], 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.* [11].

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 to generalize 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 [25]). Another example are the 7SK snRNA of *Ciona intestinalis* and *Drosophila melanogaster*, both of which were detected by `RNAz` in two different studies [40, 48], but neither one was not recognized as a 7SK RNA by any available tool until a recent systematic analysis of this family [20, 19]. 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.

2 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 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" by a newly hired PhD student (the first author of this work) with a computer science/bioinformatics background and some education in RNA bioinformatics, but without specific knowledge on 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 be more restrictive. The specificity of a descriptor can be loosened *e.g.*, by allowing more mismatches in primary sequence constraints, by reducing the minimum length of a stem, by allowing an increased number of non-standard 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 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 on-line supplementary material for a more detailed description of the search procedures (see the URL at the end). 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 at the supplementary website (see the URL at the end). 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 (suppl. tab. 3). Also in the U3 (suppl. tab. 9) and U5 (suppl. tab. 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 *H. 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 (Suppl. Tab. 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 descrip-

tors 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 pseudo-knot is the best conserved and contains the most informative patterns [45, 60]. 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 *A. mellifera* sequence, which remains undetected by all four methods. For let-7, all methods produce similar results (suppl. tab. 6), also the `RNAMotif` descriptors recovered almost all family members in most species with high specificity. In the case of vault RNAs (suppl. tab. 7), and Y RNAs (suppl. tab. 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 [51]. We note in this context that the latest release of `ENSEMBL` (v. 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 *C. intestinalis*, *B. floridae*, *S. purpuratus*, and the two teleostei with `RaveNnA`, recovering most of the candidate vault RNAs from [51].

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 find family members in all species with some misses, *e.g.*, 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, *e.g.*, the SRP and MRP results. Here, an automatic choice of suitable sub-patterns for searching would be helpful. On the other hand, the descriptors for those two families were highly specific and *e.g.*, the noise from all the SRP-derived Alu repeats was filtered out. See the supplementary material for further discussion

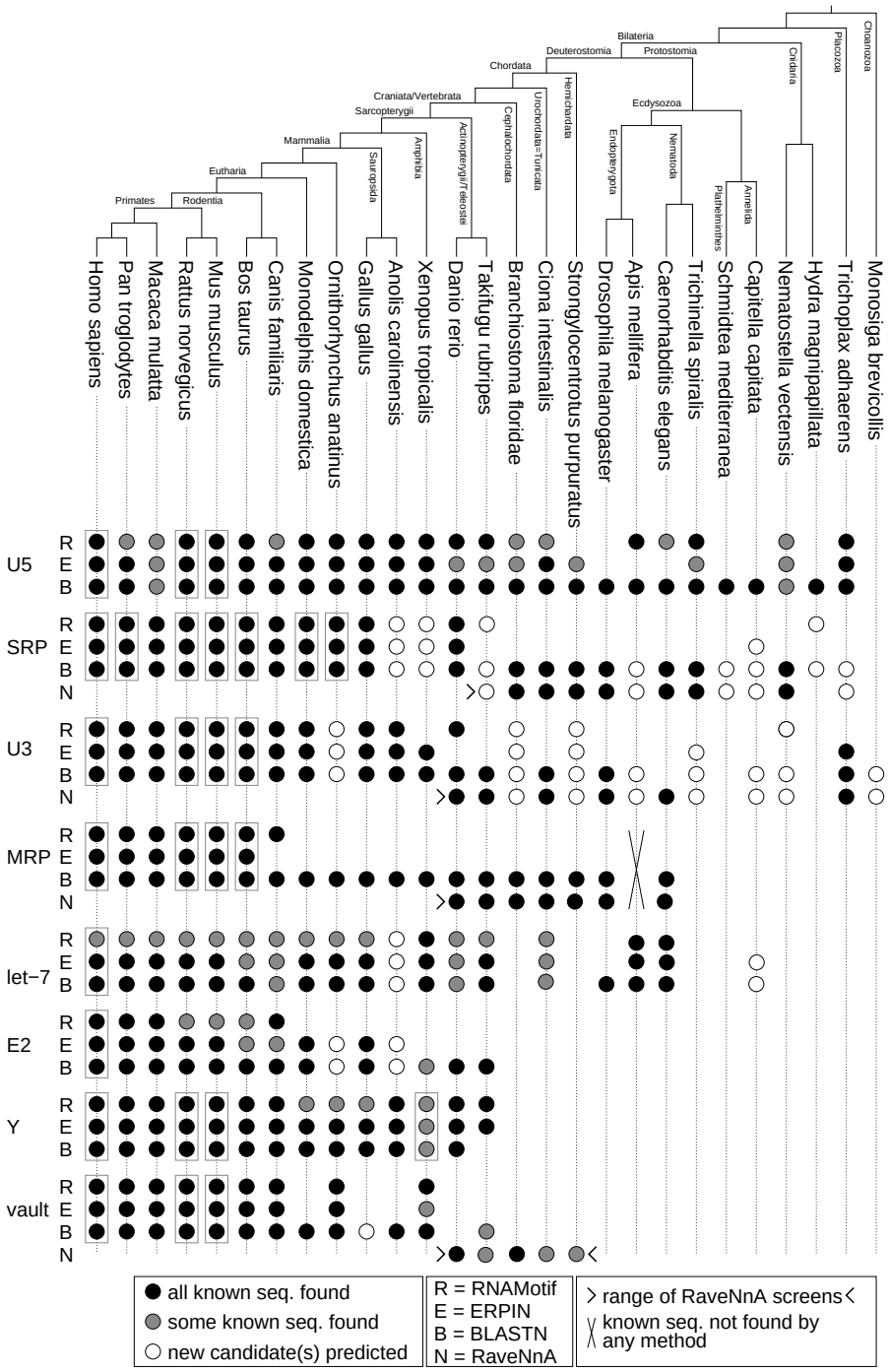

Fig. 1. Homology search results. Members of the training set are indicated by boxes: Except for Y and vault RNAs, only mammalian sequences were used to construct the search patterns. For E2 and let-7, `Rfam 8.1` provided only multiple human paralogs as seed sequences. For SRP, RNase MRP RNA, U3 and vault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

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 [39], Y RNAs [41], plant enod40 [21], telomerase RNAs [7, 61], spliceosomal snRNAs [38, 36], 7SK RNA [20, 19], nematode Sm Y RNAs [28]) 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 sub-sequences. Even when using covariance models the problem of structural variation is a non-trivial 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 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 have been observed for other RNA families, such as tmRNA [63], already a decade ago. The situation is similar for RNase P and MRP RNAs, which also present extensive structural variations. An extreme case is the the RNaseP RNA of *Candida glabrata* with a length of about 700 nt [29]. 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 the annotations provided by some well curated structural alignments, such as those stored in `tmRDB` [2], although pseudoknot annotations are included in some families.

3 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

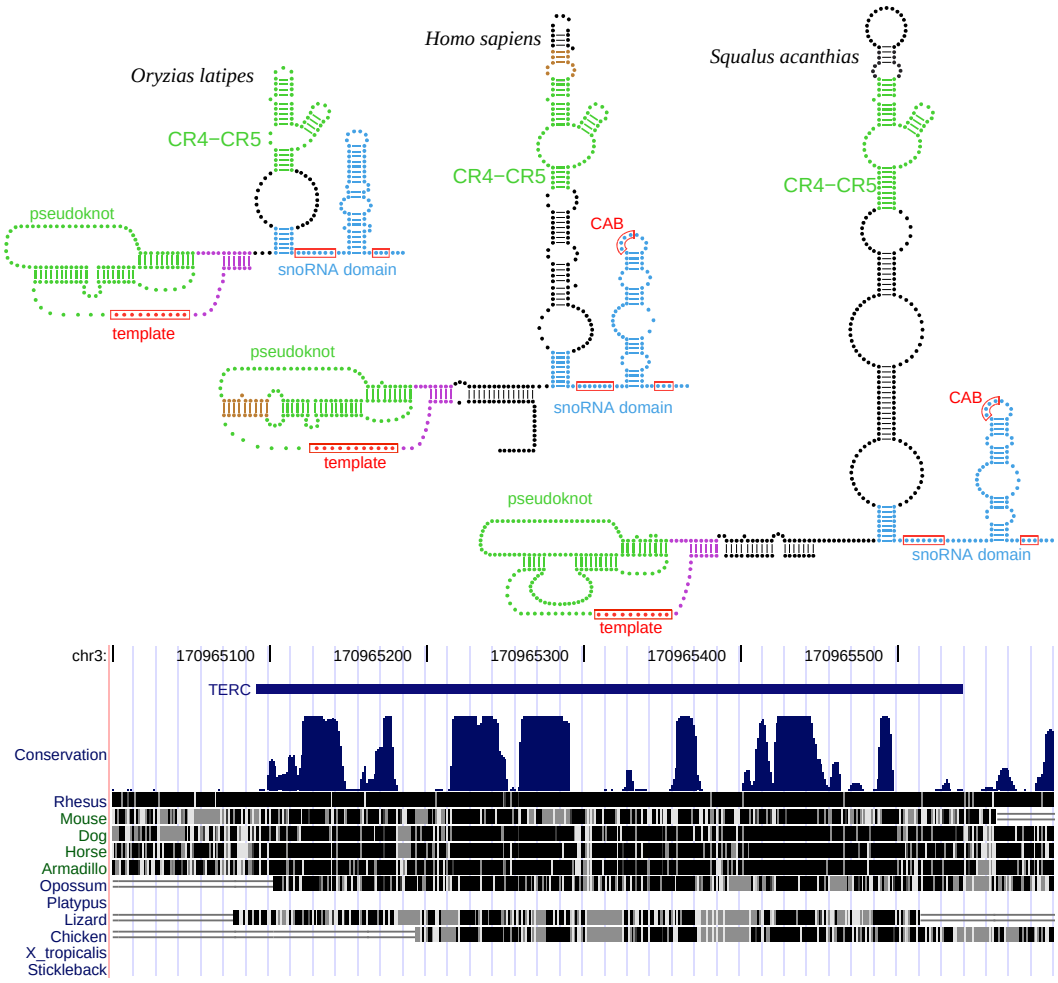

Fig. 2. Vertebrate telomerase structures. Top: secondary structures of medaka (*Oryzias latipes*, $n = 312$), human ($n = 451$), and dogfish shark (*Squalus acanthias*, $n = 559$). Adapted from [61]. Below: sequence conservation. The panel is exported from the UCSC genome browser [30], showing the `PhastCons` [50] conservation track based on the 28 vertebrate `MULTIZ` alignments [5], 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.

effort that of course is ongoing, as exemplified by the long-standing requirement to submit sequences to `GenBank` [3], and by the curation of dedicated RNA databases such as `Rfam` [12], `MirBase` [17], and a plethora of smaller endeavours specializing on 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 procaryotic — 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, *e.g.* from `Rfam`, have been observed to be non-optimal in some cases [1]. Despite the continuous updates and improvements, therefore, it is still necessary to critically review the seed data set for homology search before using them. The systematic annotation of ncRNAs in newly sequenced genomes is still a non-trivial 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 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 *e.g.* on the ENCODE data [57, 54]. 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 [59, 24, 53] 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 pre-conceptions 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 [49] framework for structural alignment of RNAs. Thus the structural variation indeed pose novel challenges in constructing efficient RNA search tools.

While this contribution was under review, a new version of `infernal` became available to the public [43]. It has an improved support for local alignments which 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 91nt 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 [22] 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 towards *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, *e.g.*, `qrna` [47], `RNAz` [56] and `EvoFold`

[44, 33, 32] 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), *e.g.*. `FOLDALIGN` [24] and `CMfinder` [62]. `Dynalign` uses a framework of local structural (re-)alignments in sliding windows over the sequence [23]. In principle, these methods can of course also be used in 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 we refer to [16].

The need for not only *de novo* search, but also for homology search is becoming apparent also when considering the strong increase of publications about particular non-coding RNAs in the recent years (see Supplementary Fig. 1). With a doubling time of 3–4 years and close to ten thousand 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 lead to the development of the RNA Family Database `Rfam` [18], 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 way that acknowledges the complexity of the task and ensures proper credit for individual annotators [13, 6]. 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.

Supplementary Material

A website containing additional documents and data is located at `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-025`.

Acknowledgements

We thank Ivo L. Hofacker for a critical reading of the manuscript. Peter Menzel is funded by the Danish research school for biotechnology through a grant from the Danish Research Council for Technology and Production.

References

[1] E. Andersen, A. Lind-Thomsen, B. Knudsen, S. Kristensen, J. Havgaard, E. Torarinsson, N. Larsen, C. Zwieb, P. Sestoft, J. Kjems, and J. Gorodkin. Semiautomated improvement of RNA alignments. *RNA*, 13(11):1850–1859, 2007. Epub 2007 Sep 5.

[2] E. S. Andersen, M. A. Rosenblad, N. Larsen, J. C. Westergaard, J. Burks, I. K. Wower, J. Wower, J. Gorodkin, T. Samuelsson, and C. Zwieb. The tmRDB and SRPDB resources. *Nucl. Acids Res.*, 34(Database issue):D163–8, 2006.

[3] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucl. Acids Res.*, 33:D34–D38, 2005.

[4] B. Billoud, M. Kontic, and A. Viari. `Palingol`: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucl. Acids Res.*, 24:1395–1403, 1996.

[5] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14:708–15, 2004.

[6] D. Butler. Publish in Wikipedia or perish. *Nature News*, page doi:10.1038/news.2008.1312, 2008.

[7] J. L. Chen, M. A. Blasco, and C. W. Greider. Secondary structure of vertebrate telomerase RNA. *Cell*, 100:503–514, 2000.

[8] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450:203–218, 2007.

[9] M. Dsouza, N. Larsen, and R. Overbeek. Searching for patterns in genomic data. *Trends Genet.*, 13:497–498, 1997.

[10] S. R. Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18, 2002.

[11] E. K. Freyhult, J. P. Bollback, and P. P. Gardner. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res*, 17:117–125, 2007.

[12] J. Gardner, P. P. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. `Rfam`: updates to the RNA families database. *Nucl. Acids Res.*, 37:D136–D140, 2009.

[13] P. G. Gardner and A. G. Bateman. A home for RNA families at *RNA Biology*. *RNA Biology*, 6:2–4, 2009.

[14] D. Gautheret and A. Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, 313:1003–1011, 2001.

[15] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.*, 6:325–331, 1990.

[16] J. Gorodkin, I. L. Hofacker, E. Torarinsson, Z. Yao, J. H. Havgaard, and W. L. Ruzzo. Advances in predicting *de novo* RNA structure from genomic data. *Submitted*, 2009.

[17] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucl. Acids Res.*, 34:1362–4962, 2006.

[18] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. `Rfam`: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33:D121–D124, 2005.

[19] A. Gruber, C. Kilgus, A. Mosig, I. L. Hofacker, W. Hennig, and P. F. Stadler. Arthropod 7SK RNA. *Mol. Biol. Evol.*, 1923-1930:25, 2008.

[20] A. R. Gruber, D. Koper-Emde, M. Marz, H. Tafer, S. Bernhart, G. Obernosterer, A. Mosig, I. L. Hofacker, P. F. Stadler, and B.-J. Benecke. Invertebrate 7SK snRNAs. *J. Mol. Evol.*, 107-115:66, 2008.

[21] A. P. Gultyaev and A. Roussis. Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. *Nucleic Acids Res.*, 35:3144–3152, 2007.

[22] S. Gunisova, E. Elboher, J. Nosek, V. Gorkovoy, Y. Brown, J.-F. Lucier, N. Laterreur, R. J. Wellinger, Y. Tzfati, and L. Tomaska. Identification and comparative analysis of telomerase RNAs from Candida species reveal conservation of functional elements. *RNA*, 15(4):546–559, Apr 2009.

[23] A. O. Harmanci, G. Sharma, and D. H. Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign. *BMC Bioinformatics*, 8:130, 2007.

[24] J. H. Havgaard, E. Torarinsson, and J. Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Computational Biology*, 3:1996–1908, 2007. (http://foldalign.ku.dk).

[25] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res.*, 37:1602–1615, 2009.

[26] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Res.*, 37:D690–D697, 2009.

[27] L. Jaeger, E. J. Verzemnieks, and C. Geary. The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res.*, 37:215–230, 2009.

[28] T. A. Jones, W. Otto, M. Marz, S. R. Eddy, and P. F. Stadler. A survey

of nematode SmY rnas. *RNA Biology*, 6:5–8, 2009.

[29] R. Kachouri, V. Stribinskis, Y. Zhu, K. S. Ramos, E. Westhof, and Y. Li. A surprisingly large RNase P RNA in *Candida glabrata*. *RNA*, 11:1064–72, 2005.

[30] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–9, 2008.

[31] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz. The kink-turn: a new RNA secondary structure motif. *EMBO J.*, 20:42144221, 2001.

[32] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucl. Acids Res.*, 31:3423–3428, Jul 2003.

[33] B. Knudsen and J. J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–454, 1999.

[34] N. B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Current Opinion Struct. Biol.*, 13:300–308, 2003.

[35] N. B. Leontis and E. Westhof. Analysis of RNA motifs. *Current Opinion Struct. Biol.*, 13:300–308, 2003.

[36] M. D. López, M. Alm Rosenblad, and T. Samuelsson. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.*, 36:3001–3010, 2008.

[37] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. `RNAMotif`, an RNA secondary structure definition and search algorithm. *Nucl. Acids Res.*, 29:4724–4735, 2001.

[38] M. Marz, T. Kirsten, and P. F. Stadler. Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, 67:594–607, 2008.

[39] M. Marz, A. Mosig, B. M. R. Stadler, and P. F. Stadler. U7 snRNAs: A computational survey. *Geno. Prot. Bioinf.*, 5:187–195, 2007.

[40] K. Missal, D. Rose, and P. F. Stadler. Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, 21 S2:i77–i78, 2005. Proceedings ECCB/JBI'05, Madrid.

[41] A. Mosig, M. Guofeng, B. M. R. Stadler, and P. F. Stadler. Evolution of the vertebrate Y RNA cluster. *Th. Biosci.*, 126:9–14, 2007.

[42] E. P. Nawrocki and S. R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput. Biol.*, 3:e56, 2007.

[43] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337, May 2009.

[44] J. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, 2(4), 2006.

[45] P. Piccinelli, M. A. Rosenblad, and T. Samuelsson. Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.*, 33:4485–4495, 2005.

[46] J. Reeder, J. Reeder, and R. Giegerich. `Locomotif`: From graphical motif description to RNA motif search. *Bioinformatics*, 23:i392–400, 2007.

[47] E. Rivas and S. Eddy. Noncoding rna gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8–8, 2001.

[48] D. R. Rose, J. Hackermüller, S. Washietl, S. Findeiß, K. Reiche, J. Hertel, P. F. Stadler, and S. J. Prohaska. Computational RNomics of drosophilids. *BMC Genomics*, 8:406, 2007.

[49] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl Math*, 45:810–825, 1985.

[50] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 2005.

[51] P. F. Stadler, J. J.-L. Chen, J. Hackermüller, S. Hoffmann, F. Horn, P. Khaitovich, A. K. Kretzschmar, A. Mosig, S. J. Prohaska, X. Qi, K. Schutt, and K. Ullmann. Evolution of vault rnas. *Mol. Biol. Evol.*, 26:1975–1991, 2009.

[52] S. L. Stricklin, S. Griffiths-Jones, and S. R. Eddy. C. elegans noncoding RNA genes. *WormBook*, doi/10.1895/wormbook.1.7.1, 2005. `http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html`.

[53] E. Torarinsson, J. H. Havgaard, and J. Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23:926–932, 2007. (http://foldalign.ku.dk).

[54] E. Torarinsson, Z. Yao, E. D. Wiklund, J. B. Bramsen, C. Hansen, J. Kjems, N. Tommerup, W. L. Ruzzo, and J. Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. *Genome research*, 18:242–251, 2008.

[55] Y. Tzfati, Z. Knight, J. Roy, and E. H. Blackburn. A novel pseudoknot element is essential for the action of a yeast telomerase. *Genes & Dev.*, 17:1779–1788, 2003.

[56] S. Washietl, I. Hofacker, and P. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA*, 102(7):2454–2459, 2005.

[57] S. Washietl, J. S. Pedersen, J. O. Korbel, A. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, C. Stocsits, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigó, M. Snyder, M. B. Gerstein,

A. Reymond, I. L. Hofacker, and P. F. Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Gen. Res.*, 17:852–864, 2007.

[58] Z. Weinberg and W. L. Ruzzo. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, 22:35–39, 2006.

[59] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3:e65, 2007.

[60] M. D. Woodhams, P. F. Stadler, D. Penny, and L. J. Collins. RNAse MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.*, 7:S13, 2007.

[61] M. Xie, A. Mosig, X. Qi, Y. Li, P. F. Stadler, and J. J.-L. Chen. Size variation and structural conservation of vertebrate telomerase RNA. *J. Biol. Chem.*, 283:2049–2059, 2008.

[62] Z. Yao, Z. Weinberg, and W. Ruzzo. CMfinder - a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, 2006.

[63] C. Zwieb, I. Wower, and J. Wower. Comparative sequence analysis of tmRNA. *Nucl. Acids Res.*, 27:2063–2071, 1999.