Strategies for Homology-Based ncRNA Gene Annotation

Axel Mosig[a,b], Liang Zhu[a,c], Peter F. Stadler[d,b,e,f,g]

[a] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, 320 Yue Yang Road, 200031 Shanghai, China*
[b] *Max-Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[c] *Graduate School of CAS, Beijing 100039, China*
[d] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[e] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*
[f] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Most non-coding RNAs are short and/or poorly conserved in sequence. Most of the longer examples, furthermore, consist of a collection of conserved structural motifs rather that a coherent globally conserved secondary structure. As a consequence, the conceptually simple problem of homology search becomes a complex and technically demanding task. Despite the best efforts of databases such as `Rfam`, the situation is complicated further by the sparsity of information on many — in particular prokaryotic — RNA families. In this contribution we review recent efforts to customize sequence-based search tools for ncRNA applications. In particular semi-global alignments and the development of methods for fragmented pattern search have brought significant practical advances. Current developments in this area focus on the integration of fragmented sequence pattern search with search algorithms for secondary structure patterns. As one example, we introduce here `fragrep3`.

1. Introduction

Non-coding RNAs are a very heterogeneous class of transcripts. It may not come as a surprise, therefore, that no single computational approach is suitable to deal with all of the diverse types. In this contribution we focus on *homology search*: given a one or more known representative sequences of a particular ncRNA the task at hand is to identify all homologous sequences in an un-annotated string of genomic DNA. For this particular purpose it is convenient to classify ncRNAs into several categories:

1. large ribosomal RNAs, i.e., the major components of the small and large RNA subunits of nuclear and organellar ribosomes.
2. small housekeeping RNAs.
   This class includes most of the classical ncRNAs, including tRNAs, 5S rRNA, snoRNAs, spliceosomal RNAs, as well as many of the small bacterial RNAs.
3. large mRNA-like ncRNAs.
   This class contains many moderate-size non-protein-coding transcripts that are spliced and polyadenylated, giving rise to processed RNAs with a size of several kb. In this class we include also giant ncRNAs.

The task of homology search can be subdivided into two phases. The first, and typically more difficult, step is to localize the ncRNA gene in the genomic DNA. In the second – refinement – step, the exact structure of the ncRNA gene needs to be determined.

The LSU and SSU rRNA of category (1) are sufficiently large and well-conserved even at kingdom level, so that their genomic locations are easily determined by a simple `blast` (1) search. Homology search of the ml-ncRNAs in category (3), on the other hand, is largely uncharted territory. We will therefore focus here on the small housekeeping RNAs.

Email addresses: `axel@picb.ac.cn` (Axel Mosig), `zhuliang@picb.ac.cn` (Liang Zhu), `studla@bioinf.uni-leipzig.de` (Peter F. Stadler)

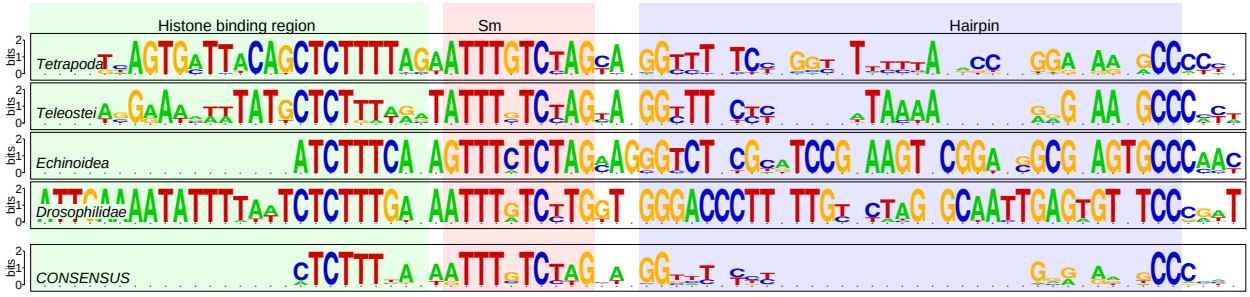

Figure 1: Conservation of the U7 sequence in four clades (Tetrapoda, dominated by mammalian sequence, teleost fishes, sea urchins, and drosophilid flies. While significant conservation is observed with each of these, fairly narrow, groups, the consensus over all four clades shows multiple in/dels and very little conservation beyond two functional sequences boxes: the histone binding motif and the Sm binding motif. Adapted from (2).

2. Conservation Patterns

The difficulty of the homology search problem is naturally determined by the amount of information that is contained in the query sequence(s), and by the level sequence conservation between query sequence(s) and subject genome. The search for house-keeping ncRNAs is in general much more difficult than the search for protein coding gene because of the much smaller size of the ncRNAs, which limits the query information. The search for conserved protein-coding sequences is furthermore simplified by the large, informative amino-acid alphabet. In contrast, the *per letter* information content is quite limited in nucleic acid queries. As a consequence, nucleotide-based `blastn` by default seed words of 7 or more. The substantial in/del rates in ncRNAs can make it impossible to meet this criterion.

Protein-coding information is localized in the three bases of the codon, and stabilizing selection at the protein level often acts to preserve contiguous peptide motifs (e.g. those specified as `prosite` patterns (3)). The need to maintain the reading frame furthermore severely restricts the distribution of in/dels between homologous protein-coding sequences. Indeed, the in/del distribution can be used to distinguish coding from non-coding regions in pairwise sequence alignments (4). A search for locally similar translations, as implemented in `tblastn` and `tblastx` thus often works very well for coding sequences.

The evolutionary constraints on ncRNAs are quite different. There is no need to maintain reading frames, thus there are not strong restrictions on the distribution of in/dels. Divergent sequences therefore do not necessarily contain gap-free substring of sufficient length to act as seeds for `blast`-like approaches. Figure 1 shows the conservation pattern of the U7 snRNA (2) as an example. In many cases, there is little constraint on local

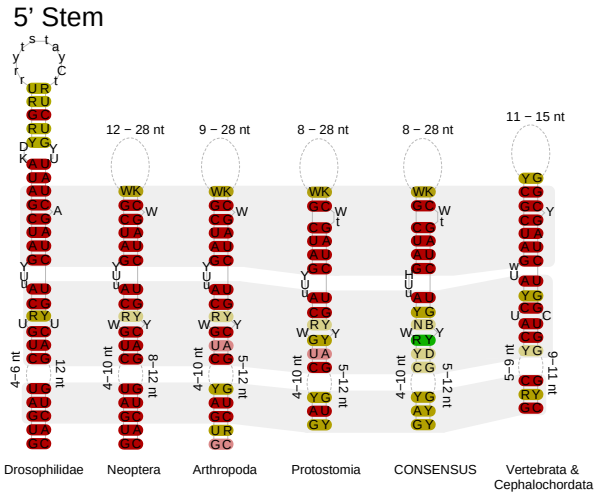

Figure 2: Comparison of the 5'stems of 7SK snRNAs. Conserved regions are color-coded to indicate conservation: conserved (red), two and three compensatory mutations (ochre, green). Pale colors indicate that a base-pair cannot be formed by all the sequences. Lower case letters denote a deletion in some sequences. Corresponding regions of the helices are highlighted by a gray background. Adapted from (5).

sequence patterns. Instead, much of the stabilizing selection may act to preserve secondary structure motifs, such as the tRNA clover leafs. Fig. 2 shows the 5' hairpins of metazoan 7SK RNA as an example.

By definition, base-pairing patterns are non-local in the sequence. Therefore, they cannot be identified by simple sequence-search techniques but require more sophisticated computational approaches — and more often than not orders of magnitudes more in terms of computational resources.

Beyond the information contained in the ncRNA itself, we can sometimes utilize additional sources of constraint's. Since many of the house-keeping ncR-

NAs are pol-III transcripts, we may include additional knowledge about the gene structure. For instance, the poly-T terminator can help to distinguish spurious hits from promising candidates. For snRNAs, we can expect a particular promoter structure with well-conserved promixal and distal sequence elements that can be included in the search patterns (6; 7). We shall return to this point in the discussion of a few individual case studies.

3. Query Data

Typically, a homology search project starts from a collection of trusted, preferably experimentally validated *seed sequences*. The most common source for them is the `Rfam` database (8), one of the family-specific RNA databases such as `mirbase` (9) for microRNAs, `snoRNABase` (10) for small nucleolar RNAs, or one of the specialized collections dealing with a single family such as the `tmRNAdb` and `SRPdb` (11), or telomerase RNA collection (12). For some RNA families, extensive data sets covering broad phylogenetic ranges are available. For less well-studied ncRNAs, however, the seed sets of often sparse and very limited in their phylogenetic range. In the latter case, it can be very hard to detect additional homologs, in particular to find homologs outside the sub-tree spanned by the available seed sequences.

In many cases, furthermore, seed sequences are not readily available but have to be retrieved *manually* from the literature. This is in particular the case for prokaryotic small RNAs, which are more often identified only by genomic coordinates in some supplementary spreadsheet file. This unfortunate state of affairs has recently been recognized as a problem. It is addressed by the *RNA Family Section* of the journal *RNA Biology*, which provides an incentive to organize such data in a form that is much more readily accessible for homology search projects and to incorporate them into the `Rfam` database (13).

4. Fragmented Pattern Search

The phenomenon of fragmentation is a common theme found in conservation patterns of ncRNA: while a few blocks are strongly conserved, large regions are not sufficiently conserved to allow unambiguous alignments or even consist of completely unrelated sequences. Fragmentation can be observed on the level of sequence as well as on the level of secondary structure, and correspondingly needs to be taken into account on both levels. For some RNA families such as telRNA (see Figure 3), this leads to highly divergent sequence lengths, contributing to the difficulty of ncRNA homology search.

4.1. Modeling Sequence Homology

As fragmentation of conserved regions is a common phenomenon in the evolution of ncRNA, basically all methods used for sequence-based homology search deal with at least some degree of fragmentation. Basic local alignment tools such as `blast` and `ssearch` (16) are limited to detecting individual blocks of conservation in pairwise comparisons. As argued above, however, the conserved regions tend to be too short and insignificant to be aligned as proper matches. Correspondingly, searching across longer phylogenetic timescales requires homology search methods with a higher degree of sensitivity. In the context of RNA homology search, profile-based approaches such as Hidden Markov Models (HMMs) have been attributed higher sensitivity than pairwise alignment based tools (17).

The log-odds scores typcially implemented in profile HMMs capture homology within conserved regions in a way that allows a much more sensitive search than simple regular expression. A shortcoming of HMMs in the presence of fragmentation is length variability in poorly conserved regions. Gap lengths in HMMs tend to follow a restricted type of distribution (such as a negative binomial distribution); correspondingly, length deviations when searching across larger evolutionary time-scales will be assigned low significance scores.

An approach where sequence fragmentation has been explicitly taken into account is implemented in the `fragrep2` tool (18): conserved blocks are modeled as *position frequency matrices* (PFMs), whereas the unconserved gaps outside these blocks are represented by lower and upper bounds on their length. Search patterns in `fragrep` are typically modelled from a multiple sequence alignment; conserved regions are labelled manually, so that the PFMs as well as bounds for the gap regions can be extracted using the `aln2pattern` tool. For querying a model against a genome, matches are reported wherever the conserved blocks match with a pre-specified score, and satisfying the upper and lower bounds on the distance between the conserved blocks.

For matching individual PFMs, `fragrep` employs a scoring scheme originally used for matching transcription factor binding sites (19), which has been additionally equipped to allow insertions and deletions. Compared with log-odd scores, `fragrep2` scores range in the unit interval, making them particularly accessible for manually adapting the search pattern.

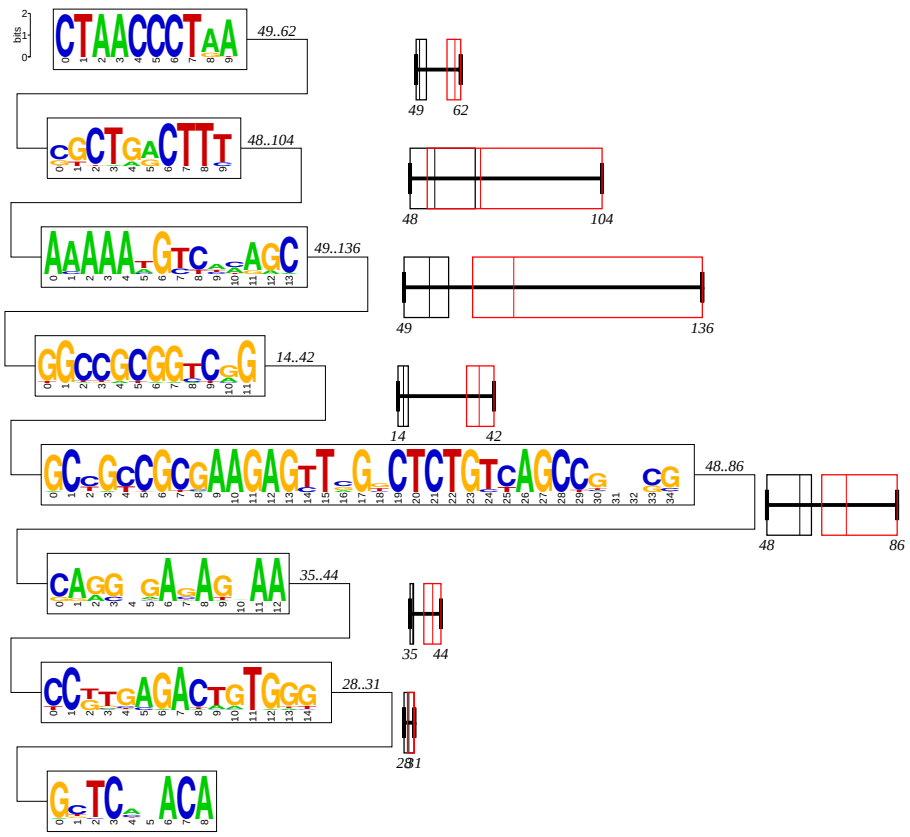

Figure 3: Vertebrate tel-RNA has been identified as fragmented into eight conserved regions (14; 15). The lengths of the non-conserved regions vary widely between teleost fish (black box plots) and eutheria (red box plots). The box plots indicate minimum, maximum, and average length of the non-conserved regions.

4.2. Modeling Structure Homology

On the level of secondary structure, fragmentation is naturally harder to capture than on sequence level. Covariance models, as implemented in `Rsearch` (20) or `infernal` (21; 22), are a natural generalization of sequence-based HMMs to SCFGs (Stochastic Context Free Grammars), which are necessary to describe secondary structures. Correspondingly, their features in terms of detecting fragmented homology are similar to HMMs: Conserved homologous blocks are represented by statistically significant patterns. As in HMMs, large irregular gap patterns are problematic in this framework, because they have to be modelled explicitly and cannot easily be approximated by simple bounds. A second issue with SCFG-based approaches is their extensive resource consumption (which is equivalent to that of the Sankoff-style structural alignments proposed in (23) for applications in homology search). In practise, one therefore typically uses efficient filters such as `RaveNnA` (24) to reduce the search space to manageable subset. `ERPIN` (25) also uses a dynamic programming approach. Instead of a full SCFG, it employs log-odds-score profiles for each of the helices and single stranded regions. An advantage is that in this way pseudoknots can be incorporated quite easily. As is the case with HMMs and SCFGs, `ERPIN` learns its search pattern from a structure-annotated input alignment.

Pattern search tools such as `RNAmotif` (26) or Sean Eddy's `rnabob`, on the other hand, utilize manually constructed search patterns that are much less detailed but much faster to search. The main problem with this approach is the quality of the search patterns. In practise, only experts on particular RNA families are capable of constructing descriptors that are sufficiently specific and nevertheless sensitive enough to beat simple `blast` searches (27). The efforts to construct good descriptors is thus likely to offset the computational efficiency of the pattern search.

The `fragrep3` tool has been designed as a hybrid of the fragmented pattern search tool `fragrep2` and the structure-search approach of `rnabob`. The resulting algorithm is conceptually similar to `ERPIN` in that it individually scores local matches of the individual sequence and structure patterns. Like its predecessors,

however, fragrep3 treats poorly conserved regions as simple distance constraints between significantly conserved blocks.

The philosophy of the fragrep tools is an intermediate between the statistical approaches and the descriptor-based methods. As with infernal or *ERPIN*, the user supplies a multiple sequence alignment. For fragrep, the user also has to provide an additional annotation line indicating the blocks that are to be converted into PFMs. Goodness-of-match parameters are then derived automatically that are adjusted so that each sequence in the input will be recognized. The user can then easily modify these parameters, e.g. in order to relax the requirements on conservation within the blocks or the inter-block distances.

5. Semi-global Alignment Strategies

Semi-global alignments can be seen as a natural approach to ncRNA homology search. Rather than Smith-Waterman-style local alignments as implemented in Ssearch, the semi-global version demands the complete query sequence to be aligned against the (long) genomic DNA. Semi-global alignments as a tool for RNA homology search have been investigated only recently. For each position k in the subject genome sequence, one computes the score s_k of the best semi-global alignment ending in k. Only local optima of s_k along the genome are of interest.

Semi-global alignments with affine gap penalties as implemented in GotohScan (29) turn out to be an extraordinarily useful tool, as lowering gap extension penalties allows to explicitly account for the phenomena of fragmented homology patterns and length variability. Fig. 4 shows an example of distribution of these locally optimal scores, which can be used readily to estimate *E*-values for candidate hits. GotohScan proved useful on a large scale for annotating ncRNA in identifying major parts of the ncRNAs in the genome of *Trichoplax adhaerens*, most notably the full repertoire of major and minor spliceosomal snRNAs, the genes for RNase P and MRP RNAs, the SRP RNA, as well as several small nucleolar RNAs (29). Similarly encouraging results were obtained for *Aspergillus fumigatus* (30).

In the context of vault RNA screens, combining GotohScan semi-global alignments proved particularly successful when combined with pre- or post-filtering with other homology search approaches, most notably fragrep.

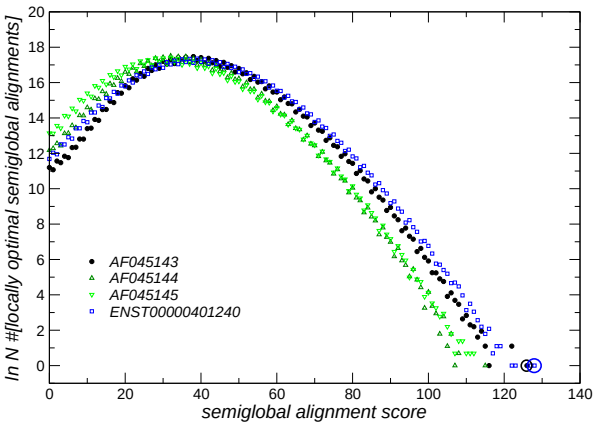

Figure 4: Distribution of GotohScan scores for the genome of the stickleback *Gasterosteus aculeatus* using four human vault RNAs as queries. In this example, the true hits (circled) are separated only marginally from the background distribution, which is estimated on flight directly from the data. For only two of the four queries (hgv1 and the ensemble ncRNA ENST00000401240, which was recently identified as a bona fide vault RNA (28)), we obtain clear hits (marked by circles). For the other two shortest query sequences, stickleback vault RNAs are among the best few dozend hits. The best hit for hgv2 (AF045144) turns out the be a false positive upon closer inspection.

6. Further Signals of ncRNA Genes

When dealing with RNA homology search problems that are not readily solved with a simple blast search, it is rarely the case that any particular search method will yields just a single or only a few candidates among which the true homologs are readily identified. Instead, further evaluation of a larger number of candidates is necessary. To this end, further evidence can be gathered from several of the following aspects:

- *Conservation scores:* When dealing with a candidate that can be spotted in a genome-wide alignment with one or several other species, it is possible to measure the evolutionary conservation of the candidate. The RNAz program (31) can be used to compute *z*-scores; also, fragrep2 allows to search genome-wide alignments rather than just single genomes.

- *Promoter sequences:* Evidence for a functional transcript may in some cases be as straightforward as a conserved TATA box at the 5' end of the putative transcript. Moreover, many polymerase III transcribed ncRNA genes have relatively well understood promoter sequences. These can enhance significance considerably in some homology models, as detailed for U7 snRNA and 7SK RNA in the case studies below.

- *Synteny:* For some ncRNA genes, vicinity to other genes is conserved. For instance, the study on vault RNA (32) found vaultRNA being part of the syntenically conserved *protocadherin cluster*, which is syntenically conserved between shark and human (33).
- *Phylogenetic coherence:* Naturally, a candidate sequence should be validated whether it fits into the phylogeny of its known homologous family members. This is typically achieved by fitting the candidate into an alignment of the known sequences, allowing to inspect a phylogenetic tree or network constructed from the alignment.
- *Functional aspects:* Some well-studied families of ncRNA contain functional elements whose homology patterns can be modelled more precisely than generic modeling and search tools would allow. Modeling of 3'P4 and 5'P4 regions in RNAse MRP (34) may be attributed as such pattern, as well as the H/ACA snoRNA domain in telRNA discussed below.

7. Case Studies

7.1. Vault RNAs in Protostomes

A scenario where `fragrep3` proves useful in combination with `GotohScan` is the annotation of protostome vault RNA. Vault RNAs are small polymerase III transcripts which are difficult to annotate due their length of only about 100 nucleotides. Until recently, they were only known in mammals, and have been found only recently in other vertebrates and basal deuterostomes (32) utilizing a combination of `blast`, `GotohScan` and `fragrep2`.

The combined search for sequence and structure homology implemented in `fragrep3` further increased the sensitivity of the search procedure and enabled us to find the first well-supported vault RNA candidates in protostomes. Candidates in *Helobdella robusta* were obtained in a two-step procedure: First, lower deuterostome sequences were aligned against the *Helobdella* genome using `GotohScan` with very low stringency. In a second step, a secondary-structure constrained `fragrep` pattern was searched against the (several tenthousands) of candidate sequences from the first step. Among the few candidates obtained this way, only one turned out to exhibit the internal *B-Box* promoter element. This candidate was searched against the *Lottia gigantea* genome, which produced a candidate with notably higher homology scores, also exhibiting the necessary secondary structure and internal *B-Box* promoter elements. A simple`blast` search of *Lottia* candidate against the the same genome, finally, revealed a paralogous locus on the same scaffold, which however lacks a descernible box B element, Fig. 5.

7.2. Telomerase RNA

Although telRNA is part of the telomerase complex in most eukaryotes, it demonstrates a surprisingly large variability in terms of both sequence and secondary structure. This is reflected by a length variation ranging from 147nt in the ciliate *Tetrahymena paravorax* to 1554nt in the fungus *Candida albicans*. Even within the mammals telRNA length stretches between 321 and 541 nucleotides. Essentially the only constant secondary structure feature is the pseudoknotted region that captures the template region, while loss or insertion of secondary structure elements is commonly observed. The challenge in homology search across longer time scales is to predict – or rather guess – which elements are conserved and which have been lost.

Some aspects of telRNA, however, contribute significantly to the specificity of search patterns. In some species, at least the template region is known precisely through sequencing the telomeric region. Although only 5 (insects) to 25 (saccharomycotina) nucleotides long, including the template region into the homology search pattern enhances the specificity significantly. Furthermore, vertebrate telRNA is known to contain a H/ACA snoRNA domain (14). This domain is known to indeed share the same function as in snoRNA, namely as a locator within the nucleus (35). This indeed legitimates to borrow strategies from snoRNA annotation tools, such as `snoReport` (36) or `snoGPS` (37), as part of telRNA homology search, and constitutes an example of how functional understanding of a non-coding RNA may not only boost homology search, but is an inevitable part of the search process.

7.3. Small nuclear RNAs

An inherent problem in annotating many of the small nuclear and nucleolar RNA their short length which does not contain homology fragments for sufficiently unambiguous identification. However, as a number of these short RNAs are transcribed by Polymerase III, their relatively well understood promoter structures can be utilized for annotating them much more reliably. This observation extends to the the small nuclear RNAs transcribed by pol-II, which share a similar promoter structure (38). These external elements were utilized in particular systematic surveys of U7 snRNA and 7SK RNA in animal genomes:

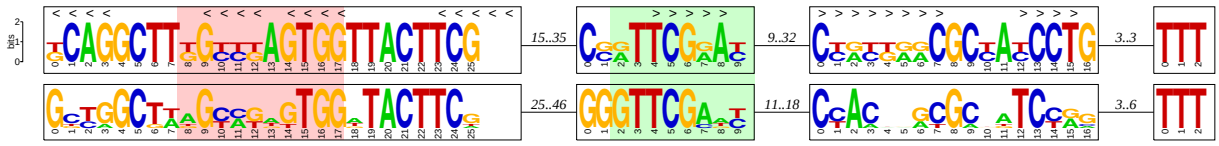

Figure 5: Consensus profiles for protostome vtRNAs (top), along with the basal deuterostome vtRNA identified in (32). The internal *Box A* and *Box B* pol-III promoters elements are highlighted red and green, respectively. Base pairs of the consensus structure are indicated by angular brackets $< \cdots >$.

U7 snRNA: The U7 snRNA is known to contain several conserved elements: beside a histone binding site and a Sm-binding motif, they are flanked by a stem-loop structure at the 3' end which is enclosed by two GC pairs. In (2), these elements were used to set up a homology search pattern, along with a species-specific model for the *proximal sequence element* (PSE). This was derived from upstream regions of U1, U2, U4, U5, U4atac, U11, and U12 spliceosomal RNAs, all of which are longer and hence typically better annotated. Assembling all these sequence models into a `fragrep` search pattern is straightforward. Unambiguous candidates were obtained through filtering the candidates obtained by `fragrep` using `RNAbob`.

7SK RNA: A similar approach as described for annotating U7 snRNA was successful to annotate 7SK RNA in invertebrate deuterostomia (39), and subsequently in arthropods (5). As insights into the functioning of 7SK RNA suggest, a GGC-GCC stem with a loop region crucial for P-TEFb binding were modelled using `fragrep`; candidates thus obtained were iteratively filtered by possessing a suitable PSE, as well as structural alignments using the `Ralee` mode (40) in the `Emacs` editor.

Similar to the U7 and 7SK RNA studies, vaultRNA candidates reported in (32) were also validated by their polIII promoter sequences. In principle, this promoter based homology might even carry to some pol II transcribed small RNAs whose transcription is activated by essentially the same PSE as many of their pol III transcribed relatives.

8. Discussion

Many ncRNA families are at present beyond the reach of automated or semi-automated pipelines for their annotation due to their rapid evolution and the resulting lack of significantly conserved features. The annotation of these families requires computational, evolutionary, and experimental approaches to go hand in hand and often require a thorough understanding of functional or regulatory aspects to separate true candidates from false positives.

In this respect, we can expect a certain relief from the increasing availability of sequenced transcriptomes and genomes, which shortens the evolutionary gaps across which homology search needs to be performed. However, major losses or gains of structural elements within relatively short evolutionary timescales are commonly observed, and still impose major challenges for homology search. Eventually, genome-wide alignments have the potential to better unveil synteny patterns. A systematic utilization of synteny will yet require a thorough and evolutionarily dense understanding of whole-genome-evolution, which are currently beyond reach. Hence, family-specific studies that meticulously assemble family-specific peculiarities into a homology model appear to be the only viable way to cover larger evolutionary gaps at present.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

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

[3] C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, P. Bucher, PROSITE: a documented database using patterns and profiles as motif descriptors, Brief Bioinform. 3 (2002) 265–274.

[4] A. Löytynoja, N. Goldman, A model of evolution and structure for multiple sequence alignment, Phil. Trans. R. Soc. B 363 (2008) 3913–3919.

[5] A. Gruber, C. Kilgus, A. Mosig, I. L. Hofacker, W. Hennig, P. F. Stadler, Arthropod 7SK rna, Mol. Biol. Evol. 1923-1930 (2008) 25.

[6] S. M. Mount, V. Gotea, C.-F. Lin, K. Hernandez, W. Makałowski, Spliceosomal small nuclear RNA genes in 11 insect genomes, RNA 13 (2007) 5–14.

[7] M. Marz, T. Kirsten, P. F. Stadler, Evolution of spliceosomal snrna genes in metazoan animals, J. Mol. Evol.10.1007/s00239-008-9149-6.

[8] 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, A. Bateman, `Rfam`: updates to the RNA families database, Nucl. Acids Res. 37 (2009) D136–D140.

[9] S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, `miRBase`: tools for microRNA genomics, Nucleic Acids Res. 36 (2008) D154–D158.
[10] L. Lestrade, M. J. Weber, snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs, Nucl. Acids Res. 34 (2006) D158–D162.
[11] E. S. Andersen, M. A. Rosenblad, N. Larsen, J. C. Westergaard, J. Burs, I. K. Wower, J. Wower, J. Gorodkin, T. Samuelsson, C. Zwieb, The tmRDB and SRPDB resources, Nucleic Acids Res. 34 (2006) D163–D168.
[12] J. D. Podlevsky, C. J. Bley, R. V. Omana, X. Qi, J. L. Chen, The telomerase database, Nucleic Acids Res. 36 (2007) D339–D343.
[13] A. home for RNA families at RNA Biology, Gardner, paul p. and bateman, alex g., RNA Biol. 6 (2009) 2–4.
[14] J. L. Chen, M. A. Blasco, C. W. Greider, Secondary structure of vertebrate telomerase RNA, Cell 100 (2000) 503–514.
[15] M. Xie, A. Mosig, X. Qi, Y. Li, P. F. Stadler, J. J.-L. Chen, Size variation and structural conservation of vertebrate telomerase RNA, J. Biol. Chem. 283 (2008) 2049–2059.
[16] W. R. Pearson, Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, Genomics 11 (1991) 635–650.
[17] E. K. Freyhult, J. P. Bollback, P. P. Gardner, Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA, Genome Res. 17 (1) (2007) 117.
[18] A. Mosig, J. L. Chen, P. F. Stadler, Homology search with fragmented nucleic acid sequence patterns, in: R. Giancarlo, S. Hannenhalli (Eds.), Algorithms in Bioinformatics (WABI 2007), Vol. 4645 of Lecture Notes in Computer Science, Springer Verlag, Berlin, Heidelberg, 2007, pp. 335–345.
[19] A. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. Kel-Margoulis, E. Wingender, MATCH: a tool for searching transcription factor binding sites in DNA sequences, Nucleic Acids Research 31 (13) (2003) 3576.
[20] R. J. Klein, S. R. Eddy, RSEARCH: finding homologs of single structured RNA sequences, BMC Bioinformatics 4 (2003) 44.
[21] S. R. Eddy, A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, BMC Bioinformatics 3 (2002) 18.
[22] E. P. Nawrocki, S. R. Eddy, Query-dependent banding for faster RNA similarity searches, PLoS Comp. Biol. 3 (2007) e56.
[23] A. F. Bompfünewerer, R. Backofen, S. H. Berhart, J. Hertel, , I. L. Hofacker, P. F. Stadler, S. Will, Variations on RNA folding and alignment: Lessons from Benasque, J. Math. Biol. 56 (2008) 129–144.
[24] Z. Weinberg, W. L. Ruzzo, Sequence-based heuristics for faster annotation of non-coding RNA families, Bioinformatics 22 (2006) 35–39.
[25] D. Gautheret, A. Lambert, Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles, J. Mol. Biol. 313 (2001) 1003–1011.
[26] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, R. Sampath, RNAMotif, an RNA secondary structure definition and search algorithm, Nucl. Acids Res. 29 (22) (2001) 4724–4735.
[27] P. Menzel, J. Gorodkin, P. F. Stadler, The tedious task of finding homologous non-coding RNA genes, RNAUnder review.
[28] C. Nandy, J. Mrázek, H. Stoiber, F. A. Grässer, A. Hüttenhofer, N. Polacek, Epstein-Barr virus-induced expression of a novel human vault RNA, J. Mol. Biol.Doi: 10.1016/j.jmb.2009.03.031.
[29] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, P. F. Stadler, Non-coding RNA annotation of the genome of *Trichoplax adhaerens*, Nucleic Acids Res. 37 (2009) 1602–1615.
[30] C. Jöchl, M. Rederstorff, J. Hertel, P. F. Stadler, I. L. Hofacker, M. Schrettl, H. Haas, A. Hüttenhofer, Small ncrna transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis, Nucleic Acids Res. 36 (2008) 2677–2689.
[31] S. Washietl, I. L. Hofacker, P. F. Stadler, Fast and reliable prediction of noncoding RNAs, Proc. Natl. Acad. Sci. USA 102 (2005) 2454–2459.
[32] P. Stadler, J.-L. Chen, J. Hackermüller, S. Hoffmann, F. Horn, P. Khaitovich, A. Kretzschmar, A. Mosig, X. Qi, K. Schutt, K. Ullmann, Evolution of vault RNAs, submitted.
[33] W.-P. Yu, V. Rajasegaran, K. Yew, W.-l. Loh, B.-H. Tay, C. T. Amemiya, S. Brenner, B. Venkatesh, Elephant shark sequence reveals unique insights into the evolutionary history of vertebrate genes: A comparative analysis of the protocadherin cluster, Proc. Natl. Acad. Sci. USA 105 (2008) 38193824.
[34] M. D. Woodhams, P. F. Stadler, D. Penny, L. J. Collins, RNAse MRP and the RNA processing cascade in the eukaryotic ancestor, BMC Evol. Biol. 7 (2007) S13.
[35] A. A. Lukowiak, A. Narayanan, Z. H. U. H. Li, R. M. Terns, M. P. Terns, The snoRNA domain of vertebrate telomerase RNA functions to localize the RNA within the nucleus, RNA 7 (2002) 1833–1844.
[36] J. Hertel, I. L. Hofacker, P. F. Stadler, `snoReport`: Computational identification of snoRNAs with unknown targets, Bioinformatics 24 (2008) 158–164.
[37] P. Schattner, W. A. Decatur, C. A. Davis, M. Ares Jr, M. J. Fournier, T. M. Lowe, Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome, Nucleic Acids Res. 32 (2004) 4281–4296.
[38] N. Hernandez, Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription, J. Biol. Chem. 276 (2001) 26733–26736.
[39] A. R. Gruber, D. Koper-Emde, M. Marz, H. Tafer, S. Bernhart, G. Obernosterer, A. Mosig, I. L. Hofacker, P. F. Stadler, B.-J. Benecke, Invertebrate 7SK snRNAs, J. Mol. Evol. 107-115 (2008) 66.
[40] S. Griffiths-Jones, RALEE-RNA ALignment editor in Emacs (2005).