Customized strategies for discovering distant ncRNA homologs

Axel Mosig, Liang Zhu and Peter F. Stadler

Advance Access publication date 24 September 2009

Abstract

A large fraction of non-coding RNAs is short and/or poorly conserved in sequence. Most of the longer examples, furthermore, consist of a collection of conserved structural motifs rather than 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 in 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. We focus here, in particular, on strategies that can be successful in the ‘twilight zone’ where generic approaches from blast to infernal start to fail.

Keywords: non-coding RNA; homology search; fragmented patterns; telomerase RNA; vault RNA

INTRODUCTION

Noncoding 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 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: the major components of the small and large RNA subunits of nuclear and organellar ribosomes are usually well conserved or at least contain large well-conserved regions.
2. Large, rapidly evolving ncRNAs: telomerase RNA, 7SK RNA and RNAse MRP RNA are the prime representatives of RNAs that evolve rapidly in both sequence and structure, exhibiting dramatic variation in size, and are hard or impossible to detect in phylogenetically distant organisms.
3. 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.
4. Large mRNA-like ncRNAs: this class contains many moderate-size nonprotein-coding transcripts that are spliced and polyadenylated, giving rise to processed RNAs with a size of several kilobase. In this class we also include giant ncRNAs.

Different strategies are required for each of these classes.

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 large subunit (LSU) and small subunit (SSU) rRNA of category (1) are sufficiently large and well conserved even at a kingdom level, so that their genomic locations are easily determined by a blast [1] search using default parameters. Homology search of the messenger RNA like non-coding RNAs (mlncRNAs) in category (4), on the other hand, is largely uncharted territory. We will therefore focus here on the small housekeeping RNAs and their larger counterparts.

The biological contexts in which ncRNA homology search is of practical relevance also come in different flavors and can be roughly categorized into genome surveys and evolutionary surveys. The first scenario corresponds to the setting typically investigated when annotating protein coding genes: here, a single genome is to be screened for a large repertoire of known ncRNA genes. The primary objective is to provide a reliable annotation, so that the focus is on specificity, because the wealth in diversity of the genomic ncRNA complement in practise precludes a very detailed individual expert analysis of each and every candidate.

In the second case, one aims at understanding the evolutionary origins and patterns of a particular RNA gene family across a large time scale in a correspondingly large number of species. A detailed expert analysis of sequence and structure is the main focus in such endeavors. The computational challenge typically arises from a lack of sensitivity, since in particular major changes in the sequence or structure are of particular interest. In principle, any RNA homology search tool is applicable to both type of studies. In practise, however, the different focus brings about subtle differences in the workflow, and in the selection of applicable tools.

In the realm of annotation, tools are meaningfully evaluated by their receiver operating characteristic (ROC) curves on standard test suites, e.g. in well-annotated genomes [2]. For the evolutionary surveys, on the other hand, such evaluation does not make too much sense. The reason is that, almost by definition, we cannot construct a test system that truly represents the task at hand, namely to learn the peculiarities of a specific RNA family and to generalize them beyond all the known examples. One could, of course, set up an evaluation procedure with a training set artificially reduced to a phylogenetically restricted set of examples and ask for ROC curves on the remaining known examples. There is a logical problem, however. We know the (overwhelming majority of) the ‘unknown’ test cases because they have previously been identified by the tools that we attempt to test, i.e. we merely assess the capability of computational methods to find what they have found already.

In this contribution our focus is homology search in the ‘twilight zone’ beyond the sensitivity limits of standard tools such as blast and infernal, and demonstrate how customizable search tools, on the one hand, and the inclusion of extended information such as genomic context and regulatory elements, on the other hand, can be applied successfully.

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 of sequence conservation between query sequence(s) and subject genome. The search for housekeeping ncRNAs is in general much more difficult than the search for protein-coding genes 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 seven or more. The substantial in/del rates as the ones shown for the telomerase RNA in Figure 3 can make it impossible to meet this criterion. The limited accuracy of blast searches has also been noted in a recent benchmark study [2] that was conducted using ncRNA families with fairly benign patterns of sequence and structure conservation.

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 noncoding 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 sub-strings of sufficient length to act as seeds for blast-like approaches. Figure 1 shows the conservation pattern of the U7 snRNA [5] as an example. In many cases, there is little constraint on local sequence patterns. Instead, much of the stabilizing selection may act to preserve secondary structure motifs, such as the tRNA clover leafs. Figure 2 shows the 5' hairpins of metazoan 7SK RNA as an example.

By definition, base-pairing patterns are nonlocal 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 constraints. Since many of the housekeeping ncRNAs are pol III transcripts, we may include additional knowledge about the gene structure. The poly-T terminator signals of these transcripts can also 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.

**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 [5].

**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 [39].

**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 are 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.

Nevertheless, it usually pays to ‘top off’ the seed set by those homologs that can be identified unambiguously by a blast search. The latest version of infernal [13] is also efficient enough to be applicable at this stage already.

Furthermore, there are many cases where 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 quite 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 [14].

**FRAGMENTED PATTERN SEARCH**

The phenomenon of fragmentation is a common theme found in conservation patterns of ncRNA: while few blocks are strongly conserved, other non-conserved regions significantly, in some cases even arbitrarily. 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 (Figure 3), this fragmentation leads to highly divergent sequence lengths, contributing to the difficulty of ncRNA homology search.

**Modeling sequence homology**

As the 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 [15] 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 [2].

The log-odds scores typically implemented in profile HMMs capture homology within conserved regions in a much more significant way. Whenever dealing with well-conserved families of ncRNAs or at least searching across a short evolutionary distance—which will typically be the case in a genome-survey setting—HMMs will be the method of choice for homology search. When dealing with fast evolving families or larger evolutionary gaps, however, one can often observe substantial length variability in unconserved regions, while conservation is restricted to a few isolated islands. 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 lower significance scores. In principle, one could of course represent each gap length as a separate state and manually set the corresponding transition probabilities.

An approach where sequence fragmentation has been explicitly taken into account is implemented in the fragrep2 tool [16]: 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 modeled from a multiple sequence alignment; conserved regions are labeled manually, so that the PFMs as well as bounds for the gap regions can be extracted using the aln2pattern tool. The scoring functions and file formats underlying fragrep2 have been kept explicitly simple, so that search patterns can easily be modified by hand. For querying a model against a genome, matches are reported wherever the conserved blocks match with a prespecified score, and satisfying the upper and lower bounds on the distance between the conserved blocks. As search patterns can be modified and refined, fragrep2
For matching individual PFMs, fragrep2 employs a scoring scheme originally used for matching transcription factor binding sites [17], which has been additionally equipped to allow insertions and deletions. Compared with log-odds scores, fragrep2 scores range in the unit interval, making them particularly accessible for manually adapting the search pattern. While HMMs certainly stand on a statistically more thorough foundation, the intuitive and easily editable scoring functions of fragrep2 has turned out to be a viable alternative whenever homology of the gene to be detected is too weak to be captured by a profile HMM.

The accuracy of PFMs can be improved further by explicitly including phylogenetic information and by adjusting them to the target genome. In [18], a maximum Likelihood procedure is employed to estimate column-wise substitution rates which in turn are used to estimate the PFM for a particular target genome.

**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 [19] or infernal [13, 20, 21], 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 modeled explicitly and cannot be easily approximated.

**Figure 3**: Vertebrate telRNA has been identified as fragmented into eight conserved regions [32, 36]. The lengths of the nonconserved 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 nonconserved regions. Inset: Phylogenetic overview of known (green circle) and unknown (red circle) telRNA genes (TR) and the TERT protein component of telomerase in animals [12]: groups where the corresponding gene is known for some, but not all species are indicated by a yellow circle. The substantial in/del rates in the unconserved regions even within vertebrates (left part of the figure) indicates the difficulty in overcoming the evolutionary gap to Urochordata (in particular *C. intestinalis*) or Protostomia such as *Caenorhabditis elegans* or the honey bee, where both the protein component and the telomere sequence are known.
by simple bounds. In practice, also efficient SCFG-based filters such as RaveNnA [22] are a popular way to reduce the search space to a smaller subset. ERPIN [23] 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 [24] 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 practice, only experts on particular RNA families are capable of constructing descriptors that are sufficiently specific and nevertheless sensitive enough to beat simple blast searches [25]. The efforts to construct good descriptors is thus likely to offset the computational efficiency of the pattern search.

Customized secondary structure modeling
Following the idea of fragrep2 to have easily editable search patterns, 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.

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 [26] 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. Figure 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 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 [41]), we obtain clear hits (marked by circles). For the other two shortest query sequences, stickleback vault RNAs are among the best few dozen hits. The best hit for hgv2 (AF045144) turns out to be a false positive upon closer inspection.
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 [26]. Similarly encouraging results were obtained for *Aspergillus fumigatus* [27].

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.

**FURTHER SIGNALS OF ncRNA GENES**

When dealing with RNA homology search problems that are not readily solved with standard search tools, 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 [28] 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 [29] found vault RNA being part of the syntenically conserved protocadherin cluster, which is syntenically conserved between shark and human [30].

- **Phylogenetic coherence**: naturally, a candidate sequence should be validated if it fits into the phylogeny of its known homologous family members. This validation 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 modeled more precisely than generic modeling and search tools would allow. Modeling of 3′P4 and 5′P4 regions in RNAse MRP [31] may be attributed as such pattern, as well as the H/ACA snoRNA domain in telRNA discussed below.

**CASE STUDIES**

**Vault RNAs in protostomes**

A scenario where fragrep3 proves useful in combination with GotohScan is the annotation of protostome vault RNA (Figure 5). Vault RNAs are small polymerase III transcripts which are difficult to annotate due to their length of only ~100 nt. Until recently, they were only known in mammals, and have been found only recently in other vertebrates and basal deuterostomes [29] 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

![Figure 5](https://academic.oup.com/bfg/article-abstract/8/6/451/215839/figure05)

**Figure 5**: Consensus profiles for protostome vault RNA (vtRNA) (top), along with the basal deuterostome vtRNA identified in [29]. The internal Box A and Box B promoters are highlighted red and green, respectively.
second step, a secondary-structure constrained \texttt{fragrep} pattern was searched against the (several ten-thousands) 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 \textit{Lottia gigantea} genome, which produced a candidate with notably higher homology scores, also exhibiting the expected secondary structure and internal B-Box promoter elements. A simple \texttt{blast} search of \textit{Lottia} candidate against the same genome, finally, revealed a paralogous locus on the same scaffold.

**Telomerase RNA**

Although telRNA is part of the telomerase complex in most of the eukaryotes, it demonstrates a surprisingly large variability in terms of both sequence and secondary structure. This variability is reflected by a length variation ranging from 147 nt in the ciliate \textit{Tetrahymena paravorax} to 1554 nt in the fungus \textit{Candida albicans}. Even within the vertebrates, telRNA length ranges from 321 nt to 541 nt. 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 timescales 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 nt (insects) to 25 nt (saccharomycotina) 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 [32]. This domain is known indeed to share the same function as in snoRNA, namely as a locator within the nucleus [33]. This observation indeed legitimates to borrow strategies from snoRNA annotation tools, such as \texttt{snoReport} [34] or \texttt{snoGPS} [35], 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 also is an useful part of the search process.

The short divergent telRNAs of teleost fishes, for instance, long escaped discovery and were found only recently using \texttt{fragrep2} [36]. Although they are not recognized by the combination of \texttt{blast} and \texttt{infernal} provided at the \texttt{Rfam} web site, they can be rediscovered with the latest version of \texttt{infernal} [13] in genome-wide searches. On the other hand, all attempts to find telRNAs in invertebrate deuterostomes, e.g. in \textit{Ciona intestinalis}, have remained unsuccessful.

**Small nuclear RNAs**

An inherent problem in annotating many of the small nuclear and nucleolar RNA arises from 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 small nuclear RNAs transcribed by pol II, which share a similar promoter structure [37]. These external elements were utilized in systematic surveys of U7 snRNA and 7SK RNA in animal genomes as given below.

**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 [5], these elements were used to set up a homology search pattern, along with a species-specific model for the \textit{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 \texttt{fragrep} search pattern is straightforward. Unambiguous candidates were obtained through filtering the candidates obtained by \texttt{fragrep} using \texttt{RNAbob}.

**7SK RNA**

A similar approach as described for annotating U7 snRNA was successful to annotate 7SK RNA in invertebrate Deuterostomia [38], and subsequently in arthropods [39]. As insights into the functioning of 7SK RNA suggest, a GCC-GCC stem with a loop region crucial for P-TEFb binding were modeled using \texttt{fragrep}; candidates thus obtained were iteratively filtered by the presence of 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, vault RNA candidates reported in [29] were also validated by their pol III 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.

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.

Two conclusions can be drawn from these observations. First, in terms of computational tools, having more customizable search tools whose search patterns can be adapted whenever standard approaches cannot recover expected genes has proven extraordinarily useful in the case studies described here. Considering that major losses or gains of structural elements within relatively short evolutionary time-scales are commonly observed, customizable search patterns appear to be a reasonable choice. Second, in terms of information utilized during the search process, we showed that including aspects such as synteny or promoter elements proved useful in several cases. The gain achieved by including such additional information in the search procedure indicates that including regulatory elements or genomic context in corresponding RNA-related resources such as Rfam may add significant value in other instances of ncRNA homology search. Eventually, a more systematic inclusion of genomic context may become more accessible with the increasing availability of genome-wide alignments. A systematic utilization of synteny will yet require a thorough and evolutionarily dense understanding of whole-genome evolution, which are currently beyond reach. As long as not available in a uniform resource, meticulously assembling family-specific peculiarities into a homology model appear to be the only viable way to cover larger evolutionary gaps.

Key Points
- The evolution of many non-coding RNA families is characterized by large variations in both sequence and structure.
- This variation is typically inhomogeneous so that the structures are composed of conserved cores and sometimes large highly variable expansion domains.
- Computational efforts to detect distant homologs of ncRNAs need to take this variability into account by allowing for fragmented patterns of conserved sequence and structure.

FUNDING
Deutsche Forschungsgemeinschaft (in parts); the European Union through the FP-6 project SYNLET (in parts).

References
15. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the


