Published in final edited form as:
Trends Biotechnol. 2010 January ; 28(1): 9–19. doi:10.1016/j.tibtech.2009.09.006.

De novo prediction of structured RNAs from genomic sequences

Jan Gorodkin[1,*], **Ivo L. Hofacker**[2], **Elfar Torarinsson**[1], **Zizhen Yao**[3], **Jakob H. Havgaard**[1], and **Walter L. Ruzzo**[3,4,*]

[1]Section for Genetics and Bioinformatics, IBHV and Center for Applied Bioinformatics, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

[2]Institut für theoretische Chemie, University of Vienna, Währingerstr. 17, A-1090 Vienna, Austria

[3]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. PO Box 19024, Seattle, Washington 98109, USA

[4]Departments of Computer Science and Engineering and Genome Sciences, University of Washington, Seattle, Washington 98195, USA

Abstract

Growing recognition of the numerous, diverse and important roles played by non-coding RNA in all organisms motivates better elucidation of these cellular components. Comparative genomics is a powerful tool for this task and is arguably preferable to any high-throughput experimental technology currently available because evolutionary conservation highlights functionally important regions. Conserved secondary structure, rather than primary sequence, is the hallmark of many functionally important RNAs, since compensatory substitutions in base-paired regions preserve structure. Unfortunately, such substitutions also obscure sequence identity and confound alignment algorithms, greatly complicating analysis. This paper surveys recent computational advances in this difficult arena, which have enabled genome-scale prediction of cross-species conserved RNA elements, suggesting that a wealth of these elements indeed exist.

Introduction

Non-coding RNAs (ncRNAs) are functional transcripts that do not encode proteins. A handful of examples, such as transfer and ribosomal RNAs, have been well known since the dawn of molecular biology, and probably have existed since the dawn of life. These few examples have critical and deeply central roles, but, ironically, elucidating the full spectrum of ncRNA activity has received relatively little attention. Within the last 10-15 years, however, a number of striking discoveries including RNA interference, microRNAs and riboswitches have demonstrated that RNAs have unexpectedly diverse, sophisticated and important roles in all living organisms, sparking renewed interest in the "modern RNA world" [1]. To hint at the scope of the issue, only 1.2% of the human genome encodes protein [2], but recent data suggest that 90% of the genome is transcribed on one or both

[*]Corresponding authors: Jan Gorodkin (gorodkin@genome.ku.dk), Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Groennegaardsvej 3, 1870 Frederiksberg C, Denmark, Phone: +45 3533 3578, Fax: +45 3533 3042; Walter L. Ruzzo (ruzzo@cs.washington.edu), Computer Science and Engineering, Box 352350, University of Washington, Seattle, WA 98195-2350, USA, Phone: (206) 543-6298, Fax: (206) 543-2969.

strands, to some extent, at some time, in some tissue [3]. The significance of this observation remains unclear and controversial, but a growing body of evidence points to the presence of many functionally important non-coding transcripts [4]. Even if the bulk of the non-coding transcription is “noise,” this still provides a vast substrate upon which natural selection may have acted to generate a multitude of biologically important non-coding RNAs. Thus, even striking ncRNA-related discoveries, such as microRNAs, may only be the “tip of the iceberg.”

Interest in conducting computational screens for functional RNAs has risen with their increased prominence [5]. However, in contrast to protein coding genes, whose regular codon structure provides strong signals aiding their computational recognition within nucleotide sequences, the signals for ncRNAs are subtler. For example, many ncRNAs are believed to be processed out of longer primary transcripts—including introns of protein-coding genes—hence lack features such as proximal promoters in addition to codon structure [1]. There is, however, one general characteristic shared by many (but not all) known ncRNAs: they fold into complex shapes that are critical to function and thus conserved. While prediction of RNA 3D structure is less well developed than protein structure prediction, the prediction of RNA *secondary* structure—the set of intramolecular, largely Watson-Crick, base pairs defining the fundamental units from which the tertiary structure is formed—is reasonably well understood and computationally tractable. Furthermore, secondary structure also tends to be conserved, while primary sequence often evolves more rapidly (see Figure 1) and is of limited use except when searching for close homologs. These facts make secondary structure the key signal to be exploited in ncRNA prediction. Although secondary structure (henceforth, simply structure) prediction is tractable, it is not trivial (*e.g.*, involving interactions between nucleotides at variable and sometimes large distances in the primary sequence), which makes the problem challenging intellectually and computationally [5].

Most RNA will fold into some secondary structure, but given the importance of secondary structure in functional ncRNAs, it is reasonable to ask whether they have more stable structures than random genomic sequences. A negative answer to this question was provided about a decade ago. That is, the folding energy of RNA genes is not easily distinguished from that of appropriately chosen background sequences, such as randomly shuffled versions of the RNA gene itself [6, 7]. Therefore, a simple screening approach that seeks unusually stable genomic segments, will generally not work. MicroRNAs, whose precursors form unusually stable stem-loops, are a notable exception [8, 9]

Although searching for possible structured elements in one nucleotide sequence is generally ineffective, as discussed above, searching in multiple (orthologous) sequences can be highly effective, since evolutionary conservation highlights functionally important regions of all kinds. Importantly, such searches leverage the rapidly increasing body of comparative genomic sequence data. A key issue, however, is that the evolutionary signature of an RNA gene is quite different from that of a protein-coding gene. In particular, as noted earlier, the nucleotide sequence of an ncRNA might evolve relatively rapidly, making identification and alignment of orthologous sequences difficult or impossible, especially when using tools that focus only on sequence. However, patterns of compensating base changes, for example, an

A–U base pair in a human RNA sequence corresponding to a C–G pair in mouse, can provide evidence supporting both the existence of a conserved RNA structure (without requiring conserved sequence; see Figure 1) and insight into the structure itself. Indeed, short of X-ray crystallography, this type of comparative analysis, done carefully by human experts [10], has been the gold standard for RNA secondary structure prediction for more than forty years. In a nutshell, this highlights the key challenges in computational prediction of ncRNAs: to find orthologous regions, expose the common structure therein, and do so both rapidly and accurately.

The comparative approach is important for an additional reason. Over the next few years, we expect that emerging technologies such as high throughput sequencing of RNA (RNAseq [11]) will reveal the transcriptomes of many organisms with unprecedented depth and precision. Yet, given the extensive breadth of genomic transcription now observed at least in mammals, evidence of transcription can no longer be taken as proof of functional importance. Any observed transcript might just be “noise” or incompletely degraded detritus arising from the expression of some nearby, functionally important RNA. Furthermore, experimental protocols will remain limited with regard to the diversity of species, cell types, states, growth and stress conditions that are probed. Consequently, lack of measured expression of a given genomic segment is not proof of lack of function. In contrast, evolutionary conservation strongly suggests functional importance, whether or not expression has already been experimentally verified. This does not deny the value of experimental evidence, of course, nor deny the existence of functionally important species-specific ncRNAs, but merely argues that detection of evolutionarily conserved ncRNAs by comparative genomics is a powerful tool, and belongs in any effort to understand living systems.

Computational search for conserved RNA structure does have certain important limitations. We highlight two of them here, since they color much of what follows. First, these searches are computationally expensive, principally due to the nature of the underlying RNA folding algorithms [12, 13] that need to be applied to these multiple sequences. Even single-sequence folding algorithms have run times that grow as the cube of the sequence length. Applied naively, however fast screening of a sequence of one kilobase is, it will be a thousand times slower for 10 kilobases and a million times slower for 100 kilobases. Hence, all successful programs in this arena are carefully engineered to control runtime, which entails some, hopefully modest, loss of accuracy on long genomic sequences. For example, one simple, widely used strategy is the “sliding window” approach, wherein the genomic sequence is cut into multiple, overlapping, fixed-length segments (“windows”) that are processed separately. This obviously limits the cubic runtime penalty to the length of the window, but, unfortunately, also limits the lengths of discoverable structures and risks arbitrarily truncating them. Even using substantially more sophisticated techniques, genome-scale ncRNA analyses often consume tens to hundreds of computer-years. These high computational costs are one reason why ncRNA gene finding is still in its infancy.

The second significant limitation of the general searches for conserved RNA secondary structures we will describe is more conceptual. It is natural to want to think of each element discovered as a “non-coding RNA gene”, but the truth is more complex since the approaches described here might generate only a partial picture for each ncRNA. For example, technical limitations related to “window” boundaries or splicing could result in partial or fragmentary predictions. More intrinsically, some ncRNAs lack conserved secondary structures, or may have only patches of conserved structure embedded in longer, largely unstructured transcripts. Additionally, conserved, functionally important RNA structures, such as the selenocysteine insertion sequence (SECIS element), are known to exist in messenger RNAs, usually in their untranslated regions. Therefore, identification of RNA structures in genomic data should trigger post-processing steps and follow up experiments to more precisely characterize transcript boundaries (and function!). For reasons of simplicity, however, in the following we will refer to individual conserved structures as “ncRNA genes.”

This review focuses explicitly on computational prediction of non-coding RNA elements by comparative genomics, *i.e*., the discovery of conserved structured elements in multiple genomic sequences. Other methods for *de novo* prediction have succeeded in some contexts (*e.g*., exploiting organism-specific differences in mono- or dinucleotide frequencies of ncRNAs *versus* background [14-16]), but the comparative approach appears to be the most broadly applicable. We will say little about the related ncRNA homology search problem—finding new instances of a particular RNA family given one or more examples—but this equally important task comes with its own set of issues [17], especially the difficulty of finding homologs outside the phylogenetic range of known examples.

From RNA folding to gene finding

Even though RNA structure cannot reliably be detected by merely folding single sequences, the principles obtained from folding single sequences are fundamental and often constitute an implicit part of more elaborate methods. For example, to date, no large-scale RNA structure screens have accounted for so-called pseudoknots, since the underlying RNA folding algorithms do not do so. Without pseudoknots, RNA secondary structure can be represented by nested parentheses. For example, the hairpin structure of the sequence `AAAAUUCGGCAAUUUU` with base pairs exactly between A's and U's can be written as `((((((….))))))`. Although most large RNA structures do contain pseudoknots, they make up a relatively small portion of the overall structure. Algorithms that account for these more general cases are dramatically more complex (and slower) for seemingly modest gains in accuracy.

Folding of sequences is typically performed through energy minimization, whereby different structural elements are associated with experimentally measured energy parameters. Well-known programs for this task are mfold [18] and RNAfold [19]. Rather than minimizing the free energy, folding can also be carried out in a probabilistic framework, wherein the structure with the highest probability is sought. This is achieved through a so-called Stochastic Context-Free Grammar (SCFG), see *e.g*., [20, 21], typically using probabilities extracted from curated rRNA and tRNA alignments [22, 23]. The probabilistic and energetic frameworks are intimately connected, since one expects energetically destabilizing substitutions to be evolutionarily disfavored. Indeed, recent work shows that statistics derived from good structural alignments can be used to improve upon experimentally measured energy parameters [24]. Both energy-based and SCFG based methods are now part of the core inventory of ncRNA gene prediction methods. These principles for folding single sequences are variously employed on multiple sequences as well, often together with schemes that simultaneously search for compensating base changes.

In silico screening for RNA structures

Predicting RNA structure in genomic sequence is of course closely related to the existing methods for RNA structure prediction. As indicated above, multiple sequences are needed to reliably predict RNA structures. Several strategies exist for RNA structure prediction based on multiple sequences [25], which can be loosely categorized as “align-first,” “fold-first” and “joint.” As the name suggests, align-first strategies start by aligning all sequences using standard multiple sequence alignment tools, followed by inference of their presumed common structure. That inference typically proceeds by some combination of folding energy prediction and detection of compensatory base changes, *e.g*., as quantified by high mutual information [20, 22] between pairs of columns in the alignment. These methods work best when sequence conservation is sufficiently low that compensatory changes are not rare, but high enough that they are generally bracketed by well-conserved patches, which constrain the changes to be correctly aligned (and hence visible as a high mutual information score between paired columns). Dual to align-first strategies, fold-first strategies fold individual sequences separately, and then align the *structures* to each other. These strategies can be expected to excel where sequence conservation is low, but structure is clear-cut, stable and well-conserved. Joint strategies, such as approaches based on Sankoff's algorithm [26], simultaneously align and fold their input sequences. In a sense, the align-first and fold-first strategies are aimed at opposite evolutionary extremes. The joint approach, with an appropriate joint model for sequence and structure evolution, subsumes the other two, as shown in [27]. In practice, however, both high computational costs and difficulties in appropriately specifying such an evolutionary model have hindered the joint methods, leaving viable niches for all three approaches. Current approaches for *de novo* genomic screening mostly utilize methods based on the fold-first and joint strategies, as exemplified in Figure 2, in which prediction is based on finding structures within relatively short sequences locally, rather than globally. Meanwhile, fold-first methods have mainly been applied in homology searches [28, 29].

Besides ignoring pseudoknots, another crucial limitation of current genomic screening methods is their inability to cope with structural variation. That is, they assume that the individual RNA sequences share a conserved structure with little variation. In particular, structural inserts, resulting in large length differences, are poorly handled by current programs. Such structural inserts are, however, frequently observed in evolutionarily old ncRNAs, such as RNaseP [30]. Even the state-of-the-art SCFG-based homology search program Infernal deals only indirectly with structural inserts through its ability to match a local substructure [31]. Recent programs for comparison of RNA structure, such as LocARNA [32], point out some ways to address the problem. Additionally, the presence of two nearby structural alignments might suggest an intervening structural insertion/deletion in some of the sequences, a situation that potentially could be detected in a post-processing step. However, none of these approaches fully overcomes this limitation of the current *de novo* screening methods.

In the following we provide an overview of the main methodologies applied for *de novo* RNA structure discovery. These are also listed in Table 1.

Exploiting sequence-based multiple alignments for ncRNA screens

A natural approach for *de novo* structure-based search is to exploit existing sequence based alignments, *i.e*., the align-first strategy outlined above. QRNA [33] was an early tool of this kind. It screened pairwise BLASTn [34] alignments using three probabilistic models: a pair-SCFG for detecting RNA structures, and two pair-HMMs (hidden Markov models) for detecting coding and background sequences, respectively. QRNA screens on sequences from *Escherichia coli* and *Saccharomyces cerevisiae* resulted in the detection of several novel ncRNA candidates, some of which were subsequently verified experimentally [35, 36].

Given the apparent importance of a stable structure, it is reasonable to speculate that the predicted folding energy of a given sequence would be a useful clue for the detection of putative ncRNAs. This appears to be true, but proved difficult to verify, requiring careful statistical analysis and the consideration of both mono- and dinucleotide frequencies, as the latter affect stacking energies [6, 7, 37]. The magnitude of the energy difference between “random” sequence and functional ncRNAs, while generally not sufficient for reliably detecting them based on the genome of a single species, becomes more pronounced in multiple alignments. AlifoldZ [38] introduced this approach. It compared the folding energy resulting from a consensus structure prediction on the native alignment with those of shuffled alignments, *i.e*., those arising from randomly shuffling the columns of the native alignment. More recently, the RNAz framework [39, 40] improved speed and sensitivity by introducing a fast regression to estimate the stability of a native RNA compared to shuffled sequences, together with a new measure for structural conservation and a support vector machine (SVM) classifier to predict whether an alignment contained an ncRNA.

None of the methods discussed so far directly exploit phylogenetic information, *e.g*., the tree structure and branch lengths that relate individual sequences in the alignment. Pfold [41, 42] and EvoFold [43] do so. These methods compare two alternative statistical models—a background model for unstructured sequence *versus* an SCFG model of RNA sequence and structure, both evaluated with respect to the given phylogeny. A section of the alignment is predicted to be an ncRNA if it is more likely under the RNA model than under the background model. In particular, a pattern of substitutions in putatively base-paired columns favoring structure-preserving compensatory changes tends to support the RNA model over the background model.

QRNA, RNAz and EvoFold all use “sliding window” strategies. The main advantage of this approach is speed; its major disadvantage is its potential for misprediction near the window boundaries, including the possibility of missing RNAs that are longer than the windows' overlap. These drawbacks might be ameliorated by adjusting window lengths dynamically, but to the best of our knowledge, this has not been attempted.

The University of California Santa Cruz (UCSC) genome browser provides whole-genome MULTIZ alignments of vertebrates [44]. Initial screens of their 8-way alignments (human, chimp, dog, mouse, rat, chicken, zebra fish and fugu) using RNAz and EvoFold resulted in an unexpectedly large number of ncRNA candidates: 36,000 [40] and 48,000 [43], respectively. In both cases, candidates were widely distributed throughout the genome, as illustrated in Figure 3 for the RNAz screen. One of the EvoFold candidates, the ncRNA gene HAR1F (Human Accelerated Region 1F), was the subject of an in-depth follow-up study [45] that demonstrated its specific expression during cortical development and accelerated evolution along the human lineage. RNAz and EvoFold have since been applied in several other screens of diverse organisms. For example, an RNAz screen on porcine EST (Expressed Sequence Tag) sequences allowed for direct comparison to expression data (ESTs) in 92 cDNA libraries from 35 tissues and different developmental stages [46]. The screen found brain and developmental tissues to host the highest relative numbers of predicted ncRNAs structures (∼2% of genes expressed at detectable levels) whereas ncRNA content for the other tissues amounts to about 1%. A summary of other screens can be found on the 'RNA Structure in Genomic Sequence' supplementary web-page: http://genome.ku.dk/resources/rsgs.

It has proven difficult to estimate the reliability of these predictions. The standard method of comparing predictions to some "gold standard" is problematic, since the available "true positives," *i.e.*, known ncRNAs, while generally well-recovered by these methods, are potentially not representative of yet undiscovered ncRNAs. Furthermore, there is no satisfactory, biologically justifiable set of "true negatives," particularly in light of the widespread transcription observed in mammals. Instead, studies generally estimate their false discovery rate (FDR) by comparison to randomly shuffled sequences or alignments. Based on this approach, the human screens described above were estimated to have FDRs of 15–50% for RNAz and up to 70% for EvoFold [47]. This emphasizes the need for continued development of ncRNA prediction methods, as well as for follow-up experiments. However, even if the true FDR in these studies were 90%, the remaining 10% still constitute thousands of novel ncRNAs in the human genome. One technical point that is worth noting is that the accuracy of the FDR estimation itself will depend on how well the randomized data reflect the complexities of real genomes. For example, the dinucleotide composition of an RNA sequence affects its stacking energy, thus randomizations that do not control for this yield biased FDR estimates [48]. Recent work [48-50] provides procedures that generate randomized multiple sequence alignments that approximately preserve dinucleotide-, gap-, and other statistics observed in the native alignments, thus providing a more appropriate null distribution for FDR estimation, but the FDR estimates reported in the ncRNA screens discussed here generally are not based on these improved methods.

Genomic screening by pairwise local structural alignments

While exploiting multiple alignments saves computational costs, there is the significant disadvantage that alignments might simply be incorrect, which will become increasingly likely as the sequence similarity drops. For example, the sensitivity of AlifoldZ drops sharply for pairwise sequence similarities below 60% [38]. This limitations can, at the expense of computational costs, be addressed within the above-mentioned joint approaches, *e.g.*, the Sankoff framework for simultaneously folding and aligning sequences [26]. Available methods using this strategy include FOLDALIGN [51, 52], Dynalign [53, 54], and LocARNA [32], which are all energy-based, as well as Consan [55] and StemLoc [56], which are SCFG-based. Furthermore, alternatives to the Sankoff framework have also been developed, such as SCARNA [57], which compares stem fragments to construct pairwise structural alignments.

The latest version of Dynalign [54] predicts ncRNAs using its own Sankoff-based pairwise structural alignments [53] and an SVM classifier trained on those alignments. In a pairwise analysis of *E. coli* and *Salmonella typhi*, its additional alignment flexibility appears to have resulted in improved sensitivity compared to RNAz, at least for regions that exhibited low sequence similarity (<50% identity). Using a sliding window approach based on genome-wise BLAST and MUMmer [58] alignments, approximately 1000 novel candidates were reported in these two genomes [54].

To detect remotely homologous ncRNAs, it is desirable to extend such searches to directly perform local structural alignments, rather than relying on purely sequence-based tools such as BLAST. To illustrate this, we have examined a 500 nucleotide region from the human genome containing one known tRNA and a 500 nucleotide region from the distant *Euglena gracilis* chloroplast genome containing three tRNAs. BLAST reports no significant match between them. FOLDALIGN, however, detects one human subregion that structurally aligns to three different locations within the chloroplast sequence, exactly corresponding to the tRNAs therein, despite the low (40–50%) sequence identities among these matching regions (Figure 4). Note the importance, and difficulty, of *local* alignment here, which allows a structure common to a small portion of each input sequence to be detected amidst extraneous flanking sequences; the algorithm must thus identify element boundaries as well as structure. The FOLDALIGN screen takes ~2 minutes, however, compared to ~0.03 seconds for BLAST, further motivating the desire to accelerate these methods.

FOLDALIGN has also been used to screen regions of low sequence similarity between human and mouse [59]. These were selected by choosing corresponding, but unaligned regions as identified via the UCSC genome browser's [60] MULTIZ alignments. A total of 37,000 such regions were screened, each in both forward and reverse directions. The screen resulted in ~1300 candidates, with an FDR of ~50% as estimated from dinucleotide-controlled shuffled data. This particular screen took 5 months of computer time using 70 CPUs (Pentium IV, 2.4 GHz) with an earlier version of FOLDALIGN [61]. The current version of FOLDALIGN is estimated to be 20 times faster (*i.e*., one week on similar hardware), largely due to its ability to discard poor intermediate alignments, but even after this improvement in speed, an all-to-all comparison of two mammalian genomes remains highly impractical. Other applications of FOLDALIGN for genomic sequence analyses can be found at: http://genome.ku.dk/resources/rsgs.

Employing local multiple structural alignments for screening

Using more than two organisms will potentially yield more accurate predictions. Extending the FOLDALIGN pairwise screen, CMfinder [62] was applied to the 44 ENCODE regions [3], which cover 1% of the human genome. CMfinder discovers structured RNA motifs in a set of unaligned sequences. It builds local alignments, *i.e*., flanking regions that do not appear to contain RNA are ignored. Furthermore, entire sequences that do not appear to contain RNA are ignored, which is valuable since the phylogenetic range of a given ncRNA is typically not known in advance. CMfinder builds an initial heuristic local alignment using both sequence- and energy-based seed structures from which it derives a covariance model [20, 22, 63] (a specialized SCFG). The alignment and model are then refined jointly, using an approach akin to that of [22], but extended to local alignments. While the approach is free of any constraints imposed by fixed window sizes and pre-computed alignments, in order to control run time, there are some *a priori*, user-specified restrictions with regard to the complexities of motifs (*e.g*., number of hairpins) that can be discovered. In the CMfinder ENCODE screen[64], precomputed MULTIZ sequence alignments for the ENCODE regions were used, but, in contrast to “align-first” methods discussed earlier, they were only used to indicate orthology, not alignment at the nucleotide level. Repetitive sequences were included, but exons and PhastCons [65] elements were omitted. The average sequence identity of the 56,017 input alignments was 50%, with an average sequence length of 155 nucleotides. The screen resulted in 6587 ncRNA candidates, with an estimated FDR of 50% [64]. About 60% of the candidates were located in non-coding parts of protein-coding genes. Of these, 83% were located in introns, 14% in 5′ UTRs and the remainder in 3′ UTRs, an interesting bias perhaps due to poor annotation of UTRs. A small number of candidates were experimentally tested. Most of these showed tissue-specific expression in humans. For one 67 nucleotide candidate in a 4KB intron of the neuron-specific Synapsin 3 gene, northern blot analysis confirmed expression of a 2.8Kb ncRNA in brain [64].

One key observation from this study was that the lower the identity of the input sequences, the more they were realigned—*i.e*., the structure-driven alignments produced by CMfinder differed more from the MULTIZ sequence-based alignments as sequence identity declined. While not surprising, the extent of this trend suggests that standard alignments might be misleading unless well-conserved in primary sequence. The issue is significant since 25% of candidates had more than 50% of their positions realigned when compared to the original MULTIZ alignments. Furthermore, many regions with identities above 70% were realigned, and even small realignments can have an impact, as illustrated in Figure 5.

The RNAZ, EvoFold and CMfinder screens processed slightly different subsets of the ENCODE regions, but their predictions on the common subset can be directly compared [47]. For example, 330 genomic loci were predicted by both RNAz and EvoFold. This degree of overlap, as well as the other overlaps quantified in Figure 6, is highly unlikely to occur purely by chance, given the lengths of the predictions in comparison to the total size of the ENCODE regions. Nevertheless, this numerically modest level of concordance among methods appears somewhat disappointing. However, further investigation has shown that this was mainly due to different sensitivities in different regimes of sequence composition and sequence similarity [47]. Whereas EvoFold candidates primarily lie in highly conserved, AU-rich regions, RNAz (and CMfinder) candidates tend to be slightly GC-rich and exhibit more sequence variation. We speculate that RNAz and EvoFold might be missing candidates due to their reliance on sequence-based alignments, particularly in diverged regions, whereas CMfinder favors such regions, while ignoring evolutionary relationships exploited by EvoFold, and energetically stable candidates lacking strong covariance patterns favored by RNAz. Thus, overall, these methods can be considered as complementary, with each of them able to predict a fraction of the ∼20,000 unique candidate loci found in total in the ENCODE regions.

Applications of CMfinder to prokaryotes (emphasizing riboswitch discovery) are also listed at: http://genome.ku.dk/resources/rsgs.

Future development of current methods for local structural RNA alignments is highly anticipated. For example, an extension of SCARNA [57], called SCARNA_LM, has been published recently [66]. It creates local multiple alignments using the same principles of comparing fixed size stem fragments that are separated into their 5′ and 3′ parts, respectively. We hope that other methods will follow as well.

Conclusions and perspectives

To date, annotation in the genome databases relates almost exclusively to protein coding genes. In contrast, the computational screens for non-coding RNAs described here suggest that a large number of functional ncRNAs still remain to be found. This is consistent with the observation that most of the non-coding mammalian genome is transcribed. Computational *de novo* discovery of non-coding RNAs within genomic sequences is still in its infancy. Nevertheless, these screens provide a valuable starting point for subsequent studies of specific genomic regions and their functional characterization.

In silico screens are useful as the information generated can be correlated to other studies, or to obtain a more general picture of the putative RNA structure content. Currently, screens essentially only address RNA secondary structure, with the numerous limitations discussed above, leaving much room for improvement. Another major issue is their high computational cost (*i.e*., their time and memory requirements). Their high false discovery rates also make them inappropriate for systematic annotation and necessitate experimental follow-up. There is clearly a demand for faster, better computational approaches. Nonetheless, present methods have already demonstrated their value and have opened an entirely new branch within comparative genomics. For instance, simple extrapolation from the ENCODE studies suggests prediction of ∼2 million RNA structures within the human genome, albeit potentially with many structures per transcript, high FDR and other caveats discussed above. Computational methods should, however, be developed beyond searching for local RNA structures, in part because some ncRNAs might either have no structure (*e.g*., piRNAs [67]), or only partial structures (*e.g*., Evf-1 [68, 69]; see also [70]). Additionally, ncRNAs are known to vary considerably in size, ranging from ∼20 nucleotides to ∼100,000 nucleotides, contributing significantly to the challenge at hand. Fusion transcripts also pose a serious computational challenge that the community has not yet addressed. An example of such a transcript (the cDNA KIAA0510) involving segments from different chromosomes was reported a few years ago [71], a story echoed by the ENCODE project [3]. Experimental approaches continue to lead to novel discoveries. The lincRNAs [72] are one recent example. Identified by a combination of ChIP-Seq and custom tiling array technologies, they include 1600 large, multiexon RNAs, evidently under purifying selection and expressed in multiple mouse cell types. Information obtained from such studies can in turn be incorporated into computational methods, which then might lead to further discoveries of novel ncRNAs. Methods to predict mRNA-like ncRNAs [73] promise to contribute to this. Another challenge is to unravel the significance of conserved structures in UTRs of protein coding genes, which constitute a sizeable fraction of predictions. For example, CMfinder candidate r-6354 ([64], see http://genome.ku.dk/resources/cmf_encode/pages/candidate.php?id=r-6354) found in a 5′ UTR in several mammals, might have a role in regulating its own mRNA, but is also a predicted microRNA precursor. In this situation, comparing to known elements in UTR databases and/or attempting *de novo* motif finding in UTRs are plausible approaches [74, 75].

Putative ncRNAs are available from various sources, including the UCSC browser [60] and (with some degree of overlap) RNAdb [76]. In our online supplement at http://genome.ku.dk/resources/rsgs, we provide links to files defining the genomic locations of many of the ncRNAs candidates described here. These files may be directly passed to the UCSC genome browser, so that the candidates may be viewed in context with other genomic features such as gene predictions and sequence-based multispecies conservation information.

Novel strategies are also emerging for post-clustering of predicted RNA structures. In particular, LocARNA [32], a Sankoff-based approach for structural alignment, grouped structurally related ncRNAs in an RNAz screen of *Ciona intestinalis* and *C. savignyi*, identifying known RNA families, such as tRNAs, microRNAs, and spliceosomal RNAs, and furthermore suggested additional groups consisting of novel, so far unannotated RNAs. Moreover, clustering based on FoldalignM detects novel relationships among microRNA family members [77]. Another clustering approach using RNAforester suggests some of the structural characteristics that distinguish microRNA precursors from the enormous number of other transcribed RNA stem-loops in mammalian genomes [78].

The increased availability of genome sequences has exposed sufficient nucleotide variation to allow the detection of patterns of RNA structure in sequence based-alignments, but has also revealed that sequence-based alignments alone might not be sufficient for detection of ncRNAs. The computational screens performed so far have pushed the limits of comparative genomics and pointed (perhaps unsurprisingly) towards structural alignments as an essential corequisite for detection of ncRNAs. As we have seen, this is not only the case for those orthologous genomic regions whose sequences are poorly conserved across related organisms, but might also be true for small portions of otherwise well conserved regions. This in turn poses new challenges in addressing comparative genomics from an RNA perspective. Furthermore, despite remarkable progress in both computer hardware and algorithms, even faster tools are highly desirable. With this in mind, tools for integrating new genomic sequence with existing *de novo* ncRNA predictions would be valuable, both for annotating the new genome and for strengthening or dismissing previously marginal predictions. As these predictions only need to be carried out in limited regions, they should be fast.

Even though further advances of methods for predicting RNA structures within genomic sequences are expected, the field is also developing tools to predict RNA interactions, with many utilizing the same principles as for RNA folding. Being able to predict RNA interactions will be an essential component in assigning a putative functional context for predicted RNA structures in the future.

Acknowledgments

We thank the anonymous referees for numerous helpful suggestions. This work was supported in part by the Danish research council for Technology and Production, The Lundbeck Foundation and Danish Center for Scientific Computation, NIEHS Grant P30ES07033, and Austrian GEN-AU project "Regulatory ncRNAs."

Bibliography

1. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2001; 2(12):919–929. [PubMed: 11733745]
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431(7011):931–945. [PubMed: 15496913]
3. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306(5696):636–640. [PubMed: 15499007]
4. Mattick JS. The genetic signatures of noncoding RNAs. PLoS Genetics. 2009; 5(4):e1000459. [PubMed: 19390609]
5. Eddy SR. Computational genomics of noncoding RNA genes. Cell. 2002; 109(2):137–140. [PubMed: 12007398]
6. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics. 2000; 16(7):583–605. [PubMed: 11038329]
7. Workman CT, Krogh A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Research. 1999; 27(24):4816–4822. [PubMed: 10572183]
8. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics. 2004; 20(17):2911–2917. [PubMed: 15217813]
9. Lindow M, Gorodkin J. Principles and limitations of computational microRNA gene and target finding. DNA Cell Biol. 2007; 26(5):339–351. [PubMed: 17504029]
10. Pace, NR.; Thomas, BR.; Woese, CR. The RNA World. CSHL Press; 1999. Probing RNA Structure, Function, and History by Comparative Analysis.
11. Cloonan N, Grimmond S. Transcriptome content and dynamics at single-nucleotide resolution. Genome Biol. 2008; 9(9):234. [PubMed: 18828881]
12. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci USA. 1980; 77(11):6309–6313. [PubMed: 6161375]
13. Eddy SR. How do RNA folding algorithms work? Nat Biotechnol. 2004; 22(11):1457–1458. [PubMed: 15529172]
14. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA. 2002; 99(11):7542–7547. [PubMed: 12032319]
15. Schattner P. Searching for RNA genes using base-composition statistics. Nucleic Acids Research. 2002; 30(9):2076–2082. [PubMed: 11972348]
16. Larsson P, Hinas A, Ardell D, Kirsebom LA, Virtanen A, Söderbom F. De novo search for non-coding RNA genes in the AT-rich genome of Dictyostelium discoideum: performance of Markov-dependent genome feature scoring. Genome Research. 2008; 18(6):888–899. [PubMed: 18347326]
17. Menzel P, Gorodkin J, Stadler PF. The Tedious Task of Finding Homologous Non-coding RNA Genes. 2009 submitted.
18. Zuker, M.; Mathews, DH.; Turner, DH. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: Barciszewski, J.; Clark, BFC., editors. RNA Biochemistry and Biotechnology. Vol. 70. Kluwer Academic Publishers; 1999. p. 11-43.NATO ASI Series

19. Hofacker IL, Fontana W, Stadler P, Bonhoeffer LS, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package). Monatshefte für Chemie (Chemical Monthly). 1994; 125:167–188.

20. Durbin, R.; Eddy, SR.; Krogh, A.; Mitchison, G. Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.

21. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004; 5:71. [PubMed: 15180907]

22. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Research. 1994; 22(11):2079–2088. [PubMed: 8029015]

23. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, Haussler D. Stochastic context-free grammars for tRNA modeling. Nucleic Acids Research. 1994; 22(23):5112–5120. [PubMed: 7800507]

24. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics. 2007; 23(13):i19–28. [PubMed: 17646296]

25. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics. 2004; 5:140. [PubMed: 15458580]

26. Sankoff DD. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math. 1985; 45:810–825.

27. Seemann S, Gorodkin J, Backofen R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Research. 2008; 36(20):6355–6362. [PubMed: 18836192]

28. Janssen S, Reeder J, Giegerich R. Shape based indexing for faster search of RNA family databases. BMC Bioinformatics. 2008; 9:131. [PubMed: 18312625]

29. Reeder J, Reeder J, Giegerich R. Locomotif: from graphical motif description to RNA motif search. Bioinformatics. 2007; 23(13):i392–400. [PubMed: 17646322]

30. Marquez SM, Harris JK, Kelley ST, Brown JW, Dawson SC, Roberts EC, Pace NR. Structural implications of novel diversity in eucaryal RNase P RNA. RNA. 2005; 11(5):739–751. [PubMed: 15811915]

31. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25(10):1335–1337. [PubMed: 19307242]

32. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Computational Biology. 2007; 3(4):e65. [PubMed: 17432929]

33. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics. 2001; 2:8. [PubMed: 11801179]

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–410. [PubMed: 2231712]

35. Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol. 2001; 11(17):1369–1373. [PubMed: 11553332]

36. McCutcheon JP, Eddy SR. Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Research. 2003; 31(14):4119–4128. [PubMed: 12853629]

37. Clote P, Ferré F, Kranakis E, Krizanc D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA. 2005; 11(5):578–591. [PubMed: 15840812]

38. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. Journal of Molecular Biology. 2004; 342(1):19–30. [PubMed: 15313604]

39. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA. 2005; 102(7):2454–2459. [PubMed: 15665081]

40. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol. 2005; 23(11):1383–1390. [PubMed: 16273071]

41. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999; 15(6):446–454. [PubMed: 10383470]

42. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research. 2003; 31(13):3423–3428. [PubMed: 12824339]

43. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent WJ, Miller W, Haussler D. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Computational Biology. 2006; 2(4):e33. [PubMed: 16628248]

44. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green E, Haussler D, Miller W. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research. 2004; 14(4):708–715. [PubMed: 15060014]

45. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel AH, Ares M, Vanderhaeghen P, Haussler D. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006; 443(7108):167–172. [PubMed: 16915236]

46. Seemann SE, Gilchrist MJ, Hofacker IL, Stadler PF, Gorodkin J. Detection of RNA structures in porcine EST data and related mammals. BMC Genomics. 2007; 8:316. [PubMed: 17845718]

47. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigo R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler P. Structured RNAs in the ENCODE selected regions of the human genome. Genome Research. 2007; 17(6):852–864. [PubMed: 17568003]

48. Babak T, Blencowe B, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics. 2007; 8:33. [PubMed: 17263882]

49. Gesell T, Washietl S. Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics. 2008; 9:248. [PubMed: 18505553]

50. Anandam P, Torarinsson E, Ruzzo WL. Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. Bioinformatics. 2009; 25(5):668–669. [PubMed: 19136551]

51. Gorodkin J, Heyer LJ, Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Research. 1997; 25(18):3724–3732. [PubMed: 9278497]

52. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Computational Biology. 2007; 3(10):1896–1908. [PubMed: 17937495]

53. Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. Journal of Molecular Biology. 2002; 317(2):191–203. [PubMed: 11902836]

54. Uzilov AV, Keegan J, Mathews DH. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics. 2006; 7:173. [PubMed: 16566836]

55. Dowell RD, Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics. 2006; 7:400. [PubMed: 16952317]

56. Holmes I. Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics. 2005; 6:73. [PubMed: 15790387]

57. Tabei Y, Tsuda K, Kin T, Asai K. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. Bioinformatics. 2006; 22(14):1723–1729. [PubMed: 16690634]

58. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2):R12. [PubMed: 14759262]

59. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Research. 2006; 16(7):885–889. Erratum: Genome Res. 816:1439. [PubMed: 16751343]

60. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser Database: update 2009. Nucleic Acids Research. 2009; 37(Database issue):D755–761. [PubMed: 18996895]

61. Havgaard JH, Lyngsø RB, Stormo GD, Gorodkin J. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics. 2005; 21(9):1815–1824. [PubMed: 15657094]

62. Yao Z, Weinberg Z, Ruzzo WL. CMfinder--a covariance model based RNA motif finding algorithm. Bioinformatics. 2006; 22(4):445–452. [PubMed: 16357030]

63. Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics. 2002; 3:18. [PubMed: 12095421]

64. Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. Genome Research. 2008; 18(2):242–251. [PubMed: 18096747]

65. Siepel A, Bejerano G, Pedersen JS, Hinrichs A, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson R, Gibbs R, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 2005; 15(8):1034–1050. [PubMed: 16024819]

66. Tabei Y, Asai K. A local multiple alignment method for detection of non-coding RNA sequences. Bioinformatics. 2009; 25(12):1498–1505. [PubMed: 19376823]

67. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. Characterization of the piRNA complex from rat testes. Science. 2006; 313(5785):363–367. [PubMed: 16778019]

68. Faedo A, Quinn JC, Stoney P, Long JE, Dye C, Zollo M, Rubenstein JL, Price DJ, Bulfone A. Identification and characterization of a novel transcript down-regulated in Dlx1/Dlx2 and up-regulated in Pax6 mutant telencephalon. Dev Dyn. 2004; 231(3):614–620. [PubMed: 15376329]

69. Kohtz JD, Fishell G. Developmental regulation of EVF-1, a novel non-coding RNA transcribed upstream of the mouse Dlx6 gene. Gene Expr Patterns. 2004; 4(4):407–412. [PubMed: 15183307]

70. Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, Prohaska SJ, Rose D, Stadler PF, Tanzer A, Washietl S, Will S. Athanasius F Bompfünewerer Consortium. RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zool B Mol Dev Evol. 2007; 308(1):1–25. [PubMed: 17171697]

71. Claverie JM. Fewer genes, more noncoding RNA. Science. 2005; 309(5740):1529–1530. [PubMed: 16141064]

72. Guttman M, Amit I, Garber M, French C, Lin M, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein B, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458(7235):223–227. [PubMed: 19182780]

73. Hiller M, Findeiβ S, Lein S, Marz M, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, Stadler PF. Conserved introns reveal novel transcripts in Drosophila melanogaster. Genome Research. 2009; 19(7):1289–1300. [PubMed: 19458021]

74. Pavesi G, Mauri G, Stefani M, Pesole G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Research. 2004; 32(10): 3258–3269. [PubMed: 15199174]

75. Mignone F, Grillo G, Licciulli F, Iacono M, Liuni S, Kersey PJ, Duarte J, Saccone C, Pesole G. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Research. 2005; 33(Database issue):D141–146. [PubMed: 15608165]

76. Pang KC, Stephen S, Dinger ME, Engström P, Lenhard B, Mattick JS. RNAdb 2.0--an expanded database of mammalian non-coding RNAs. Nucleic Acids Research. 2007; 35(Database issue):D178–182. [PubMed: 17145715]

77. Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J. Structural profiles of human miRNA families from pairwise clustering. Bioinformatics. 2009; 25(3):291–294. [PubMed: 19059941]

78. Ritchie W, Legendre M, Gautheret D. RNA stem-loops: to be or not to be cleaved by RNAse III. RNA. 2007; 13(4):457–462. [PubMed: 17299129]

79. Gorodkin J, Knudsen B. RNA Informatik. Naturens Verden. 2000; 11--12:2–9.

80. Sprinzl M, Vassilenko KS. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Research. 2005; 33(Database issue):D139–140. [PubMed: 15608164]

81. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. Rfam: updates to the RNA families database. Nucleic Acids Research. 2009; 37(Database issue):D136–140. [PubMed: 18953034]

82. Yao Z, Barrick JE, Weinberg Z, Neph S, Breaker RR, Tompa M, Ruzzo WL. A computational pipeline for high- throughput discovery of cis-regulatory noncoding RNA in prokaryotes. PLoS Computational Biology. 2007; 3(7):e126. [PubMed: 17616982]

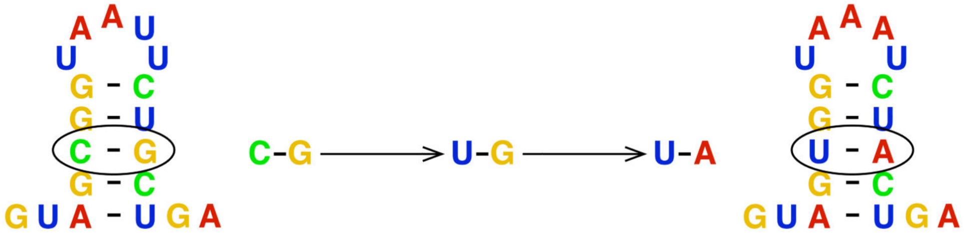


Figure 1. Compensating base changes

Changes in base pairing might preserve structure, but not the primary sequence. In addition to the usual Watson-Crick base pairs, less stable G-U pairs (sometimes called “wobble pairs”) are often seen in RNAs and are evolutionarily important since they allow single base substitutions that are not structurally disruptive. This might allow sequences to accumulate substitutions much more rapidly than would be the case if *both* nucleotides in a base pair needed to be changed more or less simultaneously. Adapted with permission from [79].

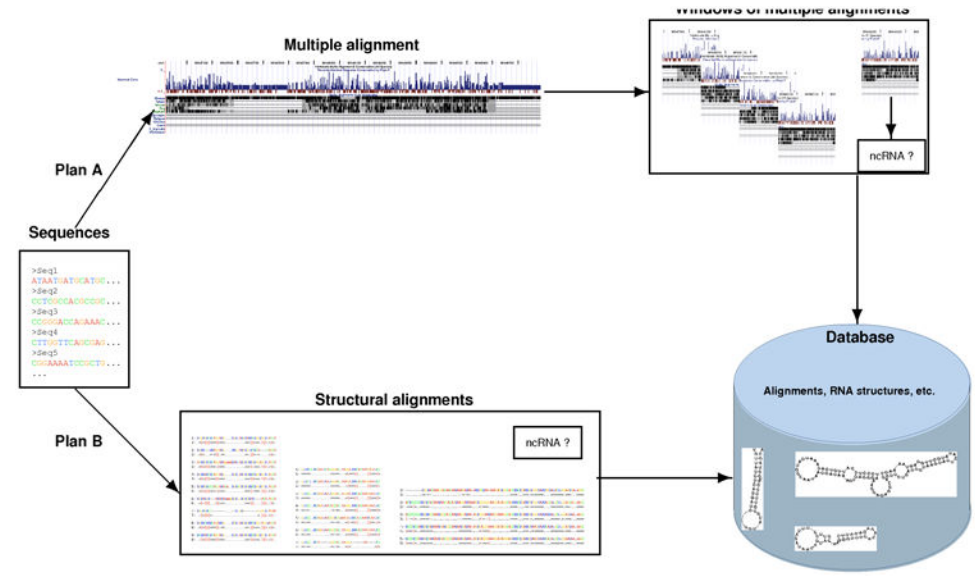


Figure 2. Strategies for ncRNA screening

The upper path illustrates RNA structure prediction using existing sequence alignments that are divided into overlapping windows (an “Align first” strategy). In contrast, in the lower path, “Fold and align,” sequence and structure alignments are performed directly from unaligned sequence data (a “joint” strategy) searching simultaneously for conserved structure and sequence, resulting in structural alignments. To date, alternative “fold-first” strategies have not been applied to genome-scale screening.

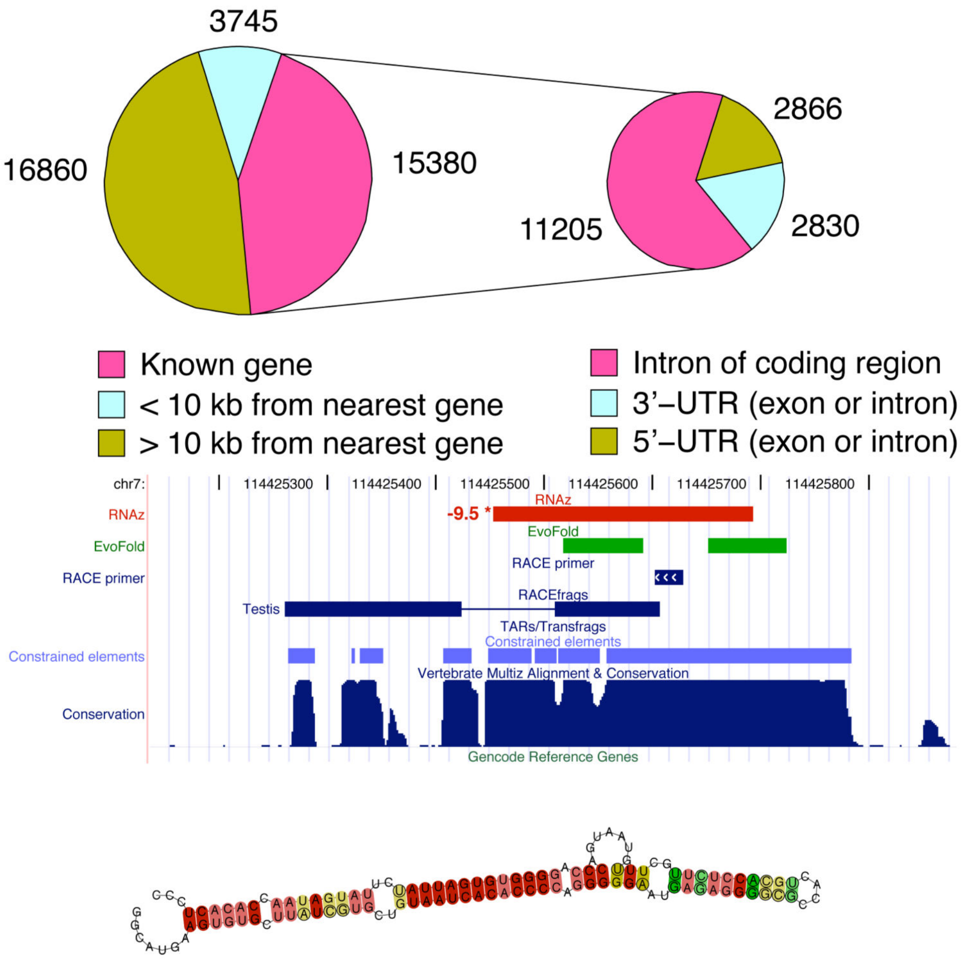
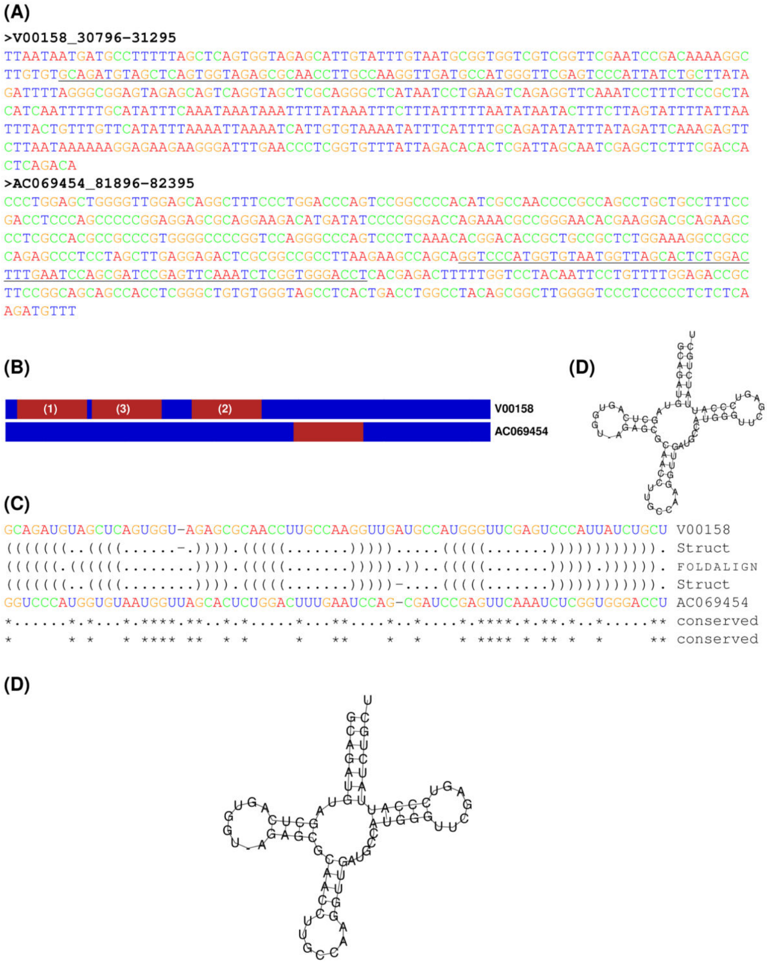


Figure 3. RNAz screen

(A). Genomic locations of RNAz predictions from [40] compared to annotated protein-coding genes. As illustrated here, ncRNA candidates are widely distributed in the genome. For example, 16860 of the RNAz predictions lie more than 10 kilobases from the nearest annotated protein-coding gene. It is plausible that the fraction of ncRNAs in this category is underestimated, since alignments tend to be less reliable in these regions, and this screen's “align-first” strategy depends on them. Of the 15380 predictions that lie within annotated protein-coding genes, the majority (11205) lie in introns flanked by coding exons, but a significant fraction lie in 5′- and 3′-untranslated regions (UTRs), or their introns. (Some candidates were counted in multiple categories due to ambiguous gene annotation. Coding exons were not screened.) (B): Detailed view of a genomic region on human chromosome 7 identified in the ENCODE screen [47]. It contained overlapping predictions from RNAz (red) and EvoFold (green), as well as an unusually high AlifoldZ Z-score (-9.5). In addition, the region corresponds to an island of high conservation. The RACEfrags track verifies that an ncRNA element from this locus is expressed in testis. (C) The predicted structure of a portion of the element is shown underneath. Adapted with permission from [40].

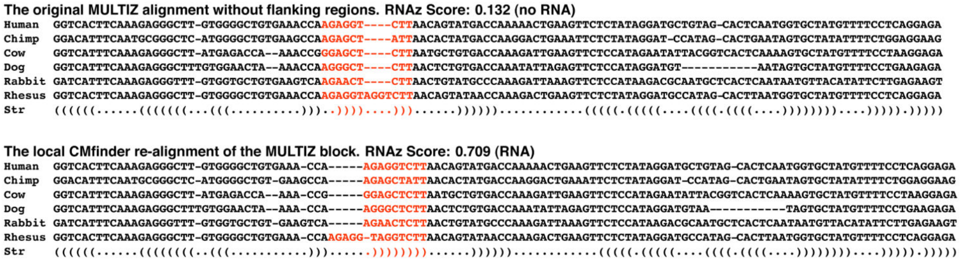


```
>V00158_30796-31295
TTAATAATGATGCCTTTTTAGCTCAGTGGTAGAGCATTGTATTTGTAATGCGGTGGTCGTCGGTTCGAATCCGACAAAAGGC
TTGTGTGCAGATGTAGCTCAGTGGTAGAGCGCAACCTTGCCAAGGTTGATGCCATGGGTTCGAGTCCCATTATCTGCTTATA
GATTTTAGGGCGGAGTAGAGCAGTCAGGTAGCTCGCAGGGCTCATAATCCTGAAGTCAGAGGTTCAAATCCTTTCTCCGCTA
CATCAATTTTTGCATATTTCAAATAAATAAATTTTATAAATTTCTTTATTTTTAATATAATACTTTCTTAGTATTTTATTAA
TTTACTGTTTGTTCATATTTAAAATTAAAATCATTGTGTAAAATATTTCATTTTGCAGATATATTTATAGATTCAAAGAGTT
CTTAATAAAAAAGGAGAAGAAGGGATTTGAACCCTCGGTGTTTATTAGACACACTCGATTAGCAATCGAGCTCTTTCGACCA
CTCAGACA
>AC069454_81896-82395
CCCTGGAGCTGGGGTTGGAGCAGGCTTTCCCTGGACCCAGTCCGGCCCCACATCGCCAACCCCGCCAGCCTGCTGCCTTTCC
GACCTCCCAGCCCCCGGAGGAGCGCAGGAAGACATGATATCCCCGGGACCAGAAACGCCGGGAACACGAAGGACGCAGAAGC
CCTCGCCACGCCGCCCGTGGGGCCCCGGTCCAGGGCCCAGTCCCTCAAACACGGACACCGCTGCCGCTCTGGAAAGGCCGCC
CAGAGCCCTCCTAGCTTGAGGAGACTCGCGGCCGCCTTAAGAAGCCAGCAGGTCCCATGGTGTAATGGTTAGCACTCTGGAC
TTTGAATCCAGCGATCCGAGTTCAAATCTCGGTGGGACCTCACGAGACTTTTTGGTCCTACAATTCCTGTTTTGGAGACCGC
TTCCGGCAGCAGCCACCTCGGGCTGTGTGGGTAGCCTCACTGACCTGGCCTACAGCGGCTTGGGGTCCCTCCCCCTCTCTCA
AGATGTTT
```

```
GCAGAUGUAGCUCAGUGGU-AGAGCGCAACCUUGCCAAGGUUGAUGCCAUGGGUUCGAGUCCCAUUAUCUGCU V00158
(((((((..((((......-.)))).(((((.......))))).....(((((.......)))))))))))). Struct
((((((.((((((........)))).(((((.......))))).))..(((((.......))))).)))))). FOLDALIGN
(((((((..((((........)))).(((((.......)))))-....(((((.......)))))))))))). Struct
GGUCCCAUGGUGUAAUGGUUAGCACUCUGGACUUUGAAUCCAG-CGAUCCGAGUUCAAAUCUCGGUGGGACCU AC069454
*......*.*...*.****.**..*.*.....*...**....*..*....*.****.*.**.*..*.....** conserved
*      * *   * **** **  * *     *   **    *  *    * **** * ** *  *     ** conserved
```

Figure 4. The value of local, structural alignments

(A) Two genomic sequences of 500 nucleotides are shown. This segment of V00158, from *Euglena gracilis* chloroplast, contains three annotated tRNAs; AC069454, from human chromosome 17, contains one. The sequences are relatively dissimilar, as is apparent from the nucleotide color-coding; *e.g*., the former has a G+C content of 35% while the human sequence is 63% G+C. When comparing the sequences *via* standard pairwise BLAST (using default parameters) the result is "No significant similarity found." (B) Comparing the sequences using FOLDALIGN [51, 52], a local, structural alignment tool, results in three pairwise structural alignments of the red region in human to the three indicated *E. gracilis* regions, each corresponding to an annotated tRNA. The numbers in parentheses give the FOLDALIGN score ranking. The sequence identities of the three pairwise alignments are (1) 45%, (2) 48% and (3) 40%, respectively. (C) Secondary structure, as annotated in the tRNA database [80], of the human tRNA. (D) The structural alignment of region (3) with the human tRNA. The "Struct" lines give the structures of these two tRNAs, again from [80], where "matching" pairs of nested parentheses indicate canonical base pairs, dots indicate unpaired nucleotides, and dashes represent gaps. The middle line shows the consensus RNA structure predicted by FOLDALIGN from this pair of regions, which is in excellent agreement with [80].



The original MULTIZ alignment without flanking regions. RNAz Score: 0.132 (no RNA)

```
Human  GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGT----CTTAACAGTATGACCAAAAACTGAAGTTCTCTATAGGATGCTGTAG-CACTCAATGGTGCTATGTTTTCCTCAGGAGA
Chimp  GGACATTTCAATGCGGGCTC-ATGGGGCTGTGAAGCCAAGAGCT----ATTAACACTATGACCAAGGACTGAAATTCTCTATAGGAT-CCATAG-CACTGAATAGTGCTATATTTTCTGGAGGAAG
Cow    GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAACCGGGAGCT----CTTAATGCTGTGACCAAAGATTGAAGTTCTCCATAGAATATTACGGTCACTCAAAAGTGCTATGTTTTCCTAAGGAGA
Dog    GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAACCAAGGGCT----CTTAACTCTGTGACCAAATATTAGAGTTCTCCATAGGATGT-----------AATAGTGCTATGTTTTCCTGAAGAGA
Rabbit GATCATTTCAAAGAGGGTTT-GTGGTGCTGTGAAGTCAAGAACT----CTTAACTGTATGCCCAAAGATTAAAGTTCTCCATAAGACGCAATGCTCACTCAATAATGTTACATATTCTTGAGAAGT
Rhesus GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGTAGGTCTTAACAGTATAACCAAAGACTGAAGTTCTCTATAGGATGCCATAG-CACTTAATGGTGCTATGTTTTCCTCAGGAGA
Str    ((((((......(((((((...(((..........)))..))))....)))......))))))............(((((.(((((....((((.((((....))))))))....))))).)))))
```

The local CMfinder re-alignment of the MULTIZ block. RNAz Score: 0.709 (RNA)

```
Human  GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCA-----AGAGGTCTTAACAGTATGACCAAAAACTGAAGTTCTCTATAGGATGCTGTAG-CACTCAATGGTGCTATGTTTTCCTCAGGAGA
Chimp  GGACATTTCAATGCGGGCTC-ATGGGGCTGT-GAAGCCA-----AGAGCTATTAACACTATGACCAAGGACTGAAATTCTCTATAGGAT-CCATAG-CACTGAATAGTGCTATATTTTCTGGAGGAAG
Cow    GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAA-CCG-----GGAGCTCTTAATGCTGTGACCAAAGATTGAAGTTCTCCATAGAATATTACGGTCACTCAAAAGTGCTATGTTTTCCTAAGGAGA
Dog    GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAA-CCA-----AGGGCTCTTAACTCTGTGACCAAATATTAGAGTTCTCCATAGGATGTAA-----------TAGTGCTATGTTTTCCTGAAGAGA
Rabbit GATCATTTCAAAGAGGGTTT-GTGGTGCTGT-GAAGTCA-----AGAACTCTTAACTGTATGCCCAAAGATTAAAGTTCTCCATAAGACGCAATGCTCACTCAATAATGTTACATATTCTTGAGAAGT
Rhesus GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCAAGAGG-TAGGTCTTAACAGTATAACCAAAGACTGAAGTTCTCTATAGGATGCCATAG-CACTTAATGGTGCTATGTTTTCCTCAGGAGA
Str    ((((((......((((((((..(((...........)))......))))))))......))))))............(((((.(((((....((((.((((....))))))))....))))).)))))
```

Figure 5. Sequence-based alignments can obscure ncRNA structure

Even small changes in a sequence-based multiple alignment can significantly affect the recognition of ncRNAs. The upper panel shows a six-species MULTIZ alignment of a region containing a box H/ACA snoRNA annotated in Rfam [81] (accession RF00402) that has been downloaded from the UCSC genome browser[1]. Using RNAz to search for ncRNAs in this alignment resulted in a low score of 0.132, suggesting "no RNA." In the De novo prediction of structured RNAs from genomic sequences CMfinder. Although only a few positions are realigned (key changes indicated in red), RNAz gives this structurally revised alignment a relatively high score of 0.709, correctly suggesting the presence of a conserved RNA structure. Matching parentheses in the structure lines (Str; bottom line in each group) indicate base pairs in RNAz's consensus secondary structure predictions.

[1] http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&db=hg18&position=chr3:180369542-180369721&rnaGene=pack&ensGene=pack. This human sequence is annotated as a snoRNA pseudogene in the UCSC browser, although both it and its cow ortholog are high-scoring members of the manually annotated Rfam "seed" alignment for this family, and the other four sequences are included in the automatically generated Rfam "full" alignment.

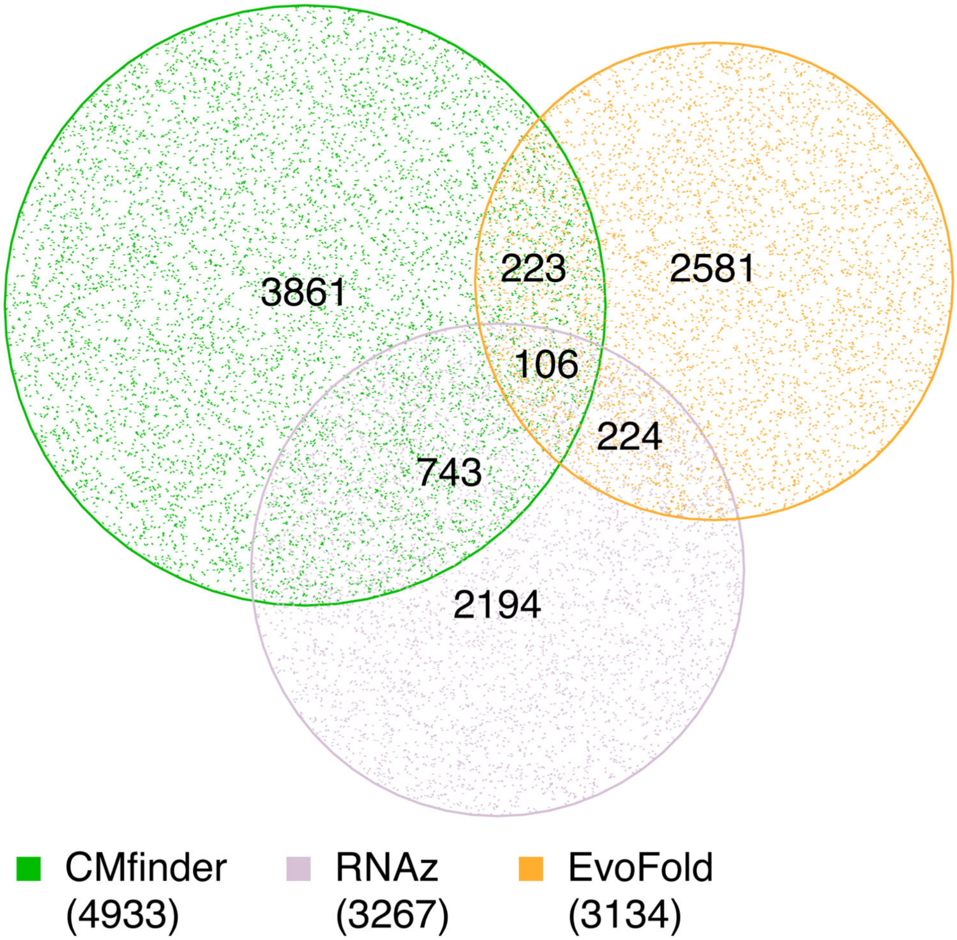


Figure 6. Comparison of ENCODE scans

The Venn diagram compares RNA elements in the ENCODE regions predicted by the three screening methods discussed here, CMfinder, RNAz, and EvoFold. Only predictions from input data that are common to all three studies are tallied, *i.e*., repeat, exon and PhastCons regions are excluded [64]. As illustrated here, of the 4933 candidates reported by CMfinder, 3861 were reported by neither of the other methods, while 106 candidates were reported by all three. Adapted with permission from [64].

Table 1

A summary of the main screens described here.

Method[2]	Organisms[3]	#Cand[4]	Experimental Verification	Size[5]	Cons[6]	Constraints
QRNA [33, 35] (pair-SCFG)	Bacteria	275	11 of 49 randomly sampled; northern	200	Seq	Sliding windows; sequence based alignments
RNAz [39, 40] (Energy, SVM)	Vertebrates	46,000	3 of top 5; northern	200	Seq	Sliding windows; sequence based alignments
	ENCODE	7,093	43 of 175; RT-PCR[6]			
EvoFold [43] (phylo-SCFG)	Vertebrates	48,500	(subsequent studies)	750 (<50)	Seq	Sliding windows; sequence based alignments
	ENCODE	9,953	43 of 175; RT-PCR[6]			
FOLDALIGN [52, 59, 61] (Sankoff)	Human-mouse	1,300	4 of top 12; Northern	200	Str	Motif size and type in local pairwise structural alignments
Dynalign [53, 54] (Sankoff)	Bacteria	1,200	none reported	150	Seq	Sliding windows; HMM based alignment constraints
CMfinder [62, 64, 82] (SCFG, EM)	Bacteria	1,466	(subsequent studies)	None (<200)	Str	Motif size and type in local multiple structural alignments
	ENCODE	6,587	10 of top 11, RT-PCR; 1 of 10, northern			

[2] Name and core methodology

[3] Genomes screened; "ENCODE" means ENCODE regions of 17-way vertebrate alignments

[4] Number of predicted RNA structures

[5] Approximate maximum size of motifs; number in parenthesis is typical size of motifs

[6] Type of conservation exploited to produce alignments (Seq means sequence-level; Str means structure-level)

[6] This number is joint for RNAz and EvoFold.