A practical guide to the art of RNA gene prediction

Irmtraud M. Meyer

Abstract

This review introduces the different strategies and computational methods that can be used in order to predict RNA genes. It discusses our current view of RNA genes as well as recent computational analyses of RNA genes and concludes with an outlook to future directions in algorithm development and data analyses.

Keywords: gene prediction; RNA genes; non-coding genes; ncRNA; RNA secondary structure; transcriptomics

INTRODUCTION

RNA genes are genes that do not encode a functional protein–product. They are thus also called non-coding genes (ncRNAs). The chemical and biophysical properties of RNA [1] make it ideal for regulating very diverse processes in the cell and for serving as a 'communication-layer' [2, 3] between a genome and the variety of its expressed products. The view that is emerging today is that RNA genes play diverse and functionally important roles in the cell and that they deserve the same attention as protein-coding genes.

The first RNA genes to be discovered were transfer RNAs (tRNAs) [4–6] and ribosomal RNAs (rRNAs). Both types of RNAs play pivotal roles in protein synthesis, both assume well-defined structures which are crucial for defining their function and both are highly conserved across bacteria, archaea and eukaryotes, not only in terms of structure, but also function. These early findings have significantly contributed to the initial view that there are few RNA genes which encode highly structured RNA molecules whose functional role is to assist the synthesis of proteins. Since then, many exciting discoveries have lead to a revision of this initial view.

The discovery of two RNAs with catalytic properties, so-called ribozymes, in the early 1980s,

namely the self-splicing group I intron in the 26S ribosomal RNA of *Tetrahymena thermophila* [7] and ribonuclease P in *Escherichia coli* [8], promoted RNA from a passive bystander to an active player in the cell. Recent, high-resolution X-ray studies of the structure of the ribosome [9–11] further support this view by showing that the ribosome's functional properties, in particular the all-important peptide bond synthesis, are due to the ribosomal RNAs and not to the protein components which serve as a structural scaffold.

Since the discovery of splicing in 1977 [12, 13], it has been shown that small nuclear RNAs (snRNAs) are essential for nuclear pre-mRNA splicing, in particular the U1 snRNA [14] and U2 snRNA [15] which bind complementary consensus sequences at the 5′ splice site and the branch site, respectively, at the beginning of several carefully orchestrated splicing steps. This type of interaction via complementary base-pairing between the U1 snRNAs and the pre-mRNA was first predicted by theoretical studies [16, 17] before it was confirmed by experiments.

Especially the last few years have seen an explosion of breakthroughs in RNA research.

Small structural RNA elements in the exons of protein-coding transcripts, so-called ribo-switches [18–20], measure the concentration of

Corresponding author. Irmtraud M. Meyer, Bioinformatics Centre and Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, BC, Canada V6T 1Z4. Tel: +1-604-8274222; Fax: +1-604-8225485; E-mail: irmtraud.meyer@cantab.net

Irmtraud M. Meyer is a faculty member at the Bioinformatics Centre and the Department of Computer Science at the University of British Columbia in Vancouver, Canada. She develops novel methods and algorithms in order to improve RNA gene prediction and our understanding of gene regulation on RNA level, both for protein-coding and RNA genes.

small metabolites by switching from one structural confirmation to another one upon binding or release of a metabolite molecule which is recognized with high specificity. Depending on the gene involved, this mechanism can cause premature termination of transcription, activate or repress translation or result in the cleavage of the mRNA by activating the mRNA's self-cleavage activity. The latter mechanism represents a clever combination of catalytic and structural activity, called a ribozyme-riboswitch [21]. So, protein-coding genes with riboswitches can control their own gene expression on RNA level.

The discovery of the RNA interference pathway (RNAi) in 1998 [22] has resulted in the discovery of many small RNA genes, so-called microRNA genes, that regulate the expression of other genes on mRNA level. The primary transcripts of these genes (pri-mRNAs) are processed into shorter, stem–loop forming transcripts of about 70 nucleotides length (pre-mRNA) which are exported from the nucleus to the cytoplasm, where they are converted into mature miRNAs of about 20 nucleotides length by the endonuclease Dicer which also initiates the formation of the RNA-induced silencing complex. A miRNA can down-regulate the expression of one or more protein–coding genes on RNA level by binding a partially complementary stretch of the mRNA which is then either cleaved and degraded or prevented from being translated. MicroRNAs were first detected in the nematode *Caenorhabditis elegans* [23–27], but have by now been also found in several organisms, including plants [28, 29], animals [30] and human [31, 32].

These experimental studies of specific families of RNA genes have recently been complemented by several genome-wide transcriptome studies employing different experimental techniques: (i) high-density tiling array studies of the human genome [33–35], (ii) cDNA studies of the human [36] and the mouse genome [37–40] and (iii) mapping studies of transcription factor binding sites in the human genome [41]. These experimental studies and their accompanying theoretical analyses, see the original papers and [42], all conclude that a large part of the genome is transcribed into transcripts which do not appear to encode proteins.

It remains to be shown which of these transcripts are functional and which ones correspond to 'transcriptional noise' [43]. More detailed experimental studies of large scale mouse cDNAs data [39],

using a combination of reverse transcriptase-dependent PCR, microarray and Northern blot analyses, 'provide strong support for the conclusion that ncRNAs are an important, regulated component of the mammalian transcriptome'. Several recent theoretical surveys that scan the human genome with computer programs in order to identify genomic locations that may overlap structural RNA genes [44–46] also provide some evidence that RNA genes may be much more abundant that previously thought.

What are RNA genes?

An RNA gene is a gene whose functional product is an RNA rather than a protein. An RNA gene thus corresponds to a contiguous sub-sequence of the genome which corresponds to the un-spliced version of its functional transcript.

The following examples and Figure 1 illustrate how the functional transcripts of RNA genes can relate to the initial transcripts of the genome. They show, in particular, that there is generally no one-to-one correspondence between the transcript that is initially transcribed from the genome and the functional transcript of an RNA gene.

RNA genes can be encoded in introns of protein-coding genes

Small nucleolar RNAs (snoRNAs) are small RNA molecules that guide the methylation or pseudo-uridylation of ribosomal RNAs. Human snoRNAs have been found to be encoded in the introns of protein-coding genes which are transcribed by polymerase II. The human U15A snoRNA, for example, resides in the most 5′ intron of the ribosomal protein S3 gene [47] thereby providing a mechanism for ensuring a balanced stoichiometry for different nucleolar components. Other examples are snoRNAs encoded in the introns of the human cell cycle regulatory gene RCC1 [48]. Further studies showed that these snRNAs are excised from already spliced and de-branched introns of the RCC1 pre-mRNA by exo-nucleolytic processing [49]. See Figure 1D.

Several RNA genes can be derived from the same transcript

Transfer RNAs (tRNAs) are short RNA molecules ranging in size from 73 to 93 nucleotides that transfer a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during
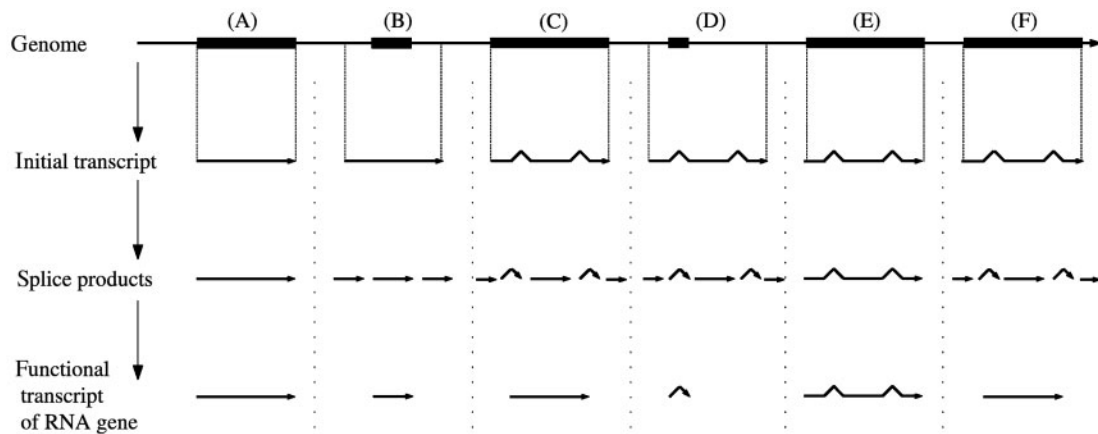
Figure I: Examples which show how the functional transcript of the RNA gene is related to the region that is initially transcribed from the genome. Black boxes in the genome indicate the RNA genes. Typically, situation (A) is assumed, i.e. the initial transcript of the genome is not spliced and corresponds to the transcript that exerts the functional role. However, we know by now of several experimentally confirmed RNA genes which correspond to more complex situations, see the text for examples. (B) shows an RNA gene whose functional transcript is excised from a longer initial transcript which encodes several RNA genes. (C) corresponds to an RNA gene with introns. (D) depicts an RNA gene which is derived from an intron of another gene. (E) and (F) correspond to RNA genes which have been derived from protein-coding genes. These genes are still transcribed, but no longer translated into a protein product. The gene may have lost (E) or retained (F) its ability to be spliced.

translation. They assume a characteristic L-shaped structure in the cell [5, 6]. The functional transcripts of many tRNA genes, both in eukaryotes and prokaryotes, are part of a larger precursor transcript, see Figure 1B. The *E. coli* genome encodes, for example, an operon that contains seven tRNA genes [50]. The functional transcripts of these tRNA genes are obtained by processing reactions that involve both the removal of nucleotides and, in some instances, the addition of nucleotides.

RNA genes can have introns
In mammals, dosage compensation between males and females is achieved by in-activating one of the two X chromosomes in females. This silencing is initiated by the expression of an RNA gene called Xist [51, 52]. The human Xist gene encodes a large (around 17 000 nucleotides), alternatively spliced and poly-adenylated transcript which does not exhibit any conserved open reading frame. The spliced transcript of Xist has been shown to coat the inactive X chromosome. The Xist gene is thus an example for an RNA gene whose spliced transcript corresponds to the functional transcript, see Figure 1C. Another example for RNA genes with introns are some tRNA genes in chloroplasts and cyanobacteria whose introns are group I introns and thus capable of self-splicing [53].

RNA genes can correspond to pseudo-genes of formerly protein-coding genes
Hirotsune *et al.* [54] studied a transgene-insertion mouse mutant and found that this insertion reduces the transcription of the pseudo-gene of Makorin1, called Makorin1-p1. The reduced expression of Makorin1-p1 de-stabilizes in turn the mRNA of Makorin1 in *trans* via a *cis*-acting RNA decay element in the 5′ side of the Makorin1 transcript that is homologous to the Makorin1-p1 transcript. The pseudo-gene thus regulates the gene expression of the corresponding protein-coding gene on mRNA level and could therefore legitimately be called an RNA gene [55]. See Figure 1E and F.

Summary
The current view of RNA genes is that they form a rather heterogeneous group of genes which fulfill diverse functional roles using diverse mechanisms. The features that are most relevant in the context of RNA gene prediction can be summarized as follows:

(i)   The functional transcript of an RNA gene need not correspond to the entire RNA sequence

that is initially transcribed from the genome. In addition, one initial transcript of the genome can give rise to several functional transcript of different RNA genes. To summarize, there is generally no one-to-one correspondence between the initial transcript of the genome and the functional transcript of an RNA gene.

(ii) RNA genes can be transcribed by polymerase II or polymerase III, their transcripts can be poly-A+ or poly-A− (or both at different stages) and these transcripts can have a wide range of lengths, from a few nucleotides to tens of thousands of nucleotides.

(iii) RNA genes can overlap protein-coding genes on the same strand in the genome. The transcript of an RNA gene may be derived from the transcript of a protein-coding gene, e.g. one of its introns.

(iv) RNA genes do not necessarily function by assuming a distinct structure, but employ a number of diverse mechanisms (which are not mutually exclusive): complementary binding to other RNA or DNA sequences in *cis* or in *trans*, sequence or structure specific binding to proteins and other molecules like metabolites, catalytic self-cleavage etc.

(v) RNA genes can correspond to remnants of protein-coding genes that can still be transcribed, but no longer translated.

Several definitions relating to RNA structure

It is important to realize that RNA structure can, but need not play a role in exerting the molecule's function in the cell, as the above examples illustrate. There are also examples, where sequence *and* structure features play functional roles. For example, both the structure and the sequence in the anti-codon loop are required for the proper functioning of tRNA molecules. In order to understand how RNA structure is modeled by computer programs, we first have to introduce a few definitions.

An RNA molecule can form RNA–RNA interaction in *cis* via hydrogen-bonds by folding back onto itself. These hydrogen bonds are weak compared to the covalent bonds that define the sequence of RNA nucleotides and involve pairs of non-consecutive nucleotides which are complementary to each other. The three so-called canonical or consensus base pairs are $\{A, U\}$, $\{G, C\}$ and $\{G, U\}$. It turns out that many important properties of the three-dimensional molecule can already be studied if we only know

the RNA sequence and the sequence positions that form base-pairs, i.e. the so-called secondary structure. This is the level of abstraction that is typically chosen to study the structures of RNA genes. The three-dimensional RNA structure is often either unknown or difficult to predict. Figure 2 shows the secondary structures of two naturally occurring RNAs. The left hand side of the figures shows a tRNA structure, once as two-dimensional figure (top) and once in the equivalent dot-bracket or Vienna notation (bottom). The right hand side of Figure 2 shows the structure of the human telomerase. Unlike the left structure, this secondary structure contains a so-called pseudo-knot, i.e. base-pairs which are not nested. In order to depict pseudo-knotted structures in the dot-bracket notation in an un-ambiguous way, several types of brackets are required, as this example shows. We include pseudo-knots into the set of secondary structures.

From the set of canonical pairs above, it is clear that a given RNA sequence has many potential structures. In fact, the number of possible structures grows exponentially with the length of the RNA sequence. The challenge for the computational biologist is to find out whether structure plays a functional role for a given RNA sequence and, if yes, to predict this functional RNA structure, i.e. the structure which is realized in the cell and which confers the observed functional property to the molecule. Some structure prediction program, e.g. the well-known programs MFOLD [56–58] and RNAFOLD of the Vienna package [58–62], take a given RNA sequence and predict the most stable structure. More precisely, they predict the pseudo-knot free secondary structure that minimizes the overall free-energy of the molecule in thermo-dynamic equilibrium, i.e. the so-called minimum free-energy (MFE) structure. As RNA sequences typically comprise nucleotides of all four types A, C, G and T, there is always the possibility of forming some consensus base-pairs and of combining these base-pairs into an RNA structure. These MFE methods thus predict an MFE structure for almost any RNA sequence. It is important to note that transcripts of RNA genes do not necessarily assume the MFE structure in the cell. Biological processes that happen while the transcript is synthesized and processed in the cell [63], the kinetics of the folding process [64–69], especially for long transcripts, and molecules binding to the RNA sequence can
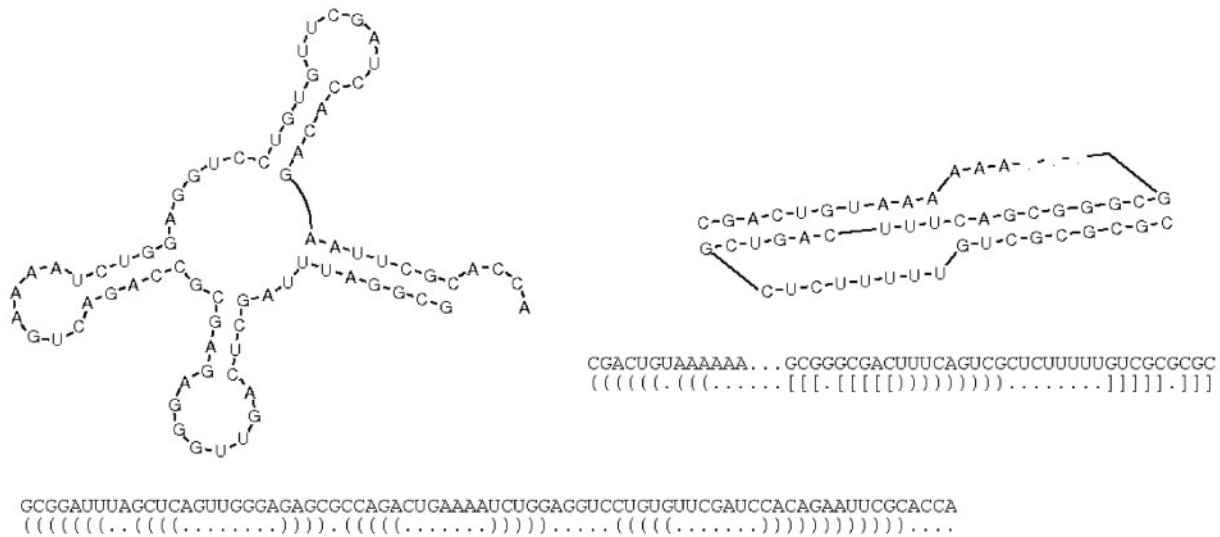
```
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAAAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
(((((((..((((.........)))).(((((.......)))))....(((((.......))))))))))))....
```

Figure 2: The left hand side of the figure shows the secondary-structure of a tRNA structure, once as two-dimensional figure (top) and once in the equivalent dot-bracket or Vienna notation, where a dot denotes an un-paired nucleotide and an opening or closing bracket a paired nucleotide (bottom). By reading the Vienna notation from left to right, one can un-ambiguously determine the base-paired sequence positions. The right hand side of the figure shows the secondary structure of the human telomerase [134]. Unlike the left structure, this secondary structure contains a so-called pseudo-knot, i.e. base-pairs which are not nested. In order to depict pseudo-knotted structures in the dot-bracket or Vienna notation in an un-ambiguous way, different types of brackets are required, as this example shows. We include pseudo-knots into the set of secondary structures. Both drawings were generated with PSEUDOVIEWER [135].

prevent an RNA sequence from assuming the thermodynamically most stable structure, i.e. the MFE structure. In those situations, MFOLD and RNAFOLD are not the appropriate methods for predicting the functional RNA structure as their implicit assumptions are not fulfilled. The comprehensive analysis by Gardner and Giegerich [70] shows that comparative methods tend to systematically outperform MFE methods in predicting the functional RNA structure.

As the examples of RNA genes above illustrate, RNA genes do not necessarily function via structure and there are also RNA genes, e.g. Xist, where a local rather than a global RNA structure plays a functional role. The term 'structural RNA gene' has been recently introduced to refer to RNA genes that have enough RNA secondary structure to be detected by some structure–searching methods [44–46]. Here, we use the term 'structural RNA gene' to denote RNA genes for which (global or local) structure plays a functional role. Likewise, 'unstructured RNA gene' denotes an RNA gene for which RNA structure plays no functional role.

GOALS OF RNA GENE PREDICTION
The main goals of RNA gene prediction are:

(i) To identify the sequence units of the genome that function as RNA genes.
(ii) Once the RNA gene and its functional transcript has been identified, to predict the function of the gene, to identify its interaction partners and to elucidate the mechanism by which it functions.

In order to achieve the first goal, we have to identify features which distinguish RNA genes from other regions in the genome. At the moment, it is not clear what the common characteristic features of all RNA genes are. In contrast to protein–coding genes, we cannot search for start and stop codons, splice sites and regions with coding potential that can be combined into a valid open reading frame.

STRATEGIES FOR PREDICTING RNA GENES
There are many different strategies for predicting RNA genes, see Figure 4 for an overview.

Homology based prediction of RNA genes (Case 1)

Many families of RNA genes appear in a wide range of evolutionarily related genomes. For example, tRNAs and rRNAs in mouse and human not only have similar RNA sequences, but also assume similar structures for exerting their function in the organism.

Some of the most powerful computational methods for predicting RNA genes today make use of this fact and employ a so-called comparative approach which simultaneously analyses several evolutionarily related input sequences. The comparative approach is the best way to detect sequence and structure features that have been conserved during evolution and that are therefore likely to play a functional role, see Figure 3.

Transcript and functional structure of RNA gene are known (Case 1a)

If we know several members of the same functional RNA gene family and their functional structures from the same of from evolutionarily related organisms, it makes a lot of sense to capture the characteristic features in a dedicated computational
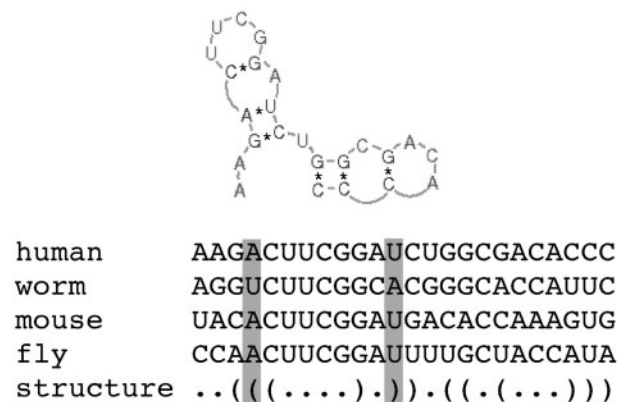


```
human     AAGACUUCGGAUCUGGCGACACCC
worm      AGGUCUUCGGCACGGGCACCAUUC
mouse     UACACUUCGGAUGACACCAAAGUG
fly       CCAACUUCGGAUUUUGCUACCAUA
structure ..(((....).)).((.(...)))
```

Figure 3: An alignment of several functionally equivalent RNA sequences from different evolutionarily related organisms with a conserved RNA structure. The base-pairs in the pairing columns have been fully conserved during evolution, while two compensatory mutations may have changed *both* nucleotides forming a pair. The resulting pairs of co-varying or co-evolving columns in the alignment where the primary sequence conservation may be low, but the functional conservation in terms of base-pairing ability is high (see highlighted pair of columns) provide a strong sequence signal for structure prediction. The drawing of the secondary structure was generated with PSEUDOVIEWER [I35].

model which we can then use to search genome sequences for members of that gene family.

Trained model for sequence and structure already exists: The data base RFAM [71] not only provides alignments for more than 500 families of non-coding RNA genes, but also dedicated computer programs which summarize the predominant structure and sequence features of one RFAM alignment. These programs can be used to search a given genome sequence for RNA genes that belong to this family. However, it should be noted that some RFAM alignments do not cover the entire transcript of the RNA gene, but only a sub-sequence. The RFAM alignments have typically been manually established by a human expert who makes sure that functionally equivalent parts of the structures and sequences are correctly aligned. Figure 3 shows a hypothetical example of such an alignment with an evolutionarily conserved RNA structure. Once the alignment has been established, so-called co-variance models (CMs) [72, 73] can be set up to capture the sequence and structure features along the alignment. CMs are probabilistic models whose parameters and predictions have a probabilistic interpretation. This is a crucial feature that renders the training of these models as well as the interpretation of their predictions fairly straight-forward. In RFAM, each alignment is used as input to the software package INFERNAL [74] which trains the parameters of a dedicated CM. Once the CM has been trained, it can then be used to search any target sequence, e.g. a genome, for regions that are similar to the ones represented by the corresponding RFAM alignment, both in terms of structure and sequence. Each match of the CM to the target sequence is assigned a score which reflects the likelihood that this region is a true match. These scores can be used to rank the matches and to discard any matches below a certain threshold score. The threshold value determines the sensitivity and specificity of the search.

In order to estimate the sensitivity of such a search and to derive a sensible threshold value, the scores of true matches can be compared to those of known false matches. It is more difficult to estimate the specificity of such a search as it can be difficult to prove that a given match, e.g. a region in a given genome, is never expressed in the living organism. This is the reason why artificial negative test sets are sometimes constructed from randomized sequences. However, the type and degree of randomness in these artificially constructed test sets does not

necessarily reflect the randomness observed in real biological data. It therefore remains difficult to estimate the true specificity of RNA gene searches.

Apart from the CMs provided for the RNA gene families in RFAM, several other specialized programs have been developed in order to detect RNA genes that belong to specific RNA gene families. TRNASCAN-SE [75] employs a CM to model tRNAs. It is the standard program to annotate tRNAs in newly sequenced genomes and has been shown to have almost perfect sensitivity as well as a low false positive rate (the estimate is one false positive per 15 gigabases). Other dedicated gene prediction programs are, for example, BRUCE [76] for detecting transfer-messenger RNA genes (tmRNA genes), SNOSCAN [77] for detecting box c/D small nucleolar RNAs (snoRNAs) which are required for methylation of eukaryotic ribosomal RNA and SNOGPS [78] and FISHER [79] for detecting box H/ACA snoRNAs. A large number of dedicated programs have also been developed to detect miRNAs, for example PROMIR [80] for human, MIR-ABELA [81] for mammals, MIR-SCAN [82] for vertebrates, MIRSEEKER [83] for drosophilids, HARVESTER [84] for plants, BAYES-MIRFINDER [85] for eukaryotes and RNA-MICRO [86] for pre-generated multiple sequence alignments. These programs use either heuristic method or a probabilistic model in order to capture the characteristic structural and sequence features of miRNAs.

Model for sequence and structure needs to be established: If we have an RNA gene for which no dedicated prediction program or a CM in RFAM already exists, it is possible to create a new one.

If the function of the RNA gene relies on well-defined structural elements and if we have orthologous sequences from related organism, we can employ INFERNAL or CMFINDER [87] in order to derive a dedicated CM which we can then use to search long target sequences for RNA genes of this type. For INFERNAL, we first need to establish a high quality sequence alignment that correctly aligns functionally equivalent sequence and structure elements. Often, this is best done manually with the help of a visualization program such as RALEE [88] that highlights consensus base-pairs in base-pairing alignment columns. For this, we can for example use an alignment computed by CLUSTALW [89] as the starting point. If we happen to know the functional

structure for each individual sequence in our data set, we can alternatively employ structure alignment tools such as RNAFORESTER [90, 91] or RNA-DISTANCE [58, 92] in order to create a global sequence and structure alignment. Once the alignment has been fixed, it is used as input to INFERNAL in order to train the parameters of a dedicated CM. CMFINDER takes un-aligned sequences as input and outputs a trained CM. Both, for INFERNAL and CMFINDER, the resulting parameters of the trained CM capture the sequence and structure variation in the training sequences. One potential drawback of CMs is their limited ability to detect target sequences whose primary sequence or structure differs significantly from those in the training alignment. Ideally, the degree of structure and sequence variation in the training alignment should reflect the sequence and structure diversity that we aim to detect with the CM. INFERNAL addresses these limitations by employing more sophisticated techniques for CM training and for searching databases. It utilizes Dirichlet mixture priors and an effective sequence weighting method to extend the model's ability to also recognize more diverged target sequences. In addition, a local search mode can be also used to detect target sequences with a high degree of structural variation. CMs cannot model pseudo-knotted RNA structures because they employ essentially variants of stochastic context-free grammars (SCFGs).

The computational complexity for searching a target sequence of length L with a CM of INFERNAL that captures an alignment of N nucleotides length is $\mathcal{O}(L\,N^{1.3})$ time [93] and $\mathcal{O}(N^2 \log(N))$ memory [74] using several heuristic (time) and exact (memory) tricks in order to reduce the nominal time requirement $\mathcal{O}(L\,N^3)$ and memory requirement $\mathcal{O}(N^3)$ of CMs.

It is possible to further reduce the memory and time requirements of CM-based analyses by using a two-step approach. In the first step, the potentially long target sequence is searched with a computationally less expensive method that captures only sequence features. In the second step, the matches returned in the first step are analyzed with the CM. For the first step, Weinberg and Ruzzo [94] introduced heuristic filters, i.e. profile-HMMs which capture only sequence rather than structural features explicitly. Similar to CMs, profile-HMMs can be automatically trained with a given sequence alignment in order to capture its characteristic sequence

features. For tRNAs genes, Weinberg and Ruzzo show that the performance of their automatically trained heuristic filter is about as accurate as the custom-tailored method TRNASCAN-SE [75] which cannot be readily adjusted to model other RNA families.

If we have only a single RNA gene whose functional structure we know, but no functionally equivalent sequences from evolutionarily related genomes, we can still attempt to search a long target sequence for regions that are similar to the known RNA gene, both in terms of sequence and structure. RSEARCH [95] is an SCFG-based program that can be used to search one long target sequence with an RNA sequence whose known, pseudo-knot free secondary structure is explicitly taken into account. It assigns a score to each match which quantifies the reliability of the prediction. These scores can be used to rank the matches. The main drawback of this strategy is that it is computationally costly. The computational requirements for searching a long target sequence of length L with a shorter query sequence of length N are $\mathcal{O}(L\,N^3)$ time and $\mathcal{O}(N^3)$ memory, if no heuristic tricks are used. As this method compares the target sequence only to a single RNA sequence rather than a set of evolutionarily related sequences that represent the RNA gene family, its sensitivity is typically lower than that of CM-based methods. Alternatively, we can use the program RNAMOTIF [96]. This programs allows the user to *manually* define a motif. The motif definition can capture sequence features, secondary and even tertiary structure features, including pseudo-knots. The user can also define custom scores that are then used to score every match of the motif's model, called a descriptor, to a potentially long target sequence. As long as the user is willing (and knowledgeable enough) to manually specify both the motif and the scoring scheme, RNAMOTIF can be used for highly specific and sensitive sequence searches. Coming up with a motif description and a good scoring scheme requires a fairly high degree of expertise and biological insight because both define what the characteristic and important features of the corresponding RNA family should be.

Only transcript of RNA gene is known (Case 1b)
Sequence based homology search: Some RNA genes do not act by assuming a well-defined RNA structure or we simply do not know whether or not RNA structure plays a functional role. In those cases, it is possible to attempt a homology search that is based on sequence similarity only rather than sequence *and* structure similarity. This approach can also be justified for structural RNA genes whose sequences have been so well conserved during evolution that sequence similarity suffices to identify them in genomes, e.g. tRNA and rRNA genes.

Profile hidden Markov models (profile-HMMs) are probabilistic models that capture the sequence conservation along a given alignment, in the same way that CMs capture sequence *and* structure conservation along an alignment. As for CMs, the parameters of profile-HMMs can be trained with a fixed alignment of evolutionarily related sequences where functionally equivalent regions are aligned. For profile-HMMs, this training can be done with the software package HMMER [97]. Once the profile-HMM has been established, it can then be used to search a long target sequence for regions that resemble the sequences in the training alignment. Profile-HMMs are used by the PFAM [98] data base of protein families in the same way that CMs are used by the RFAM data base of RNA families. Profile-HMM are faster than CMs because they do not consider the long-range correlations along the sequence that arise through the base-pairs of the conserved structure, but they are slower than well-known heuristic local sequence alignment methods such as BLAST [99].

The computational complexity for searching a target sequence of length L with a profile-HMM that models an alignment of N nucleotides length is $\mathcal{O}(L\,N)$ time and $\mathcal{O}(L\,N)$ memory. For long genome sequences L and profile-HMMs that represent long RNA genes, i.e. large N values, this can be computationally too costly and a fast pre-processing step with heuristic similarity search programs like BLAST is required to quickly narrow down the search space before more costly methods are employed.

Pair-wise homology search by simultaneously predicting structure and alignment: If a profile-HMM for a similarity based homology search cannot be established for the known RNA gene, either because functionally equivalent sequences are not known or because they are too diverged to establish the high-quality alignment required for model training, it is possible to employ methods that take the single RNA sequence and search it against a target sequence by simultaneously investigating sequence and

structural similarities, for example DYNALIGN [100, 101] (MFE approach), FOLDALIGN [102, 103] (MFE approach), STEMLOC [104] (pair-SCFG approach) or CONSAN [105] (pair-SCFG approach). These programs are computationally very costly as they take two un-aligned sequences as input and simultaneously predict and align the pseudo-knot free secondary structures of the two sequences. Without playing any heuristic tricks, it takes $\mathcal{O}(L_1^2 L_2^2)$ memory and $\mathcal{O}(L_1^3 L_2^3)$ time to analyze two input sequences of length L_1 and L_2 with a pair-SCFG.

DYNALIGN and FOLDALIGN can both be used to scan a long target sequence with a shorter sequence. It is important to note that FOLDALIGN has been devised to detect local regulatory structures rather than global structures with multi-loops. It computes a score which can be used to rank putative conserved structural elements. DYNALIGN reduces the computational complexity of the pair-SCFG approach by limiting the search space that is explored to align the two sequences and by limiting the size of internal loops in RNA structures. The advantage of these programs is that they do not require a fixed input alignment which can be hard to establish if the primary sequence identity between the two RNA sequences is low, typically below 40%. Their disadvantage is that they do not explicitly model unstructured regions in the two input sequences. The predicted results may therefore strongly depend on the chosen sequence window. If the score of the predicted common RNA structure is high and the structure is similar to the known RNA structure, the region is likely to overlap the desired RNA gene. Because methods that simultaneously align and fold are computationally costly, they should only be applied if global RNA structure is likely to play a functional role and if the RNA sequence that is searched against the longer sequence is fairly short.

Analyzing sets of potentially homologous RNA genes

The sequence based homology search against one or several genomes results in sets of potentially homologous RNA genes. These sets can be analyzed in more detail in order to test the hypothesis that they actually correspond to RNA genes, see Figure 3.

Classifying a given sequence alignment: Often, sets of potentially homologous sequences are presented as multiple sequence alignments. We can use these multiple sequence alignments as input to classification programs like QRNA [106], EVOFOLD [46] and RNAZ [45] which test whether the sequences in the alignment contain similar structures or not. QRNA [106] takes a fixed alignment of two sequences as input and classifies it into RNA structure-containing, protein-coding or other. It uses an SCFG-based approach and assigns a score to each input alignment which quantifies the reliability of the predicted classification. RNAZ (MFE approach) and EVOFOLD (SCFG approach) can handle input alignments of more than two RNA sequences. Both programs classify a fixed input alignments as either structure encoding or non-encoding based on its observed pattern of mutations, see Figure 3, and assign a reliability score to their classification. Both programs evaluate the structure-encoding potential of the alignment by considering pseudo-knot free secondary structures. In cases where the input alignment does not exhibit the characteristic pattern of co-evolving columns, RNAZ relies on the assumption that RNA genes are thermodynamically more stable than expected by chance, an assumption which has been shown not to hold in general [107]. RNAZ predictions should depend to a smaller extent than those of EVOFOLD on the quality of the input alignment as RNAZ evaluates the similarity of the encoded structures based on the similarity of their *minimum free energies* rather than the corresponding individual MFE structures themselves. For the same reason, RNAZ should be better at handling RNA structure variation. However, EVOFOLD depends to a much smaller extent than RNAZ on the chosen sequence window as its underlying model can explicitly model non-structural regions in the input alignment, whereas RNAZ forces each sequence in the input alignment to assume its MFE structure which can strongly depend on the chosen sequence window. An added benefit of EVOFOLD is that it takes the evolutionary tree relating the input sequences in the alignment explicitly into account.

Folding a fixed multiple sequence alignment: Rather than classifying the fixed multiple-sequence alignment as described before, we can also use programs like PFOLD [108, 109], RNAALIFOLD [61] and RNA-DECODER [110, 111] in order to predict a common, evolutionarily conserved RNA structure. These methods take as input a fixed multiple sequence alignment (as well as an evolutionary tree relating the sequences in the alignment, in case of PFOLD and RNA-DECODER) and predict a conserved secondary
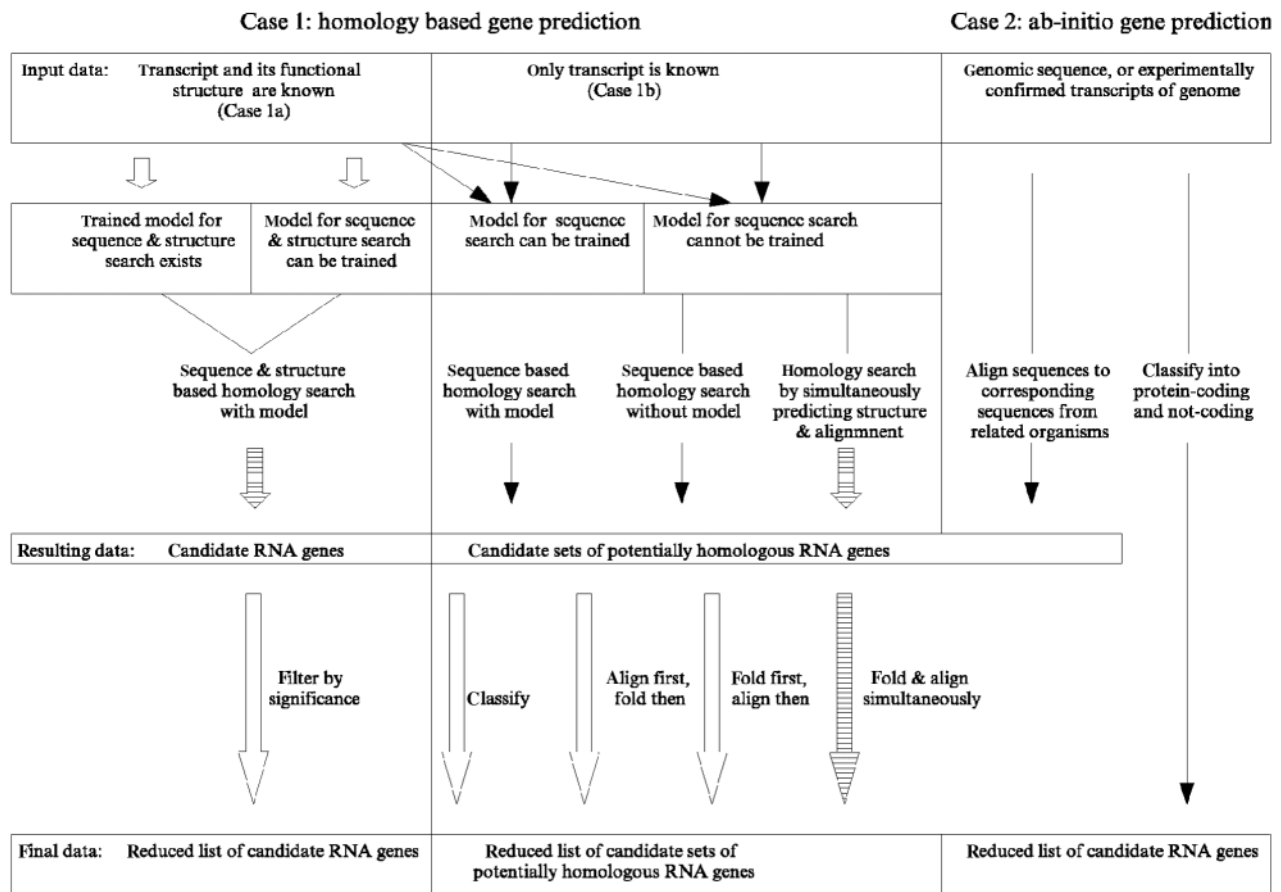
Figure 4: Current strategies for predicting RNA genes. Empty arrows indicate steps that investigate or rely on RNA structure. Of those, hatched arrows indicate steps that are computationally expensive.

structure without pseudo-knots. RNAALIFOLD and RNA-DECODER can also use a structure as additional input constraint. All three programs estimate the reliability of the predicted consensus structure by calculating the base-pairing probabilities for different pairs of columns in the alignment (RNAALIFOLD) or the base-pairing probability for each individual column in the alignment (PFOLD and RNA-DECODER). This extra information is valuable for highlighting regions in the predicted RNA structure that are particularly well or poorly supported by the fixed sequence alignment. All three methods look for co-varying columns in the input alignment, see Figure 4, as strong evidence for particular base-pairs in the predicted structure. As the primary sequence conservation in the co-varying columns can be low, structure prediction methods that take a fixed input alignment only work well if the sequences in the input alignment have a minimum pair-wise sequence identity, typically at least 60%, which allows them

to be reliably aligned based on sequence similarity alone.

RNAALIFOLD uses an extension of the MFE folding algorithm employed by the MFE method RNAFOLD to compute a pseudo-knot free consensus structure. This is done by minimizing the overall free energy and simultaneously taking primary sequence conservation and co-varying columns in the alignment into account. This optimization, implemented in a dynamic programming procedure, combines free energy parameters with conservation scores; two very different types of information whose respective contribution to the overall optimization is *a priori* not clear. The algorithm that is used to predict the consensus structure is the same as for the non-comparative MFE program RNAFOLD and the computational complexity of the two programs is the same.

PFOLD and RNA-DECODER both employ a stochastic context-free grammar (SCFG) in order

to derive a consensus structure and to compute base-pairing probabilities for each column in the input alignment. SCFGs are probabilistic models that can capture the long-range constraints imposed by the base-pairing sequence positions of the RNA structure. In contrast to programs like MFOLD, RNAFOLD and RNAALIFOLD which use free energies as parameters, SCFGs use probabilities rather than physical quantities as parameters. SCFG-based methods are thus not limited to predicting the MFE structure. Once the parameters of an SCFG have been specified, efficient dynamic programming algorithms can be used to derive the pseudo-knot free secondary structure that has the highest overall probability [CYK-algorithm [97]] or to calculate the base-pairing probability for each sequence position [inside-outside algorithm [97]]. As for the MFE approach, SCFG-based methods make use of the fact that the overall probability of a structure can be expressed as the product of probabilities for smaller parts of the RNA structure. The parameters of an SCFG can be derived from large datasets of known RNA structures and reflect the observed frequencies of base-pairs and structural features. Pseudo-knotted RNA structures cannot be modeled with SCFGs because pseudo-knots are context-dependent. Both, MFE-based and SCFG-based methods can be adapted to take comparative information from several RNA sequences into account. In RNAALIFOLD, PFOLD and RNA-DECODER this is done by using a fixed alignment of several RNA sequences as input. Because the alignment is fixed, it can be viewed as a hyper-sequence where each position corresponds to a column of aligned nucleotides and gaps rather than a single nucleotide. PFOLD and RNA-DECODER use parameters that score alignment columns rather than single nucleotides. For example, the score for pairing two columns (rather than two nucleotides) reflects how well the nucleotides and gaps in the two columns could pair. SCFGs provide a consistent probabilistic framework for combining different sources of information, in this case structure information and conservation information.

The computational complexity of RNAALIFOLD, PFOLD and RNA-DECODER is $\mathcal{O}(L^3)$ time and $\mathcal{O}(L^2)$ memory for an input alignment of length L. For a given alignment, RNAALIFOLD is typically the least and RNA-DECODER the computationally most expensive choice. This is due to the fact that the models employed by RNA-DECODER are more complex than those of the other two programs. Of the three programs, RNA-DECODER is the only method that can take known protein–coding regions in the input alignment explicitly into account. It is also the only one that explicitly models un-structure regions in the input alignment. RNA-DECODER is thus particularly well suited for detecting local consensus RNA structures that do not involve the entire alignment.

None of the above three programs can handle pseudo-knotted structures. There exist computationally more expensive programs, e.g. HXMATCH [112], KNETFOLD [113] and ILM [114] which take a fixed input alignment and predict a consensus RNA structure which may also contain pseudo-knots. As these programs explore a much larger search space than methods that only investigate pseudo-knot free secondary structures, their prediction accuracy on typical test sets (with no or few pseudo-knotted structures) tends to be lower than for programs that do not model pseudo-knots.

Aligning folded sequences: In exceptional cases, for example if the sequences correspond to the biological transcript units *and* if we know (or suspect) that a global RNA structure plays a functional role, we can first predict an RNA structure for each individual sequence, e.g. with MFOLD [56–58] or RNAFOLD of the Vienna package [58–62], and then align the predicted structured sequences using programs like RNAFORESTER [90,91], RNADISTANCE [58,92] or MARNA [115].

MFOLD and RNAFOLD take $\mathcal{O}(L^2)$ memory and $\mathcal{O}(L^3)$ time to predict the pseudo-knot-free MFE structure for an RNA sequence of length L. Both programs express the MFE of the entire structure as the sum of MFE-contributions from smaller structural elements and employ a dynamic programming procedure in order to determine the global structure that minimizes the overall sum of corresponding MFE-contributions. As they also employ the same set of thermodynamic parameters, the difference between both programs and their respective structure predictions is minor.

RNAFORESTER and MARNA can both be used to compute a global alignment of several un-aligned input sequences whose pseudo-knot free secondary structures are already known, whereas RNADISTANCE calculates only pair-wise structural alignments. RNADISTANCE takes only the structures

as input, whereas RNAFORESTER and MARNA also use the RNA sequences. RNAFORESTER and RNADISTANCE present the individual structures as trees and compute a structural alignment by aligning the trees. Both programs calculate a score for the resulting structural alignments. In contrast to RNAFORESTER and RNADISTANCE, MARNA is a method that employs a progressive pair-wise alignment strategy which takes the known structures indirectly into account when calculating the global alignment. MARNA is the only one of the three programs that is capable of proposing a pseudo-knot free secondary structure for input sequences whose structure is not known. In the extreme case, it can even derive a consensus structure when none of the individual structures are known. Gardner and Giegerich [70] conclude, in a comparative analysis of RNAFORESTER and MARNA, that 'MARNA is generally less dependant upon the accuracy of the input structures' than RNAFORESTER. However, one advantage of RNAFORESTER is that is can single out poorly predicted input structures. Overall, the quality of this two-step approach is limited by the quality of the individually predicted input structures. This can be a serious limitation for long sequences.

Simultaneously aligning and folding sequences: If the pair-wise sequence identity in the set of potentially homologous RNA sequences is too low to come up with a reliable sequence alignment, typically below 40%, it becomes difficult to predict a common RNA structure.

The idea of co-estimating pseudo-knot free RNA secondary structures and multiple sequence alignments (and evolutionary trees) was first suggested in a theory paper by David Sankoff in 1985 [116]. It is possible, to define extensions of SCFGs that take N rather than just a single RNA sequence as input and simultaneously predict their pseudo-knot free secondary structures as well as a global alignment. These so-called N-SCFGs are computationally very expensive and require $\mathcal{O}(L^{3N})$ time and $\mathcal{O}(L^{2N})$ memory to analyze a set of N RNA sequences which each have length L. It is possible to keep the memory and time requirements at bay, either by analyzing only two input sequences with a pair-SCGF or by analyzing more than two sequences with a heuristic method. So far, there are only few alignment-free structure prediction methods, e.g. CARNAC [117, 118] (no pseudo-knots),

COMRNA [119] (pseudo-knots), STEMLOC [104] (no pseudo-knots) and CONSAN [105] (no pseudo-knots). However, these programs tend to detect only very conserved local structures. COMRNA, CARNAC and STEMLOC can analyze several input sequences (STEMLOC achieves this by a progressive pair-wise method), whereas CONSAN is limited to only two input sequences. COMRNA is the only among these programs that can predict pseudo-knotted structures. The predictions of COMRNA rely on the calculation of maximal cliques, a problem which is known to be NP-complete. In the general case, it thus requires exponential time to run analyses, but may be fast enough to analyze short sequences.

Gardner and Giegerich [70] compare different comparative RNA structure prediction approaches and conclude that 'structure-prediction-algorithms vary widely in terms of both sensitivity and specificity across different lengths and homologies'. It is thus difficult to recommend a specific program that is bound to outperform all others irrespective of the data presented to the program. One feature that has been noted in a number of studies is that CARNAC predictions tend to have a high specificity, but medium to low sensitivity.

Summary: homology based prediction of RNA genes

In a recent paper, Freyhult, Bollback and Gardner [120] provide the first critical comparative assessment of the performance of different homology search methods on non-coding RNA and come to the conclusion that 'the most popular homology search methods are often the least accurate' and that 'many studies have used inappropriate tools for their analysis'. They provide a detailed analysis of different strategies and programs for several typical data sets which enables the user to design new homology searches. They show that computationally more expensive, RNA-specific probabilistic methods like INFERNAL and RSEARCH are particularly good at discriminating signal from noise. A decision for or against a particular strategy and program is inevitably a decision between high accuracy and high speed which each user has to make for himself.

Ab-initio RNA gene prediction (Case 2)

Ab-initio gene prediction constitutes the most challenging case of RNA gene prediction. In the most

difficult case, we are given a single genome sequence without any annotation and have to find the encoded RNA genes. In the general case, *ab-initio* RNA gene prediction is still a more or less unsolved problem and will not be discussed here. However, for specific genomes, we can sometimes play surprising tricks.

One such example are the genomes of AT-rich hyperthermophiles *Methanococcus jannschii* and *Pyrococcus furiosus*. These two bacteria have to stabilize their double-stranded DNA genome as well as structural RNA genes against thermal denaturation. One way of achieving this is by increasing the strength of the hydrogen bonds, i.e. by increasing the number of {G, C} base-pairs which form the most stable base-pairs. Galtier and Lobry [121] observed already in 1997 that there is a strong correlation between the GC contents of rRNA and tRNA genes and their optimal growth temperature in hyperthermophiles. By searching for GC-rich regions, Klein *et al.* [122] and, independently, Schattner [123], managed to efficiently detect known and new structural RNA genes in *M. jannschii* and *P. furiosus* (Klein *et al.* only).

There also exist computationally efficient algorithms which aim to predict locally stable RNA structures on a genome-wide scale, for example the program RNAPLFOLD [124, 125]. RNAPLFOLD computes the average base-pairing probability for any pair of sequence positions (i, j) by considering the statistical ensemble of all pseudo-knot free secondary structures in thermodynamic equilibrium for a sub-sequence of length L. This fixed sized window of length L is scanned along a potentially long target sequence. This method is implemented in an efficient sliding window-approach which requires $\mathcal{O}(NL^2)$ time and $\mathcal{O}(N + L^2)$ memory to calculate the MFE structure and the matrix of base pairing probabilities for each sequence window of length L in a target sequence of length N. So far, the potential of RNAPLFOLD for detecting RNA genes has not been systematically investigated.

In these days, we usually have additional data to help us find the RNA genes in a given genome. One important source of information are genome sequences from evolutionarily related organisms. In that case, we can start the search for RNA genes by mapping genome sequences from related organisms to different regions of the target genome. This type of information is often readily available in genome browsers like ENSEMBL [126] and the UCSC genome browser [127] or it can be established using fast alignment tools like BLAT [128].

The first genome-wide screen for RNA genes of this type was done for *E. coli* [129]. Comparative sequence data from four related bacteria were mapped to the *E. coli* genome, resulting in more than 23 000 pair-wise alignments which were classified into three mutually exclusive classes (structure-containing, protein-coding and other) using QRNA [106]. Of the 275 loci in the *E. coli* genome that were predicted to be structure-containing, 49 were assayed experimentally and 11 of 49 were found to express small transcripts of unknown function.

So, once we have established sets of potentially homologous sequences, we can, in principle, analyze them in the same way as we would analyze sets of potentially homologous RNA genes, see text above and Figure 4. However, once crucial difference is that the sets of sequences do not necessarily correspond to transcribed units of the respective genomes. Unless we have additional evidence from cDNA data (or weaker evidence from tiling array experiments), we simply do not know whether the sequences in our sets are ever transcribed and whether their boundaries correspond to the biological transcript units. This complicates the interpretation of the results that we get when analyzing the sets with the methods described before, see in 'Analyzing sets of potentially homologous RNA genes'. As all of the methods investigate to some degree RNA structure, i.e. the potential of different positions along the sequences to form base-pairs, their predictions can be biased in unknown ways if the methods are presented with sequences that do not correspond to biological transcripts. For example, if the sequences are shorter than the real transcripts, the methods cannot consider base-pairs that involve nucleotides outside the sub-sequence. If, on the other hand, the sequences are longer than the real transcripts, the methods will also take base-pairs into account that cannot form in the cell. These examples illustrate why it is difficult to interpret the predictions of two recent studies of the human genome based on RNAZ [44, 45] and EVOFOLD [46] and explain why it is almost impossible to convert the number of structure-encoding regions that these studies predict into a reliable RNA gene count.

Due to several genome-wide cDNA and tiling array experiments, for example for the human

genome [33–36] and the mouse genome [37–40], we know regions of the genome which are transcribed. In case of full length cDNA sequences, we even know the exact transcript boundaries. We can use these transcripts as input to a theoretical analysis which determines which of the transcripts correspond to RNA genes. This is currently done by investigating the protein-coding potential of each transcript and by discarding any transcripts with coding potential. Typical measures for coding potential are the length of the longest putative open reading frame as well as the ratio of synonymous to non-synonymous mutations in the putative open reading frame, see the afore-mentioned publications for details. By now, there already exists dedicated computer programs in order to distinguish protein-coding from non-coding transcripts, e.g. CONC [130] which employs support vector machines. However, depending on the nature of the transcripts (splice, un-spliced, poly-A+, poly-A-, length bias), these measures may be too crude to reliably distinguish the transcripts of protein-coding genes from those of RNA genes. As was explained earlier, see 'What are RNA genes?' above, RNA genes can, for example, be found in the introns of pre-mRNA transcripts. This explains why a binary sorting scheme into protein-coding versus non-coding transcripts is probably too coarse to get an unbiased and comprehensive view of all RNA genes.

The study by Cawley *et al.* [41] shows how different types of annotation, in their case predicted transcription factor binding sites along human chromosomes 21 and 22, can help to narrow down the location of potential RNA genes.

SUMMARY, DISCUSSION AND OUTLOOK

RNA gene prediction is an exciting field of research, where much has already been achieved, but where there is also ample scope to make important novel contributions.

We already have a wealth of computer programs that predict RNA structures or investigate the structure formation potential of RNA sequences. It has been shown in numerous studies that comparative approaches provide the best way of identifying potentially functional sequence and structure features because they allow us to detect sequence and structure features that have been conserved during evolution.

Several already promising methods could be further improved by extending them to deal with un-aligned input sequences. The requirement for a fixed input alignment currently imposes a considerable limitation on several popular programs which needs to be overcome with new conceptual ideas and computational tricks in order to make these methods applicable to a wider range of interesting data and to increase their sensitivity. Several comparative methods would probably also benefit from explicitly modeling the known evolutionary relationship of the input sequences. This would allow these methods to dynamically re-adjust their parameters according to the range of evolution in the input sequences. Another limitation is due to the fact that most RNA gene prediction method employ computational techniques which make the investigation of pseudo-knotted structures computationally very costly. This is the main reason why most methods completely ignore pseudo-knotted structures. Pseudo-knotted structures constitute only a minority in current structural data bases. However, this may—at least in part—be more a reflection of our difficulty to detect them than their true abundance in nature. As pseudo-knots are known to play diverse and important functional roles [131] and as ignoring pseudo-knots can bias the structure predictions in unknown ways, it would be good to invest some effort into developing novel algorithms that can model pseudo-knotted structures in a conceptually more elegant and computationally more efficient way.

Right now, all RNA gene prediction methods aim to detect structural RNA genes and essentially ignore unstructured RNA genes, i.e. genes which do not exert their function via a well-defined RNA structure. Moreover, many programs silently assume a global RNA structure that spans the entire transcript rather than one or several local RNA structures, i.e. RNA structures that span only a sub-sequence of the entire transcript and that are separated by potentially long, unstructured spacer regions. In addition, none of these methods attempt to identify the transcripts of RNA genes. The overall effect is that most RNA gene prediction methods effectively search of structure-containing regions in the genome which do not necessarily correspond to transcripts of the genome. This is typically done by scanning a sequence window of fixed length

along the entire genome sequence. However, an accurate assessment of potential RNA structure and structure-formation potential can only be made if the investigated sequences correspond to the relevant biological sequence units that are present in the cell. Future developments in RNA gene prediction methods should therefore address the first goal in RNA gene prediction more explicitly and aim to identify regions of the genome that are transcribed. One structure-independent approach for detecting tRNA genes by scanning for polymerase III transcription sites has already been described in 1994 by Pavesi *et al.* [132]. More recently, Glusman *et al.* [133] detected sequence signals that can distinguish transcribed from un-transcribed regions of the genome and that may be employed by future gene prediction programs to predict both, structural and non-structural RNA genes. While we learn how to predict the transcribed regions of a genome, we should reduce the window dependency of the existing methods and pragmatically integrate experimental data (e.g. cDNA and tiling array–data) as well as theoretical predictions (e.g. prediction transcription factor binding sites) into genome-wide analyses.

Much can probably be gained by caring even more about the details, both on the theoretical and in experimental side. We know from numerous examples, that RNA genes constitute a diverse group of genes that go about very different tasks in the cell using very diverse mechanisms. We are likely to deprive ourselves of the opportunity to discover novel and maybe unexpected classes of RNA genes if we use a too general one-fits-all approach. Finding and exploring new sequence signals, like those discovered by Glusman *et al.* [133], will not only increase our insight into the underlying biology and will also lead to improved methods that give appropriate weight to the different signals in RNA genes. It may also be possible to improve the detection of some RNA gene families by searching for potential interaction partners that are known to interact via a known mechanism. This strategy has already been successfully employed to detect miRNAs. These novel theoretical approaches, together with unbiased and comprehensive transcriptome data from genome-wide experiments have the potential to significantly improve our understanding of how genomes regulate themselves.

Key Points

- RNA gene prediction remains a challenge.
- So far, only the prediction of highly structured RNA genes has been attempted.
- The computational approach to be chosen strongly depends on the available data.
- The most successful approaches today employ a comparative approach in which sets of homologous sequences are analyzed simultaneously.

References

1. Higgs PG. RNA secondary structure: physical and computational aspects. *Quaterly Rev of Biophys* 2000;**33**:199–253

2. Mattick JS, Gagen MJ. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Molecular Biol and Evol* 2001;**18**:1611–30

3. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 2003;**25**:930–39

4. Holley RW, Apgar J, Everett GA, *et al.* Structure of a ribonucleic acid. *Science* 1965;**147**:1462–65

5. Klug A, Robertus JD, Ladner JE, *et al.* Conservation of the molecular structure of yeast phenylalanine transfer RNA in two crystal forms. *Proc Natl Acad Sci USA* 1974;**71**:3711–15

6. Kim SH, Sussman JL, Suddath FL, *et al.* The general structure of transfer RNA molecules. *Proc Natl Acad Sci USA* 1974;**71**:4970–74

7. Kruger K, Grabowski PJ, Zaug AJ, *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* 1982;**31**:147–57

8. Guerrier-Takada C, Gardiner K, Marsh T, *et al.* The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 1983;**35**:849–57

9. Ban N, Nissen P, Hansen J, *et al.* The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 2000;**289**:905–20

10. Nissen P, Hansen J, Ban N, *et al.* The structural basis of ribosome activity in peptide bond synthesis. *Science* 2000;**289**:920–30

11. Yusupov MM, Yusupova GZ, Baucom A, *et al.* Crystal structure of the ribosome at 5.5 A resolution. *Science* 2001;**292**:883–96

12. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 1977;**74**:3171–5

13. Chow LT, Gelinas RE, Broker TR, *et al.* An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 1977;**12**:1–8

14. Padgett RA, Mount SM, Steitz JA, *et al.* Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. *Cell* 1983;**35**:101–7

15. Parker R, Siliciano PG, Guthrie C. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell* 1987;**49**:229–39

16. Lerner MR, Boyle JA, Mount SM, *et al.* Are snRNPs involved in splicing? *Nature* 1980;**283**:220–4

17. Rogers J, Wall R. A mechanism for RNA splicing. *Proc Natl Acad Sci USA* 1980;**77**:1877–79

18. Nudler E, Mironov AS. The riboswitch control of bacterial metabolism. *Trends Biochem Sci* 2004;**29**:11–17

19. Winkler WC. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol* 2005;**9**:594–602

20. Tucker BJ, Breaker RR. Riboswitches as versatile gene control elements. *Curr Opin Chem Biol* 2005;**15**:342–8

21. Winkler WC, Nahvi A, Roth A, *et al.* Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* 2004;**428**:281–6

22. Fire A, Xu S, Montgomery MK, *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998;**391**:806–11

23. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with anti-sense complementarity to lin-14. *Cell* 1993;**75**:843–54

24. Reinhart BJ, Slack FJ, Basson M, *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 2000;**403**:901–6

25. Lee RC, Ambros V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* 2001;**294**:862–4

26. Lau NC, Lim LP, Weinstein EG, *et al.* An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 2001;**294**:858–62

27. Lagos-Quintana M, Rauhut R, Lendeckel W, *et al.* Identification of novel genes coding for small expressed RNAs. *Science* 2001;**294**:853–8

28. Reinhart BJ, Weinstein EG, Rhoades MW, *et al.* MicroRNAs in plants. *Gene Dev* 2002;**13**:1616–26

29. Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAS and their regulatory roles in plants. *Ann Rev Plant Biol* 2006;**57**:19–53

30. Berezikov E, Plasterk RH. Camels and zebrafish, viruses and cancer: a microRNA update. *Hum Mol Genet* 2005;**14**:R183–90

31. Cullen BR. Transcription and processing of human microRNA precursors. *Mol Cell* 2004;**16**:861–65

32. Sharp PA. *RNA World.* Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY, 2006

33. Bertone P, Stolc V, Royce TE, *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* 2004;**306**:2242–6

34. Kampa D, Cheng J, Kapranov P, *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004;**14**:331–42

35. Cheng J, Kapranov P, Drenkow J, *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005;**308**:1149–54

36. Imanishi T, Itoh T, Suzuki Y, *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2004;**2**:e162

37. Okazaki Y, Furuno M, Kasukawa T, *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-lengths cDNAs. *Nature* 2002;**420**:563–73

38. Carnini P, Kasukawa T, Katayama S, *et al.* The transcriptional landscape of the mammalian genome. *Science* 2005;**309**:1559–63

39. Ravasi T, Suzuki H, Pang KC, *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res* 2006;**16**:11–9

40. Maeda N, Kasukawa T, Oyama R, *et al.* Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2006;**2**:e62

41. Cawley S, Bekiranov S, Ng HH, *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004;**116**:499–509

42. Numata K, Kanai A, Saito R, *et al.* Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res* 2003;**13**:1301–06

43. Hüttenhofer A, Schattner P, Polaceka N. Non-coding RNAs: hope or hype? *Trends in Genet* 2005;**21**:289–97

44. Washietl S, Hofacker IL, Lukasser M, *et al.* Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 2005;**23**:1383–90

45. Washietl S, Hofacker IL, Stadler PF, Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 2005;**102**:2454–59

46. Pedersen JS, Bejerano G, Siepel A, *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2006;**2**:e33

47. Tycowski KT, Shu MD, Steitz JA. A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Gene Dev* 1993;**7**:1176–90

48. Kiss T, Filipowicz W. Small nucleolar RNAs encoded by introns of the human cell cycle regulatory gene RCC1. *EMBO J* 1993;**12**:2913–20

49. Kiss T, Filipowicz W. Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. *Gene Dev* 1995;**9**:1411–24

50. Nakajima N, Ozeki H, Shimura Y. Organization and structure of an E. coli tRNA operon containing seven tRNA genes. *Cell* 1981;**23**:239–49

51. Chow JC, Yen Z, Ziesche SM, *et al.* Silencing of the mammalian X chromosome. *Annu Rev Genom Hum G* 2005;**6**:69–92

52. Chang SC, Tucker T, Thorogood NP, *et al.* Mechanisms of X-chromosome inactivation. *Front Biosci* 2006;**11**:852–66

53. Reinhold-Hurek B, Shub DA. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* 1992;**357**:173–6

54. Hirotsune S, Yoshida N, Chen A, *et al.* An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 2003;**423**:91–6

55. Yano Y, Saito R, Yoshida N, *et al.* A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J Mol Med* 2004;**82**:414–22

56. Mathews DH, Sabina J, Zuker M, *et al.* Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999;**288**:911–40

57. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**:3406–15

58. Hofacker IL, Fontana W, Stadler PF, *et al.* Fast folding and comparison of RNA secondary strutures. *Monatshefte für Chemie (Chemical Monthly)* 1994;**125**:167–88

59. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res* 1981;**9**:133–48

60. Wuchty S, Fontana W, Hofacker IL, *et al.* Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 1999;**49**:145–65

61. Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 2002;**319**:1059–66

62. Hofacker IL. The Vienna RNA Secondary structure server. *Nucleic Acids Res* 2003;**31**:3429–31

63. Neugebauer KM. On the importance of being co-transcriptional. *J Cell Sci* 2002;**115**:3865–71

64. Morgan SR, Higgs PG. Evidence for kinetic effects in the folding of large RNA molecules. *J Chem Phys* 1996;**105**:7152–57

65. Repsilber D, Wiese S, Rachen M, *et al.* Formation of metastable RNA structures by sequential folding during transcription: Time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis. *RNA* 1999;**5**:574–84

66. Ro-Choi TS, Choi YC. Structural elements of dynamic RNA strings. *Mol Cells* 2003;**16**:201–210

67. Lewicki BTU, Margus T, Remme J, *et al.* Coupling of rRNA transcription and ribosomal assembly *in vivo* – formation of active ribosomal-subunits in Escherichia coli requires transcription of RNA genes by host RNA polymerase which cannot be replaced by T7 RNA polymerase. *J Mol Biol* 1993;**231**:581–93

68. Chao MY, Kan MC, S. Lin-Chao. RNAII transcribed by IPTG-induced T7 RNA polymerase is non-functional as a replication primer for ColE1-type plasmids in Escherichia coli. *Nucleic Acids Res* 1995;**23**:1691–95

69. Meyer IM, Miklós I. Co-transcriptional folding is encoded within RNA genes. *BMC Mol Biol* 2004;10:5

70. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 2004;5:140

71. Griffiths-Jones S, Moxon S, Marshall M, *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005;**33**:D121–24

72. Eddy SR, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Res*, 1994;**22**:2079–2088

73. Yasubumi Sakakibara, Michael Brown, Rebecca Underwood, *et al.* Stochastic context-free grammars for modeling RNA. In *Proceedings of the 27th Hawaii International Conference on System Sciences*, 1994; p. 284–3, IEEE Computer Society Press, Honolulu, 1994

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

75. Lowe T, Eddy SR. tRNAscan-SE: a Program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;**25**:955–64

76. Laslett D, Canback B, Andersson S. BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res* 2002;**30**:3449–53

77. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. *Science* 1999;**283**:1168–71

78. Schattner P, Decatur WA, Davis CA, *et al.* Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Res* 2004;**32**:4281–96

79. Edvardsson S, Gardner PP, Poole AM, *et al.* A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics* 2003;**19**:865–73

80. Nam JW, Shin KR, Han J, *et al.* Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res* 2005;**33**:3570–81

81. Sewer A, Paul N, Landgraf P, *et al.* Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* 2005;**6**:267

82. Lim LP, Glasner ME, Yekta S, *et al.* Vertebrate microRNA genes. *Science* 2003;**299**:1540

83. Lai EC, Tomancak P, Williams RW, *et al.* Computational identification of Drosophila microRNA genes. *Genome Biol* 2003;**4**:R42

84. Dezulian T, Remmert M, Palatnik JF, *et al.* Identification of plant microRNA homologs. *Bioinformatics* 2006;**22**:359–60

85. Yousef M, Nebozhyn M, Shatkay H, *et al.* Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics* 2006;**22**:1325–34

86. Hertel J, Stadler PF. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* 2006;**22**:e197–202

87. Yao Z, Weinberg Z, Ruzzo WL. CMfinder - a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006;**22**:445–52

88. Griffiths-Jones S. RALEE–RNA ALignment editor in Emacs. *Bioinformatics* 2005;**21**:257–9

89. Higgins D, Thompson J, Gibson T, *et al.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673–80

90. Höchsmann M, Töller T, Giegerich R, Kurtz S. Local similarity in RNA secondary structures. *Proceedings of the IEEE Bioinformatics Conference*, 2003; p. 159–168

91. Höchsmann M, Voss B, Giegerich R. Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE/ACM T Comput Biol Bioinfo* 2004;**1**:53–62

92. Fontana W, Konings DAM, Stadler PF, *et al.* Statistics of RNA secondary structures. *Biopolymers* 1993;**33**:1389–404

93. Nawrocki EP, Eddy S. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* 2007;**3**:e56

94. Weinberg Z, Ruzzo WL. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* 2006;**22**:35–9

95. Klein RJ, Eddy SR. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* 2003;**4**:44

96. Macke TJ, Ecker DJ, Gutell RR, *et al.* RNAMotif – a new RNA secondary structure definition and discovery algorithm. *Nucleic Acids Res* 2001;**29**:4724–35

97. Richard Durbin, Sean Eddy, Anders Krogh, *et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, 1998

98. Finn RD, Mistry J, Schuster-Bockler B, *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;**34**:D247–51

99. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10

100. Mathews DH, Turner DH and Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* 2002;**317**:191–203

101. Mathews DH. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* 2005;**21**:2246–53

102. Havgaard JH, Lyngsø BR, Stormo GD, *et al.* Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* 2005;**21**:1815–24

103. Havgaard JH, Lyngsø BR, Gorodkin J. The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res* 2005;**33**:W650–3

104. Holmes I. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinfomatics* 2005;**6**:73

105. Dowell RD, Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment con-straints. *BMC Bioinfomatics* 2006;**7**:400

106. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001;**2**:8

107. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding rnas. *Bioinformatics* 2000;**16**:583–605

108. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 1999;**15**:446–54

109. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 2003;**31**:3423–8

110. Pedersen JS, Meyer IM, Forsberg R, *et al.* A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* 2004;**32**:4925–36

111. Pedersen JS, Forsberg R, Meyer IM, *et al.* An evolutionary model for protein-coding regions with conserved RNA structure. *Mol Biol Evol* 2004;**21**:1913–22

112. Witwer C, Hofacker IL, Peter F. Stadler. Prediction of consensus RNA structures including pseudoknots. *IEEE/ACM Trans Comp Biol Bioinf* 2004;**1**:66–77

113. Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k–nearest neighbor classifiers. *RNA* 2006;**12**:342–52

114. Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 2004;**20**:58–66

115. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 2005;**21**:3352–9

116. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* 1985;**45**:810–25

117. Perriquet O, Touzet H, Dauchet M. Finding the common structure shared by two homologous RNAs. *Bioinformatics* 2003;**19**:108–16

118. Touzet H, Perriquet O. CARNAC: folding families of related RNAs. *Nucleic Acids Res* 2004;**32**:W142–45

119. Ji Y, Xu X, Stormo GD. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* 2004;**20**:1591–602

120. Freyhult EK, Bollback JP, Gardner PP. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* 2007;**17**:117–25

121. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol* 1997;**44**:632–36

122. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT–rich hyperthermophiles. *Proc Natl Acad Sci USA* 2002;**99**:7542–47

123. Schattner P. Searching for RNA genes using base–composition statistics. *Nucleic Acids Res* 2002;**30**:2076–82

124. Hofacker IL, Priwitzer B, Stadler PF. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* 2004;**20**:186–190

125. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. *Bioinformatics* 2006;**22**:614–15

126. Hubbard TJP, Aken BL, Beal K, *et al.* Ensembl 2007. *Nucleic Acids Res* 2007;**35**:D610–17

127. Kent WJ, Sugnet CW, Furey TS, *et al.* The human genome browser at UCSC. *Genome Res* 2002;**12**:996–1006

128. Kent WJ. BLAT - The BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64

129. Rivas E, Klein RJ, Jones TA, *et al.* Computational identification of noncoding RNAs in E. coli by compara-tive genomics. *Curr Biol* 2001;**11**:1369–73

130. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet* 2006;**2**:e29

131. Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* 2005;**3**:e213

132. Pavesi A, Conterio F, Bolchi A, *et al.* Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of

transcriptional control regions. *Nucleic Acids Res* 1994;**22**:1247–56

133. Glusman G, Qin S, El-Gewely MR, *et al.* A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLoS Comput Biol* 2006;**2**:e18

134. Chen JL, Greiger CW. Functional analysis of the pseudoknot structure in human telomerase RNA. *Proc Natl Acad Sci USA* 2005;**102**:8080–85

135. Byun Y, Han K. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res* 2006;**34**:416–22