Computational analysis of noncoding RNAs

Published in final edited form as:
Wiley Interdiscip Rev RNA. 2012 November ; 3(6): 759–778. doi:10.1002/wrna.1134.

Computational Analysis of Noncoding RNAs

Stefan Washietl[a,b,e], **Sebastian Will**[a,e], **David A. Hendrix**[a], **Loyal A. Goff**[a,d], **John L. Rinn**[b,d], **Bonnie Berger**[a,b], and **Manolis Kellis**[a,b]

[a]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology Cambridge, MA 02139, USA

[b]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge Massachusetts 02142, USA

[c]Mathematics Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[d]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA

Abstract

Noncoding RNAs have emerged as important key players in the cell. Understanding their surprisingly diverse range of functions is challenging for experimental and computational biology. Here, we review computational methods to analyze noncoding RNAs. The topics covered include basic and advanced techniques to predict RNA structures, annotation of noncoding RNAs in genomic data, mining RNA-seq data for novel transcripts and prediction of transcript structures, computational aspects of microRNAs, and database resources.

1. Introduction

Noncoding RNAs (ncRNAs) are transcripts that are not translated to proteins but act as functional RNAs. Several well-known ncRNAs such as transfer RNAs or ribosomal RNAs can be found throughout the tree of life. They fulfill central functions in the cell and thus have been studied for a long time.

However, over the past years a few key discoveries have shown that ncRNAs have a much richer functional spectrum than anticipated[1]. The discovery of microRNAs for example changed our view of how genes are regulated[2,3]. Another surprising observation revealed by high-throughput methods is that in human 90% of the genome is transcribed at some time in some tissue[4]. Although the full extent and functional consequences of this pervasive transcription remains highly controversial[5,6], the vast amount of transcripts produced suggests that many important ncRNA functions are yet to be discovered.

In particular, long noncoding RNAs (lncRNAs) – transcripts that can be several kilobases in length, spliced and processed like mRNAs but lack obvious coding potential – seem to be a rich source of novel functions[7]. All these ncRNAs have been suggested to form a hidden layer of regulation that is necessary to establish the complexity of eukaryotic genomes[8].

Prokaryotic genomes also contain many surprises. Riboswitches[9], small regulatory RNAs[10], or completely unknown structured RNAs[11] suggest that ncRNAs also form an important functional layer in bacteria.

[e]Correspondence: Stefan Washietl (wash@mit.edu) and Sebastian Will (wills@mit.edu) .

Understanding the function of ncRNAs – in particular in the age of high-throughput experiments – is clearly not possible without computational approaches. Algorithms to annotate, organize and functionally characterize ncRNAs are of increasing relevance. In this paper, we give a broad overview of programs and resources to analyze many different aspects of ncRNAs (Fig. 1).

2. Structural analysis

For many RNAs there is a close connection between structure and function. Having a good model of the structure of an RNA is thus critical and often the first clue towards elucidating its function. Determining the complete three dimensional structure ("tertiary structure") of an RNA is a tedious and time-consuming undertaking. Computational methods – either completely *de novo* or assisted by experimental data – are therefore routinely used to *predict* structure models. A strong focus lies on the prediction of the secondary structure, i.e. the pattern of intramolecular base-pairs (A·U, G·C and G·U) typically formed in RNAs.

2.1. Secondary structure

2.1.1. Thermodynamic folding of single sequences—In RNAs, the secondary structural elements are responsible for most of the overall folding energy and can be seen as a coarse-grained approximation of the tertiary structure. This important biophysical property in combination with the fact that secondary structure can easily be formalized as a simple graph (Fig. 2), led to secondary structure being widely studied early on. One of the first attempts to approach the RNA folding problem (i.e. predicting the secondary structure from the primary sequence) was by Nussinov and Jacobson[12]. They proposed an algorithm to find the secondary structure with the maximum number of base pairs. It is one of the classical examples of dynamic programming algorithms in computational biology and all modern variants of folding algorithm essentially use the same principle (Box 1).

In practice, however, finding the structure with the maximum number of pairs does not give accurate results. Ideally, we want to find the structure of minimum folding energy. Since most of the folding energy in RNAs is contributed by stacking interactions between neighbouring base pairs, counting single base pairs is not sufficient. Therefore, current folding algorithms use a "nearest neighbour" or also called "loop-based" energy model. A structure is uniquely decomposed into substructural elements (stacked bases, hairpin-loops, bulges, interior-loops, and multi-way-junctions, Fig. 2A). The structural elements are assigned energies which add up to the total folding energy of the structure (Fig. 2B). The energy values are established empirically and typically come from systematic melting experiments on small synthetic RNAs. An up-to-date set of energy parameters is maintained by Douglas Turner's lab[13,14].

A dynamic programming algorithm to find the minimum free energy (MFE) for this more complex energy model was proposed by Zuker and Stiegler[15] and forms the basis for modern prediction programs. The most common implementations used are UNAFold[16], RNAfold of the Vienna RNA package[17] and RNAstructure[18]. The accuracy of MFE predictions depend on the type of RNA. Although some RNAs can be predicted with high accuracy, in general one has to expect that roughly a third of the predicted base pairs are wrong and one third of true base pairs are missed. It is thus important to keep in mind that even the best currently available prediction methods only give a rough model of the structure. Tertiary interactions, protein context and other inherent limitations of the energy model are all sources of potential errors.

At room temperature, RNAs usually exist in an ensemble of different structures and the MFE structure is not necessarily the biologically relevant structure. There are different algorithms to predict suboptimal structures close to the MFE structure[19,20]. McCaskill's algorithm[21] allows one to calculate the partition function over *all* possible structures and subsequently the probability of a particular base pair in the thermodynamic ensemble. Considering the pair-probability matrix of all possible base pairs gives a more comprehensive view of the structural properties of an RNA than just the MFE prediction. It is also possible to obtain individual structure predictions from the pair-probability matrix, either by sampling[22] or by finding a structure that maximizes the expected accuracy considering a weighting factor between sensitivity and specificity[23,24].

2.1.2. RNA folding using probabilistic models—An alternative to the thermodynamic approach of RNA folding is a probabilistic approach based on machine-learning principles. Instead of using energy parameters, folding parameters can be estimated from a training set of known structures and are used to predict structures of unknown sequences. There are several probabilistic frameworks to accomplish parameter estimation and prediction. Stochastic context-free grammars (SCFGs) are a generalization of Hidden Markov models that are widely used in bioinformatics (Box 2). SCFGs allow one to consider nested dependencies, a prerequisite to model RNA structure. They have been used successfully for homology search problems and consensus structure prediction (see below). Although SCFGs can be used for single sequence structure prediction[25] they are not widely used for this problem. CONTRAFOLD[23], an alternative machine learning approach based on conditional random fields, however, could establish itself as a serious alternative to thermodynamic methods. There are also hybrid approaches that try to enhance the thermodynamic parameters by training on known structures[26,27].

2.1.3. Incorporating structure probing data into folding algorithms—Structure probing experiments typically use enzymatic or chemical agents that specifically target paired or unpaired regions[28]. Most implementations of thermodynamic folding like UNAfold or RNAfold allow for the incorporation of this type of information by restricting the folding to structures consistent with the experimental constraints. As an alternative, experimental information can also be incorporated as "pseudo-energies" into the folding algorithm enforcing regions to be preferentially paired or unpaired reflecting the experimental evidence[29–31]. Recently, high-throughput sequencing techniques were used to scale up structure probing experiments massively[32–34]. Dealing with inherently noisy data of this type turned out to be challenging and is still an active field of research.

2.1.4. Secondary structure prediction for homologous sequences—Another way to improve secondary structure prediction is to consider homologous sequences from related species. If two or more sequences share a common structure but have diverged on the sequence level, typical base substitution patterns that maintain the common structure can be observed (Fig. 3A). A consistent mutation changes one base (e.g. A·U ↔ G·U) while compensatory mutations change both bases in the base pair (e.g. A·U ↔ G·C or C·G ↔ G·C). Clearly, these patterns provide useful information to infer a secondary structure.

The simplest way to exploit this signal is to calculate a mutual information score to find columns that show highly correlated mutation patterns. This method led to surprisingly accurate structures for rRNAs as early as 30 years ago[35]. Since we rarely have such large datasets of RNAs with extremely conserved structures it is natural to combine co-variance analysis with classical folding algorithms. RNAalifold[36] extends Zuker's folding algorithm to multiple sequence alignments by averaging the energy contribution over the sequences and adding co-variance information in the form of "pseudo-energies". A probabilistic alternative is Pfold[37], which uses a simple stochastic context free grammar combined with an evolutionary model of sequence evolution to infer a consensus structure. PetFold[38] extends Pfold by incorporating pair probabilities from thermodynamic folding and thus unifies evolutionary and thermodynamic information. A more recent program, TurboFold[39] also uses thermodynamic folding and iteratively refines the energy parameters by incorporating pair probabilities from homologous sequences.

2.1.5. Structural alignment—

Even unaligned RNAs can provide more information about their common structure than a single sequence. Low sequence homology below of 60% sequence identity[40] prohibits the sequence alignment-based approach of the previous section (Fig. 3B), since correct alignment requires information about the structure. Since structure predictions for single sequences are unreliable, folding the sequences followed by structure-based alignment can also fail.

Therefore, the most successful strategies fold and align the RNAs *simultaneously*. The first such algorithm[41], by Sankoff, simultaneously optimizes the alignment's edit distance and the free energies of both RNA structures applying a loop-based energy model. However, only recent advances made this strategy applicable to practical RNA analysis.

The first complete pairwise Sankoff-implementation Dynalign[42] implements a loop-based energy model, but employs a simple banding technique for increasing efficiency. A further pairwise Sankoff-like tool is Foldalign[43]. Computing a loop-based energy model during the alignment, these algorithms are accurate but computationally expensive; in practice, they compensate this by strong sequence-based heuristic restrictions.

Several less expensive Sankoff-like algorithms are based on simplifications introduced by PMcomp[44]. PMcomp replaces the loop-based energy model by assigning "pseudo-energies" to single base pairs. This reduces the computational overhead significantly. By computing the pseudo-energies of base pairs from their probabilities in the structure ensembles of the single RNAs, accurate information from the loop-based energy model is fed back into the light-weight algorithm.

Approaches following this idea are LocARNA[45], foldalignM[46], RAF[47], and LocARNA-P[48]; All these tools additionally employ sparsity in the structure ensemble of the single sequences.

The sparsified PMcomp-like approaches are sufficiently fast for multiple alignment and large scale studies, performing e.g. clustering[45,46] and *de novo* prediction of structural RNA[49].

2.2. Prediction with pseudoknots

Pseudoknotted structures follow the same rules as other secondary structures, but allow non-tree like configurations, e.g. due to an additional level of nested base pairs or pairing between different hairpin loops (kissing hairpin) (Fig. 4). Pseudoknots are prevalent in many ncRNAs; still, most algorithms ignore them for technical reasons: pseudoknot-folding is computationally expensive and accurate empirical energy models are missing.

2.2.1. Algorithmic challenges—

The run-time of pseudoknot-free structure prediction grows only with the cube of the sequence length. Unfortunately, when general pseudoknots are allowed, the computation time grows much faster, namely exponentially with the sequence length[50]. Consequently, finding exact solutions is intractable for all but very short RNAs. Note that the "principle of optimality", which allows dynamic programming (Box 1) in the pseudoknot-free case, is not applicable in the case of general pseudoknots. By this principle, every optimal structure can be composed from optimal structures of its subsequences.

In practice, often heuristic methods are applicable. Among the numerous approaches are ILM[53], HotKnots[54], KnotSeeker[55], and IPknot[56]. ILM[53] applies the classic principle of “hierarchic folding”; it constructs a pseudoknotted structure by iteratively predicting a most likely stem, using pseudoknot-free prediction, which is then added to the structure. HotKnots[54] refines the construction of the pseudoknotted structure from likely more general secondary structure elements. TurboKnot[57] even predicts conserved pseudoknots from a set of homologous RNAs. Applying a topological classification of RNA structures[58], TT2NE[59] guarantees to find the best RNA structure regardless of pseudoknot complexity; however, this limits the length of treatable sequences.

Other algorithms restrict the types of pseudoknots, such that dynamic programming can be applied[50–52,60–63]. These algorithms differ in their computational complexity and the complexity of considered structures. Fig. 4 shows pseudoknots of different complexity.

Rivas and Eddy[51] proposed the most general such algorithm. It predicts three-knots (Fig. 4C), but cannot predict more complex pseudoknots such as the one shown in Fig. 4D. Its large time and space requirements prohibit its application to large scale data analysis. The most efficient such algorithm[52] has only the same space requirements as pseudoknot-free prediction; its run-time grows with the fourth power of the sequence length, adding only a linear factor over pseudoknot-free folding. However, it predicts only *canonical pseudoknots* (Fig. 4E), which are formed by two canonical stems: such stems are (i) composed of canonical base pairs (ii) “*perfect*”, i.e. they do not contain interior loops or bulges, and (iii) *maximally extended*, i.e. they cannot be extended by canonical base pairs. Further such algorithms are tailored for specific interesting pseudoknot-classes (Fig. 4A and Fig. 4B[63]). Möhl *et al.*[64] recently managed to speed up such algorithms non-heuristically.

2.2.2. Energy models for pseudoknots—A further challenge of pseudoknot prediction is to find an accurate energy model. The established loop-based energy models for RNA are tailored for pseudoknot-free structures; to date, there are no empirical energy parameters for pseudoknotted structure elements.

Consequently, some algorithms consider only the simplistic case of base-pair maximization[65,53]. Although some authors argue that important entropy contributions in pseudoknots cannot be covered by a loop-based energy model[66], most approaches extend the loop-based energy model for pseudoknot-loops[51,67].

2.3. Tertiary structure

While secondary structure is strongly stabilizing the three-dimensional structure, the tertiary structure depends on stabilizing non-canonical base pairs and van-der-Waals interactions. Furthermore, pseudoknots impact the tertiary structure. Therefore, deriving the tertiary structure in a hierarchic way from predicted secondary structure is not straightforward.

There are two main ways to model tertiary RNA structure. One, template-based modeling, employs homology to other RNAs with known structures. The other, *de novo* prediction, computes structures from physical and knowledge based rules. For example, the MC-fold/MC-sym pipeline[68] (Fig. 5) assembles fragments of experimentally determined three-dimensional structures. Based on such structures, the approach builds a library of frequent small secondary structure loop-motifs, called Nucleic Cyclic Motifs (NCMs), together with their three-dimensional configurations. Given a RNA sequence, MC-Fold constructs probable secondary structures by merging NCMs; in this process it assigns likely NCMs to subsequences. MC-sym assigns concrete 3D-structures to the NCMs to generate a consistent three-dimensional structure. The approach has been tested on 13 RNAs of an average size of 30 nucleotides. Running several hours per prediction, the known 3D-structures have been reproduced within 2.3Å at average[68]. Similar to MC-Fold, the recent RNAwolf[69] predicts extended secondary structures considering non-canonical base pairs; the same work reports a more efficient, dynamic programming-based reimplementation of MC-Fold, which improves parameter estimation. A more detailed review and a systematic performance comparison of RNA tertiary structure prediction programs is provided by Laing *et al.*[70].

2.4. RNA/RNA interactions

Many ncRNAs interact with other RNAs by specific base-pairing; most prominently, microRNAs target the untranslated regions of mRNAs. Predicting RNA/RNA interactions can thus elucidate RNA interaction partners and potential functions.

Most generally, one aims to predict the secondary structure of the interaction complex of two RNAs consisting of *intra*molecular and *inter*molecular base-pairs (Fig. 6). Alkan *et al.*[71] formalized the problem and showed that – similar to pseudoknot prediction – it cannot be solved efficiently. Therefore, several simplifications and heuristics have been proposed.

Most approaches restrict the possible structures of the interaction complex to enable efficient algorithms using dynamic programming. Fig. 6 shows interaction complexes from several restriction classes. The simplest approaches ignore intramolecular base-pairs and predict only the best set of interacting base-pairs (Fig. 6A); examples are RNAhybrid[72] and RNAduplex[73].

A more general approach optimizes intra- and intermolecular base-pairs simultaneously in a restricted structure space. “Co-folding” of RNAs, for example implemented by RNAcofold[73], concatenates the two RNA sequences and predicts a pseudoknot-free structure for the concatenation. Co-folding leads to a very efficient algorithm but strongly restricts the space of possible structures, such that only external bases can interact (Fig. 6B).

The dynamic-programming algorithms[71,74] that predict more general structures (Fig. 6C), forbidding only pseudoknots, crossing interaction, and zig-zags (Fig. 6D), are computationally as expensive as the most complex efficient pseudoknot prediction algorithm[51]; they are therefore rarely used in practice, albeit their efficiency has been improved recently[75].

Several fast methods[71,76,77] assume that interactions form in two steps: First, the RNA unfolds partially, which requires certain energy to open the intramolecular base-pairs. Second, the unfolded, now accessible, RNA hybridizes with its partner forming energetically favorable intermolecular base-pairs. RNAup[76] computes the energies to unfold each subsequence in the single RNAs and combines the unfolding energies with the hybridization energies to approximate the energy of the interaction complex. IntaRNA[77] optimizes this approach and extends it to screen large data sets for potential interaction targets.

Finally, several approaches predict conserved interactions between multiple sequence alignments[78,79].

2.5. Kinetic folding

Common structure prediction methods assume that the functional RNA structure can be identified solely based on the thermodynamic equilibrium without considering the dynamics of the folding process. Although the true impact of kinetics on functional RNA structures is still unknown, for example RNA-switches[80] show the importance of understanding the RNA folding process.

Several groups have studied the folding process of RNAs (reviewed in[66,83]). The RNA folding process is commonly modeled using energy landscapes[84]. Such landscapes assign energies to single structures, or states, and define neighborship between states. RNAlocopt[85] enables studying the Boltzmann ensemble of local optima in an RNA energy landscape. The folding process iteratively moves from one state to a neighbor; the move probability depends on the energy difference. Studying folding by simulation[83] is expensive since it requires averaging over many trajectories. Because the exact model of folding as a Markov process can be solved only for small systems, many methods coarse-grain the energy landscape to enable the analysis of the process. For example "barrier-trees" represent the energy landscape as a tree of local minima connected by their saddle points[82]. BarMap[81] generalizes coarse-graining to non-stationary scenarios like temperature changes or co-transcriptional folding.[86] predicts RNA folding pathways based on motion planning techniques from robotics. Kinefold[87] simulates single folding paths over seconds to minutes for sequences up to 400 bases.

3. Annotating ncRNAs in genomic data

Another major challenge in understanding the function of ncRNAs is to find and annotate them in complete genomes. We distinguish homology search, i.e. trying to identify new members of already known classes of ncRNAs, and *de novo* prediction with the aim to discover novel ncRNAs.

3.1. *De novo* prediction

Although a general *de novo* ncRNA finder remains elusive, some progress has been made in the identification of structural RNAs, i.e. ncRNAs that rely on a defined secondary structure for their function.

As a first attempt, one could use normal folding algorithms such RNAfold and hope to find structural RNAs to be thermodynamically more stable than the genomic background. However, although on average structural RNAs are more stable than expected this approach is generally not significant enough to reliably distinguish true structural RNAs from the rest of the genome[88,89]. Comparative approaches that make use of evolutionary signatures in alignments of related sequences can improve the signal considerably.

The first program that used pairwise alignments to find structured RNAs was QRNA[90]. Based on stochastic context free grammars it could successfully identify novel ncRNAs in bacteria[91,92]. MSARI was the first algorithm applying the idea of finding conserved RNA structures to multiple sequence alignments[93].

To screen larger genomes higher accuracy was necessary. RNAz[94] analyzes multiple sequence alignments and combines evidence from structural conservation and thermodynamic stability. EvoFold[95] searches for conserved secondary structures in multiple alignments using a phylogenetic stochastic context-free grammar. Both programs were used to map potential RNA secondary structures in the human[95,96] and many other genomes (e.g.[97]).

Another approach that was used to detect conserved RNA secondary structures in bacteria[98] is implemented in CMFinder[99]. CMFinder builds a covariance model from a set of unaligned sequences by iterative optimization.

3.2. Homology search

Pure sequence based search algorithms like BLAST quickly reach their limits when used to identify distant homologues of RNAs[100,101].

A solution is to include structure information in the search. Several motif description languages have been developed that allow one to manually specify sequence and structure properties and subsequently use these patterns to search databases or genomic data. Examples of such descriptor based search algorithm are RNAMOT[102] and RNAmotif[103].

Another class of programs automatically create a description of a structural RNA from a structure annotated alignment. The most commonly used program of this class is INFERNAL that uses covariance models, a full probabilistic description of an RNA family based on stochastic context-free grammars[104]. The Rfam database (see below) is based on INFERNAL and provides a curated collection of such covariance models.

In addition to these generic homology search tools, there are several specialized programs for finding ncRNAs of a particular family such as tRNAs[105], rRNAs[106], snoRNAs[107–109], tmRNAs[110], signal recognition particle RNAs[111].

3.3. Coding potential

A complication during ncRNA annotation is the fact that many transcripts appear to be noncoding but in fact have the potential to code for a protein[112]. For example, short open reading frames can be easily missed and biological ambiguities of transcripts that act both on the level of the RNA and protein can make the annotation difficult[113,114]. A good overview of different methods to assess the coding potential of RNAs is given by Frith *et al*. The benchmark study[115] found comparative analysis to be one of the most promising approaches. Purifying selection on the protein sequence turned out to be a reliable indicator of coding potential and several programs were developed to exploit this feature[90,116,117].

4. Mining RNA-seq data for noncoding RNA transcripts

The advent of high-throughput RNA sequencing has provided a robust platform for the development and expansion of several transcriptome-level analyses. RNA-seq is the highly parallelized process of sequencing individual cDNA fragments created from a population of RNA molecules. Here we discuss three challenges that must be addressed to mine RNA-seq data for noncoding transcripts: (i) read mapping to a reference genome (or transcriptome) (ii) transcriptome reconstruction from mapped reads, and (ii) quantification of transcript levels.

4.1. Short read mapping

The first stage in short-read sequencing data analysis is the alignment of the sequenced reads to a reference genome. The algorithmic details of short read mapping is beyond the scope of this review. Here, it is important to note that transcript reconstruction, requires so-called “spliced aligners”. Spliced aligners such as TopHat[118], GSNAP[119], and SpliceMap[120] identify and map short reads that span exon-exon junctions. From these spliced alignments novel splicing events and subsequently new transcript models can be identified.

4.2. Reconstruction of transcript models

A key advantage of RNA-seq over traditional forms of RNA expression analysis is the fact that little to no *a priori* information on the presence of an RNA sequence is required and, in principle, all required information can be learned directly from the data. This advantage, however, is dependent on the ability to re-construct a transcriptome fragmented into millions of short reads. Common approaches to solving this jigsaw puzzle-like problem focus on one of two different strategies: (i) reference-guided assembly, or (ii) *de novo* assembly (Fig. 8).

With a reference-guided assembly, reads are initially aligned using a spliced aligner to a reference genome sequence. The requirement for gapped alignments allows for discovery of putative splice junctions at the locations in which a read maps to the reference with a gap across an appropriately sized genomic interval. The two most popular reference-guided transcriptome assembly tools, Scripture[121] and Cufflinks[122], both treat these gaps as candidate splice junctions, and use this information to construct a graph representation of the transcriptome. In the case of Scripture, the graph represents the exonic connectivity potential of the reference genome. Cufflinks creates independent graph models for each independent genomic interval assumed to be a putative “gene”. In eithercase, the various paths through these connectivity graphs represent independent transcript isoforms. Scripture will attempt to identify all possible paths through the graph that can be explained by the mapped reads for a given gene and in this regard is useful for identifying lowly expressed isoforms, but tends to produce more noise in highly spliced structures. In contrast, Cufflinks produces a set of isoforms that represent the most parsimonious paths that can explain the given mapped reads, which may not report some redundant (but true) isoforms, but does not overburden the results with false positives. Additionally, Cufflinks estimates the read coverage across the paths to assist the selection of the most parsimonious isoforms. The result of either of these two approaches is a reconstructed transcriptome, the detail of which is supported by the read sequences, abundance, and mappability to a reliable reference.

In contrast to these reference-guided approaches, Velvet[123] and transABySS[124] use the short-read sequences directly and attempt to construct contig-like transcripts[124]. This approach tends to be significantly more computationally intensive, but is essential in species that do not have a reliable reference genome, or in the case when the expected transcriptome can deviate significantly from the reference genome due to rearrangements.

4.3. Quantification and differential expression

In RNA-Seq, the number of individual sequenced fragments from a given transcript is used as a proxy for its abundance. Determinations of expression level can be coarsely determined at the gene-level[125,126] using a pseudo-model that consists of either the most-abundant isoform model, an intersection model quantifying only the regions present in all predicted isoforms, or a union model. The intersection model has been shown to to reduce the ability to accurately determine differential expression and the union model can under-estimate expression for those genes with alternative splicing[127,122]. More accurately, gene-level estimates can be determined as the sum of isoform-level abundance estimates[122,128] involving a likelihood function to model the various effects encountered in the sequencing process[129]. The result of fitting these models to the data is a maximum likelihood estimate of the isoform-level abundances for each gene. Gene-level abundance estimates are easily determined by summing the expression levels of individual isoforms.

RNA-seq expression values must be normalized to correct for inherent biases in the data. The “Reads Per Kilobase of transcript per Million mapped” (RPKM) has emerged as a standard metric for reporting of estimated abundance levels. This metric has the advantage of correcting for the two main sources of variability in RNA-seq data: the length of the transcript, and the depth of the libraries.

The robust quantification of transcript levels also allows one to study differential expression. Since most gene-level projections of abundance estimation result in a single RPKM value for each gene, it would be reasonable to directly use most of the many differential expression tests that have been developed for microarray analysis over the past few years. There are, however, additional benefits that can be gained from using RNA-seq data such as the ability to derive a distribution of abundance estimates from a given sample or set of samples. Short read mapping to a given genomic interval can be considered a counting problem, and many differential expression analyses initially attempted to fit read counts to either Poisson or Binomial distributions to determine enriched transcripts. These methods, however, fail to incorporate any information about biological variability. Several more recent applications such as Cuffdiff[122], DESeq[130], and EdgeR[131] incorporate variance information from biological replicates in their differential expression models leading to more rigorous statistics.

5. MicroRNAs

MicroRNAs (miRNAs/miRs) are short endogenous regulatory non-coding RNAs found in eukaryotic cells, whose primary function is to post-transcriptionally repress genes[132]. miRNAs inhibit translation and promote mRNA degradation via sequence-specific binding to the 3′ UTR regions and coding sequences[133–135]. They are produced from hairpin precursors (pri-miRNAs) that are processed by Drosha to form a pre-miR hairpin and then by Dicer to generate one or more 18- to 23-nt mature microRNAs[136]. Mature microRNAs are then incorporated into RISC where they hybridize with target sites of the mRNA, which are complementary to the microRNA seed (positions 2-8), leading to post-transcriptional repression. Since their discovery, there has been much interest in the computational identification of miRNAs at a genome wide level using some combination of evolutionary conservation, hairpin structure, thermodynamic stability, genomic context, and more recently the presence of the mature miRNA in sequencing data[137].

5.1. Identification of miRNAs

5.1.1. Evolutionary conservation—Some of the first approaches such as miRScan and snarloop use conservation and similarity to known microRNAs for the prediction new examples[138,139]. Other approaches such as miRSeeker incorporate microRNA-specific patterns of conservation, such as stronger conservation in the hairpin stem compared to the loop[140]. Additionally, patterns of conservation of target sites have been used to identify novel microRNAs[141] and to refine annotations of known microRNAs[142]. Other approaches combine both sequence and structural alignments to find microRNA homologs[143,144].

5.1.2. Structural properties—The secondary structure and thermodynamic stability are important features for the prediction of miRNAs, especially when they are not conserved or orthologs do not exist in known species. Because miRNAs need to form stable hairpins for their processing, many studies have used structural features for their prediction. It has been demonstrated that miRNAs are significantly more stable than randomized sequences of the same nucleotide or dinucleotide composition[145], and many studies have used programs like RNAz to predictnovel microRNAs based on this characteristic feature[146,147]. Other studies have developed machine learning approaches that train classifiers on known miRNAs and subsequently identify novel, and in many cases nonconserved, microRNAs[148–150].

5.1.3. Genomic context—Other approaches have looked for features in the surrounding genomic context for the prediction of novel miRNAs and for refining other predictions. Some early work helped to filter predictions with a characteristic motif upstream of and patterns of conservation flanking the pre-miR[151]. Other studies have used the fact that microRNAs tend to reside within polycistronic clusters of more than one miR to identify novel miRNAs[3,152]. Some approaches also make use of the fact that regions proximal to microRNAs tend to be devoid of other non-miR small RNAs and when they are flanked by other small RNAs such as miRNA offset RNAs (moRs) the separation from mature microRNA sequences is minimal[153].

5.1.4. Next generation sequencing—The analysis of the sequencing of size-selected cDNA libraries has proved to be the most reliable method for the identification of novel microRNAs, in most instances coupled with other features such as structure and conservation to enhance predictions, because it provides validation that the mature sequence is expressed. In addition to novel miRNAs, the analysis of high-throughput sequencing data of small RNAs has led to the elucidation of many other classes of small RNAs including endogenous siRNAs[154], piRNAs[155,156], and moRs[157] among others. There are now a few publicly available software tools for the prediction of microRNAs from high-throughput sequencing data such as miRDeep, MIReNA, and miRTRAP[158,153,159].

5.2. miRNA target prediction

MicroRNA *target* prediction is another lively area of computational analysis related to microRNAs. Early approaches such as targetScan identify evolutionary conserved seed matches and later approaches such as PicTar have incorporated target site stability[160,155]. The topic is related to the problem of predicting RNA/RNA interactions, discussed above. For a comprehensive review on target prediction, see Bartel[161].

6. Databases

There are many databases related to ncRNAs and we cannot cover all of them here. A more specialized review[162] and the yearly database issue of *Nucleic Acids Research*[163] are good resources to get a more detailed overview.

There are many highly specialized databases that collect RNAs of a specific class. Basically for all well known "classical" RNAs like tRNAs, rRNAs, snoRNAs, SRP RNAs, tmRNAs, group I or II introns a database is available[162].

All newly identified ncRNA sequences are usually deposited in general sequence database such as Genbank. However, typically they are not systematically annotated in these databases and consequently a few other databases have emerged that systematically collect ncRNAs (NONCODE[164], RNAdb[165] fRNAdb[166] lncRNAdb[167]).

Rfam is an important resource for structured RNAs and also includes structured regulatory elements in mRNAs[168]. It collects hand curated covariance models (see above) that are used to systematically search sequence databases for new members. As of writing this review, Rfam contains 1973 families with a total of 2,756,313 members (Fig. 9). All RNA families in Rfam are manually annotated.

The most extensive database for microRNA sequences, hairpins, and target sites is miRBase (http://www.mirbase.org)[169]. miRBase has seen rapid growth over the past few years (Fig. 9) and is the official repository and naming authority for newly discovered miRNAs. Other related databases include Tarbase, which is a database of experimentally verified target sites[170], and miR2Disease, which is a database that maintains a manually curated set of disease associated microRNA target sites[171].

6.1. Conclusions and outlook

The wide variety of topics covered in this paper reflects the increasing complexity of the field. It also clearly demonstrates the interdisciplinary effort that is necessary to address these problems. It is safe to predict that computational problems related to ncRNAs will remain challenging for the coming years. In particular elucidating the functions of lincRNAs will require new approaches. Many methods for structural analysis, for example, were developed for rather short structured ncRNAs and cannot be directly applied to lncRNAs that can be kilobases in length. Prediction of long range intramolecular interactions within lncRNAs or prediction of intermolecular RNA/RNA or RNA/DNA interactions of lncRNAs will require extensions and improvements of established algorithms. Also the problem of predicting protein-RNA interactions will be of high relevance given the increasing number of examples of lincRNAs that act as scaffolds for protein complexes. Also more accurate and efficient analysis of high-throughput data will be a challenge for the field. We have mentioned analysis of RNA-seq data, but next generation sequencing can also used for a variety of other ncRNA related problems such as high throughput RNA secondary structure probing or mapping RNA/protein interactions. Also new approaches to organize ncRNA data will be important and there is need for new centralized databases and specialized resources[172].

Acknowledgments

This work was supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung [Schrödinger Fellowship J2966-B12 to StW] and by the German Research Foundation (grant WI 3628/1-1 to SW) and NIH award 1RC1CA147187 to DAH.

References

1. Amaral P, Dinger M, Mercer T, Mattick J. The eukaryotic genome as an RNA machine. Science. 2008; 319:1787–9. [PubMed: 18369136]
2. Pasquinelli A, Reinhart B, Slack F, Martindale M, Kuroda M, Maller B, Hayward D, Ball E, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 2000; 408:86–9. [PubMed: 11081512]
3. Lau N, Lim L, Weinstein E, Bartel D. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science. 2001; 294:858–62. [PubMed: 11679671]
4. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146):799–816. [PubMed: 17571346]
5. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most ”dark matter” transcripts are associated with known genes. PLoS Biol. 2010; 8(5):e1000371. [PubMed: 20502517]
6. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS. The reality of pervasive transcription. PLoS Biol. 2011; 9(7):e1000625. [PubMed: 21765801]
7. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. 2011; 477(7364):295–300. [PubMed: 21874018]
8. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays. 2003; 25(10):930–9. [PubMed: 14505360]
9. Breaker RR. Riboswitches and the RNA World. Cold Spring Harb Perspect Biol. 2010
10. Frhlich KS, Vogel J. Activation of gene expression by small RNA. Curr Opin Microbiol. 2009; 12(6):674–82. [PubMed: 19880344]
11. Weinberg Z, Perreault J, Meyer MM, Breaker RR. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. Nature. 2009; 462(7273):656–9. [PubMed: 19956260]
12. Nussinov R, Jacobson A. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A. 1980; 77(11):6309–13. [PubMed: 6161375]
13. Mathews D, Sabina J, Zuker M, Turner D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999; 288:911–40. [PubMed: 10329189]

14. SantaLucia J Jr, Burkard M, Kierzek R, Schroeder S, Jiao X, Cox C, Turner D. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998; 37:14719–35. [PubMed: 9778347]
15. Zuker M, Stiegler P. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research. 1981; 9:133–148. [PubMed: 6163133]
16. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008; 453:3–31. [PubMed: 18712296]
17. Gruber A, Lorenz R, Bernhart S, Neuböck R, Hofacker I. The Vienna RNA websuite. Nucleic Acids Res. 2008; 36:W70–4. [PubMed: 18424795]
18. Reuter J, Mathews D. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010; 11:129. [PubMed: 20230624]
19. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999; 49(2):145–65. [PubMed: 10070264]
20. Zuker M. On finding all suboptimal foldings of an RNA molecule. Science. 1989; 244:48–52. [PubMed: 2468181]
21. McCaskill J. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29:1105–19. [PubMed: 1695107]
22. Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA. 2005; 11(8):1157–66. [PubMed: 16043502]
23. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006; 22(14):e90–8. [PubMed: 16873527]
24. Lu Z, Gloor J, Mathews D. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA. 2009; 15:1805–13. [PubMed: 19703939]
25. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004; 5:71. [PubMed: 15180907]
26. Andronescu M, Condon A, Hoos H, Mathews D, Murphy K. Computational approaches for RNA energy parameter estimation. RNA. 2010; 16:2304–18. [PubMed: 20940338]
27. Andronescu M, Condon A, Hoos H, Mathews D, Murphy K. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics. 2007; 23:i19–28. [PubMed: 17646296]
28. Weeks K. Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol. 2010; 20:295–304. [PubMed: 20447823]
29. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci USA. 2004; 101:7287–7292. [PubMed: 15123812]
30. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA. 2009; 106:97–102. [PubMed: 19109441]
31. Washietl S, Hofacker IL, Stadler PF, Kellis M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res. 2012; 40(10): 4261–4272. [PubMed: 22287623]
32. Kertesz M, Wan Y, Mazor E, Rinn J, Nutter R, Chang H, Segal E. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010; 467:103–7. [PubMed: 20811459]
33. Underwood J, Uzilov A, Katzman S, Onodera C, Mainzer J, Mathews D, Lowe T, Salama S, Haussler D. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods. 2010; 7:995–1001. [PubMed: 21057495]
34. Lucks J, Mortimer S, Trapnell C, Luo S, Aviran S, Schroth G, Pachter L, Doudna J, Arkin A. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci U S A. 2011
35. Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, Herr W, Stahl DA, Gupta R, Woese CR. Secondary structure model for 23S ribosomal RNA. Nucleic Acids Research. 1981; 9(22):6167–6189. [PubMed: 7031608]
36. Bernhart S, Hofacker I, Will S, Gruber A, Stadler P. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics. 2008; 9:474. [PubMed: 19014431]

37. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003; 31:3423–8. [PubMed: 12824339]

38. Seemann S, Gorodkin J, Backofen R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res. 2008; 36:6355–62. [PubMed: 18836192]

39. Harmanci AO, Sharma G, Mathews DH. TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences. BMC Bioinformatics. 2011; 12:108. [PubMed: 21507242]

40. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Research. 2005; 33(8):2433–9. [PubMed: 15860779]

41. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. 1985; 45(5):810–825.

42. 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]

43. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol. 2007; 3(10):1896–908. [PubMed: 17937495]

44. Hofacker IL, Bernhart SH, Stadler PF. Alignment of RNA base pairing probability matrices. Bioinformatics. 2004; 20(14):2222–7. [PubMed: 15073017]

45. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. PLoS Computational Biology. 2007; 3(4):e65. [PubMed: 17432929]

46. Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics. 2007; 23(8):926–32. [PubMed: 17324941]

47. Do CB, Foo CS, Batzoglou S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics. 2008; 24(13):i68–76. [PubMed: 18586747]

48. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. RNA. 2012; 18(5):900–14. [PubMed: 22450757]

49. Will, S.; Yu, M.; Berger, B. Structure-based Whole Genome Realignment Reveals Many Novel Non-coding RNAs. Proceedings of the 16th International Conference on Research in Computational Molecular Biology (RECOMB 2012); Springer-Verlag; 2012. p. 341

50. Lyngso, RB.; Pedersen, CNS. Pseudoknots in RNA Secondary Structures. Proc. of the Fourth Annual International Conferences on Computational Molecular Biology (RE-COMB'00); ACM Press; 2000. [BRICS Report Series RS-00-1]

51. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. Journal of Molecular Biology. 1999; 285(5):2053–68. [PubMed: 9925784]

52. Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics. 2004; 5:104. [PubMed: 15294028]

53. 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. [PubMed: 14693809]

54. Ren J, Rastegari B, Condon A, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA. 2005; 11(10):1494–504. [PubMed: 16199760]

55. Sperschneider J, Datta A. KnotSeeker: heuristic pseudoknot detection in long RNA sequences. RNA. 2008; 14(4):630–40. [PubMed: 18314500]

56. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011; 27(13):i85–93. [PubMed: 21685106]

57. Seetin MG, Mathews DH. TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots. Bioinformatics. 2012; 28(6):792–8. [PubMed: 22285566]

58. Bon M, Vernizzi G, Orland H, Zee A. Topological classification of RNA structures. J Mol Biol. 2008; 379(4):900–11. [PubMed: 18485361]

59. Bon M, Orland H. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. Nucleic Acids Research. 2011; 39(14):e93. [PubMed: 21593129]

60. Uemura Y, Hasegawa A, Kobayashi S, Yokomori T. Tree adjoining grammars for RNA structure prediction. Theoretical Computer Science. 1999; 210:277–303. [Paper as Print Copy].

61. Akutsu T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. Discrete Applied Mathematics. 2000; 104:45–62.

62. Dirks RM, Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J Comput Chem. 2003; 24(13):1664–77. [PubMed: 12926009]

63. Chen HL, Condon A, Jabbari H. An O(n(5)) Algorithm for MFE Prediction of Kissing Hairpins and 4-Chains in Nucleic Acids. Journal of Computational Biology. 2009; 16(6):803–15. [PubMed: 19522664]

64. Mohl M, Salari R, Will S, Backofen R, Sahinalp SC. Sparsification of RNA structure prediction including pseudoknots. Algorithms Mol Biol. 2010; 5:39. [PubMed: 21194463]

65. Tabaska JE, Cary RB, Gabow HN, Stormo GD. An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics. 1998; 14(8):691–9. [PubMed: 9789095]

66. Chen SJ. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu Rev Biophys. 2008; 37:197–214. [PubMed: 18573079]

67. Andronescu MS, Pop C, Condon AE. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. RNA. 2010; 16:26–42. [PubMed: 19933322]

68. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008; 452(7183):51–5. [PubMed: 18322526]

69. zu Siederdissen CH, Bernhart SH, Stadler PF, Hofacker IL. A folding algorithm for extended RNA secondary structures. Bioinformatics. 2011; 27(13):i129–36. [PubMed: 21685061]

70. Laing C, Schlick T. Computational approaches to 3D modeling of RNA. J Phys Condens Matter. 2010; 22(28):283101. [PubMed: 21399271]

71. Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. RNA-RNA interaction prediction and antisense RNA target search. Journal of Computational Biology. 2006; 13(2):267–82. [PubMed: 16597239]

72. Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Research. 2006; 34(Web Server issue):W451–4. [PubMed: 16845047]

73. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte Chemie. 1994; 125:167–188.

74. Pervouchine DD. IRIS: intermolecular RNA interaction search. Genome Inform. 2004; 15(2):92–101. [PubMed: 15706495]

75. Salari, R.; Möhl, M.; Will, S.; Sahinalp, S. Backofen R: Time and Space Efficient RNA-RNA Interaction Prediction via Sparse Folding. In: Berger, B., editor. Proc. of RECOMB 2010. Vol. Volume 6044 of Lecture Notes in Computer Science. Springer Berlin; Heidelberg: 2010. p. 473-490.

76. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics. 2006; 22(10):1177–82. [PubMed: 16446276]

77. Busch A, Richter AS, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics. 2008; 24(24):2849–56. [PubMed: 18940824]

78. Li AX, Marz M, Qin J, Reidys CM. RNA-RNA interaction prediction based on multiple sequence alignments. Bioinformatics. 2011; 27(4):456–63. [PubMed: 21134894]

79. Seemann SE, Richter AS, Gesell T, Backofen R, Gorodkin J. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. Bioinformatics. 2011; 27(2):211–219. [PubMed: 21088024]

80. Nagel JHA, Pleij CWA. Self-induced structural switches in RNA. Biochimie. 2002; 84(9):913–23. [PubMed: 12458084]

81. Hofacker IL, Flamm C, Heine C, Wolfinger MT, Scheuermann G, Stadler PF. BarMap: RNA folding on dynamic energy landscapes. RNA. 2010; 16(7):1308–16. [PubMed: 20504954]

82. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF. Efficient computation of RNA folding dynamics. Journal of Physics A: Mathematical and General. 2004; 37(17):4731–4741. [http://stacks.iop.org/0305-4470/37/4731].

83. Flamm C, Hofacker I. Beyond energy minimization: approaches to the kinetic folding of RNA. Chemical Monthly. 2008; 139:447–457.

84. Chen SJ, Dill KA. RNA folding energy landscapes. Proc. Natl. Acad. Sci. USA. 2000; 97(2):646–51. [PubMed: 10639133]

85. Lorenz WA, Clote P. Computing the partition function for kinetically trapped RNA secondary structures. PLoS One. 2011; 6:e16178. [PubMed: 21297972]

86. Tang X, Thomas S, Tapia L, Giedroc DP, Amato NM. Simulating RNA folding kinetics on approximated energy landscapes. Journal of Molecular Biology. 2008; 381(4):1055–67. [PubMed: 18639245]

87. Xayaphoummine A, Bucher T, Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Research. 2005; 33(Web Server issue):W605–10. [PubMed: 15980546]

88. 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]

89. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol. 2004; 342:19–30. [PubMed: 15313604]

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

91. del Val C, Rivas E, Torres-Quesada O, Toro N, Jimnez-Zurdo JI. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. Mol Microbiol. 2007; 66(5):1080–91. [PubMed: 17971083]

92. Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol. 2001; 11(17):1369–73. [PubMed: 11553332]

93. Coventry A, Kleitman DJ, Berger B. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc Natl Acad Sci U S A. 2004; 101(33):12102–7. [PubMed: 15304649]

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

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

96. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermller 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, Guig R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res. 2007; 17(6):852–64. [PubMed: 17568003]

97. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, Ivens A, Newbold C, Pain A. Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. Genome Res. 2008; 18(2):281–92. [PubMed: 18096748]

98. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res. 2007; 35(14): 4809–19. [PubMed: 17621584]

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

100. 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. [PubMed: 17151342]

101. Menzel P, Gorodkin J, Stadler PF. The tedious task of finding homologous noncoding RNA genes. RNA. 2009; 15(12):2075–82. [PubMed: 19861422]

102. Gautheret D, Major F, Cedergren R. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. Comput Appl Biosci. 1990; 6(4):325–31. [PubMed: 1701686]

103. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 2001; 29(22):4724–35. [PubMed: 11713323]

104. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25(10):1335–7. [PubMed: 19307242]

105. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25(5):955–64. [PubMed: 9023104]

106. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35(9):3100–8. [PubMed: 17452365]

107. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science. 1999; 283(5405):1168–71. [PubMed: 10024243]

108. Schattner P, Decatur WA, Davis CA, Fournier MJ, Lowe TM. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. Nucleic Acids Res. 2004; 32(14):4281–96. [PubMed: 15306656]

109. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics. 2008; 24(2):158–64. [PubMed: 17895272]

110. 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(15):3449–53. [PubMed: 12140330]

111. Regalia M, Rosenblad MA, Samuelsson T. Prediction of signal recognition particle RNA genes. Nucleic Acids Res. 2002; 30(15):3368–77. [PubMed: 12140321]

112. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25(18):1915–27. [PubMed: 21890647]

113. Findei S, Engelhardt J, Prohaska SJ, Stadler PF. Protein-coding structured RNAs: A computational survey of conserved RNA secondary structures overlapping coding regions in drosophilids. Biochimie. 2011; 93(11):2019–2023. [PubMed: 21835221]

114. Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and non-coding RNA: challenges and ambiguities. PLoS Comput Biol. 2008; 4(11):e1000176. [PubMed: 19043537]

115. Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S, Furuno M, Bult CJ, Quackenbush J, Kai C, Kawai J, Carninci P, Hayashizaki Y, Pesole G, Mattick JS. Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol. 2006; 3:40–48. [PubMed: 17114936]

116. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011; 27(13):i275–i282. [PubMed: 21685081]

117. Washietl S, Findeiss S, Mller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA. 2011; 17(4):578–94. [PubMed: 21357752]

118. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. [PubMed: 19289445]

119. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26(7):873–81. [PubMed: 20147302]

120. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010; 38(14):4570–8. [PubMed: 20371516]

121. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28(5):503–10. [PubMed: 20436462]

122. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5):511–5. [PubMed: 20436464]

123. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18(5):821–9. [PubMed: 18349386]

124. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010; 7(11):909–12. [PubMed: 20935650]

125. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5(7):621–8. [PubMed: 18516045]

126. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA. Alternative expression analysis by RNA sequencing. Nat Methods. 2010; 7(10):843–7. [PubMed: 20835245]

127. Wang X, Wu Z, Zhang X. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. J Bioinform Comput Biol. 2010; 8(Suppl 1):177–92. [PubMed: 21155027]

128. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010; 7(12):1009–15. [PubMed: 21057496]

129. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011; 12(3):R22. [PubMed: 21410973]

130. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10):R106. [PubMed: 20979621]

131. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40. [PubMed: 19910308]

132. Ambros V. The functions of animal microRNAs. Nature. 2004; 431:350–5. [PubMed: 15372042]

133. Bartel D. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116:281–97. [PubMed: 14744438]

134. Schnall-Levin M, Zhao Y, Perrimon N, Berger B. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′UTRs. Proc Natl Acad Sci U S A. 2010; 107:15751–6. [PubMed: 20729470]

135. Schnall-Levin M, Rissland OS, Johnston WK, Perrimon N, Bartel DP, Berger B. Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. Genome Res. 2011; 21(9):1395–403. [PubMed: 21685129]

136. Grishok A, Pasquinelli A, Conte D, Li N, Parrish S, Ha I, Baillie D, Fire A, Ruvkun G, Mello C. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell. 2001; 106:23–34. [PubMed: 11461699]

137. Berezikov E, Cuppen E, Plasterk R. Approaches to microRNA discovery. Nat Genet. 2006; 38(Suppl):S2–7. [PubMed: 16736019]

138. Lim L, Lau N, Weinstein E, Abdelhakim A, Yekta S, Rhoades M, Burge C, Bartel D. The microRNAs of Caenorhabditis elegans. Genes Dev. 2003; 17:991–1008. [PubMed: 12672692]

139. Grad Y, Aach J, Hayes G, Reinhart B, Church G, Ruvkun G, Kim J. Computational and experimental identification of C. elegans microRNAs. Mol Cell. 2003; 11:1253–63. [PubMed: 12769849]

140. Lai E, Tomancak P, Williams R, Rubin G. Computational identification of Drosophila microRNA genes. Genome Biol. 2003; 4:R42. [PubMed: 12844358]

141. Xie X, Lu J, Kulbokas E, Golub T, Mootha V, Lindblad-Toh K, Lander E, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 2005; 434:338–45. [PubMed: 15735639]

142. Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon G, Kellis M. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. Genome Res. 2007; 17:1865–79. [PubMed: 17989255]

143. Nam J, Shin K, Han J, Lee Y, Kim V, Zhang B. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res. 2005; 33:3570–81. [PubMed: 15987789]

144. Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y. MicroRNA identification based on sequence and structure alignment. Bioinformatics. 2005; 21:3610–4. [PubMed: 15994192]

145. 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; 20:2911–7. [PubMed: 15217813]

146. Washietl S, Hofacker I, Lukasser M, Hüttenhofer A, Stadler P. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol. 2005; 23:1383–90. [PubMed: 16273071]

147. Hertel J, Stadler P. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics. 2006; 22:e197–202. [PubMed: 16873472]

148. Bentwich I. Prediction and validation of microRNAs and their targets. FEBS Lett. 2005; 579:5904–10. [PubMed: 16214134]

149. Xue C, Li F, He T, Liu G, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005; 6:310. [PubMed: 16381612]

150. Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grässer F, van Dyk L, Ho C, Shuman S, Chien M, Russo J, Ju J, Randall G, Lindenbach B, Rice C, Simon V, Ho D, Zavolan M, Tuschl T. Identification of microRNAs of the herpesvirus family. Nat Methods. 2005; 2:269–76. [PubMed: 15782219]

151. Ohler U, Yekta S, Lim L, Bartel D, Burge C. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA. 2004; 10:1309–22. [PubMed: 15317971]

152. Seitz H, Royo H, Bortolin M, Lin S, Ferguson-Smith A, Cavaillé J. A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res. 2004; 14:1741–8. [PubMed: 15310658]

153. Hendrix D, Levine M, Shi W. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol. 2010; 11:R39. [PubMed: 20370911]

154. Lu C, Tej S, Luo S, Haudenschild C, Meyers B, Green P. Elucidation of the small RNA component of the transcriptome. Science. 2005; 309:1567–9. [PubMed: 16141074]

155. Ruby J, Jan C, Player C, Axtell M, Lee W, Nusbaum C, Ge H, Bartel D. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell. 2006; 127:1193–207. [PubMed: 17174894]

156. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein M, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo J, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006; 442:203–7. [PubMed: 16751777]

157. Shi W, Hendrix D, Levine M, Haley B. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. Nat Struct Mol Biol. 2009; 16:183–9. [PubMed: 19151725]

158. Friedländer M, Mackowiak S, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2011

159. Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics. 2010; 26:2226–34. [PubMed: 20591903]

160. Stark A, Brennecke J, Bushati N, Russell R, Cohen S. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell. 2005; 123:1133–46. [PubMed: 16337999]

161. Bartel D. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136:215–33. [PubMed: 19167326]

162. Washietl S, Hofacker IL. Nucleic acid sequence and structure databases. Methods Mol Biol. 2010; 609:3–15. [PubMed: 20221910]

163. Galperin MY, Cochrane GR. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res. 2011; 39(Database issue):D1–6. [PubMed: 21177655]

164. He S, Liu C, Skogerb G, Zhao H, Wang J, Liu T, Bai B, Zhao Y, Chen R. NONCODE v2.0: decoding the non-coding. Nucleic Acids Res. 2008; 36(Database issue):D170–2. [PubMed: 18000000]

165. Pang KC, Stephen S, Engstrm PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS. RNAdb–a comprehensive mammalian noncoding RNA database. Nucleic Acids Res. 2005; 33(Database issue):D125–30. [PubMed: 15608161]

166. Mituyama T, Yamada K, Hattori E, Okida H, Ono Y, Terai G, Yoshizawa A, Komori T, Asai K. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. Nucleic Acids Res. 2009; 37(Database issue):D89–92. [PubMed: 18948287]

167. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011; 39(Database issue):D146–51. [PubMed: 21112873]

168. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A. Rfam: Wikipedia, clans and the ”decimal” release. Nucleic Acids Res. 2011; 39(Database issue):D141–5. [PubMed: 21062808]

169. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011; 39:D152–7. [PubMed: 21037258]

170. Sethupathy P, Corda B, Hatzigeorgiou A. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA. 2006; 12:192–7. [PubMed: 16373484]

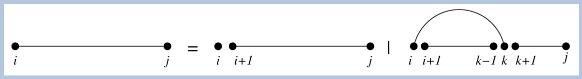
171. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009; 37:D98–104. [PubMed: 18927107]

172. Bateman A, Agrawal S, Birney E, Bruford EA, Bujnicki JM, Cochrane G, Cole JR, Dinger ME, Enright AJ, Gardner PP, Gautheret D, Griffiths-Jones S, Harrow J, Herrero J, Holmes IH, Huang HD, Kelly KA, Kersey P, Kozomara A, Lowe TM, Marz M, Moxon S, Pruitt KD, Samuelsson T, Stadler PF, Vilella AJ, Vogel JH, Williams KP, Wright MW, Zwieb C. RNAcentral: A vision for an international database of RNA sequences. RNA. 2011; 17(11):1941–6. [PubMed: 21940779]

Box 1

Dynamic programming

The Dynamic Programming (DP) paradigm is used for many algorithms related to RNA folding. DP breaks down a problem in smaller sub-problems to find the overall solution efficiently. Nussinov's algorithm is a classical example of a DP algorithm. Let's assume we want to find the minimum free energy $E_{i,j}$ between the positions i and j of a sequence and already know the solution for a sequence from $i + 1$ to j, i.e. a sequence that is one base shorter. The new base i can either be unpaired or forms a base-pair with some position k:



The base-pair k divides the problem into two smaller sub-problems, namely finding the solution for $E_{i+1,k-1}$ and $E_{k+1,j}$. We thus can find the solution using a recursive algorithm:

$$E_{ij}=\min\left\{E_{i+1,j},\ \min_{i+1\le k\le j}\left\{E_{i+1,k-1}+E_{k+1,j}+\beta_{i,k}\right\}\right\}$$

$\beta_{i,k}$ is the energy contribution for the base-pair i, k in this simplified energy model.

Box 2

Stochastic context free grammars

Context free grammars (CFG) are concepts from formal language theory. In most simple terms, a CFG is a set production rules $V \to w$ where V represent a so-called nonterminal symbol that produces a string of terminal or non-terminal symbols w. An example of a simple RNA grammar would be $S \to aS\hat{a}|aS|Sa|SS|\epsilon$. The grammar has one type of non-terminal symbol S and one type of terminal symbols $a \in A, C, G, T$ representing the bases. The grammar consists of production rules for unpaired and paired bases ($a\hat{a}$ represent two complementary bases). This simple rules allow to produce all possible RNA secondary structures. A stochastic context free grammar (SCFG) extends CFGs by assigning probabilities to all production rules. In the case of our RNA grammar, a full parametrized SCFG would thus describe the probability distribution over all structures and sequences.

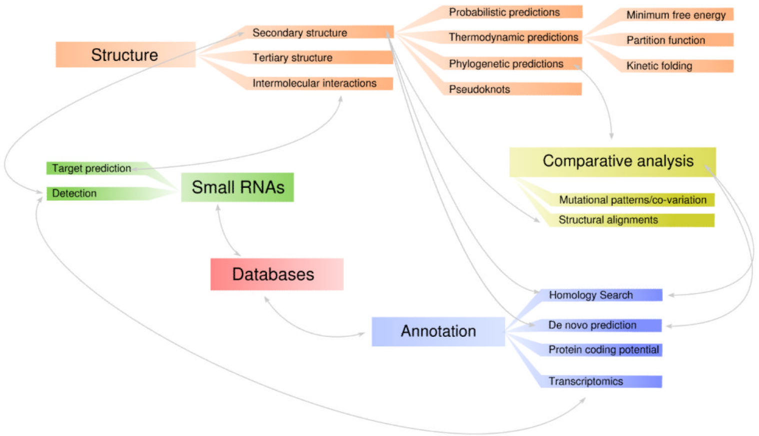


Figure 1.

Outline of the main topics covered in this review. Many topics overlap, depend on each other or share similar concepts. The most important of these interconnections are shown by arrows.

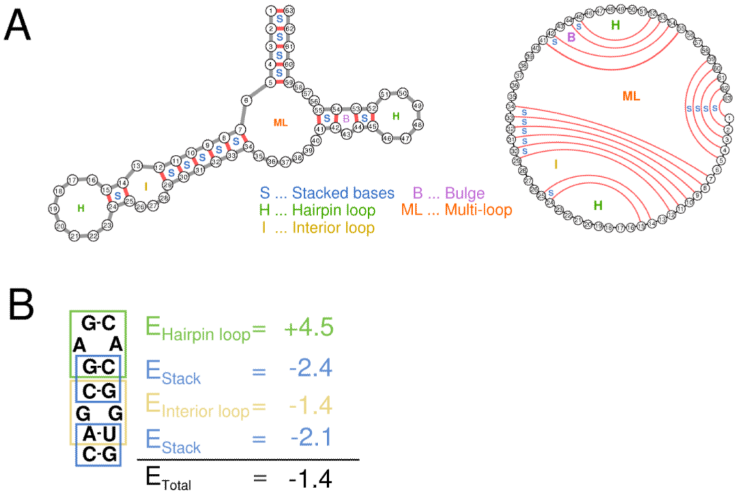


Figure 2.

Principles of RNA structure prediction (**A**) RNA secondary structure can be represented as an *outerplanar* graph (right). The backbone is arranged as a circle and base pairs are represented as arcs. The faces of this graph correspond to different structural elements. This formalization is the basis for most structure prediction algorithms. Any structure can be uniquely decomposed into these basic elements which are independent from each other. This allows for efficient folding algorithms based on the “dynamic programming” principle that breaks down the problem into smaller sub-problems (see also Box 1). (**B**) Example of energy evaluation of a small RNA structure. Thermodynamic folding algorithms assign free energies to the structural elements. In the example shown, two stacks and a symmetric interior loop stabilize the structure (negative free energy) while the hairpin loop destabilizes the structure (positive free energy). The total free energy of the structure is the sum of the energy of all structural elements.

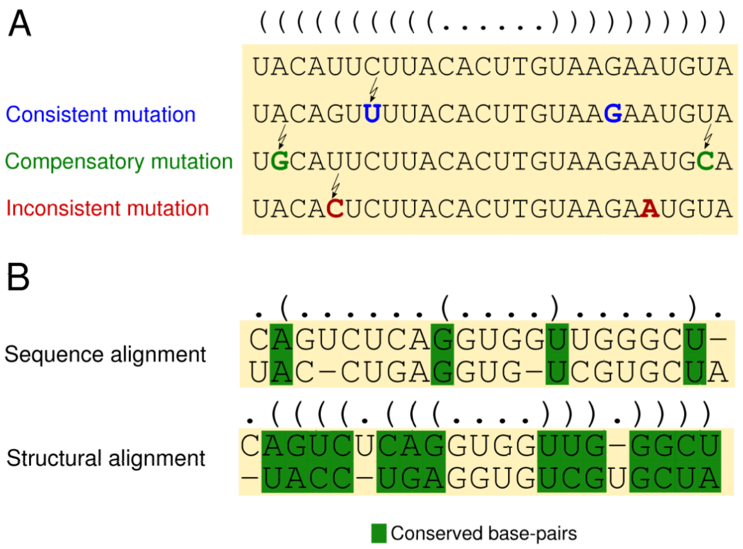


Figure 3.

Principles of comparative analysis for RNA structure prediction. (**A**) A short sequence that can fold into a hairpin is aligned to three other sequences with different mutation patterns. Mutated bases are indicated by a lightning symbol. The affected base pair is shown in blue, green and red for the case of a consistent mutation, a compensatory double mutation, or a inconsistent mutation that destroys the base pair. (**B**) Sequence based alignment vs. structural alignment. Consensus structure predicted for two aligned sequences. First, the alignment is optimized to match the sequences resulting in a poor consensus structure with few conserved base-pairs (green). Second, the two sequences are aligned to optimize a common structure, resulting in a much better consensus structure with more conserved pairs. (All structures are shown in “dot/bracket” notation, in which base-pairs are indicated by brackets and unpaired positions are shown as dots.)

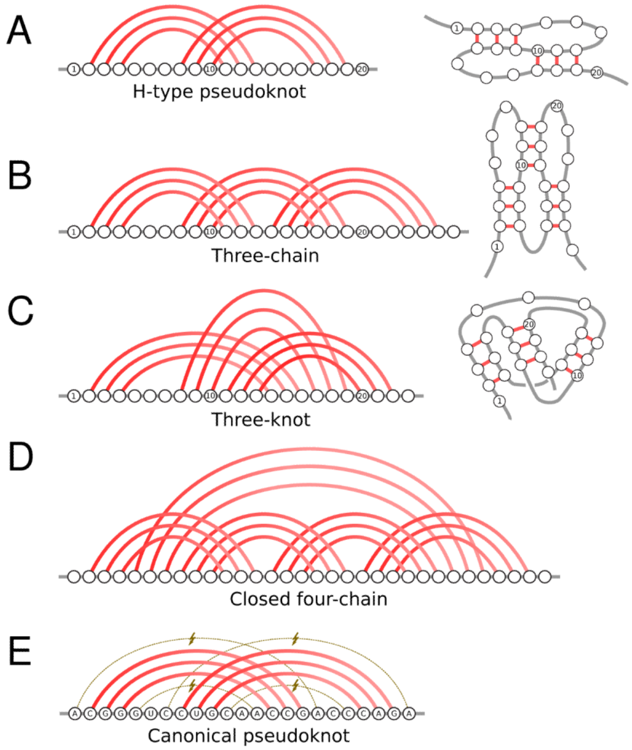


Figure 4.

Pseudoknot types. (**A**) The simplest type of pseudoknot (H-type) formed by two crossing stems. The most efficient algorithms predict only this most common form of pseudoknot. (**B**) Three-chain or kissing hairpin. Two hairpin loops are connected by one or more base pairs. (**C**) Three-knot. Three stems cross each other. This configuration is predicted only by the expensive algorithm by Rivas and Eddy[51] (**D**) Four-chain, closed by a fifth stem. This complex motif cannot be predicted by the algorithm of Rivas and Eddy, but would require an even more costly algorithm. (**E**) Canonical pseudoknot. A pseudoknot formed by two perfect stems of canonical base pairs that are *maximally extended*, i.e. they cannot be extended further by canonical base pairs; the figure indicates the latter by the dashed "conflict"-arcs between non-canonical base pairs AA, GA, CA, and CC (from left to right). The most space-efficient pseudoknot prediction algorithm[52] predicts only canonical pseudoknots.

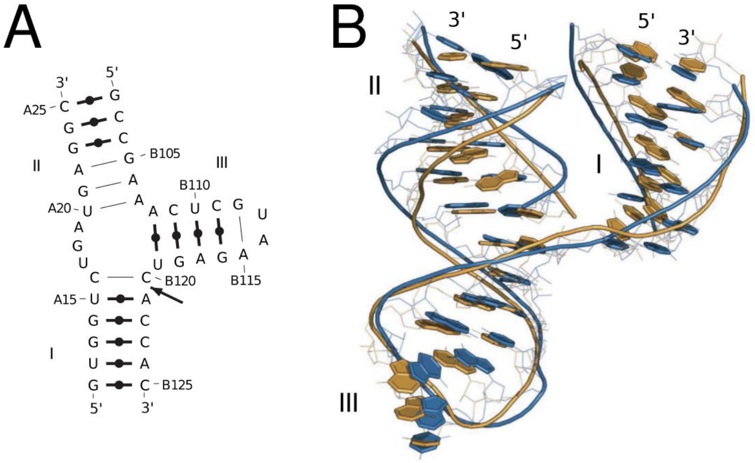


Figure 5.

Tertiary structure prediction. Example prediction from the MC-Fold and MC-Sym pipeline[68]. (**A**) Secondary structure including canonical (bold lines) and non-canonical base-pairs (non-bold lines) as predicted by MC-Fold. (**B**) Tertiary structure predicted from secondary structure (A) by MC-Sym. The prediction (blue) is compared to the experimental structure (gold).

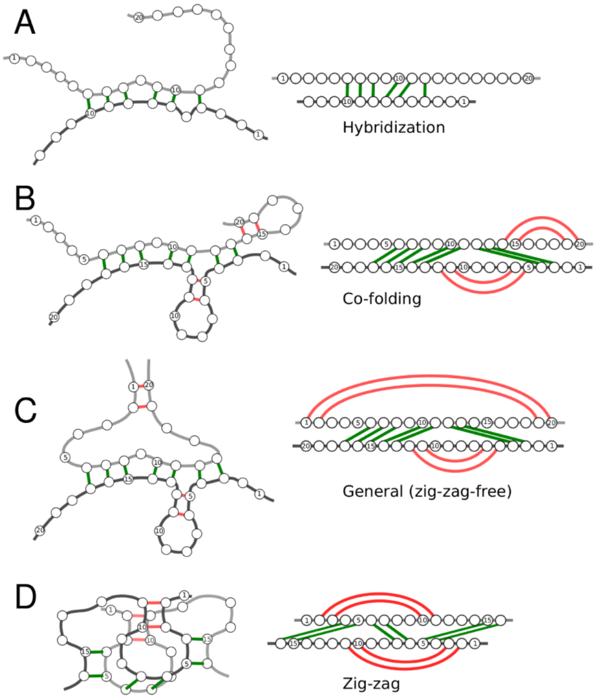


Figure 6.

RNA/RNA interactions. (A) Simple hybridization, no internal structure of RNAs. The simplest interaction prediction approaches predict only the hybridization at a single site without considering internal structure. (B) Hybridization and restricted internal structure as in the co-folding model. Interactions can occur at several sites, however only between external bases. When concatenating the two interacting RNAs, the structure of all inter- and intramolecular base-pairs is pseudoknot-free, such that it can be predicted from the concatenation by (a variant of) Zuker's algorithm. (C) Interaction structure as predictable by the most complex dynamic programming algorithms. Such structures are free of pseudoknots, crossing interactions, and zig-zags (see D). (D) Zig-zag. Intramolecular stems in each RNA cover a common interaction as well as interactions to the outside of the stems.

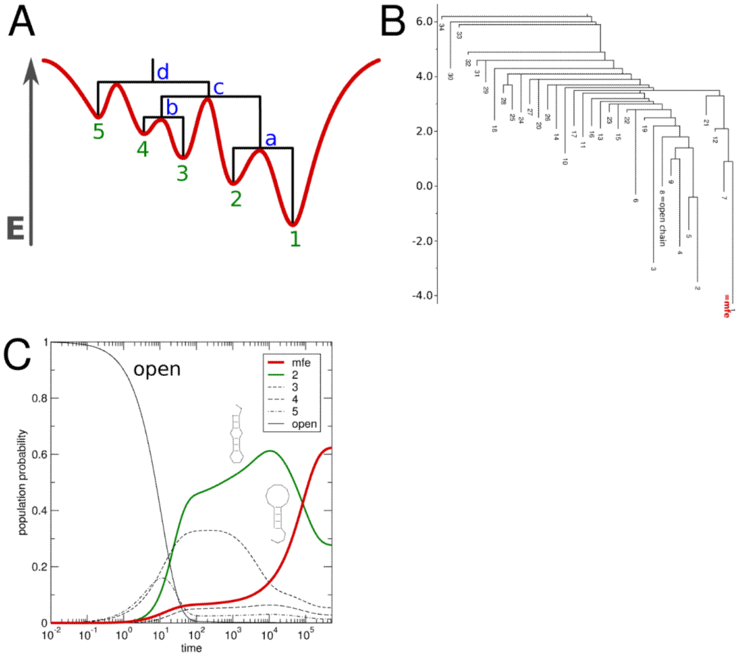


Figure 7.

Kinetic folding pathways. (**A**) Schematic energy landscape and associated barrier tree. A barrier tree shows the local minima and the minimum energy barriers between them (adapted from[81]). (**B**) Barrier tree of the small RNA xbix. (**C**) Exact folding kinetics of xbix starting from the open chain. Probability of local minima over time. While the minimum free energy (MFE) structure is finally most prominent, other "intermediary" structures (2, 3, and 4) are temporarily more probable (adapted from[82]).

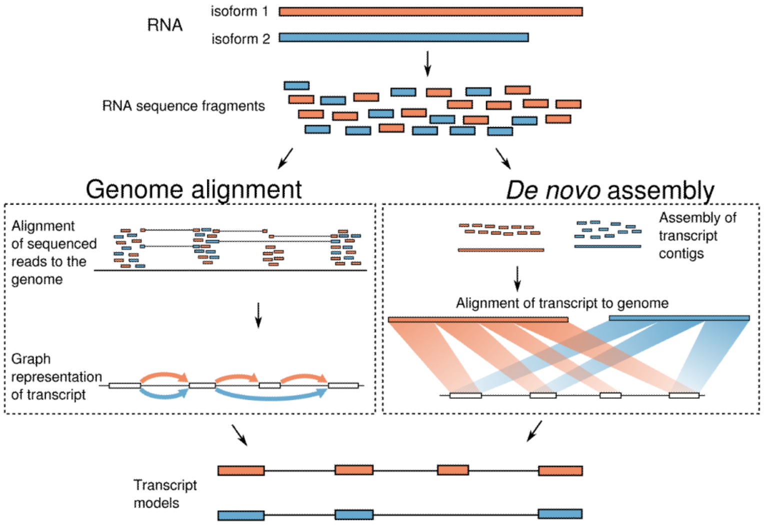

RNA
isoform 1
isoform 2
RNA sequence fragments
Genome alignment
De novo assembly
Alignment of sequenced reads to the genome
Graph representation of transcript
Assembly of transcript contigs
Alignment of transcript to genome
Transcript models

Figure 8.

Reconstructing transcript models from RNA-seq data. Two splice isoforms of RNAs are shown for which the RNA-seq experiment generated short sequence fragments. One approach (left) to reconstruct the transcript is mapping the fragments to a reference genome. Spliced reads that span exons boundaries can be used to infer the connectivity graph. The paths through this graph correspond to the different isoforms. Alternatively, the transcripts can be re-constructed by *de novo* assembly of the reads into transcripts (right). If available, the assembled RNA transcripts can be mapped to a reference genome afterwards to obtain the intron-exon structure of the isoforms.

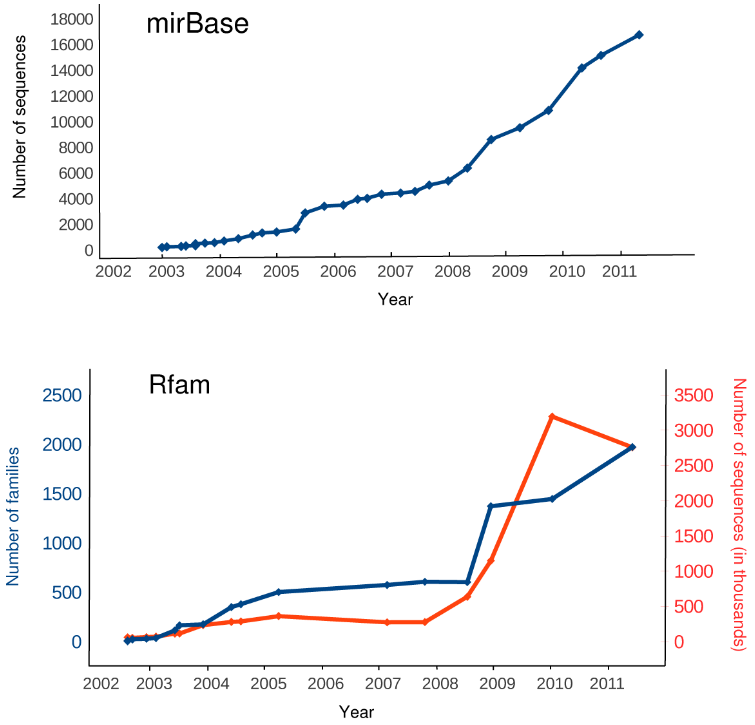


Figure 9.

Growth of miRBase and Rfam over the past 8 years. For miRBase the number of microRNAs are shown, for Rfam the number of structure families and the number of sequences found to be member of an Rfam family. (The drop of the number of Rfam sequences in 2011 is the result of the re-organization of some large families and the elimination of pseudogenes.)