J. Math. Biol. (2008) 56:15–49
DOI 10.1007/s00285-007-0122-6

Mathematical Biology

Computational methods in noncoding RNA research

Ariane Machado-Lima · Hernando A. del Portillo · Alan Mitchell Durham

Received: 1 January 2007 / Published online: 4 September 2007

Abstract Non protein-coding RNAs (ncRNAs) are a research hotspot in bioinformatics. Recent discoveries have revealed new ncRNA families performing a variety of roles, from gene expression regulation to catalytic activities. It is also believed that other families are still to be unveiled. Computational methods developed for protein coding genes often fail when searching for ncRNAs. Noncoding RNAs functionality is often heavily dependent on their secondary structure, which makes gene discovery very different from protein coding RNA genes. This motivated the development of specific methods for ncRNA research. This article reviews the main approaches used to identify ncRNAs and predict secondary structure.

Keywords Review · Noncoding RNAs · Secondary structure prediction · Structure comparison · Gene finding

Mathematics Subject Classification (2000) 92C40

During the execution of this work, AML was supported by CAPES fellowship.

A. Machado-Lima (✉) · A. M. Durham
Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, SP, Brazil
e-mail: ariane@ime.usp.br

A. M. Durham
e-mail: durham@ime.usp.br

H. A. del Portillo
Institute of Biomedical Sciences, University of Sao Paulo, Sao Paulo, SP, Brazil
e-mail: hernando@icb.usp.br

Present Address:
H. A. del Portillo
Barcelona Centre for International Health Research CRESIB, Barcelona, Spain
e-mail: hernandoa.delportillo@cresib.cat

1 Introduction

In a very recent past, RNAs were considered mere intermediates between the genome and the proteins. Recent discoveries involving a variety of new ncRNA genes [85], biological roles and action mechanisms have shown that the diversity and importance of ncRNAs were underestimated [65]. Nowadays, it is known that functional RNAs that do not code to proteins perform important roles.

They are involved in several cellular activities such as gene silencing [53], replication [48], gene expression regulation [79], transcription [155], chromosome stability [9], protein stability [103], translocation [72] and localization [120] and RNA modification [88], processing [13] and stability [135]. New long antisense ncRNAs have been found in many genomes, that seem to be involved in gene expression regulation [65,98,113,114]. In addition, many more ncRNAs are expected to be unveiled [37,65]. It is even speculated that they can better explain the differences in complexity of the several organisms than protein genes can do [86].

These new discoveries have motivated the ncRNA research in many aspects. For instance, once the RNA structure and function are closely related, it is desirable to know the common structure of homologous RNAs in order to discover functional signatures. It is also desirable to scan a genome looking for ncRNAs. However, due to the exponential number of possible solutions, RNA structure prediction is a complex problem. In addition, strategies used in protein coding gene identification often fail when searching for ncRNAs. As a result, in silico identification of ncRNAs is still an open problem in bioinformatics [38,89,99].

Different approaches have their own strengths and weaknesses, each one being more suitable for specific problem domains. In this article, we present a review of computational methods for ncRNA-related problems. Due to the wide literature on the area, our first goal is not to do an exhaustive survey. Instead, we intend to review the main approaches applied so far by commenting some methods, and point the main strengths and weaknesses of each one depending on the problem domain. In Sect. 2 we briefly describe the main components of an RNA secondary structure. In Sect. 3 we present the main problems involving ncRNAs and the main approaches used to address them. In Sects. 4–6 we discuss each problem separately. Finally, in Sect. 7 we discuss some perspectives and make concluding remarks. In Appendix we present a list of all available programs or web servers of the methods cited here.

2 RNA secondary structure

Most of the RNAs are single-stranded molecules and can fold forming base pairings. These pairings often occur between the bases G and C, A and U, and, occasionally between G and U.

Formally, let $x = x_1 x_2 \ldots x_n$ be an RNA sequence, where $x_i \in \{\text{A, C, G, U}\}$ for $i = 1, \ldots, n$.

Definition 1 A *secondary structure* of x is a set of base pairs $P = \{(i, j) | i < j\}$ with the following constraints [21]:

1. if $(i, j) \in P$ then $(x_i, x_j) \in \{\text{(G, C), (C, G), (A, U), (U, A), (G, U), (U, G)}\}$

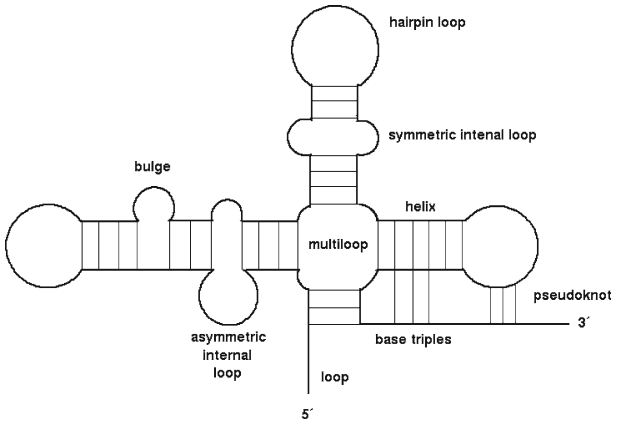

Fig. 1 Structural components of an RNA secondary structure

2. if $(i, j) \in P$ and $(i, l) \in P$ then $j = l$
3. if $(i, j) \in P$ and $(k, j) \in P$ then $i = k$
4. if $(i, j) \in P$ then $j - i < \theta$
5. if $(i, j) \in P$ and $(k, l) \in P$ and $i < k < j$ then $i < k < l < j$.

These pairings form structural components (Fig. 1) known as:

– *helix* or *stem*: a contiguous stacking of base pairs (two stacking base pairs are (i, j) and $(i + 1, j - 1)$);
– *loop*: a region of unpaired bases;
– *hairpin loop*: a loop enclosed by a helix;
– *multi-loop*: a loop region from which three or more helices arise;
– *internal loop*: a loop inside a helix; an *internal loop* is *asymmetric* if the number of nucleotides in each side of the helix is different, and *symmetric* otherwise;
– *bulge*: a loop inside a helix but occurring at only one side of it.

In addition to the secondary structure defined above, base pairs can be part of two other structures: *pseudoknots* (that violate constraints 2 and 3) and *base triples* (that violate constraint 5). These other base interactions are considered part of the tertiary structure.

3 Problems and general approaches in ncRNA research

3.1 Three main problems

An ncRNA often requires a specific three-dimensional structure to perform its function. Since the three-dimensional structure is determined by the secondary structure, the last is used as an approximation in the study structure–function. The secondary structure in turn is defined by the primary sequence. Therefore, tools to predict the secondary structure from an RNA sequence are useful to study its function. When, instead of one sequence, a set of homologous RNAs is known, its consensus secondary structure can

be more reliably predicted. Moreover, the conservation of structural domains across different species are additional evidence that they are related with the specific function of these sequences. Therefore, prediction of conserved structures are useful to discover and characterize signatures for a specific RNA family.

Structure comparison can serve many purposes. For instance, it can be used to classify an unknown RNA as member of a known family by comparing its structure with the consensus structure of the several known families. In addition, if the function of a single RNA or of a family is not known, it can be inferred by comparing the RNA structure (or consensus in the case of a family) with a database of functionally annotated structural signatures. In addition, structural comparison can be used to detect the occurrence of different stable structures for the same molecule (which may indicate possible structural switching related to its role), to predict mutations in an RNA sequence that causes rearrangements in the secondary structure and to compare a set of structures to choose a representative.

Finally, we can use structure prediction and comparison to search whole genomes for ncRNA sequences, either searching RNAs homologous to a specific candidate or family of candidates, or looking for all ncRNAs, including new families still unknown.

In summary the ncRNA research involves three main types of problems:

– secondary structure prediction;
– secondary structure comparison and
– noncoding RNA identification.

The methods described in this paper are classified according to these three problems. For each one of them, most of the computational methods do not consider pseudoknots, since they are considered a tertiary interaction. However, the prediction of these structures may be desirable. The general problem of pseudoknot prediction is computationally unfeasible yet, due to the NP-completeness of RNA secondary structure prediction with general pseudoknots [91]. Methods that deal with them often pose constraints in the pseudoknot structure in order to make the problem tractable, but they may be still impractical for long sequences [110,117]. Throughout the paper, we point the methods that consider pseudoknots and their time and memory complexities.

Time and memory complexities are also mentioned when this information is an important issue that differentiates alternative methods for a specific problem (Sect. 4.2.3).

3.2 Three main approaches

When dealing with a specific RNA problem, ab initio and comparative strategies can be applied. Ab initio methods, in the context of this review, deal with a single sequence, whereas comparative methods analyze a set of sequences. Three main approaches can be applied to the problem, thermodynamic, probabilistic and covariation-based. Independently of the strategy, a method may use one or more approaches to model the RNA structure problem. Thermodynamic and probabilistic approaches can be explored by both ab initio and *comparative* strategies. Covariation analysis, however, can be applied only in comparative methods. Each one has different assumptions about ncRNA problems. Therefore, methods based on the same approach share strengths and

weaknesses. Since these approaches are not mutually exclusive, a combination of them can decrease their individual limitations.

Thermodynamic approach This approach is based on the Gibbs free energy value of RNA structures [141]. The free energy of an RNA structure is computed under the nearest-neighbor model. This model considers that the energy associated with a structural motif is only dependent on the nucleotides of this motif and on the adjacent nucleotide interactions [141]. The nearest-neighbor model is then composed of a set of parameters defining the energy associated to a variety of neighbor interactions. These thermodynamic parameters have been experimentally estimated since 1971 [138] and remain being improved [93]. However, these measures still carry experimental and precision errors that limit the structure prediction accuracy to approximately 50–75% [40]. Efforts to estimate these thermodynamic parameters statistically using RNA structure databases have shown promising results [29].

Probabilistic approach This approach assumes that there is a probability distribution over the set being characterized, such as sequences, structures or alignments. Probabilistic models are built estimating parameters from a set of known examples called *training sample*. The advantage of this approach is that it uses a well-defined theoretical framework for doing statistics over the solution space. Furthermore, different probabilistic models can be designed to characterize different features. This flexibility allows the building of different models for specific RNA families [35], which is particularly important for sequence classification. Sophisticated models can be designed, but wrong assumptions may lead to a model that supplies wrong results. In addition, the increase in the sophistication level can be accompanied of an increase in the number of parameters to be estimated. This may demand larger training samples, sometimes not available. Finally, a biased training sample may lead to a biased model. Therefore, the quality of the estimated model depends not only on the model design but also on the size and quality of the available training sample.

Covariation analysis Homologous RNAs have common ancestry and function and are expected to have similar structures and some similarity in sequence. When comparing two or more homologous RNA sequences, if a hypothetical base pair in one sequence (G–C, for instance) is different in the other sequences but is still a valid base pair (such as A–U), it can be considered evidence of selective pressure to maintain the base pair. Therefore, it can be used as support for the existence of that pairing [66]. Double mutations preserving a base pair are known as *compensatory mutations* and the process of detecting them is called *covariation analysis*. To detect these compensatory mutations, some methods model each sequence position as a random variable and calculate the mutual information of each pair of random variables. Note, however, that it is not necessary the two bases covary, since some point mutations, such as G–C to G–U, are still evidence for base pairing. Therefore, methods that just search for covariation miss valuable information. The main advantage of covariation analysis is the use of information of a specific gene family to gain in accuracy. However, compensatory mutations are often searched by scrutinizing columns of a multiple alignment. To achieve success, the sequences being analysed must be conserved enough to

allow an accurate multiple alignment and distinct enough to show covariations [66]. Errors in the multiple alignment might affect the accuracy of the methods. Another obvious disadvantage is that covariation analysis can not deal with isolated sequences. Other drawbacks are specific to each specific RNA problem, being described in Sects. 4–6.

4 Secondary structure prediction

Given an RNA sequence, the number of possible secondary structures grows exponentially with the sequence length [149]. The issue is how to search a structure in this exponential solution space in order to choose the best structure. When the secondary structure of just one RNA sequence needs to be predicted, only ab initio methods can be used. If a set of homologous RNAs is available, comparative methods can predict the consensus structure more accurately [66].

4.1 Ab initio methods: prediction from a single sequence

4.1.1 Thermodynamic predictors

Thermodynamic predictors explore the hypothesis that an RNA molecule is folded in the most thermodynamically stable structure, that is, the one having the minimum free energy (MFE). A straightforward approach is to enumerate all possible structures and then select the one with the minimum value for the free energy [109], but the exponential time complexity spent in the enumeration step is unfeasible but for the smallest sequences. To deal with this complexity issue, all current methods use a dynamic programming method first proposed by Nussinov et al. [104] that reduces the time complexity to $O(n^3)$. However, the straightforward MFE structure may not be the correct one. Not only there are errors in the thermodynamic parameters and unknown thermodynamic rules but, most important, the MFE structure may not be the one adopted by the RNA. The correct structure may be among the sub-optimal free energy ones [56]. In this case, the analysis of the space of possible structures (folding space) can supply clues about the most probable structures and motifs. Finally, algorithms can be based on the hypothesis that an RNA molecule may assume the easiest structure to be formed due to kinetic traps in the folding process, that is, due to the foldings that occur incrementally when the RNA molecule is synthesized.

Minimum free energy

The current methods that compute secondary structure are non-trivial extensions of the Nussinov's algorithm, but keeping the time complexity of $O(n^3)$ when pseudoknots are not allowed. They also consider several possible features of the structural components such as interior loop symmetry/asymmetry and coaxial stacking of helices.

Some of the programs that compute the secondary structure with minimum free energy are MFOLD [158,160], RNAfold [59], RNASTRUCTURE [93], PKNOTS [117] and pknotsRG [110,130].

In particular, the last two methods include the prediction of pseudoknots. Both impose restrictions on the kind of pseudoknots found in order to reduce the theoretical complexity of the problem, PKNOTS being more general (time complexity of $O(n^6)$ and memory complexity of $O(n^4)$) and pknotsRG more restrictive ($O(n^4)$ for time and $O(n^2)$ for memory).[1]

Folding space analysis

Methods that perform folding space analysis explore sub-optimal structures in order to create a more accurate profile of what type of structures (or sub-structures) are more likely to occur in real situations.

In order to efficiently analyze the structure space, most methods use the Boltzmann distribution to model the probability of a structure. This probability is given by the equation:

$$P_S = \frac{\exp(-E_S/RT)}{Z(T)} \tag{1}$$

where E_S is the free energy of the structure S, R is the gas constant, T is the temperature in Kelvin and $Z(T)$, known as *partition function*, is:

$$Z = \sum_{S'} \exp(-E_{S'}/RT) \tag{2}$$

that is calculated using dynamic programming, i.e., without enumerating all possible structures.

This structure probability P_S can be used to extract other information. McCaskill [96] proposed a formula to compute the probability of any two nucleotides in different positions to be paired in a secondary structure. The probability of each base pair (i, j) in a molecule is given by:

$$P_{ij} = \sum_{(i,j)\in S} P_S \tag{3}$$

These values compose a thermodynamic base pair probability matrix, that can be analysed in order to detect well-defined helices. This matrix is calculated by RNAfold [59] and RNASTRUCTURE [93].

The NUPACK program [32,33] uses a more general partition function that is able to include a class of physically relevant pseudoknots, but demanding O(n^5) time.[2]

The Boltzmann distribution can also be used to partition the folding space by ranges of energy values in an approach known as *density of states* [19,21,27]. States[3] that

[1] It is noteworthy that the success of thermodynamic methods that attempt to deal with pseudoknots is limited, specially because experimental data is still scarce for a good estimation of pseudoknots energy parameters.

[2] The McCaskill's algorithm demands $O(n^3)$ time without considering pseudoknots [96].

[3] A state is a part corresponding to a single energy range.

are dense in structures may indicate possible stable intermediate structures. In particular some of these structures can be "kinetic traps" formed during folding/unfolding process [18].

If all structures in a specific energetic range are enumerated, different analysis can be performed, such as calculation of the most probable motifs or structure clusterings. `RNAsubopt` [153] pioneered such an exhaustive enumeration. It generates all suboptimal structures having energy in an user-defined range from the MFE. To compensate for the excessively large number of such structures for longer sequences, `Sfold` [16,30,31] produces a sampling of the complete structure space using the Boltzmann distribution. Analyses can now be performed over this sample, once it has, theoretically, the same distribution of the whole folding space. This program also identifies clusters of the sampled structures, based on structural similarity, and selects a representative of each cluster, called *centroid*. The centroid is the structure having the minimum base pair distance to all other structures in the cluster. The program returns the MFE structure, the cluster centroids and the *ensemble centroid* (that is, the centroid for all sampled structures).

Many alternative structures share the same structural pattern, or shape.[4] Therefore, the shape can also be used to produce a partition of the folding space of a particular RNA molecule. `RNAshapes` [134,144] performs shape partition and ascribes to each shape a probability value consisting of the sum of the Boltzmann probabilities of all structures in the corresponding part. `RNAshapes` also reports the MFE structure for each part.[5]

A last approach is used by `MFOLD` [157] which, instead of using the Boltzmann distribution, computes the *h-num*(i,j) quantity, a measure applied to pairs of positions in the RNA molecule that indicates the level of pairing promiscuity between the bases at positions i and j. Well-determined pairings are expected to have low *h-num* values [159].

Kinetic folding

Kinetic folding algorithms are based on the hypothesis that the final structure of an RNA molecule depends on local rather than on global dependencies. In other words, during the folding process, optimal sub-structures may be formed, acting as kinetic traps and precluding the globally optimal structure to be achieved.

A common approach is to build the composite structure starting from individual helices. Abrahams et al. [1] developed an algorithm that first finds all possible helices, then incrementally builds the final structure by selecting, at each step, the helix that minimizes the free energy of the current structure, allowing the introduction of pseudoknots. Schmitz and Steger developed a similar algorithm [129], but allowing the removal of current helices of the structure when the newly formed structure leads to pseudoknots or overlaps (bases participating in different helices). `RDfolder` [156]

[4] We can informally define the shape as the visual appearance of the structure, such as a clover-leaf format, a particular succession of helices, etc.

[5] These two components, representative energy and shape probability, are not redundant, since the shape having the optimal representative is not necessarily the most probable one.

uses a Monte Carlo approach to implement a variation of the second algorithm: the basic structure building cycle is performed many times, randomly selecting the helices to be added, but not allowing helix removal. If the sequence is small (up to 150 bases), the predicted structure is the most frequent one among all runs. For larger sequences, to avoid an exponential increase in the number of simulations, the program counts instead the frequency of each helix and then builds the final structure incrementally selecting the most frequent helices that are compatible with the partial structure.

`RNAKinetics` [28] takes this process further by taking into account the fact that RNA molecules may start the folding process during transcription [97]. So, the rate and direction of transcription are taken into account when calculating the helices to be added or removed from the structure. The folding simulation is parameterized by assigning random variables to compute the number of simulation steps and the time increment used in the folding simulation. The program presents the list of most likely structures sorted by their lifetimes.

A different Monte Carlo approach is to work at a smaller granularity level, simulating formation and disruption of single base pairs instead of complete helices. This strategy is adopted in `Kinfold` [43], which models RNA folding as a Markov process in the folding space.

`HotKnots` [115] adopts the strategy of pruning the initial helix space used to incrementally build the RNA structures. In this approach only sets of *promising hotspots* are used. A hotspot is a helix-like substructure comprised of stacked pairs, 1-base bulges and 2-bases symmetric interior loops. A set of hotspots is promising if the difference of the MFE of the sequence constrained by the hotspots is at most 80% higher than the MFE of the unconstrained sequence. This program also predicts pseudoknots.

4.1.2 A probabilistic model

Different from the thermodynamic approach, where parameters are experimentally estimated, with probabilistic models we can estimate the parameter values for different RNA families, using data from RNA structure databases. Since this estimation is automatic and fast, different models can be built and tested in order to explore alternative features in RNA structure modelling.

`CONTRAfold` [34] is a program that uses a Conditional Log Linear Model (CLLM), a generalization of Stochastic Context Free Grammars (SCFGs). This model is marked by three main innovations: discriminative training, flexible parameterization and an accuracy-adjustable optimization. CLLMs parameterize the conditional probability of a structure as a log linear function of the model parameters. This discriminative model is claimed to have a prediction power superior to generative models, which describes the joint probability of sequences and structures. In addition, the `CONTRAfold` parameters are not restricted to probabilities of traditional SCFG rules. Instead, they are scores for 13 structural features such as base pairs, lengths of hairpins, helices, bulge loops and internal loop and internal loop asymmetry. Finally, the predicted structure is not the one that maximizes the structure score, but the one that maximizes the expected accuracy. This accuracy is a user-defined trade-off between sensitivity and specificity of base pair predictions.

4.1.3 Other approaches

Different strategies can be combined to try to improve the quality of structural predictions.

Graph theory can be used to maximize numerical scores used to evaluate possible secondary structures for different scoring systems. `Maximum Weighted Matching (MWM)` [136] is a graph-based program that can be used to predict both single sequence structure and consensus structure of multiple sequences, including pseudoknots.[6] A graph is built with nodes representing bases and edges connecting possible pairings. The edges are weighted according to some score, such as thermodynamic or probabilistic values. Candidate secondary structures are *matching* subgraphs, i.e., subgraphs having nodes connected to at most one other node. The optimal structure can be found searching for the matching subgraph that has the highest total edge weight. `MWM` can also be used for detecting base triples.

Nussinov's dynamic programming algorithm was developed to maximize the number of base pairs in a structure prediction [104]. The same simple scoring scheme can be used with the ideas of Kinetic Analysis and Fold Space Analysis, out of the context of thermodynamics.[7] `Iterated Loop Matching (ILM)` [121,122] performs the original Nussinov's algorithm iteratively. In each iteration the best helix is selected for the final structure and cut out of the sequence for the next iteration. Since hairpin loops are never removed the algorithm can detect pseudoknots.[8] Although there is no guarantee of optimality, it has shown better results when compared to MWM. `RNALOSS` [22] also uses maximization of the number of base pairs. These parameters are used with the Boltzmann distribution to perform a density of states analysis.

4.2 Comparative methods

When trying to find the best consensus structure for a family of ncRNA molecules, the ideal situation would be to have the biologically validated secondary structures for each one of them, and then calculate the consensus. Unfortunately, the number of validated structures is very small. Most of the available structures are the result of predictions. In this context, the best situation is when the set of sequences is similar enough for a multiple alignment to be produced, and variant enough to show covariations. If such adequate alignment cannot be found but we have structures for these molecules, this information can be used in the search for a consensus. An alternative is just to consider the plain sequences and try to build the consensus structure from scratch. We will classify comparative structure predictors into three groups according to the input they receive:

1. aligned and unfolded sequences,

[6] Time and memory complexities of $O(n^3)$ and $O(n^2)$, respectively.

[7] In fact these are related measures, since the structure energy depends, among other characteristics, on the number of base pairs of a structure.

[8] The worst and average case time complexity are $O(n^4)$ and $O(n^3)$.

2. unaligned and individually folded sequences and
3. unaligned and unfolded sequences.

4.2.1 Prediction from aligned and unfolded sequences

Since a multiple alignment is available, the straightforward approach is searching for covariations. In addition, other information can be used to assign weight to the this covariation data, such as phylogenetic and thermodynamic data.

Covariation only methods

If a score system such as mutual information is used to describe covariations, methods such as Nussinov's algorithm [104], `Maximum Weighted Matching` [136] or `Iterated Loop Matching` [121,122]—all used to predict the secondary structure of a single molecule (Sect. 4.1)—can be applied to detect the optimal consensus secondary structure [41]. Then, the algorithms need to consider columns of the multiple alignment instead of individual nucleotide positions of a single sequence.

Including phylogenetic information

Score systems based only on covariations do not take into account evolutionary information involving these covariations. This information is used in the approach known as `Tree Model` [51]. A phylogenetic tree and mutation rate matrices (for base pairs and unpaired nucleotides) are used to compute the posterior probability of two columns being paired or unpaired. Phylogenetic trees can better discriminate between column pairs having strong or weak evidence of base pairing. This discrimination comes from the evolutionary ordering of the sequences imposed by the tree. This ordering shows the minimum number of events of compensatory mutations in each column pair. If, for instance, two column pairs have the same number of G–C and A–U pairs, the tree may indicate different numbers of paired mutation events. The larger the number of required mutation events, the stronger is the pairing evidence. Because of this extra sensitivity, the `Tree Model` performs better than other methods based exclusively on mutual information.

Akmaev et al. [2] also used phylogenetic models to propose a set of statistics to decide if two columns of a multiple alignment are paired or not. Akmaev's approach, however, is extended also to base-triples.

A third phylogenetic approach is adopted by `Pfold` [76,77], which combines an evolutionary model of RNA sequences with a probabilistic model of secondary structures. These are used to calculate a phylogenetic tree and a consensus structure from a multiple alignment. The evolutionary model consists of a set of nucleotide probabilities and mutation rates for paired and unpaired bases, all estimated from a large sample of RNAs. The probabilistic model is a fixed stochastic context free grammar estimated from the same (folded) RNA sample used to estimate the mutation rates. `Pfold` achieves reasonable results even using two sequences.

Based on the same ideas of `Pfold`, `RNA-Decoder` [106] goes further and models phylogenetic and probabilistic models for coding and noncoding sequences. The

phylogenetic model is built using different sub-models for noncoding/coding regions, in particular characterizing the different mutation rates due to selective pressure for aminoacid conservation in coding regions. In particular, `RNA-Decoder` can be used to scan alignments of whole genomes (viral, for instance) or even messenger RNAs (mRNAs) to identify putative conserved secondary structures with fewer false positives than when using other methods [106].

Including thermodynamic information

Some methods also explore thermodynamic information in their analyses. `X2` [70] uses a weighted linear combination of covariation, thermodynamic and heuristic values to sort column-pairs from the multiple alignment. The thermodynamic term is based on the energy value of the most stable base pair in a column-pair.[9] The heuristic term controls the distance between paired bases close to an optimal value, favoring local pairings. The high-scoring paired regions are included in the final structure progressively, according to the score sorting and avoiding overlapping. This assembly allows the formation of pseudoknotted structures.

`ConStruct` [90], combines all individual thermodynamic probability matrices (Sect. 4.1.1, Eq. 3) in a consensus matrix, which is used to supply the most thermodynamically probable consensus structure. To calculate this consensus matrix, gaps are inserted in individual matrices guided by the gaps in the multiple alignment. These matrices, now having the same dimension, are combined using an equation that takes into account sequence weights, in order to avoid over-representation of highly similar sequences. To alleviate the impact of alignment errors, the consensus structure and the alignment are graphically presented to the user, who can manually adjust the alignment. If the alignment is edited, the new alignment generates a new matrix alignment, restarting the process. Since covariations are not analysed, it can detect a consensus structure even in highly conserved regions.

`RNAalifold` [61] combines covariation and thermodynamic information in a score system to be used directly in the folding algorithm. That is, the dynamic programming matrix is calculated only once for the entire alignment. If the input sequences do not have a common folding, no structure is output.

Since both `RNAalifold` and `ConStruct` have a consensus matrix, they can be used to calculate suboptimal structures, base pair probabilities and perform folding space analysis, instead of predicting only the optimal structure.

Each of `X2` and `RNAalifold` explicitly combine specific sources of information in a single score system using pre-determined weights. `BayesFold` [75] proposes a bayesian strategy to combine any kind and number of data sources, without being dependent of arbitrary choices about weights.[10] Initially `RNAsubopt` is used to produce a first list of candidate structures (H_k). A uniform probability distribution is ascribed to these first candidates (D_0). Then, to each structure (H_i), a combined probability is ascribed using the different sources (D_j) of information by sequentially

[9] Optionally, it can be calculated as the average energy value of all pairings of the column-pair.

[10] Although it has a flexible schema for data source combination, it is initially implemented to combine thermodynamic value, mutual information and chemical maps.

applying the Bayes formula:

$$P(H_i|D_{j+1}) = P(H_i|D_j)P(D_{j+1}|H_i)/\left(\sum_k P(H_k|D_j)P(D_{j+1}|H_k)\right), \quad (4)$$

In other words, the probability of a structure given a source of information is combined with the probability of the same structure using another source by considering the first an a priori information for the calculation of the second, and assuming all data sources are independent.

4.2.2 Prediction from unaligned and folded sequences

When the input is a set of folded but unaligned sequences, the challenge is to produce a structural alignment of these sequences, detecting the consensus structure of the RNA family. For multiple structural alignment, the problem is similar to the standard multiple alignment of DNA sequences: computing the optimal multiple alignment is a problem where the solution has exponential time complexity[11] [71]. In this spirit, all approaches have two algorithms, the algorithm to align two sequences and the heuristic approach to perform multiple alignments incrementally aligning sequences or partial consensus two at a time.

An RNA structure can be represented as a tree.[12] Then, aligning two structures can be translated into the problem of aligning two trees [69]. If we describe each local secondary structure as a separate tree, the problem of aligning two structures is then translated into the problem of aligning a forest of trees. This is the approach used by `RNAForester` [57]. The advantage of the forest approach is that alignments now have a "local" flavor, permitting the program to find similar sub-structures when the global structures are divergent, and therefore detected conserved motifs in two RNA structures. `RNAforester` also performs global multiple forest alignment [58,111].

Another possible RNA structure representation is arc-annotated sequences where arcs connect paired bases. Aligning two structures in this representation means aligning both bases and arcs. `RNA_align` [68] performs this alignment by maximizing the score given by eight possible events: three events for unpaired bases (*base-match, base-mismatch* and *base-deletion*) and five events for arcs and their base pairs (*arc-match, arc-mismatch, arc-removing, arc-altering* and *arc-breaking*, the last three being particular cases of disruption of an arc).

`MARNA` [133] performs multiple structural alignment using `RNA_align` to perform the pairwise alignments and T-Coffee [102] to perform the progressive multiple alignment. Once the multiple alignment is built, a variant of Nussinov's algorithm [104] is used to predict the consensus structure that maximizes the number of base pairs conserved across a specific number of sequences. `MARNA` also accepts unfolded input sequences. In this case, it can predict one structure for each sequence

[11] The optimal dynamic programming solution has time $O(n^k)$ for k sequences, but solutions with optimal SP-scores are proven to be NP-hard [71].

[12] In this tree we have nodes to represent pairings, and leaves to represent unpaired nucleotides.

(MFE structure) or several suboptimal ones (using RNAshapes [134] or a stochastic backtracking of RNAsubopt [59]).

4.2.3 Prediction from unaligned and unfolded sequences

Some methods receive only the plain sequences as input. Therefore they have to fold and align the input sequences. Some of these methods perform this task in two steps, first fold then align,[13] and others perform the folding and the alignment simultaneously. Two-step methods, although they seem similar to those of group 2 (prediction from unaligned and folded sequences), have the advantage that they do not consider the information of only one structure of each input sequence to build the consensus structure. Instead, they take into account alternative structures. Simultaneous folding and alignment methods have the most attractive proposal. However, their drawback is the high time and memory complexities, exponential in the number of sequences. Although polynomial when applied for only two sequences, they are still time and memory expensive for real applications. Simplifications and heuristics are needed to deal with it.

Two-step methods

Briefly, these methods perform two phases: first, the prediction of a set of structures for each input sequence, usually using thermodynamic values; second, the selection of one consensus structure based on the sets of structural folds obtained in the first phase. The main difference among the methods is found in the similarity measure used to select the conserved structure and on the granularity of the initial folds (global folds vs. local folds).

Some predictors consider only the structural similarity, not taking into account the primary sequence. Therefore, they are suitable when comparing remote homologous sequences that present little primary sequence conservation. Examples are the method proposed by Bouthinon and Soldano [12] and RNAGA [17].

The method described by Bouthinon and Soldano [12] finds the individual sets of structures by detecting palindromes. Each palindrome represents a possible helix in a final folding, and a combination of compatible palindromes represents a possible fold. For each input sequence the only combinations with free energy values below an user-defined threshold are computed, and the combination with the lowest free energy value is selected. The set of the largest structural patterns (palindrome combinations) that occur in at least q sequences ($q \leq n, n$ = number of input sequences) is computed, representing the candidates for the consensus structure. The algorithm selects the structural pattern of this final set that occurs less frequently in in random sequences generated by shufflings of the input sequences.

The program RNAGA [17] uses two genetic algorithm: the first one produces, for each sequence, a set of global structures; the second one generates a list of

[13] A method can also first align then fold the sequences. However, they often merely uses a third party software (CLUSTALW [44]) to produce a single initial multiple alignment. Therefore this case was included in Sect. 4.2.1.

candidate consensus structure. Both algorithms use operators that add and remove helices according some criteria. The first algorithm prefers helices that decrease the energy of the new individual, whereas the second one prefers helices with a high structural conservation score.

Since the number of possible structures grows with the sequence length, the number of spurious common structures grows as well [12]. This effect can be decreased by taking into account the primary sequence during the structure comparison. The programs CARNAC [139] and comRNA [67] implement this strategy by using regions of primary sequence conservation as anchors to constrain the pairwise structural alignment of all pairs of the input sequences. From these alignments, they select a set P of pairwise conserved helices, i.e., a set

$$P = \{(h_i^A, h_j^B) | h_i^A \text{ and } h_j^B \text{ are conserved in the sequences } A \text{ and } B\}.$$

CARNAC and comRNA use two different criteria for this selection, but both require that two conserved helices must fall between the same anchors. CARNAC requires at least one compensatory mutation in the helix pair [107], whereas comRNA compares the similarities in helix length, helix energy, helix sequence and loop sequence (unpaired bases enclosed by the helix). Both programs build a graph where each selected helix h_k^X is represented as a node and is connected to the node h_l^Y iff $(h_k^X, h_l^Y) \in P$. In the graph produced by CARNAC, connected components[14] represent conserved helices that are candidates to participate in the final consensus structure. These components are scored based on topological features such as number of nodes and edges, since these features indicate the conservation level of the helices among all sequences. The final structure is built in a greedy approach by adding the compatible helices[15] represented by the best scored connected components. Therefore, helices do not have to be conserved in all sequences to be detected. The comRNA graph is an m-partite[16] graph, where each partition is made up of nodes from the same sequence and m is the number of input sequences. A helix conserved in at least k sequences is a clique having at least k nodes.[17] Therefore, the program searches all maximal cliques with size at least k. The larger cliques are selected to compose the final structure, allowing formation of pseudoknots.[18]

If on one hand similarity based only on structure leads to spurious consensus structures, on the other hand the dependence on the primary sequence conservation is a trap when dealing with very divergent sequences. Attacking this issue, RNAscf [4] uses conserved helices as anchors, and primary sequence only as additional informational.

[14] The connected components of a graph are the equivalence classes of vertices under the "is reachable from" relation [25].

[15] Two helices are considered compatible if they do not overlap or form a pseudo-knotted structure.

[16] An m-partite graph is a graph where the set of nodes can be divided into m disjoint sets such that any edge of the graph has endpoints in different sets.

[17] A clique is a complete subgraph, i.e., a subgraph in which all nodes are connected to all nodes.

[18] Although the algorithm for clique searching is exponential in the number of input sequences, its implementation in comRNA has an acceptable average processing time, allowing analysis of 18 sequences up to 300nt each [67].

The initial step consists in filtering spurious short helices by selecting, for each sequence, all possible helices with a minimum length with a thermodynamic value not exceeding a pre-defined threshold. Compatible helices conserved in all sequences are used as anchors for the multiple alignment. Each anchor helix generates a set of alignment blocks: blocks for the helices and blocks for the unpaired regions between the helix sides. Each block is globally aligned according to a mixture of thermodynamic and similarity scores, but using different score values for blocks coming from paired or unpaired regions.

Simultaneous folding and alignment

Simultaneously folding and aligning sequences is an optimization problem, where the optimal solution in the joint problem may not coincide with the optimal solutions for each sub-problem individually. This optimization can be deterministic or stochastic. Deterministic optimization produces a single solution for each input and parameters. Stochastic optimization has as a goal to explore alternatives in the solution space.

Deterministic optimization

Sankoff proposed a dynamic programming algorithm to optimize the folding and alignment using a score system composed by thermodynamic parameters and alignment scores [125]. However, the $O(n^{3m})$ time and $O(n^{2m})$ memory requirements (n = typical sequence length and m = number of sequences) of the unconstrained version limit the application to only two short sequences. Since then, many Sankoff variants have been developed by adding heuristics or simplifications to speed up the original algorithm. Examples are `Dynalign` [92,94], `PMComp` and `PMMulti` [60], `Stemloc` [63], `Consan` [36], `FoldAlign` [46] and `SLASH` [47].

`Dynalign` [92,94] is based on the original Sankoff algorithm. Imposing a limit w on the distance between two aligned nucleotides, time and memory complexities are reduced, respectively, to $O(n^3w^3)$ and to $O(n^2w^2)$. The score system is purely structural, which makes it attractive for application on divergent sequences.

`PMMulti` [60] performs a structural multiple alignment of a set of sequences using a companion program, `PMComp` to perform pairwise alignments.[19] `PMComp` uses the same distance limiting strategy of `Dynalign` to compute pairwise structural alignments. However, instead of implementing the thermodynamic rules, it uses the thermodynamic information from the base pair probability matrices computed for each input sequences. Dynamic programming recursions involving unpaired bases use an unpaired substitution score. Recursions involving base pairs (i, j) in the sequence A and (k, l) in the sequence B combine a paired substitution score, a covariation score and a thermodynamic-based probability score based on the joint probability of (i, j) and (k, l) being paired. More specifically, given two probability matrices P_A and P_B for the sequences A and B, respectively, the probability score is calculated as

[19] Actually, for the sake of speed, the first phase of calculation, where the initial unfolded sequences are aligned two at a time, `PMMulti` uses a coarser alignment of probabilities using an algorithm described in [10].

$\log(P_A(i, j)/p_{\min}) + \log(P_B(k, l)/p_{\min})$, where $p_{\min}$ is the minimum pair probability that is deemed significant. This strategy, together with the alignment distance constrain, allows PMComp to run in $O(n^4)$ time and $O(n^3)$ memory.

Two methods, Stemloc [63] and Consan [36], use a Pair Stochastic Context Free Grammars (Pair SCFGs) as probabilistic models and impose restrictions in the parsing algorithm to obtain gains in efficiency. Stemloc provides a flexible schema to impose user-defined constrains over the alignment, over the folding of one sequence and over the folding of two sequences. The program uses parsers adapted to restrict the syntactic analysis to *envelopes* defined by the constrains, reducing time and memory usage [64].

Consan [36] uses *pins*, a specific type of alignment anchor. A pin is a pair of positions from the two input sequences that have a high probability of being aligned. Two consecutive pins define alignment boundaries, where the alignment/folding is performed. The pins are selected from a probabilistic pairwise non-structural alignment obtained using a Pair Hidden Markov Model. Even few pins can decrease the time and memory requirements significantly, reaching best performance when they are evenly spaced. Depending on the constrains, STEMLOC can reach $O(n^2)$ time and memory complexity, and CONSAN can reach $O(n^3)$ time and $O(n^2)$ memory complexity for two sequences.

FOLDALIGN [54] performs local instead of global alignment, which can detect common motifs in the input sequences. The algorithm uses a sliding window on each sequence to limit the search for motifs. Windows have a limited size parameterized by the maximum size of a motif, λ. Further speedup is achieved by limiting the difference in size of two subsequences using a second input parameter, δ. The total time complexity is $O(n^2\lambda^2\delta^2 + N\lambda^4\delta^2)$, where N is the maximum number of structural local alignments to be extracted from the sequences.

The SLASH(Stem-Loop Align SearcH) system [47] combines FOLDALIGN [46] and COVE[20] [41]. FOLDALIGN is applied on a small subset of the sequences, producing a set of local structural alignments. These local structural alignments are combined to form a set of multiple alignments, each one describing a local motif. SLASH then uses COVE to search for occurrences of each of these motifs in the remaining sequences. SLASH detects only stem-loop structures, not branch structures, due to the use of an older version of FOLDALIGN.

Stochastic optimization

RAGA [101] and COFOLGA [137] use genetic algorithms to produce global structural alignments from unfolded pairs of sequences. In both cases, the population of individuals consists of putative pairwise alignments of the input sequences. Mutation operators change the alignments performing localized changes such as gap insertion or gap shift. Crossover operators generate descendants that keep the common alignment blocks of the parent alignments with a combination of the remaining alignment blocks.

[20] See Sect. 6.

RAGA (RNA Alignment by Genetic Algorithm) [101] and its parallel version PRAGA (Parallel RAGA) both require one of the sequences to be previously folded. Compatibility with this sequence's structure is used to evaluate the population of alignments. Both programs accept pseudoknots in the alignments.

COFOLGA (COmmon FOLding by Genetic Algorithm) [137], does not require any previous folding. Instead, it implements a variation of the simulated annealing algorithm described by Schmitz [129] to fold the pair of aligned sequences. In this variation, helices that are not compatible with the alignment of the sequences are discarded (that is, helices that occur only on one of the sequences). The score of the alignment and the free energy value of each structure are combined to obtain the individual score used in the selection process. In addition to the common structure, a post-processing step also predicts structures specific for each sequence.

5 Structural comparison

Structure comparison calculates how different two structures are. We can measure this difference by computing an *edit distance* between these two structures. The edit distance depends on how many edit operations we need to transform one of the structures into the other and on the cost of each type of edit operation. The computation of the edit distance is directly related to the way that structures are represented and at which resolution level the comparison is performed. Three common ways of representing structures are trees, bracket strings and generic graphs. Resolution levels range from base pairs to structural patterns like helices, loops and multi-loops.

5.1 Using trees and strings

When using trees to represent structures, the edit operations used to compute distance are the insertion and removal of nodes. Some programs label the tree nodes with numerical information, affecting the weight of each operation.

RNAdistance [62] defines three resolutions: full, coarse grained or weighted coarse grained. Full resolution uses two representations: bracket strings and trees. Both coarse grained and weighted coarse grained use trees. Bracket strings use matched open and close brackets (or parenthesis) to indicate paired bases, while dots indicate unpaired bases [62]. The distance between two structures in this representation is computed as the number of gaps needed to produce an alignment of the respective strings. The trees used to represent full resolution, homeomorphically irreducible trees (*HITs*) are composed of two types of nodes, P for paired bases and U for unpaired ones. Each node is labeled by the number of consecutive paired or unpaired bases, respectively. For instance, a helix having three consecutive base pairs is represented by the node P3. The trees used to represent coarse grained and weighted coarse grained resolution have five types of nodes [132]: "stem", "hairpin", "bulge", "internal loop" and "multi-loop". The difference is that in weighted coarse grained, the nodes are labeled with their size (base pairs for helices, and single bases for all others).

MiGal [3] uses trees with only two types of (unlabeled) nodes: "helix" and "loop". On the other hand two new edit operations are used: "edge fusion" and "node fusion".

These new operations enable to associate, for instance, one long helix to two smaller helices separated by an internal loop or bulge, instead of performing an one-to-one association (one helix to just one of the smaller helices) demanded by the traditional tree edit operations described above.

5.2 Using graphs

A secondary structure can also be represented as a generic graph (i.e. not a tree). This allows the application of graph theory techniques to perform topological graph comparison. One possible mapping is modeling loops as nodes and helices as edges. The second smallest eigenvalue of the Laplacian matrix of a graph is a measure of its connectivity, which indicates the branching pattern of the secondary structure [6]. This value, together with the number of vertices, can be used to cluster similar structures.

`RNAMute` [20] uses graphs along with tree edit metrics to predict structural changes caused by point mutations.

The same principle is used to characterize and search RNA structures in the `RAG` database [45]. This database is a catalog of real and hypothetical secondary structures in graph format, accumulating information about topological features (including pseudoknots) and RNA families presenting specific structures. Given the secondary structure of an RNA of interest, `RAG` outputs structurally isomorphic RNAs from its database.

5.3 Using the entire folding space

`RNApdist` [62] compares the whole folding space of two sequences to compute a distance value. The computation of this distance takes into account the average length of the two sequences and the similarity score S of the alignment of their folding spaces. The folding space of each sequence is represented by three vectors, $p^<$, $p^>$ and p^o, where $p_i^<$, $p_i^>$ and p_i^o are the probability of a position i of a sequence being upstream paired, downstream paired and unpaired, respectively. These values are calculated using the thermodynamic base pair probability matrix of that sequence:

$$p_i^< = \sum_{j>i} p_{ij} \qquad p_i^> = \sum_{j<i} p_{ij} \qquad p_i^o = 1 - p_i^< - p_i^> \tag{5}$$

Given two positions i and k from the sequences A and B, respectively, the score $\gamma(i,k)$ to align these two positions is defined as

$$\gamma(i,k) = \sqrt{p_i^<(A)p_k^<(B)} + \sqrt{p_i^>(A)p_k^>(B)} + \sqrt{p_i^o(A)p_k^o(B)} \tag{6}$$

The score of a particular alignment of A and B is

$$\hat{\gamma}(\overrightarrow{i}, \overrightarrow{k}) = \sum_{i \text{ aligned } k} \gamma(i,k) \tag{7}$$

and the score of the optimal alignment is

$$S(A, B) = \max_{(\vec{i}, \vec{k})} \hat{\gamma}(\vec{i}, \vec{k}) \tag{8}$$

Finally, a distance $\delta(A, B)$ between the two sequences A and B is defined as

$$\delta(A, B) = \frac{|A| + |B|}{2} - S(A, B) \tag{9}$$

where $|A|$ and $|B|$ are the lengths of the sequences A and B, respectively.

6 Noncoding RNA identification

With the exponential growth of the sequencing data being generated, the task of scanning a new genome to find candidate ncRNA genes is becoming increasingly important. Computational detection of ncRNAs in general is a challenge and considered an open problem. A sensible approach is either the development of identification programs targeted to very specific ncRNA families, using as much as possible peculiarities of these families, or to create more general programs that can be trained to identify characteristics of a specific family or even a single input sequence. Still, the development of general ncRNA gene-finders is still a very important challenge, since it is also desirable to search for new gene families.

Gene finders can be implemented as genome scanners or as a classifiers. Genome scanning is often performed by sliding a window and analyzing its sequence. Classifiers receive an input and output a label to it. This input can be a sequence in the case of an ab initio gene finder or an alignment in the case of a comparative gene finder. In the last case, input sequences can be obtained by using a sliding window over the genome or from transcript sequences.

6.1 General ncRNA gene-finders

6.1.1 Ab initio methods

Using only thermodynamic information

Le et al. [83] proposed that structured RNAs have a lower free energy than random sequences with the same mononucleotide frequency. This characteristic could be used to detect putative ncRNAs by sliding a window over a genome and comparing its MFE with shufflings of the same sequence. The program `NCRNASCAN` [118] was implemented to test this strategy. However, the authors showed that, although most of the ncRNAs present this energetic difference, this difference is not significant enough to be used as a general ncRNA discriminator.

Considering the fact that the MFE calculation is based on base stackings, Workman and Krogh [152] argued that the MFE of the original sequence should be compared with

shufflings that preserve the dinucleotide frequency, and not only the mononucleotide one. The program RANDFOLD [11] was implemented based on this argument. It was used to show that, although some ncRNAs do not have a significant lower energy, precursor microRNAs do have. Posterior work [23] showed that other ncRNAs can also be detected by the significance of their folding energy.

When performing thermodynamic calculations, the minimum free energy value is affected if extra nucleotides are included in the upstream and/or downstream of a real RNA. Similarly, missing nucleotides also have a deleterious effect. This has an important impact when scanning a genome with a sliding window, since the size of the window cannot be guaranteed to match exactly that of the ncRNAs that are being searched. The problem of missing nucleotides can be overcome using a window large enough to accommodate the any putative ncRNA, but the effect of additional nucleotides remains. As a consequence, a single folding of each window will produce an unreal energy value. To decrease this problem, RNAplfold [8] combines the information obtained from all individual windows from a genome. Given a window length L, the stable structures of the genome are detected by computing, for all positions i and j the average probability of two bases at positions i and j being paired. This probability is calculated considering all possible structures in all L-size windows having the bases i and j.

Using compositional information

Noncoding RNAs are reported to have on average, a G+C content of 50% [118].[21] This fact inspired ncRNA search using compositional statistics. Success was achieved when searching GC-rich islands in some AT-rich organisms [74]. Similar strategy using other mono and dinucleotide statistics was also performed in other organisms, being the G+C one of the most significant statistics in the successful tests [126].

Using machine learning techniques

Machine learning approaches are an interesting option when designing adaptable gene finders. Once a model is designed it can be trained to recognize a specific gene family from samples of existing genes. CONC [86] is a program developed to analyze transcribed sequences using Support Vector Machine technology (SVM). It was implemented to classify a sequence either as protein-coding or noncoding RNA. It uses a set of features to characterize protein-coding sequences such as aminoacid composition, peptide length and other more specific features. The positive training sample is a set of protein-coding sequences whereas the negative one is a set of noncoding RNAs.

[21] G+C content is the percentage of bases that are either G or C.

6.1.2 Comparative methods

Given two or more related species, it is possible to perform alignments of their whole genomes to find the maximum regions of similarity between them. In the case of multiple genomes, these alignments can be built comparing two genomes at a time and then building a multiple alignment by extracting aligned regions common to all genomes. Once a multiple alignment is available, either thermodynamic information or covariation analysis can supply evidence for a ncRNA labeling. This means, however, that almost all programs depend heavily on the quality of these initial alignments.

The program `ddbRNA` [7] computes the number of compensatory mutations in a multiple alignment and its z-score by shuffling the alignment. In spite of being very fast ($O(n^2)$) this program has very low reliability (average sensitivity of $\sim$22% for pairwise blast alignments), and has not been used in real-life applications [148].

`MSARI` [26] decreases the dependence on the quality of the initial alignment by allowing misalignments with up two characters of distance. In addition, the significance of covariations is calculated taking into account the possibility of random nucleotide substitutions causing covariation. Still, to achieve high accuracy, `MSARI` needs a deep alignment of 10–15 sequences.

`QRNA` [119] uses three different probabilistic models to classify pairwise alignments: a pair-SCFG model for ncRNAs, a pair-HMM model for protein coding RNAs and a position-independent pair-HMM for "other" sequences.[22] Originally, `QRNA` had low reliability for alignments outside the optimal identity range of 65–85% [148], however, a newer version, `eQRNA` [116], includes the evolutionary distance between the sequences in the model parameters, achieving better results.

`RNAZ` [148], a development of `RNAalifoldz` [147], is a Support Vector Machine classifier that uses two scores to classify multiple alignments as ncRNA or not: z-scores[23] of the individual sequences from the input alignment and a structure conservation index (SCI). The last is given by $E_A/\overline{E}$, where E_A is the alignment thermodynamic score given by the `RNAalifold` and $\overline{E}$ is the average free energy of the individual sequences (given by the `RNAfold`). The idea is that, the higher the structure conservation and covariations, the higher the SCI is.

As we have mentioned above, all comparative methods described above depend of the alignment accuracy. Even `MSARI` only reports good results for alignments having identity superior to 50%. Recent work by Uzilnov et al. [143] proposed a method that does not require an initial alignment by using `Dynalign`, a tool for conserved secondary structure prediction.[24] `Dynalign` simultaneously aligns and folds two sequences considering only thermodynamic information. Since the program outputs the free-energy of the alignment, this can be used to calculate the alignment's

[22] We can say that "other" means "I have no idea".

[23] The z-score computation needs the average and standard deviation of the scores of negative sequences. Often, these measures are calculated using shufflings of the input sequence. Instead, `RNAZ` pre-computed these values for specific sequence lengths and base compositions using a Support Vector Machine regression. It speeds up the `RNAZ` runtime.

[24] See Sect. 4.2.3.

thermodynamic z-score.[25] In tests involving low-identity sequences (less than 50%), `Dynalign` outperformed `RNAZ` both in sensitivity and specificity.

6.2 ncRNA gene-finders based on a set of known sequences

These gene-finders address the problem of, given a set of genes from the same RNA family, analyze another sequence set and classify each one as belonging to that first family or not.

Stochastic context free grammars were originally proposed to characterize RNA families and search homologous by the independent and simultaneous work of Sakakibara et al. [124] and Eddy and Durbin [41].

Sakakibara et al. [123] designed grammars for tRNAs and snoRNAs [142] by hand. Each grammar was used to parse sequences, computing a probability value. Sequence classification was performed based on z-score calculation (details in [78]).

In contrast, Eddy and Durbin developed the `INFERNAL` package [41] (previously called `COVE`) using a specific type of SCFGs called *Covariance Models*. `INFERNAL` includes programs to automatically infer a grammar from a structural multiple alignment (which will characterize an ncRNA family) and to scan a genome using one such grammar, searching for candidate homologous sequences. One important characteristic of `INFERNAL` is using a search algorithm more memory efficient than other SCFG-based ones ($O(n^2 \times \log n)$ against $O(n^3)$) [39], which allows the search for long ncRNA sequences such as ribosomal RNAs.

`Rfam` [49,50] is a database system that stores covariance models for 503 different ncRNA families (release 7.0) and that uses `INFERNAL` to identify candidate ncRNAs in an arbitrary input sequence (possibly a whole genome). Since analyzing the input data using all of the 503 models would be too slow, a previous Blast-based filter is applied to select promising models. This similarity filter imposes restrictions on the degree of variation that is acceptable for finding ncRNAs. Weinberg et al. [150,151] proposes a different filter, based on profile HMMs, to minimize this effect. These filters, however, are not part of the `Rfam` database.

Analyzing a sequence, even by a single SCFG, can be slow. The $O(n^4)$ time complexity means a long time waiting for the analysis of longer sequences. To speed up this process, parsing constrains were implemented in the SCFG-based `RNACAD` [14] package, a software used in the Ribosomal Database Project (RDP-II) [24]. Each constrain limits the subsequences that may be recognized by a given non-terminal. These constrains are generated automatically by building an HMM that approximates the SCFG, and identifying subsequences associated with an HMM state with high probability. These associations constitute the SCFG constrains.

Context-free grammars cannot model crossed dependences, and therefore are also unable to model pseudoknots [131]. We can model a pseudoknot, however, by modelling each of the two helices separately. One such model based on intersections of SCFGs is described by Brown and Wilson [15]. This work describes how

[25] Another option explored in this work was the use of the alignment energy as a feature of a Support Vector Machine classifier (faster but slightly less sensitive).

to estimate the probabilities of the grammar combinations and also how to parse sequences keeping the time and complexities of $O(n^3)$ and $O(n^2)$, respectively.

PSoL [145] proposes a very flexible manner of parameterizing classification. It receives as input feature vectors of the training sequences (the RNA family to characterize) and of the search data base and then selects a set of significant features to be used in the classification process. PSoL uses SVM technology and, therefore, needs positive and negative training samples. While the positive sample is a set of ncRNAs of interest, a proper definition of a negative sample is not trivial. Therefore, PSoL (Positive Sample only Learning) implements a strategy to select this negative sample directly from the unlabeled sample (sequences to be classified). An initial set of negative sequences are selected in order to maximize the distance between the the positive and the current negative sample as well as the mutual distance between negative training sequences.[26] After an initial training using this sample, the classification is started and the negative sample is incrementally expanded, retraining the SVM. This process is performed until the unlabeled sample reaches a threshold size. PSoL was developed to perform a search for any ncRNA in a database, having as a training sample a mixed set of ncRNAs families. However, it is clear that the approach is flexible enough to also be applied to specific gene families.

There are two ncRNA gene finders that do not perform previous statistical training, RSEARCH [73] and FastR [5]. Both accept as input a RNA sequence and its secondary structure and search a database to find sequences similar to the input by performing local pairwise structural alignments. To build this alignment, both use RIBOSUMs, RNA-specific substitution matrices developed for RSEARCH. Both methods, in addition to the alignment score, output the score significance. Scanning a database, however, may be a time consuming task. For a query sequence of length n and a database of length m, the worst case time complexity for scanning is $O(mn^3)$, plus an additional $O(n^4)$ for statistical significance calculations. RSEARCH tries to circumvent this difficulty by also offering a parallel implementation. FastR, in contrast, uses structural filters to pre-select promising sequences.

6.3 Family-specific ncRNA gene finders

In order to pursue better sensitivity and specificity, some methods are tailored to deal with specific families of RNAs, exploring features specific to each of them. Some of these methods were implemented and are available as standalone programs or web services. Others methods were never made available as integrated programs, being searching methodologies that were used together with heuristics, manual refinements and expert inspection, sometimes using more general tools such as those mentioned throughout this paper [108,112,140]. A complete listing of all of them would be inappropriate. Therefore, we selected some methods from the first group in order to exemplify how the a priori knowledge about an RNA family can be used to build a family-specific ncRNA gene finder.

[26] The last requirement intends to minimize the redundancy and improve the cover of the negative sample.

Transfer RNAs (tRNAs) are sequences having between 74 and 90 bases, transcribed by RNA polymerase III, that fold in a clover-leaf structure. The program `tRNAscan-SE` [87] is considered one of the most accurate tRNA predictors [81]. It combines three programs: two tRNA predictors that search for RNA polymerase III promoters and characteristic secondary structure [42,105] and a core Covariance Model [41] trained with tRNAs sequences. The first two programs are fast and, when combined, have a sensitivity superior to ∼99%. However, such combination implies roughly ∼1.85 false positives per Mb, which is acceptable for small genomes, but it means ∼5,500 false positives in the human genome. The Covariance Model is very sensitive and specific, but too slow. Therefore, the first two tRNA predictors are used with low stringency as pre-filters in order to get promising tRNA candidates from a genome. Then the candidates are analyzed by the highly stringent covariance model. The result is a tRNA finder presenting higher sensitivity (99–100%) and selectivity (with a rate lower than 0.00007 false positives per Mb) with reasonable speed (30 Kb/s).

Transfer-messenger RNAs (tmRNAs) are sequences having between 350 and 400 bases[27] that are able to liberate defective mRNAs from stalled ribosomes [52]. The 5′ and 3′ ends of this molecule form a tRNA-like domain that surrounds an internal region consisting of stem-loops, pseudoknots and a messenger RNA domain. The last codes for a tag peptide that signals the defective mRNA for degradation [82].

`ARAGORN` [81], an improved version of `BRUCE` [82], is a program that identifies tmRNAs by selecting candidates having specific tRNA sequence motifs involving the 5′ and 3′ ends and inspecting the predicted secondary structure. In addition, `ARAGORN` analyzes the regions around the tag peptide to analyze if they can fold in specific structural motifs. `ARAGORN` can also detect tRNAs with a sensitivity comparable to `tRNAscan-SE`, but does not match its specificity.

Small nucleolar RNAs (snoRNAs) are sequences having between 60 and 300 bases that are related with site-specific modifications of other RNAs [95]. They are divided in *guide* and *orphan* snoRNAs, depending on the presence or absence of a known RNA target. Depending on their secondary structure, they are classified as *C/D box* and *H/ACA box*. The secondary structure of C/D box snoRNAs is basically a hairpin-like structure with a large hairpin-loop, where some regions located in this loop (C and D boxes) have a conserved sequence. This loop also has regions that are complementary to the target RNA. The secondary structure of H/ACA box snoRNAs consists in two consecutive helices, each one followed by a conserved loop region (H box after the first helix and ACA box after the second one). In addition, each helix has an internal loop whose sequences are complementar to the target RNA.

The program `snoscan` [88] is a guide C/D box snoRNA finder that uses a SCFG to model the hairpin, one HMM to each conserved region and a length distribution model to characterize the distances between the conserved regions. The combined model is then used to score snoRNA candidates. This program can only detect guide snoRNAs due to the fact that one of the HMMs models the target-binding region for

[27] *C. merolae* tmRNA is exceptionally smaller, having only 235 bases [103].

known targets. Combination of different models for distinct regions is also employed by snoGPS [128] to identify guide H/ACA snoRNAs. Weight matrices are used to model several regions of primary sequence (boxes, internal loops and separate helix sides) and length distribution models are used to connect them. Recently, a snoRNA finder for both guide and orphan snoRNAs was developed. The snoSeeker system [154] consists of two programs: CDseeker and ACAseeker. They first search for all conserved regions (boxes) and then search for the structural evidences (helices) of the corresponding snoRNA type. The last step is to search for targets matching the complementar target-binding region.[28] If the target is found, the candidate is classified as guide, otherwise it is classified as orphan.

Micro RNAs (miRNAs) are initially transcribed as sequences between 60 and 90 bases that fold in a stem-like structure (pre-miRNA). This sequence is processed, producing two molecules of ∼21 bases, each one coming from one stem side. The sequence originated from the 5′ side will anneal with a target mRNA in order to avoid its translation.[29] This processed sequence is well conserved across species, the initial portion being the most conserved.

Micro RNA predictors, in addition to searching for miRNA's common features, also search for evolutionary conservation. For instance, miRseeker [80] and mirScan [84] use the sequence of two genomes of different species to identify candidate miRNAs conserved in both. These genomes are searched for potential stem-loop structures that, if conserved, are selected as pre-candidates for posterior analysis. The programs miralign [146] and ProMirII [100] use a set of known miRNAs and compare them with miRNA candidates in an input genome. Finally, RNAmicro [55] does not search a genome, but instead classifies multiple sequence alignments.

7 Perspectives and conclusion

There is no tool that solves a problem for any kind of RNA. A good practice is to carefully choose a set of applicable methods which different approaches and compare the results. In addition, insertion of a priori knowledge is a powerful strategy, which is particularly relevant in noncoding RNA research where the solution space is often too large. For secondary structure prediction, alternative structures should be analysed instead of just the optimal one, constrains should be inserted if partial information about the structure is available, the predicted folding process should be analysed by kinetic tools, and comparative methods should be also used if homologous sequences are available. For ncRNA searching, different suitable methods should be used in order to build a combined set of candidates.

The thermodynamic approach has been shown to yield the most accurate results for ab initio secondary structure prediction. Therefore, most of the methods that perform a folding space analysis, such as density of states, are based on this approach. Recently, a novel method using probabilistic models, CONTRAfold [34], outperformed the best

[28] The target RNAs are assumed to be rRNA or snRNAs.

[29] In some miRNAs both sequences can be used into different mRNA targets.

thermodynamic predictors for the first time. A next step is to explore the use of this model in the folding space analysis.

Since there may exist unknown ncRNA families, tools for searching general ncRNAs are desirable. Until few years ago, no satisfactory method was available. Recently, new methods were developed and achieved results that has been bringing some light at the end of the tunnel.

Methods incorporating evolutionary models have presented promising results. One particular general ncRNA finder that performs a comparative strategy, EQRNA [116], parameterized the time divergence between the two organisms being compared. The same approach could be adopted by trainable ncRNA finders, in order to adapt its parameters according to the phylogenetic distance between the target organism and those used in the training sample.

The ncRNA research has a long road ahead. Current methods are based on our knowledge about ncRNAs. But probably we know only a small fraction of the RNA world. Adding insights and speculations to our current knowledge may help us to discover new facets about this world and, in turn, improve computational tools.

Finally, we would like to reinforce that all these methods make predictions. Therefore, their results can not be considered the supreme truth even when more than one method are in agreement. A biological experimentation is needed for a reliable validation of in silico predictions.

Acknowledgments All authors idealized this review. AML and AMD elaborated the manuscript. HAP was responsible for biological input and for help in the revision process.

Appendix

Tables 1 to 3 present a list of all available programs or web servers.

Table 1 Some available programs and web servers for secondary structure prediction

Method	URL
Secondary structure prediction	
Ab initio	
CONTRAfold [34]	http://contra.stanford.edu/contrafold/
HotKnots [115]	http://www.cs.ubc.ca/labs/beta/Software/HotKnots/
ILM [121,122]	http://cic.cs.wustl.edu/RNA/
Kinfold [43]	http://www.tbi.univie.ac.at/∼xtof/RNA/Kinfold
MFOLD [157,158,160]	http://www.bioinfo.rpi.edu/applications/mfold/
MWM [136]	ftp://ftp.cshl.org/pub/science/mzhanglab/tabaska/
NUPACK [32,33]	http://www.acm.caltech.edu/ niles/software.html
PKNOTS [117]	http://selab.janelia.org/software
pknotsRG [110,130]	http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/
RNAfold [59]	http://www.tbi.univie.ac.at/∼ivo/RNA/
RDfolder [156]	http://rna.cbi.pku.edu.cn/
RNAKinetics [28]	http://bioinf.fbb.msu.ru/RNA/kinetics/

Table 1 continued

Method	URL
RNALOSS [21,22]	http://clavius.bc.edu/∼clotelab/RNALOSS/
RNAshapes [134,144]	http://bibiserv.techfak.uni-bielefeld.de/rnashapes/
RNASTRUCTURE [93]	http://rna.urmc.rochester.edu
RNAsubopt [153]	http://www.tbi.univie.ac.at/∼ivo/RNA
Sfold [30,31,16]	http://sfold.wadsworth.org/index.pl
Comparative	
BayesFold [75]	http://jaynes.colorado.edu/Bayes/
Bouthinon et al. [12]	Available upon request
CARNAC [139]	http://bioinfo.lifl.fr/carnac
COFOLGA [137]	Available upon request
comRNA [67]	http://ural.wustl.edu/∼yji/comRNA/
Consan [36]	http://selab.wustl.edu/people/robin/consan
ConStruct [90]	http://www.biophys.uni-duesseldorf.de/local/ConStruct
Dynalign [92,94]	http://rna.urmc.rochester.edu
FOLDALIGN [54]	http://foldalign.kvl.dk
ILM [121,122]	http://cic.cs.wustl.edu/RNA/
MARNA [133]	http://biwww2.informatik.uni-freiburg.de/Software
MWM [136]	ftp://ftp.cshl.org/pub/science/mzhanglab/tabaska/
Pfold [76,77]	http://www.daimi.au.dk/∼compbio/rnafold/
PMComp/PMMulti [60]	http://www.tbi.univie.ac.at/∼ivo/RNA/PMcomp/
RAGA/PRAGA [101]	http://igs-server.cnrs-mrs.fr/∼cnotred/
RNAalifold [61]	http://www.tbi.univie.ac.at/∼ivo/RNA/
RNA_align [68]	http://www.csd.uwo.ca/∼bma/rna_align
RNA-Decoder [106]	http://www.ebi.ac.uk/∼meyer/rnadecoder/
RNAforester [57]	http://bibiserv.techfak.uni-bielefeld.de/rnaforester
RNAGA [17]	ftp://ftp.ncifcrf.gov/pub/users/chen/rnaga.tar.Z
RNAscf [4]	Available upon request
Stemloc [63]	http://biowiki.org
SLASH [47]	http://www.bioinf.au.dk/slash
X2 [70]	http://tyrant.ucsc.edu/X2s

Table 2 Some available programs and web servers for structure comparison

Method	URL
Structural comparison	
MiGal [3]	http://igm.univ-mlv.fr/ allali/migal/
RAG database [45]	http://monod.biomath.nyu.edu/rna
RNAdistance [62]	http://www.tbi.univie.ac.at/∼ivo/RNA
RNAMute [20]	http://www.cs.bgu.ac.il/ RNAMute/
RNApdist [10]	http://www.tbi.univie.ac.at/∼ivo/RNA

Table 3 Some available programs and web servers for ncRNA identification

Method	URL
NcRNA identification	
General	
NCRNASCAN [118]	http://selab.janelia.org/software
RANDFOLD [11]	http://bioinformatics.psb.ugent.be/software.php
RNAplfold/RNALfold [8]	http://www.tbi.univie.ac.at/∼ivo/RNA
ddbRNA [7]	http://www.tigem.it/Research/Personal%20Web%20Page_files/dibernardo/links.htm
Dynalign [143]	http://rna.urmc.rochester.edu
EQRNA [116,119]	http://selab.janelia.org/software
MSARI [26]	http://theory.csail.mit.edu/MSARi
Rfam database [49,50]	http://rfam.wustl.edu
RNAZ [148]	http://www.tbi.univie.ac.at/*sim*wash/RNAz
Query-based	
INFERNAL [39]	http://selab.janelia.org/software
RNACAD [14]	http://www.cse.ucsc.edu/∼mpbrown/rnacad
RSEARCH [73]	http://selab.janelia.org/software
Family-specific	
ARAGORN [81]	http://bioinfo.thep.lu.se
miralign [146]	http://166.111.201.26/miralign
mirScan [84]	http://genes.mit.edu/mirscan
miRseeker [80]	http://www.fruitfly.org/seq_tools/miRseeker.html
ProMirII [100]	http://cbit.snu.ac.kr/∼ProMiR2
RNAmicro [55]	http://www.bioinf.uni-leipzig.de/Software
snoscan [88,127]	http://lowelab.ucsc.edu/snoscan
snoGPS [127,128]	http://lowelab.ucsc.edu/snoGPS
tRNAscan-SE [87]	http://lowelab.cse.ucsc.edu/tRNAscan-SE

References

1. Abrahams, J.P., van den Berg, M., van Batenburg, E., Pleij, C.: Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. Nucleic Acids Res. **18**(10), 3035–3044 (1990)
2. Akmaev, V.R., Kelley, S.T., Stormo, G.D.: Phylogenetically enhanced statistical tools for RNA structure prediction. Bioinformatics **16**(6), 501–512 (2000)
3. Allali, J., Sagot, M.F.: A new distance for high level RNA secondary structure comparison. Trans. Comput. Biol. Bioinform. **2**(1), 3–14 (2005)
4. Bafna, V., Tang, H., Zhang, S.: Consensus folding of unaligned RNA sequences revisited. J. Comput. Biol. **13**(2), 283–295 (2006)
5. Bafna, V., Zhang, S.: Fast, R.: Fast database search tool for non-coding RNA. In: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB2004) (2004)
6. Barash, D.: Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation. Bioinformatics **20**(12), 1861–1869 (2004)

7. di Bernardo, D., Down, T., Hubbard, T.: ddbRNA: detection of conserved secondary structures in multiple alignments. Bioinformatics **19**(13), 1606–1611 (2003)
8. Bernhart, S.H., Hofacker, I.L., Stadler, P.F.: Local RNA base pairing probabilities in large sequences. Bioinformatics **22**(5), 614–615 (2006)
9. Blackburn, E.H.: Telomerase (1993) The RNA World. Cold Spring Harbor Laboratory Press, New York
10. Bonhoeffer, S., McCaskill, J.S., Stadler, P.F., Schuster, P.: RNA multi-structure landscapes—a study based on temperature dependent partition functions. Eur. Biophys. J. **22**(1), 13–24 (1993)
11. Bonnet, E., Wuyts, J., Rouze, P., de Peer, Y.V.: Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20**(17), 2911–2917 (2004)
12. Bouthinon, D., Soldano, H.: A new method to predict the consensus secondary structure of a set of unaligned RNA sequences. Bioinformatics **15**(10), 785–798 (1999)
13. Brown, J.W.: The ribonuclease P database. Nucleic Acids Res. **27**(1), 314 (1999)
14. Brown, M.P.S.: Small subunit ribosomal RNA modeling using stochastic context-free grammars. Proc. Int. Conf. Intell. Syst. Mol. Biol. **8**, 57–66 (2000)
15. Brown, M.P.S., Wilson, C.: RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. Pacif Symposium on Biocomputing, pp. 109–125 (1996)
16. Chan, C.Y., Lawrence, C.E., Ding, Y.: Structure clustering features on the Sfold web server. Bioinformatics **21**(20), 3926–3928 (2005)
17. Chen, J.H., Le, S.Y., Maizel, J.V.: Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Res. **28**(4), 991–999 (2000)
18. Chen, S.J., Dill, K.A.: RNA folding energy landscapes. Proc. Natl Acad. Sci. **97**(2), 646–651 (2000)
19. Chiang, D., Joshi, A.K.: Formal grammars for estimating partition functions of double-stranded chain molecules. In: Proceedings of HLT 2002, San Diego, March, pp. 63–67 (2002)
20. Churkin, A., Barash, D.: RNAmute: RNA secondary structure mutation analysis tool. BMC Bioinformatics **7**(221) (2006). doi:10.1186/1471-2105-7-221
21. Clote, P.: An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov–Jacobson energy model. J. Comput. Biol. **12**(1), 83–101 (2005)
22. Clote, P.: RNALOSS: a web server for RNA locally optimal secondary structures. Nucleic Acids Res. **33**, 600–604 (2005)
23. Clote, P., Ferre, F., Kranakis, E., Krizanc, D.: Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**, 578–591 (2005)
24. Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarell, D.M., Schmidt, T.M., Garrity, G.M., Tiedje, J.M.: The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res. **31**(1), 442–443 (2003)
25. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms. MIT Press, Cambridge (1990)
26. Coventry, A., Kleitman, D.J., Berger, B.: MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc. Natl Acad. Sci. **101**(33), 12, 102–12, 107 (2004)
27. Cupal, J., Hofacker, I.L., Stadler, P.F.: Dynamic programming algorithm for the density of states of RNA secondary structures. Comput. Sci. Biol. **96**, 184–186 (1996)
28. Danilova, L.V., Pervouchine, D.D., Favorov, A.V., Mironov, A.A.: RNAKINETICS: a web server that models secondary structure kinetics of an elongating RNA. J. Bioinform. Comput. Biol. **4**(2), 589–596 (2006)
29. Ding, Y.: Statistical and bayesian approaches to RNA secondary structure prediction. RNA **12**, 323–331 (2006)
30. Ding, Y., Chan, C.Y., Lawrence, C.E.: Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. **32**(Web Server issue), W135–W141 (2004)
31. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31**(24), 7280–7301 (2003)
32. Dirks, R.M., Pierce, N.A.: A partition function algorithm for nucleic acids secondary structure including pseudoknots. J. Comput. Chem. **24**(13), 1664–1677 (2003)
33. Dirks, R.M., Pierce, N.A.: An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J. Comput. Chem. **25**, 1295–1304 (2004)

34. Do, C.B., Woods, D.A., Batzoglou, S.: CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics **22**(14), e90–e98 (2006)
35. Dowell, R.D.: RNA structural alignment using stochastic context-free grammars. Ph.D. Thesis (2004)
36. Dowell, R.D., Eddy, S.R.: Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics **7**, 400 (2006)
37. Eddy, S.R.: Non-coding RNA genes and the modern RNA world. Nat. Rev. **2**, 919–929 (2001)
38. Eddy, S.R.: Computational genomics of noncoding RNA genes. Cell **109**, 137–140 (2002)
39. Eddy, S.R.: A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics **3**(1), 18 (2002)
40. Eddy, S.R.: How do RNA folding algorithms work. Nat. Biotechnol. **22**(11), 1457–1458 (2004)
41. Eddy, S.R., Durbin, R.: RNA sequence analysis using covariance models. Nucleic Acids Res **22**(11), 2079–2088 (1994)
42. Fichant, G.A., Burks, C.: Identifying potential tRNA genes in genomic DNA sequences. J. Mol. Biol. **220**, 659–671 (1991)
43. Flamm, C., Fontana, W., Hofacker, I.L., Schuster, P.: RNA folding at elementary step resolution. RNA **6**, 325–338 (2000)
44. Higgings, D.G., Thompson, J.D., Gibson, T.J.: Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**, 383–402 (1996)
45. Gan, H.H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., Kim, N., Schlick, T.: RAG: RNA-As-Graphs database—concepts, analysis, and features. Bioinformatics **20**(8), 1285–1291 (2004)
46. Gorodkin, J., Heyer, L.J., Stormo, G.D.: Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res. **25**(18), 3724–3732 (1997)
47. Gorodkin, J., Stricklin, S.L., Stormo, G.D.: Discovering common stem-loop motifs in unaligned RNA sequences. Nucleic Acids Res. **29**(10), 2135–2144 (2001)
48. Greider, C.: Telomerase biochemistry and regulation (1995) In: Telomeres. Cold Spring Harbor Laboratory Press, New York
49. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R.: Rfam: an RNA family database. Nucleic Acids Res. **31**(1), 439–441 (2003)
50. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. **33**, D121–D124 (2005)
51. Gulko, B., Haussler, D.: Using multiple alignment and phylogenetic trees to detect RNA secondary structure. Pacific Symposium on Biocomputing, pp. 350–367 (1996)
52. Haebel, P., Gutmann, S., Ban, N.: Dial tm for rescue: tmRNA engages ribosomes stalled on defective mRNAs. Curr. Opin. Struct. Biol. **14**, 58–65 (2004)
53. Hannon, G.J.: RNA interference. Nature **418**, 244–251 (2002)
54. Havgaard, J.H., Lyngso, R., Stormo, G.D., Gorodkin, J.: Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics **21**(9), 1815–1824 (2005)
55. Herbel, J., Stadler, P.F.: Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics **22**(14), 197–202 (2006)
56. Higgs, P.G.: RNA secondary structure: physical and computational aspects. Q. Rev. Biophys. **33**(3), 199–253 (2000)
57. Hochsmann, M., Toller, T., Giegerich, R., Kurtz, S.: Local similarity in RNA secondary structures. In: Proceedings of the Computational Systems Bioinformatics **(CSB 2003)**, 159–168 (2003)
58. Hochsmann, M., Voss, B., Giegerich, R.: Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE Trans. Comput. Biol. Bioinform. **1**(1), 53–62 (2004)
59. Hofacker, I.L.: Vienna RNA secondary structure server. Nucleic Acids Res. **31**(13), 3429–3431 (2003)
60. Hofacker, I.L., Benhart, S.H.F., Stadler, P.F.: Alignment of RNA base pairing probability matrices. Bioinformatics **20**(14), 2222–2227 (2004)
61. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319**, 1059–1066 (2002)
62. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**, 167–188 (1994)
63. Holmes, I.: Accelerated probabilistic inference of RNA struture evolution. BMC Bioinformatics **6**(73) (2005). doi:10.1186/1471-2105-6-73
64. Holmes, I., Rubin, G.M.: Pairwise RNA structure comparison with SCFGs. Pacif Symposium on Biocomputing, pp. 163–174 (2002)

65. Huttenhofer, A., Schattner, P., Polacek, N.: Non-coding RNAs: hope or hype. Trends Genet. **21**(5), 289–297 (2005)
66. James, B.D., Olsen, G.J., Pace, N.R.: Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol. **180**, 227–239 (1989)
67. Ji, Y., Xu, X., Stormo, G.D.: A graph theoretical approach for predicting common RNA secondary structure motifs including psudoknots in unaligned sequences. Bioinformatics **20**(10), 1591–1602 (2004)
68. Jiang, T., Lin, G., Ma, B., Zhang, K.: A general edit distance between RNA structures. J. Comput. Biol. **9**, 371–388 (2002)
69. Jiang, T., Wang, L., Zhang, K.: Alignment of trees—an alternative to tree edit. Theor. Comput. Sci. **143**, 137–148 (1995)
70. Juan, V., Wilson, C.: RNA secondary structure prediction based on free energy and phylogenetic analysis. J. Mol. Biol. **289**, 935–947 (1999)
71. Just, W.: Computational complexity of multiple sequence alignment with SP-score. J. Comput. Biol. **8**(6), 615–623 (2001)
72. Keenan, R.J., Freymann, D.M., Stroud, R.M., Walter, P.: The signal recognition particle. Annu. Rev. Biochem. **70**, 755–775 (2001)
73. Klein, R.J., Eddy, S.R.: RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics **4**(1), 44 (2003)
74. Klein, R.J., Misulovin, Z., Eddy, S.E.: Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc. Natl Acad. Sci. **99**(11), 7542–7547 (2002)
75. Knight, R., Birmingham, A., Yarus, M.: BayesFold: rational 2^o folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. RNA **10**, 1323–1336 (2004)
76. Knudsen, B., Hein, J.: RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics **15**(6), 446–454 (1999)
77. Knudsen, B., Hein, J.: Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. **31**(13), 3423–3428 (2003)
78. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D.: Hidden markov models in computational biology—applications to protein modeling. J. Mol. Biol. **235**, 1501–1531 (1994)
79. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T.: Identification of novel genes coding for small expressed RNAs. Science **294**, 853–858 (2001)
80. Lai, E.C., Tomancak, P., Williams, R.W., Rubin, G.M.: Computational identification of Drosophila microRNA genes. Genome Biol. **4**, R42.1–R42.20 (2003)
81. Laslett, D., Canback, B.: ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. **32**(1), 11–16 (2004)
82. Laslett, D., Canback, B., Andersson, S.: BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. Nucleic Acids Res. **30**(15), 3449–3453 (2002)
83. Le, S.V., Chen, J.H., Currey, K.M., Maizel, J.V.J.: A program for predicting significant RNA secondary structures. Comput. Appl. Biosci. **4**(1), 153–159 (1988)
84. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., Bartel, D.P.: The microRNAs of *Caenorhabditis elegans*. Genes Dev. **17**, 991–1008 (2003)
85. Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., Chen, R.: NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res. **33**, D112–D115 (2005)
86. Liu, J., Gough, J., Rost, B.: Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet. **2**(4), e29 (2006)
87. Lowe, T.M., Eddy, S.R.: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**(5), 955–964 (1997)
88. Lowe, T.M., Eddy, S.R.: A computational screen for methylation guide snoRNAs in yeast. Science **283**, 1168–1171 (1999)
89. Lowe, T.M.J.: Combining new computational and traditional experimental methods to identify tRNA and snoRNA gene families. Master's thesis, Washington University (1999)
90. Luck, R., Graf, S., Steger, G.: ConStruct: a tool for thermodynamic controlled prediction of conserved structure. Nucleic Acids Res. **27**(21), 4208–4217 (1999)
91. Lyngso, R.B., Pedersen, C.N.: RNA pseudoknot prediction in energy-based models. J. Comput. Biol. **7**, 409–427 (2000)
92. Mathews, D.H.: Predicting a set of minimal free energy RNA secondary structures common to two sequences. Bioinformatics **21**(10), 2246–2253 (2005)

93. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H.: Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl Acad. Sci. **101**(19), 7287–7292 (2004)
94. Mathews, D.H., Turner, D.H.: Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. **317**, 191–203 (2002)
95. Mattick, J.S., Makunin, I.V.: Non-coding RNA. Human Mol. Genet. **15**(1), 17–29 (2006)
96. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**, 1105–1119 (1990)
97. Meyer, I.M., Miklos, I.: Co-transcriptional folding is encoded within RNA genes. BMC Mol. Biol. **5**(10) (2004). doi:10.1186/1471-2199-5-10
98. Militello, K.T., Patel, V., Chessler, A.D., Fisher, J.K., Kasper, J.M., Gunasekera, A., Wirth, D.F.: RNA polymerase II synthesizes antisense RNA in *Plasmodium falciparum*. RNA **11**(4), 365–370 (2005)
99. Moulton, V.: Tracking down noncoding RNAs. Proc. Natl Acad. Sci. **102**(7), 2269–2270 (2005)
100. Nam, J.W., Kim, J., Kim, S.K., Zhang, B.T.: ProMIR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res. **34**, 455–458 (2006)
101. Notredame, C., Brien, E.A.O., Higgins, D.G.: RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res. **25**(22), 4570–4580 (1997)
102. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**(1), 205–217 (2000)
103. de Novoa, P.G., Williams, K.P.: The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. Nucleic Acids Res. **32**, D104–D108 (2004)
104. Nussinov, R., Pieczenik, G., Griggs, J.R., Kleitman, D.J.: Algorithms for loop matchings. SIAM J. Appl. Math. **35**(1), 68–82 (1978)
105. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., Ottonello, S.: Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. Nucleic Acids Res. **22**(7), 1247–1256 (1994)
106. Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P., Hein, J.: A comparative method for finding and folding RNA secondary structures withing protein-coding regions. Nucleic Acids Res. **32**(16), 4925–4936 (2004)
107. Perriquet, O., Touzet, H., Dauchet, M.: Finding the common structure shared by two homologous RNAs. Bioinformatics **19**(1), 108–116 (2003)
108. Piccinelli, P., Rosenblad, M.A., Samuelsson, T.: Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. Nucleic Acids Res. **33**(14), 4485–4495 (2005)
109. Pipas, J.M., McMahon, J.E.: Method for predicting RNA secondary structure. Proc. Natl Acad. Sci. **72**(6), 2017–2021 (1975)
110. Reeder, J., Giegerich, R.: Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics **5**(104) (2004). doi:10.1186/1471-2105-5-104
111. Reeder, J., Hochsmann, M., Rehmsmeier, M., Voss, B., Giegerich, R.: Beyond mfold: recent advances in RNA bioinformatics. J. Biotechnol. **124**(1), 41–55 (2006)
112. Regalia, M., Rosenblad, M.A., Samuelsson, T.: Prediction of signal recognition particle RNA genes. Nucleic Acids Res. **30**(15), 3368–3377 (2002)
113. Reis, E.M., Louro, R., Nakaya, H.I., Verjovski-Almeida, S.: As antisense RNA gets intronic. OMICS **9**(1), 2–12 (2005)
114. Reis, E.M., Nakaya, H.I., Louro, R., Canavez, F.C., Flatschart, A.V., Almeida, G.T., Egidio, C.M., Paquola, A.C., Machado, A.A., Festa, F., Yamamoto, D., Alvarenga, R., da Silva, C.C., Brito, G.C., Simon, S.D., Moreira-Filho, C.A., Leite, K.R., Camara-Lopes, L.H., Campos, F.S., Gimba, E., Vignal, G.M., El-Dorry, H., Sogayar, M.C., Barcinski, M.A., da Silva, A.M., Verjovski-Almeida, S.: Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene **23**(39), 6684–6692 (2004)
115. Ren, J., Rastegari, B., Condon, A., Hoos, H.: HotKnots: heuristic prediciton of RNA secondary structures including pseudoknots. RNA **11**, 1419–1504 (2005)
116. Rivas, E.: Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics **6**, 63 (2005)
117. Rivas, E., Eddy, S.R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. **285**, 2053–2068 (1999)

118. Rivas, E., Eddy, S.R.: Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics **16**(7), 583–605 (2000)
119. Rivas, E., Eddy, S.R.: Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics **2**(1), 8 (2001)
120. Rosenblad, M.A., Gorodkin, J., Knudsen, B., Zwieb, C., Samuelsson, T.: SRPDB: signal recognition particle database. Nucleic Acids Res. **31**(1), 363–364 (2003)
121. Ruan, J., Stormo, G.D., Zhang, W.: ILM: a web server for predicting RNA secondary structures with pseudoknots. Nucleic Acids Res. **32**(Web Server issue), W146–W149 (2004)
122. Ruan, J., Stormo, G.D., Zhang, W.: An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatic **20**(1), 58–66 (2004)
123. Sakakibara, Y., Brown, M.: The application of stochastic context-free grammars to folding, aligning and modeling homologous RNA sequences (1993). Techn. Rep. UCSC-CRL-94-14
124. Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C., Haussler, D.: Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res. **22**(23), 5112–5120 (1994)
125. Sankoff, D.: Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. **45**(5), 810–825 (1985)
126. Schattner, P.: Searching for RNA genes using base-composition statistics. Nucleic Acids Res. **30**(9), 2076–2082 (2002)
127. Schattner, P., Brooks, A.N., Lowe, T.M.: The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33**, 686–689 (2005)
128. Schattner, P., Decatur, W.A., Davis, C.A., Ares, M.J., Fournier, M.J., Lowe, T.M.: Genome-wide searching for pseudouridylation guide snoRNAs: analysis of *Saccharomyces cerevisiae* genome. Nucleic Acids Res. **32**(14), 4281–4296 (2004)
129. Schmitz, M., Steger, G.: Description of RNA folding by simulated annealing. J. Mol. Biol. **255**, 254–266 (1996)
130. Sczyrba, A., Kruger, J., Mersch, H., Kurtz, S., Giegerich, R.: RNA-related tools on the Bielefeld bioinformatics server. Nucleic Acids Res. **31**(13), 3767–3770 (2003)
131. Searls, D.B.: The language of genes. Nature **420**, 211–217 (2002)
132. Shapiro, B.A., Zhang, K.: Comparing multiple RNA secondary structures using tree comparisons. Comput. Appl. Biosci. **6**(4), 309–318 (1990)
133. Siebert, S., Backofen, R.: MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics **21**(16), 3352–3359 (2005)
134. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., Giegerich, R.: RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics **22**(4), 500–503 (2006)
135. Storz, G.: An expandind universe of noncoding RNAs. Science **296**, 1260–1263 (2002)
136. Tabaska, J.E., Cary, R.B., Gabow, H.N., Stormo, G.D.: An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics **14**(8), 691–699 (1998)
137. Taneda, A.: Cofolga: a genetic algotithm for finding the common folding of two RNAs. Comput. Biol. Chem. **29**, 111–119 (2005)
138. Tinoco, I.J., Uhlenbeck, O.C., Levine, M.D.: Estimation of secondary structure in ribonucleic acids. Nature **230**(5293), 362–367 (1971)
139. Touzet, H., Perriquet, O.: CARNAC: folding families of related RNAs. Nucleic Acids Res. **32**, W142–W145 (2004)
140. Tsui, V., Macke, T., Case, D.A.: A novel method for finding tRNA genes. RNA **9**, 507–517 (2003)
141. Turner, D.H., Sugimoto, N.: RNA structure prediction. Annu. Rev. Biophys. Biophys. Chem. **17**, 167–192 (1988)
142. Underwood, R.C.: Stochastic Context-Free Grammars for Modeling Three Spliceosomal Small Nuclear Ribonucleic Acids. Master's thesis, Baskin Center for Computer Engineering and Information Sciences, University of California (1994)
143. Uzilov, A.V., Keegan, J.M., Mathews, D.H.: Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics **7** (2006). doi:10.1186/1417-2105-7-173
144. Voss, B., Giegerich, R., Rehmsmeier, M.: Complete probabilistic analysis of RNA shapes. BMC Biology **4**(5) (2006). doi:10.1186/1741-7007-4-5
145. Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R.: PSoL: a positive sample only learning algorithm for finding ncRNA genes. Bioinformatics **22**(21), 2590–2596 (2006)
146. Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., Li, Y.: MicroRNA identification based on sequence and structure alignment. Bioinformatics **21**(18), 3610–3614 (2005)

147. Washiet, S., Hofacker, I.L.: Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J. Mol. Biol. **342**, 19–30 (2004)
148. Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding RNAs. Proc. Natl Acad. Sci. **102**(7), 2454–2459 (2005)
149. Waterman, M.S., Smith, T.F.: RNA secondary structure: a complete mathematical analysis. Math. Biosci. **42**, 257–266 (1978)
150. Weinberg, Z., Ruzzo, W.L.: Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. Bioinformatics **20**(suppl 1), i334–i341 (2004)
151. Weinberg, Z., Ruzzo, W.L.: Sequence-based heuristics for faster annotation of non-coding RNA families. Bioinformatics **22**(1), 35–39 (2006)
152. Workman, C., Krogh, A.: No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res. **27**(24), 4816–4822 (1999)
153. Wuchty, S., Fontana, W., Hofacker, I.L., Schuster, P.: Complete suboptimal folding of RNA and the stability of secondary structure. Biopolymers **49**, 145–165 (1999)
154. Yang, J.H., Zhang, X.C., Huang, Z.P., Zhou, H., Huang, M.B., Zhang, S., Chen, Y.Q., Qu, L.H.: snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res (2006). doi:10.1093/nar/gkl672
155. Yang, Z., Zhu, Q., Luo, K., Zhou, Q.: The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature **414**, 317–322 (2001)
156. Ying, X., Luo, H., Luo, J., Li, W.: RDfolder: a web server for prediction of RNA secondary structure. Nucleic Acids Res. **32**(Web Server issue), W150–W153 (2004)
157. Zuker, M.: On finding all suboptimal foldings of an RNA molecule. Science **244**, 48–52 (1989)
158. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**(13), 3406–3415 (2003)
159. Zuker, M., Mathews, D.H., Turner, D.H.: Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski, J., Clark, B.F.C., (eds.) RNA biochemistry and biotechnology. NATO ASI Series, Kluwer (1999)
160. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**(1), 133–148 (1981)