RNA Secondary Structures

Ivo L. Hofacker[1] *and Peter F. Stadler*[1,2]
[1] *University of Vienna, Wien, Austria*
[2] *University of Leipzig, Leipzig, Germany*

1 Introduction 583

2 Basic Concepts 585
2.1 Representations of Secondary Structures 585
2.2 Combinatorics and Typical Structures 585
2.3 Energy Model 587
2.4 Pseudoknotted Structure 588

3 RNA Structure Prediction 589
3.1 Structure Prediction by Energy Minimization 589
3.2 Suboptimal Folding and Pair Probabilities 590
3.2.1 Available Programs and Web Services 591
3.2.2 Prediction Accuracy 591
3.3 Well-defined Regions and Reliability 591
3.4 Structure Prediction using Sequence Covariation 592
3.4.1 Search for ncRNA 594

4 Current Research in RNA Secondary Structures 595

Bibliography 597
Books and Reviews 597
Primary Literature 598

Encyclopedia of Molecular Cell Biology and Molecular Medicine, 2nd Edition. Volume 12
Edited by Robert A. Meyers.

ISBN: 3-527-30649-8

Keywords

Base Pair
Base pairs are formed by specific hydrogen bonds between two nucleobases. The canonical Watson–Crick pairs between guanosine and cytosine (GC), and between adenosine and uracil (AU), as well as the wobble pairs between guanosine and uracil (GU) regularly appear in RNA structures. The nucleobases of the two paired nucleotides form a planar structure.

Dynamic Programming
A general computational technique that is based on solving a combinatorial optimization problem recursively, while tabulating intermediate results. In the case of RNA folding, this means composing large structures from the foldings of shorter subsequences.

Helix or Stem
Helices or stems are formed by stacking base pairs on top of each other. The interaction of the parallel planar arrangements of the nucleobases stabilizes the three-dimensional structure. The geometry of helices is nearly sequence independent, as long as they contain only canonical base pairs, whence energy contributions of stacked pairs are nearly independent of each other.

Loops
Loops are formed by those parts of the molecule that are not part of a helix. Different types of loops (bulges, interior loops, and multibranch loops) can be distinguished depending on their structure. Loops destabilize the structure. In the context of RNA folding algorithms, one considers adjacent stacked pairs as a special type of (stabilizing) loop. In this language, each secondary structure can be decomposed completely into loops, which form the basis of the empirical energy model.

Noncoding RNA
An RNA molecule with any function other than encoding proteins.

Pseudoknot
A structural feature composed of crossing base pairs. Pseudoknots are in most cases excluded from computational approaches for technical reasons.

RNomics
The emerging science of RNA molecules and their interactions in the cell.

Secondary Structure
The secondary structure is the collection of *base pairs* or, equivalently, of *loops* and *helices*. In a stricter sense, secondary structures are defined as crossing-free collections of base pairs. All other base pairs are then considered as part of a higher-order (tertiary) structure.

The three-dimensional structures of nucleic acids, RNA, and DNA, are dominated by double-helical regions that are formed by canonical Watson–Crick and wobble (GU) base pairs, which collectively are referred to as the *secondary structure of the molecule*. In the case of RNA, this level of description is of particular interest because it captures the thermodynamics of RNA folding quite well and allows a self-consistent description of folding dynamics. Furthermore, secondary structures are often evolutionarily well conserved, evolving much slower than the underlying sequences. From a bioinformatics perspective, RNA secondary structures are a convenient representation because efficient exact algorithms are known to enumerate the structures that can be formed by a given sequence, to solve the folding problem, and to compute the partition function (and hence any thermodynamic quantity of interest) from a well-measured set of empirical energy parameters.

1 Introduction

Secondary structures, that is, patterns of base pairing, not only form the most important distinguishing characteristics of the various classes of RNA molecules but they also provide us with a unique coarse-grained level of description that naturally lends itself to computational studies of RNA structure and evolution.

RNA has moved from a fringe existence to the center stage of research in molecular biology following the discovery of novel classes of small regulatory RNAs. While textbooks still speak of *genes and their encoded protein products*, thousands of human genes produce transcripts that exert their function without ever producing proteins. The list of functional noncoding RNAs (ncRNAs) includes key players in the biochemistry of the cell. Many of them have characteristic secondary structures that are highly conserved in evolution. Databases (referenced in the Table 1) collect the most important classes.

In addition, there is a diverse and growing list of ncRNAs with sometimes enigmatic function. We give just a few examples: The 17-kb *Xist* RNA of humans and the smaller roX RNAs of Drosophila play a key role in dosage compensation and X-chromosome inactivation. Several large ncRNAs are expressed from imprinted regions. Many of these are cis-antisense RNAs that overlap coding genes on the other genomic strand. An RNA (meiRNA) regulates the onset of meiosis in fission yeast. Human vaults are cytoplasmic ribonucleoprotein particles believed to be involved in multidrug resistance. The complex contains several small untranslated RNA molecules. No precise function is known at present for the human H19 transcript, the hrsω transcript induced by heat shock in Drosophila, or the *Escherichia coli* 6S RNA.

Even though the sequence of the human DNA is known by now, the contents of about half of it remains unknown. The diversity of sequences, sizes, structures, and functions of the known ncRNAs strongly suggests that we have seen only a small fraction of the functional RNAs. Most of the ncRNAs are small, they do not have translated ORFs, and they are not polyadenylated. Unlike protein-coding genes, ncRNA gene sequences do

Tab. 1 Major classes of functional RNAs

Type	*Size*	*Organisms*	*Function* –
miRNA	22	Multicellular?	RNA interference
tRNA	70–80	All	Transfer RNA
rRNA 5S	130	All	Part of ribosome
RNase P	260–360	All	tRNA maturation
tmRNA	300–400	Bacteria	Tags protein for proteolysis
snRNA	50–300	Eukaryotes	Part of spliceosome
snoRNA	80–600	Eukaryotes Archea	RNA modification
SRP RNA	300–400	All	Signal recognition particle
rRNA 16S	1500	All	Part of ribosome
rRNA 23S	3300	All	Part of ribosome

not seem to exhibit a strong common statistical signal; hence, a reliable general-purpose computational genefinder for ncRNA genes has been elusive. It is quite likely therefore that a large class of genes has gone relatively undetected so far because they do not encode proteins.

Another level of RNA function is presented by functional motifs within protein-coding RNAs. We list a few of the best-understood examples of structurally conserved RNA motifs in viral RNAs:

- An *IRES* (internal ribosomal entry site) region is used instead of a CAP to initialize translation by Picornaviridae, some Flaviviridae including Hepatitis C virus, and a small number of mRNAs.
- The TAR hairpin structure in HIV and related retroviruses is the target for viral transactivation.
- The RRE structure of retroviruses serves as a binding site for the Rev protein and is essential for the viral replication. The RRE is a characteristic five-fingered structural motif.
- The CRE hairpin in Picornaviridae is vital for replication.

A textbook example of a functional RNA secondary structure is the *Rho*-independent termination in *E. coli*. The newly synthesized mRNA forms a hairpin in the 3′UTR that interacts with the RNA polymerase causing a change in conformation and the subsequent dissociation of the enzyme–DNA–RNA complex.

Only part of the mature mRNA is translated into protein. At the 5′ terminus of the mRNA, just behind the cap, is a noncoding sequence, the so-called leader sequence (10–200 nt) that may be followed by another noncoding sequence of up to 600 nt. An increasing number of functional features in the untranslated regions of eukaryotic mRNA have been reported in recent years.

An extreme example is the Early Noduline gene. Enod40, which is expressed in the nodule primordium developing in the root cortex of leguminous plants after infection by symbiotic bacteria, codes for an RNA of about 700 nt that gives rise to two short peptides of 13 and 27 amino acids, respectively. The RNA structure itself exhibits significant conservation of secondary structure motifs, and might take part in localization of mRNA translation,

as in the case of the bicoid gene *bcd* of Drosophila.

2 Basic Concepts

2.1 Representations of Secondary Structures

A *secondary structure* Ψ is a special type of contact structure, represented by a list of base pairs (i, j) with $i < j$ on a sequence x, such that for any two base pairs $[i, j]$ and $[k, l]$ with $i \leq k$ holds:

1. $i = k$ if and only if $j = l$, and
2. $k < j$ implies $i < k < l < j$.

The first condition simply means that each nucleotide can take part in at most one base pair. The second condition forbids knots and pseudoknots. While pseudoknots are important in many natural RNAs, they can be considered part of the tertiary structure for our purposes. We will therefore neglect them for the most part of this presentation. The restriction to knot-free structures is necessary for the efficient dynamic programming algorithms discussed below.

The two conditions above imply that secondary structures form a special type of graph. In particular, a secondary structure graph is *subcubic* (i.e. the vertex degree is at most three) and *outer-planar*. The latter property means that the structure can be drawn in the plane in such a way that all vertices (which represent the nucleotides) are arranged on a circle (the molecule's backbone), and all edges (which represent the bases pairs) lie inside the circle and do not intersect, see Fig. 1.

Secondary structures can also be stored compactly in strings consisting of dots and matching brackets: For any pair between positions i and j ($i < j$) we place an open bracket "(" at position i and a closed bracket ")" at j, while unpaired positions in the molecule are represented by a dot ("."), see Fig. 1c. Since base pairs may not cross, the representation is unambiguous.

Although secondary structure graphs are always planar, it is not trivial to find nonoverlapping and visually pleasing layouts for large structures. A representation that works well for large structures and is well suited for comparing structures is the so-called mountain representation. In the mountain representation, a single secondary structure is represented in a two-dimensional graph, in which the x-coordinate is the position k of a nucleotide in the sequence and the y-coordinate is the number $m(k)$ of base pairs that enclose nucleotide k.

Often it is desirable to present not just one structure, but an ensemble of structural possibilities. Mountain plots can be adapted for structural ensembles by using the *mean* number of base pairs $m(k)$. A more detailed representation is given by so-called dot plots, where each possible pair (i, j) is represented by a dot on square grid. The size or color of the dot is used to indicate the probability p_{ij} of the pair, or the best possible energy of a structure containing this pair. Dot plots can provide an excellent overview of possible alternative foldings. Fig. 4 gives examples of dot plots and mountain plots.

2.2 Combinatorics and Typical Structures

The basic studies into the combinatorics of RNA secondary structures goes back to the work of M. Waterman in the late 1970s.

We begin our exposition by counting the secondary structures that can be formed by a given sequence $x = (x_1, x_2, \ldots, x_n)$

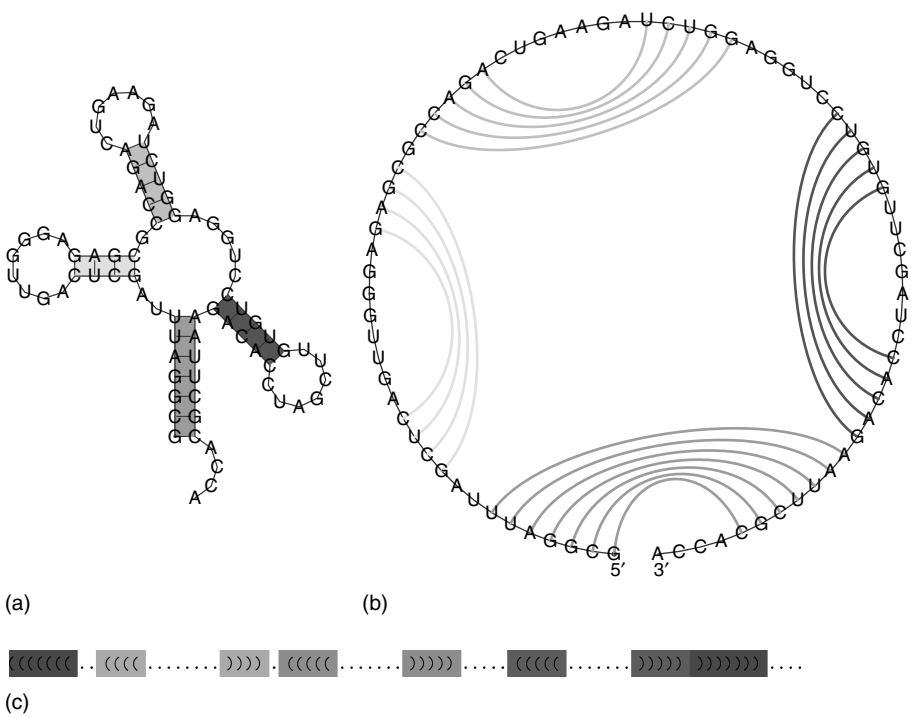
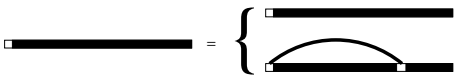


Fig. 1 Secondary structure of phenylalanine-tRNA from yeast as (a) conventional drawing, (b) in circular representation, and (c) in dot-bracket notation. Colors are used to highlight pairs belonging to the same helix. Note that the chords in the circular representation must not cross in secondary structure graphs. (See color plate p. xl).

of length n. We will simply write "(i, j) pairs" to mean that the nucleotides x_i and x_j *can* form a Watson–Crick or a wobble pair, that is, $x_i x_j$ is one of **GC**, **CG**, **AU**, **UA**, **GU**, or **UG**. The basic idea behind all dynamic programming algorithms for RNA folding is the observation that a structure on n nucleotides can be formed in only two distinct ways from shorter structures: Either a structure on $n-1$ nucleotides is extended by an unpaired base, or the nth nucleotide is paired. In the latter case, it has a pairing partner, say j such that the (j, n) pair encloses a secondary structure on the subsequence from $j+1$ to $n-1$ since base pairs must not cross by definition. The remainder, the interval from 1 to $j-1$ is of course also a secondary structure:



It is now easy to compute the number N_{ij} of secondary structures on the subsequence $x[i..j]$ from positions i to j:

$$N_{ij} = N_{i+1,j} + \sum_{k,(i,k)\text{pairs}} N_{i+1,k-1} N_{k+1,j} \tag{1}$$

The first term accounts for the case in which position i is unpaired, the

terms in the sum consider the base pairs from i to some position k. Because of the "no-pseudoknots" condition, both the part of the sequence that is enclosed by the pair (i, k) and the part beyond the base pair form secondary structures that are completely independent of each other: thus, we may simply multiply their numbers.

The counting recursion can also be used to produce uniformly distributed random structures, or adapted to derive typical structural features of RNA molecules such as expected helix length or distribution of loop types. One should note, however, that such random structures differ significantly from the typical structures obtained from folding random or natural sequences. In contrast to folded structures, random structures contain many loops connected by short helices or even isolated base pairs. Qualitative features, such as an average length of helical regions independent of the sequence lengths or the distribution of branching orders of loops is shared between random and biological structures.

2.3 Energy Model

The physicochemical basis for the coarse-grained secondary structure model is the fact that a dominating part of the energy of structure formation is contained in the stacking of the aromatic nucleobases that gives rise to helical base-paired regions with a spatial structure that can be considered as almost sequence independent. This distinction between paired and unpaired regions allows us to approximate the free energy of structure formation given the sequence and the list of base pairs. It is important to note that a secondary structure as defined in the previous section corresponds to an *ensemble* of conformations of the molecule at atomic resolution restricted to a certain base pairing (hydrogen bonding) pattern. For example, no information is assumed about the spatial conformation of unpaired regions. The entropic contributions of these restricted conformations have to be taken into account, and hence we are dealing with (temperature dependent) free energies.

This free energy of an RNA secondary structure is assumed to be the sum of the energy contributions of all "loops," that is, the faces of the planar drawing of the structure. This decomposition has a solid graph theoretical foundation: the loops form the unique minimal cycle basis of the secondary structure graph. More importantly, however, a large number of careful melting experiments have shown that the energy of structure formation (relative to the random coil state) is indeed additive to a good approximation. Usually, only Watson–Crick (**AU**, **UA**, **CG**, and **GC**) and wobble pairs (**GU**, **UG**) are allowed in computational approaches since nonstandard base pairs have, in general, context-dependent energy contributions that do not fit into the "nearest-neighbor model". Individual nonstandard base pairs are therefore treated as special types of interior loops in the most recent parameter sets.

Qualitatively, there are two major energy contributions: stacking of base pairs and loop entropies. Stacking energies can be computed for molecules in the vacuum by means of standard quantum chemistry approaches. The secondary structure model, however, considers only energy differences between folded and unfolded states in an aqueous solution with rather high salt concentrations. As a consequence, one has to rely on empirical energy parameters. Loops are destabilizing: the closing base pair restricts the possible conformations

of the sequence in the loop relative to the conformations that could be formed by the same sequence segments in a random coil, resulting in an entropy loss and thus an increase in free energy.

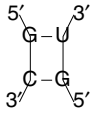
A collection of energy parameters is maintained by the group of David Turner. These standard parameters are measured in a buffer of 1 M NaCl at 37 °C. Since both entropies and enthalpies are available, the parameters can be extrapolated to other temperatures. As examples we list the free energies for stacked pairs in Table 2. Note that a single additional base pair can stabilize a structure by up to −3.4 kcal mol^{-1}.

2.4 Pseudoknotted Structure

The definition of secondary structure relegates pseudoknots to the realm of a tertiary structure. Often this is justified purely by convenience, since the simple dynamic programming algorithms, which will be described in some detail in the following section, cannot handle pseudoknotted structures. It is worth mentioning, however, that many pseudoknotted base pairing patterns are not sterically feasible, while any knot-free secondary structure can be realized in 3D.

Tab. 2 Free energies for stacked pairs in kcal mol^{-1}.

	CG	GC	GU	UG	AU	UA
CG	–2.4	–3.3	–2.1	–1.4	–2.1	–2.1
GC	–3.3	–3.4	–2.5	–1.5	–2.2	–2.4
GU	–2.1	–2.5	–1.3	–0.5	–1.4	–1.3
UG	–1.4	–1.5	–0.5	0.3	–0.6	–1.0
AU	–2.1	–2.2	–1.4	–0.6	–1.1	–0.9
UA	–2.1	–2.4	–1.3	–1.0	–0.9	–1.3



–2.5
kcal mol^{-1}

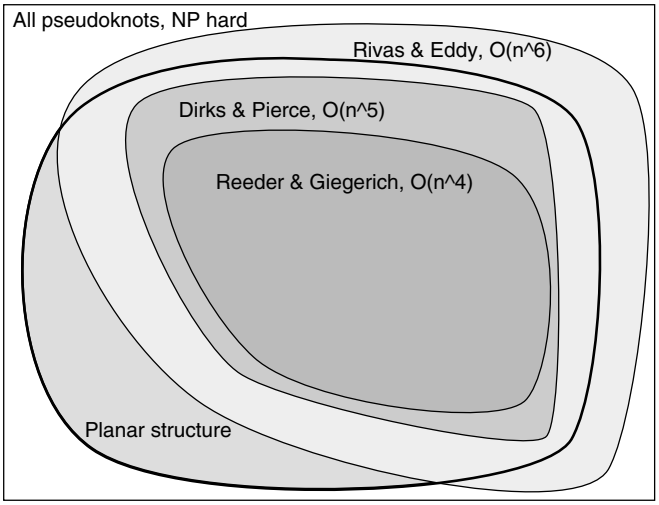


Fig. 2 Classes of allowed pseudoknots in different algorithms.

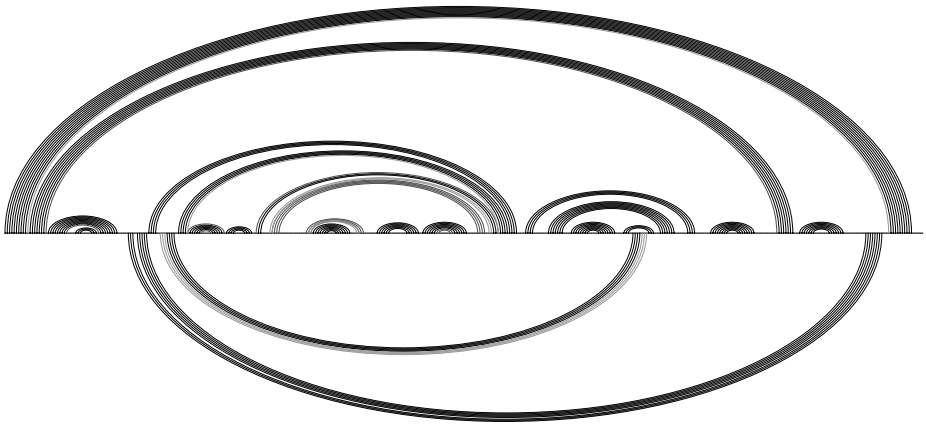


Fig. 3 Bisecondary structure of RNase P RNA as predicted by HXMATCH compared to the reference structure. Black: correct predicted base pairs; green: base pairs not predicted; red: predicted pairs not present in reference structure (See color plate p. xxxix).

Comparative sequence analysis revealed conserved pseudoknots in many important classes of RNA, such as rRNAs, RNase P RNAs, and tmRNA. Consequently, there is significant interest in computational methods for their prediction. While thermodynamic structure prediction with a loop-based energy model is NP-complete, in general, restricted classes of pseudoknots can be dealt with by polynomial algorithms (Fig. 2). Fortunately, most known pseudoknots are relatively simple. With few exceptions, they can be drawn as a superposition of just two knot-free secondary structures, that is, as a bisecondary structure. Recently a number of polynomial time algorithms have been proposed that predict pseudoknotted structures. Depending on how restrictive the set of allowed pseudoknots is chosen, their time complexity ranges between $\mathcal{O}(n^6)$ and $\mathcal{O}(n^4)$.

The practical applicability of these algorithms, however, is limited not only by their time and memory requirements, but most severely by our poor knowledge of pseudoknot energetics. Useful energy parameters have so far only been collected for simple H-type pseudoknots. In the absence of good scoring functions, expanding the set of possible structures will often lead to predictions that are worse than knot-free predictions.

The problem can be somewhat mitigated in methods that work with multiple sequence and exploit the covariance information in an alignment, such as ILM and HXMATCH. An example is shown in Fig. 3.

3 RNA Structure Prediction

3.1 Structure Prediction by Energy Minimization

Given an energy model the simplest approach to structure prediction is to determine the optimal structure with respect to its free energy, that is, the minimum free energy (MFE) structure. For pseudoknot free RNA secondary structures, the optimal structure can

be computed efficiently using dynamic programming.

If the energy of a structure could be modeled as a sum of contributions from base pairs, an optimal structure could be computed from a simple variant of the counting recursion 1,

$$E_{ij} = \min\{E_{i+1,j} + \min_{k,\,(i,k)\ \text{pairs}} \times \{E_{i+1,k-1} + E_{k+1,j} + \beta_{ik}\}\}, \quad (2)$$

where β_{ik} is the energy contribution of a pair between positions i and k. The algorithm requires $\mathcal{O}(n^2)$ memory for storing the E matrix while cpu timescales as $\mathcal{O}(n^3)$. While this was the first approach to structure prediction used by Nussinov, a loop-based energy model as described above is necessary to obtain reasonable accuracies. Using a loop-based energy model requires somewhat more involved recursions, but still exhibits the same time and space complexity.

As is typical for dynamic programming, the algorithm first computes the best possible energy of a secondary structure, then ''backtracks'' to find a structure with this optimal energy. In the simplest case, this procedure returns a single solution, the MFE structure. This is unsatisfactory not only because inaccuracies in the energy parameters will lead to errors in the predicted structure but also because significantly different structures may be needed to represent the molecule in thermodynamic equilibrium.

3.2 Suboptimal Folding and Pair Probabilities

The most common strategy for generating additional suboptimal structures is the algorithm Michael Zuker, which considers for each possible base pair the best structure containing that pair. The number of structures in the output is further reduced by considering only structures within some energy interval of the mfe and filtering out structures that are too similar to others. The method usually returns a short list of possible foldings that form a representative sample. Occasionally, however, important alternatives will be missed.

A more rigorous approach is the computation of the partition function and base pairing probabilities using McCaskill's algorithm. For every possible base pair (i, j), the algorithm yields the probability p_{ij} that the base pair will be formed, that is, the sum of the probabilities of all structures containing that pair. The partition function can also be used to calculate heat capacities and thus characterize melting transitions.

Base pair probabilities can be nicely represented in dot plots, where we plot a square with area p_{ij} for each pair (i, j). Similarly, Zuker's suboptimal folding algorithm can produce energy dot plots, where instead of the probability, the best possible energy in structures containing (i, j) is plotted.

The *complete* suboptimal folding algorithm of Wuchty et al. can generate *all* suboptimal structures in a predefined energy range above the mfe. For small molecules, it can be illuminating to look at the exhaustive list of structural possibilities. For larger molecules, the information quickly becomes overwhelming. With further postprocessing, however, these data can be used for detailed analysis of RNA energy landscapes and the dynamics of the folding process.

Finally, the partition function algorithm can be enhanced with a stochastic backtracking procedure, which samples suboptimal structures from the Boltzmann distribution. This can be used to compute the average value of any structural feature

simply by sampling a sufficiently large number of suboptimal structures.

3.2.1 Available Programs and Web Services

Zuker's suboptimal folding algorithm is implemented in his popular `mfold` program for UNIX, as well as in David Mathew's `RNAstructure` for Windows. Partition function folding, complete suboptimal folding, and stochastic backtracking are all available as part of the `Vienna RNA Package`.

Users who need to do structure predictions only occasionally will find several of the above algorithms as a Web service on Michael Zuker's mfold server at http://bioinfo.math.rpi.edu/~zuker/rna/, as well as the Vienna RNA server http://rna.tbi.univie.ac.at.

3.2.2 Prediction Accuracy

Prediction accuracies can be evaluated by comparison to structures inferred using phylogenetic methods (see the following). For short RNAs such as tRNAs, one can expect accuracies (in terms of known base pairs that are correctly predicted) of around 70%. As pointed out recently by Doshi et al., long-range pairs are generally predicted poorly, resulting in a mean accuracy of only 41% for the large 16S and 23S rRNAs. While predicted mfe structure may have less then 20% correct pairs in unfortunate cases, good structures are still found in the vicinity of the mfe structure by suboptimal folding. Moreover, even when overall prediction accuracy is low, pairs that are predicted with high probability are usually correct.

Often a small number of constraints can dramatically improve prediction accuracy. Most folding programs allow constraints, such as specifying positions as (un)paired, to be specified. Such constraints can be obtained without too much effort from chemical probing experiments.

3.3 Well-defined Regions and Reliability

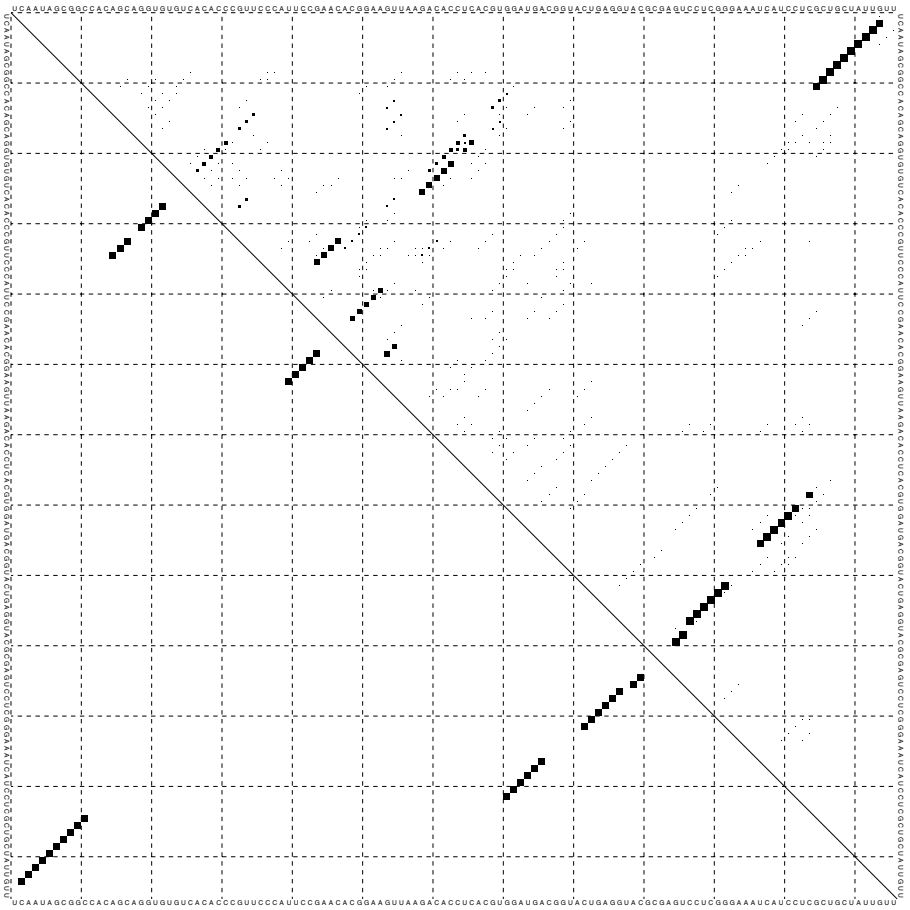
Pair probability and energy dot plots can also give a good visual impression of the quality of prediction and well-defined regions. A dot plot cluttered by many alternatives may not only indicate structural flexibility but may also make the prediction of a single mfe structure less reliable. Well-defined structures are likely to be correctly predicted, since they will be robust with respect to small variations of the energy parameters.

Several quantitative measures of well definedness are being used. In the simplest case, the well definedness of the prediction can be quantified by the difference between the MFE and the free energy of the best suboptimal structure. A more robust measure is the difference between the mfe and the ensemble free energy $G = -RT\ln(Q)$ where Q is the partition function. The latter is equivalent to the probability of the mfe structure in the ensemble given by Boltzmann's law $p(mfe) = \exp(-E_{\min}/RT)/Q = \exp(-(E_{\min} - G)/RT)$.

Even more useful are position-wise measures that help identify credible parts of the prediction. One can, for example, compute from the pair probabilities p_{ik} the positional entropy

$$S_k = -\sum_i p_{ik} \ln p_{ik} \qquad (3)$$

where p_{ii} is defined as the probability that i does not pair $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. A useful application of this measures is the annotation of structure drawings (Fig. 4).



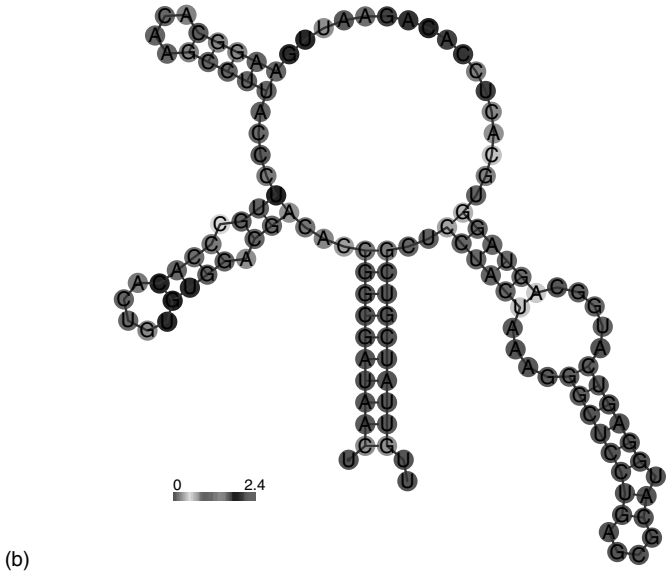
(a)

Fig. 4 Predicted structure for an 5S rRNA. (a) base pair probabilities as computed by `RNAfold -p -noLP`; (b) predicted mfe structure annotated with positional entropy; (c) Mountain plot. Note that all regions with low entropy (red) are correctly predicted, while high entropy regions (blue) deviate from the reference structure (See color plate p. xli).

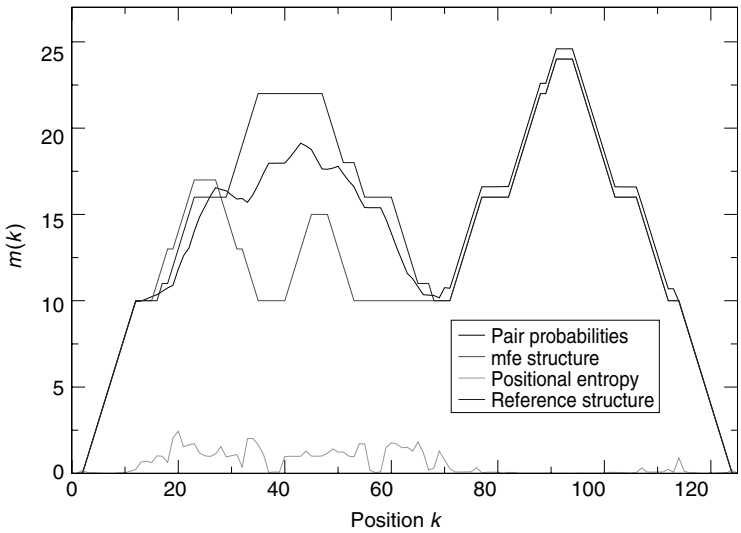
3.4 Structure Prediction using Sequence Covariation

If several sequences are known to fold into (almost) the same structure, their common structure can be inferred from sequence covariation, typically measured as mutual information between two columns of a multiple sequence alignment. The approach requires a large number of related sequences and has therefore been mostly limited to ribosomal RNAs and tRNAs. However, where applicable, these phylogenetic methods produce secondary structure models of high quality, and can even elucidate some tertiary interactions.

Recently, a number of methods have appeared that combine thermodynamic prediction with covariation analysis, in order to achieve accurate predictions with only a few related sequences. Usually, these



(b)



(c)

Fig. 4 (*Continued*)

methods start from a conventional multiple sequence alignment. One approach, as exemplified by the `alidot` and `ConStruct` tools, is to predict structures of individual sequences, combine them using the sequence alignment, and then find significantly conserved structure motifs. This is particularly useful when searching for locally conserved structure motifs in larger sequences.

Another approach that is suitable for predicting globally conserved structures

works by modifying the folding algorithm itself so that an optimal structure is predicted for a sequence profile, or a multiple sequence alignment, instead of a single sequence. The two best-known implementations of this approach are `pfold` and `RNAalifold`. `pfold` is based on a stochastic context-free grammar, and thus uses parameters derived from a training set; it also makes explicit use of a predicted phylogenetic tree. `RNAalifold`, on the other hand, uses the standard energy model for RNA secondary structures, augmented with a covariation term that rewards consistent and compensatory mutations. Thus, for identical sequences, it gives the same result as the single sequence prediction from `RNAfold`. With a few (or even just two) related sequences, these programs achieve prediction accuracies much higher than prediction methods for single sequences.

Methods starting from a multiple sequence alignment are of course limited by the accuracy of the input alignment. Typically, this becomes a problem when pairwise sequence identities drop below 60%. An alternative is to start by aligning predicted mfe structures using `RNAforester` or `MARNA`. The most rigorous way is to use a variant of the Sankoff algorithm, which computes the alignment and consensus structure simultaneously. Notable implementations are `foldalign`, `dynalign`, `pmcomp/pmmulti`, and `dart`. The Sankoff algorithm is, however, computationally very expensive ($\mathcal{O}(n^6)$ in the unrestricted case). The above algorithms therefore use various restrictions to improve speed.

3.4.1 Search for ncRNA

With the realization that a significant fraction of the genomic DNA of higher organisms might code for functional RNAs rather than protein-coding genes, the development of techniques for the detection and classification of RNAs in large sequences has become an active field of research. Generally, we can distinguish between experimental and computational methods to search for ncRNA. Experimental methods, as ligation of 5′ and 3′ adapters to size fractionated isolated RNA, followed by reverse transcription, cDNA cloning and sequencing, or simply using Northern blot analysis for sequence-specific detection of ncRNAs are at present restricted to the detection of abundant ncRNA. In this chapter, we will therefore concentrate on computational methods.

Some ncRNAs can be found by searching for likely transcripts that do not contain an open reading frame. A survey of the *E. coli* genome for DNA regions that contain a $\sigma 70$ promotor within a short distance of a *Rho*-independent terminator, for instance, resulted in 144 novel possible ncRNAs. This approach is limited, however, to functional RNAs that are transcribed in the ''usual'' manner by means of pol-II. For many ncRNAs, however, the mode of transcription is unknown.

Neural networks or support vector machines have also been successfully trained to recognize all known ncRNA in the *E. coli* genome. The survey conducted resulted in 370 new candidates, which have not been verified experimentally. The number seems to be quite reasonable, in particular when we keep in mind that false positives are a common problem with most computational searches for ncRNA so far.

Some structure-based searches use the known secondary structure of the major classes of functional RNAs. Programs such as `RNAmot`, `tRNAscan`, `HyPa`, `RNAMotif`, `bruce`, and many others exploit this

avenue. An interesting variant that makes use of evolutionary computation is described by Fogel. Nevertheless, all these approaches are restricted to searching for new members of the few well-established families. Especially for miRNAs, being quite short with a characteristic stem-loop structure ending up in a hairpin, several genome-wide surveys have been undertaken. The initially very large number of candidates is filtered in a postprocessing process according to other structural and sequential elements.

Comparative approaches such as the program `QRNA` can detect novel structural RNA genes in a pair of aligned homologous sequences by deciding whether the substitution pattern fits better with (1) synonymous substitutions, which are expected in protein-coding regions; (2) the compensatory mutations consistent with some base-paired secondary structure; or (3) uncorrelated mutations. Genome-wide scans for new ncRNA have already been accomplished with `QRNA`.

Another approach tries to determine functional RNAs by means of structure prediction. The basic assumption is that functional and hence conserved structures will be thermodynamically more stable. While such procedures are capable of detecting some particularly stable features, a recent study concludes that ''although a distinct, stable secondary structure is undoubtedly important in most noncoding RNAs, the stability of most noncoding RNA secondary structures is not sufficiently different from the predicted stability of a random sequence to be useful as a general genefinding approach.'' Nevertheless, in some special cases such as hyperthermophilic organisms, **GC**-content (and hence thermodynamic stability) proved sufficient.

When thermodynamic stability is combined with comparative methods, structural ncRNAs can indeed be recognized with high specificity and sensitivity. Two such approaches were developed recently building on consensus structure predictions using `RNAalifold`. The `alifoldz` method compares the folding energy of the native alignment with those of randomized alignments, using a somewhat costly shuffling procedure. The even faster `RNAz` method combines a measure for RNA secondary structure conservation on the basis of comparing the consensus folding energies to folding energies of individual sequences, with a measure for thermodynamic stability, which is normalized with respect to both sequence length and base composition. This method is fast and reliable enough to scan even large vertebrate genomes.

4 Current Research in RNA Secondary Structures

Despite the recent progress in detection algorithms for ncRNAs in comparative genomics approaches, a reliable *annotation* of functional RNAs in genomic data is still a largely unsolved problem for a variety of reasons. Most importantly, annotation is almost always limited to comparative methods; at present there are no generally applicable methods that could be used to determine the function of an RNA from its sequence and/or (predicted) structure. So far, structure and function of only a very small fraction of ncRNAs and other functional RNA motifs have been characterized. In fact, the functions of a number of very well conserved, evolutionary ancient ncRNAs such as Y

RNAs or vault RNAs have remained in the dark.

Even the recognition of members of the large classes of known ncRNAs is less straightforward than one might expect: the sequences of snoRNAs, snRNAs, and many others evolve relatively rapidly, their secondary structures are thus the main features that can be used to identify them. Note that it is much easier to identify additional members of a family of homologous snoRNA and microRNAs than to recognize a new class of homologous ncRNAs, for which no member has been described experimentally, as a novel family of microRNAs or snoRNAs. For the first class of problems, simple sequence comparison or general pattern matching techniques that combine sequence and secondary structure information, such as `ERPIN`, `HyPa`, or `PATSearch`, can be used in many cases. While search tools exist for such pattern description languages, the inference of characteristic patterns from a set of aligned (or, even more demanding, nonaligned) sequences is still an open problem.

In the case of microRNAs, a characteristic pattern of mutations in the precursor hairpin is a reliable signal provided a sufficient number of homologous sequences can be found. The motifs of box H/ACA and box C/D snoRNAs, as well as the short sequences for some snRNAs, are less informative, however. In this case, knowledge about their potential targets, rRNAs, snRNAs, and some mRNAs, can be used to improve the quality of the prediction. In the case of snRNAs, conserved promoter and enhancer sequences could potentially be used. However, efficient tools for RNA classes besides tRNAs that could be used in genomic surveys are at present not available.

The identification of the target genes of microRNAs is a hot topic at present. While microRNAs bind their target mRNAs almost complementarily in plants, the mechanism appears to be much more complicated in animals. As a consequence, none of the approaches published to date, such as `miRanda`, `targetscan`, or `RNAhybrid` provide a complete solution to the problem. From an algorithmic point of view, the microRNA target problem has stimulated research into methods for computing RNA–RNA interactions, and to consider RNA–protein interactions in a more systematic way.

A related open problem concerns the classification of the ncRNAs that do *not* belong to one of the known classes. While it is unlikely that the functions of these molecules can be elucidated without further experimental data, it is of interest to determine whether there are additional larger classes of ncRNAs that have escaped our attention. Of course, the same question arises for functional RNA motifs in mRNAs.

Structure-function studies of small RNAs as well as the analysis of artificial RNA sequences obtained from SELEX experiments depend upon the ability to compute structure-based alignments. As mentioned earlier, Sankoff's algorithm solves the RNA folding and pairwise alignment problem simultaneously, but is often computationally too expensive. The alternative, obtaining pairwise alignments through tree editing or tree alignment, is problematic since it depends on predicted, possibly incorrect, structures. In all cases, heuristics are used to obtain multiple alignments from pairwise alignments. Open questions in this area concern both algorithmic improvements and more realistic cost functions; for instance, tree alignments are currently available only with linear gap costs.

RNA structure design can be regarded as the inverse of the RNA folding problem: for a given secondary structure graph, find one or more sequences that have this structure as their MFE structure. An enumeration of sequences that fold into a given structure is infeasible in general because of their large number. A heuristic that turns this problem into a combinatorial optimization problem has already been described. There is, however, no exact algorithm that could determine whether there is at least one sequence that folds into a prescribed secondary structure graph.

RNA switches exhibit two competing conformations, whose equilibrium can be shifted easily by molecular events such as the binding of another molecule. Such elements have recently been identified as important regulators of gene expression, in particular, in bacteria. A theoretical study shows that RNAs that have very different secondary structures with near-groundstate energy, that is, potential riboswitches, are relatively frequent and easily accessible in evolution. The secondary structure model furthermore allows estimates of energy barriers and even entire folding trajectories. An impressive experimental study demonstrates that a single sequence can have two distinct functional ribozyme structures. Heuristic algorithms can be used to design sequences with two near groundstates. The case of more than two alternative structures is only partially understood.

Small subunit ribosomal RNAs are probably the most frequently used data sets in molecular phylogenetics. Most studies assume that individual sequence positions evolve at least approximately independently of each other. This is, however, at odds with the fact that rRNAs form stable secondary structures that are very well preserved over very long timescales. Selection for stabilizing RNA structure is seen in many types of structural RNA sequences. Nucleotides in stem regions evolve in strong correlation with their pairing counterpart. More elaborate scoring schemes and even associated maximum likelihood techniques acting directly on RNA secondary structures are algorithmically relatively unproblematic but will require detailed knowledge about the dynamics of RNA structure evolution. While the structures of many RNA families change very little, for example within vertebrates, significant changes of the structure, however, are evident at timescales of the divergence of the major metazoan phyla. As a consequence, not only fixed secondary structures but the evolution of the structures themselves need to be taken into account. Secondary structures have rarely been used in molecular phylogenetics so far. An exception is the investigation into the history of RNase P and RNase MRP RNAs by Davin Penny and coworkers, which demonstrates that ''RNA secondary structure is useful for evaluating evolutionary relatedness, even with sequences that cannot be aligned with confidence''. More recently, cladistic analyses based on RNA secondary structure have demonstrated this point convincingly, in particular at the level of deep phylogenies.

See also RNA Methodologies; RNA Three-Dimensional Structures, Computer Modeling of.

Bibliography

Books and Reviews

Eddy, S.R. (2002) Computational genomics of noncoding RNA genes, *Cell* **109**, 137–140.

Flamm, C., Hofacker, I.L., Stadler, P.F. (1999) RNA in silico: the computational biology of RNA secondary structures, *Adv. Complex Syst.* **2**, 65–90.

Higgs, P.G. (2000) RNA secondary structure: physical and computational aspects, *Q. Rev. Biophys.* **33**, 199–253.

Zuker, M. (2000) Calculating nucleic acid secondary structure, *Curr. Opin. Struct. Biol.* **10**, 303–310.

Primary Literature

Abfalter, I., Flamm, C., Stadler, P.F. (2003) Design of multi-stable nucleic acid sequences, in: Mewes, H.-W., Heun, V., Frishman, D., Kramer, S. (Eds.) *Proceedings of the German Conference on Bioinformatics. GCB 2003*, Vol. 1, Belleville Verlag Michael Farin, München, Germany, pp. 1–7.

Accardo, M.C., Giordano, E., Riccardo, S., Digilio, F.A., Iazzetti, G., Calogero, R.A., Furia, M. (2004) A computational search for box C/D snoRNA genes in the D. melanogaster genome, *Bioinformatics* **20**, 3293–3301.

Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V., Sundaresan, V. (2005) Computational prediction of miRNAs in Arabidopsis thaliana, *Genome Res.* **15**, 78–91.

Akutsu, T. (2001) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, *Discr. Appl. Math.* **104**, 45–62.

Andronescu, M., Fejes, A.P., Hutter, F., Hoos, H.H., Condon, A. (2004) A new algorithm for RNA secondary structure design, *J. Mol. Biol.* **336**, 607–624.

Argaman, L., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., Altuvia, S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli, *Curr. Biol.* **11**, 941–950.

Avner, P., Heard, E. (2001) X-chromosome inactivation: counting, choice, and initiation, *Nat. Rev. Genet.* **2**, 59–67.

Brown, J. (1999) The ribonuclease P database, *Nucleic Acids Res.* **27**, 314–314.

Brown, J.W., Nolan, J.M., Haas, E.S., Rubio, M.A.T., Major, F., Pace, N.R. (1996) Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 3001–3006.

Caetano-Anollés, G. (2002a) Evolved RNA secondary structure and the rooting of the universal tree, *J. Mol. Evol.* **54**, 333–345.

Caetano-Anollés, G. (2002b) Tracing the evolution of RNA structure in ribosomes, *Nucleic Acids Res.* **30**, 2575–2587.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs, *BMC Bioinformatics* **3**, 2, http://www.rna.icmb.utexas.edu.

Carter, R., Dubchak, I., Holbrook, S. (2001) A computational approach to identify genes for functional RNAs in genomic sequences, *Nucleic Acids Res.* **29**, 3928–3938.

Chen, S., Lesnik, E., Hall, T., Sampath, R., Griffey, R., Eker, D., Blyn, L. (2002) A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome, *Biosystems* **65**, 157–177.

Collins, L.J., Moulton, V., Penny, D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP, *J. Mol. Evol.* **51**, 194–204.

d'Aubenton Carafa, Y., Brody, E., Thermes, C. (1990) Prediction of rho-independent Escherichia coli transcription terminators. a statistical analysis of their RNA stem-loop structures, *J. Mol. Biol.* **216**, 835–858.

Dayton, E., Konings, D., Powell, D., Shapiro, B., Butini, L., Maizel, J., Dayton, A. (1992) Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein, *J. Virol.* **66**, 1139–1151.

Dimitrov, R.A., Zuker, M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids, *Biophys. J.* **87**, 215–226.

Ding, Y., Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction, *Nucleic Acids Res.* **31**, 7180–7301.

Dirks, R.M., Pierce, N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots, *J. Comput. Chem.* **24**, 1664–1677.

Doshi, K., Cannone, J., Cobaugh, C., Gutell, R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary

structure prediction, *BMC Bioinformatics* **5**(1), 105.

Eddy, S. (2001) Non-coding RNA genes and the modern RNA world, *Nat. Genet.* **2**, 919–929.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.S. (2003) MicroRNA targets in Drosophila, *Genome Biol.* **5**(1), Article R1.

Erdmann, V., Szymanski, M., Hochberg, A., de Groot, N., Barciszewski, J. (1999) Collection of mRNA-like non-coding RNAs, *Nucleic Acids Res.* **27**, 192–195.

Erdmann, V., Barciszewska, M., Hochberg, A., de Groot, N., Barciszewski, J. (2001) Regulatory RNAs, *Cell. Mol. Life Sci.* **58**, 960–977.

Farris, A.D., Koelsch, G., Pruijn, G.J., van Venrooij, W.J., Harley, J.B. (1999) Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis, *Nucleic Acids Res.* **27**, 1070–1078.

Flamm, C., Fontana, W., Hofacker, I., Schuster, P. (2000a) RNA folding kinetics at elementary step resolution, *RNA* **6**, 325–338.

Flamm, C., Hofacker, I.L., Maurer-Stroh, S., Stadler, P.F., Zehl, M. (2000b) Design of multi-stable RNA molecules, *RNA* **7**, 254–265.

Fogel, G., Porto, V., Weekes, D., Fogel, D., Griffey, R., McNeil, J., Lesnik, E., Ecker, D., Sampath, R. (2002) Discovery of RNA structural elements using evolutionary computation, *Nucleic Acids Res.* **30**, 5310–5317.

Fontana, W., Konings, D., Stadler, P., Schuster, P. (1993) Statistics of RNA secondary structures, *Biopolymers* **33**, 1389–1404.

Franke, A., Baker, B. (2000) Dosage compensation rox! *Curr. Opin. Cell Biol.* **12**, 351–354.

Gardner, P.P., Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches, *BMC Bioinformatic* **5**, 140.

Gautheret, D., Lambert, A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles, *J. Mol. Biol.* **313**, 1003–1011.

Gautheret, D., Major, F., Cedergren, R. (1990) Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA, *Comput. Appl. Biosci.* **6**, 325–331.

Gong, H., Liu, C.-M., Liu, D.-P., Liang, C.-C. (2005) The role of small RNAs in human diseases: potential troublemaker and therapeutic tools, *Med. Res. Rev.* **25**, 361–381, Epub 6 Jan 2005.

Gorodkin, J., Heyer, L., Stormo, G. (1997) Finding Common Sequences and Structure Motifs in a Set of RNA Molecules, in: Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C., Valencia, A., eds. *Proceedings of the ISMB-97*, AAAI Press, Menlo Park, CA, pp. 120–123.

Gorodkin, J., Stricklin, S.L., Stormo, G.D. (2001b) Discovering common stem-loop motifs in unaligned RNA sequences, *Nucleic Acids Res.* **29**(10), 2135–2144.

Gorodkin, J., Knudsen, B., Zwieb, C., Samuelsson, T. (2001a) SRPDB (signal recognition particle database), *Nucleic Acids Res.* **29**, 169–170.

Grad, Y., Aach, J., Hayes, G., Reinahrt, B., Church, G., Ruvkun, G., Kim, J. (2003) Computational and experimental identification of C. elegans microRNAs, *Mol. Cell* **11**, 1253–1263.

Gräf, S., Strothmann, D., Kurtz, S., Steger, G. (2001) HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns, *Nucleic Acids Res.* **29**, 196–198.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. (2003) Rfam: an RNA family database, *Nucleic Acids Res.* **31**, 439–441.

Gultyaev, A.P., van Batenburg, F., Pleij, C. (1999) An approximation of loop free energy values of RNA H-pseudoknots, *RNA* **5**, 609–617.

Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J., Stormo, G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, *Nucleic Acids Res.* **20**, 5785–5795.

Hackermüller, J., Meisner, N.-C., Auer, M., Jaritz, M., Stadler, P.F. (2005) The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model, *Gene* **345**, 3–12. doi:10.1016/j.gene.2004.11.043.

Harris, J.K., Haas, E.S., Williams, D., Frank, D.N. (2001) New insight into RNase P RNA structure from comparative analysis of the archaeal RNA, *RNA* **7**, 220–232. Hernandez, N. (2001) Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription, *J. Biol. Chem.* **276**, 26733–26736.

Hobza, P., Šponer, J. (2002) Towards true DNA base-stacking energies: MP2, CCSD(T), and

complete basis set calculations, *J. Am. Chem. Soc.* **124**, 11802–11808.

Höchsmann, M., Töller, T., Giegerich, R., Kurtz, S. (2003) Local similarity in RNA secondary structures, Proceedings of the Computational Systems Bioinformatics Conference, Stanford, CA, Aug. 2003 (CSB 2003), pp. 159–168.

Hofacker, I.L. (2003) The Vienna RNA secondary structure server, *Nucleic Acids Res.* **31**, 3429–3431.

Hofacker, I., Stadler, P. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes, *Comput. Chem.* **23**, 401–414.

Hofacker, I.L., Schuster, P., Stadler, P.F. (1998b) Combinatorics of RNA secondary structures, *Discr. Appl. Math.* **88**, 207–237.

Hofacker, I., Fekete, M., Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences, *J. Mol. Biol.* **319**, 1059–1066.

Hofacker, I.L., Bernhart, S.H.F., Stadler, P.F. (2004) Alignment of RNA base pairing probability matrices, *Bioinformatics* **20**, 2222–2227.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M., Schuster, P. (1994) Fast folding and comparison of RNA secondary structures, *Monatsh. Chem.* **125**, 167–188.

Hofacker, I., Fekete, M., Flamm, C., Huynen, M., Rauscher, S., Stolorz, P., Stadler, P. (1998a) Automatic detection of conserved RNA structure elements in complete RNA virus genomes, *Nucleic Acids Res.* **26**, 3825–3836.

Holmes, I. (2004) A probabilistic model for the evolution of RNA structure, *BMC Bioinformatics* **5**, 166.

Huez, I., Créancier, L., Audigier, S., Gensac, M., Prats, A., Prats, H. (1998) Two independent internal ribosome entry sites are involved in translation initiation of vascular endothelial growth factor mRNA, *Mol. Cell. Biol.* **18**, 6178–6190.

Hüttenhofer, A., Kiefmann, M., Neier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J., Brosius, J. (2001) Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse, *EMBO J.* **20**, 2943–2953.

Huynen, M., Perelson, A., Viera, W., Stadler, P. (1996) Base pairing probabilities in a complete HIV-1 RNA, *J. Comput. Biol.* **3**, 253–274.

Jacobs, G., Rackham, O., Stockwell, P., Tate, W., Brown, C.M. (2002) Transterm: a database of mRNAs and translational control elements, *Nucleic Acids Res.* **30**, 310–311.

Klein, R., Misulovin, Z., Eddy, S. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7542–7547.

Knudsen, B., Hein, J. (1999) Using stochastic context free grammars and molecular evolution to predict RNA secondary structure, *Bioinformatics* **15**, 446–454.

Knudsen, B., Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars, *Nucleic Acids Res.* **31**, 3423–3428.

Kool, E.T. (1997) Preorganization of DNA: design principles for improving nucleic acid recognition by synthetic oligonucleotides, *Chem. Rev.* **97**, 1473–1487.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs, *Science* **294**, 853–857.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., Tuschl, T. (2002) Identification of tissue specific microRNAs from mouse, *Curr. Biol.* **12**, 735–739.

Laslett, D., Canback, B., Andersson, S. (2002) BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences, *Nucleic Acids Res.* **30**, 3449–3453.

Lau, N., Lim, L., Weinstein, E., Bartel, D. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, *Science* **294**, 858–862.

Le, S.-Y., Chen, J.-H., Currey, K., Maizel, J. (1988) A program for predicting significant RNA secondary structures, *Comput. Appl. Biosci.* **4**, 153–159.

Lee, R.C., Ambros, V. (2001) An extensive class of small RNAs in Caenorhabditis elegans, *Science* **294**, 862–864.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., Burge, C.B. (2003) Prediction of mammalian microRNA targets, *Cell* **115**, 787–798.

Leydold, J., Stadler, P.F. (1998) Minimal cycle basis, outerplanar graphs, *Elec. J. Comb.* **5**, R16, See http://www.combinatorics.org.

Lim, L., Lau, N., Weinstein, E., Abdelhakim, A., Yekta, S., Rhoades, M., Burge, C., Bartel, P. (2003) The microRNAs of Caenorhabditis elegans, *Genes Dev.* **17**, 991–1008.

Lowe, T., Eddy, S. (1997) tRNAscan-SE: a program for improved detection of transfer

RNA genes in genomic sequence, *Nucleic Acids Res.* **25**, 955–964.

Lück, R., Steger, G., Riesner, D. (1996) Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein, *J. Mol. Biol.* **258**, 813–826.

Lück, R., Gräf, S., Steger, G. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure, *Nucleic Acids Res.* **27**, 4208–4217.

Lyngsø, R.B., Pedersen, C.N.S. (2000) RNA pseudoknot prediction in energy based models, *J. Comput. Biol.* **7**(3/4), 409–428.

Macdonald, P. (1990) Bicoid mRNA localization signal: phylogenetic conservation of function and RNA secondary structure, *Development* **110**, 161–171.

MacIntosh, G., Wilkerson, C., Green, P. (2001) Identification and analysis of arabidopsis expressed sequence tags characteristic of non-coding RNAs, *Plant Physiol.* **127**, 765–776.

Macke, T., Ecker, D., Gutell, R., Gautheret, D., Case, D., Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm, *Nucleic Acids Res.* **29**, 4724–4735.

Maidak, B., Cole, J., Lilburn, T., Parker, C. Jr., Saxman, P., Farris, R., Garrity, G., Olsen, G., Schmidt, T., Tiedje, J. (2001) The RDP-II (ribosomal database project), *Nucleic Acids Res.* **29**, 173–174.

Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization, *RNA* **10**, 1178–1190.

Mathews, D., Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences, *J. Mol. Biol.* **317**, 191–203.

Mathews, D., Sabina, J., Zuker, M., Turner, D. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure, *J. Mol. Biol.* **288**, 911–940.

Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7287–7292.

McCaskill, J. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers* **29**, 1105–1119.

McCutcheon, J., Eddy, S. (2003) Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics, *Nucleic Acids Res.* **31**, 4119–4128.

Merino, E., Yanofsky, C. (2002) *Regulation by Termination-Antitermination: A Genomic Approach*, ASM Press, pp. 323–336.

Nussinov, R., Piecznik, G., Griggs, J., Kleitman, D. (1978) Algorithms for loop matching, *SIAM J. Appl. Math.* **35**(1), 68–82.

Ohno, M., Mattaj, I. (1999) Meiosis: MeiRNA hits the spot, *Curr. Biol.* **28**, R66–R69.

Oleynikov, Y., Singer, R. (1998) RNA localization: different zipcodes, same postman? *Trends Cell Biol.* **8**, 381–383.

Omer, A., Lowe, T., Russel, A., Ebhardt, H., Eddy, S., Dennis, P. (2000) Homologs of small nucleolar RNAs in Archaea, *Science* **288**, 517–522.

Ornstein, R.L., Rein, R., Breen, D.L., MacElroy, R.D. (1978) An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking, *Biopolymers* **17**, 2341–2360.

Otsuka, J., Sugaya, N. (2003) Advanced formulation of base pair changes in the stem regions of ribosomal RNAs; its application to mitochondrial rRNAs for resolving the phylogeny of animals, *J. Theor. Biol.* **222**, 447–460.

Pesole, G., Liuni, S., D'Souza, M. (2000) PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance, *Bioinformatics* **16**(5), 439–450.

Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., Sabino, L. (2001) Structural and functional features of eukaryotic mRNA untranslated regions, *Gene* **276**, 73–81.

Peterson, K., Eernisse, D.J. (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S DNA gene sequences, *Evol. Dev.* **3**, 170–205.

Reeder, J., Giegerich, R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics, *BMC Bioinformatics* **5**(1), 104.

Rehmsmeier, M., Steffen, P., Höchsmann, M., Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes, *RNA* **10**, 1507–1517.

Rivas, E., Eddy, S. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots, *J. Mol. Biol.* **285**, 2053–2068.

Rivas, E., Eddy, S. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, *Bioinformatics* **16**, 583–605.

Rivas, E., Eddy, S. (2001) Noncoding RNA gene detection using comparative sequence analysis, *BMC Bioinformatics* **2**(8), 19.

Rivas, E., Klein, R., Jones, T., Eddy, S. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics, *Curr. Biol.* **11**, 1369–1373.

Ruan, J., Stormo, G.D., Zhang, W. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots, *Bioinformatics* **20**, 58–66.

Rueckert, R. (1996) Picornaviridae: The Viruses and their Replication, in: Fields, N., Knipe, D., Howley, P. (Eds.) *Virology*, Vol. 1, 3rd edition, Lippincott-Raven, Philadelphia, NY, pp. 609–654.

Samarsky, D., Fournier, M. (1999) A comprehensive database for the small nucleolar RNAs from Saccharomyces cerevisiae, *Nucleic Acids Res.* **27**, 161–164.

Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment, and proto-sequence problems, *SIAM J. Appl. Math.* **45**, 810–825.

Savill, N.J., Hoyle, D.C., Higgs, P.G. (2001) RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods, *Genetics* **157**, 399–411.

Schattner, P. (2002) Searching for RNA genes using base composition statistics, *Nucleic Acids Res.* **30**, 2076–2082.

Schöninger, M., von Haeseler, A. (1999) Towards assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models, *J. Mol. Evol.* **49**, 691–698.

Schultes, E.A., Bartel, D.P. (2000) One sequence, two ribozymes: Implications for the emergence of new ribozyme folds, *Science* **289**, 448–452.

Shapiro, B.A. (1988) An algorithm for comparing multiple RNA secondary structures, *Comput. Appl. Biosci.* **4**, 387–393.

Siebert, S., Backofen, R. (2003) MARNA a server for multiple alignment of RNAs, in: Mewes, H.-W., Heun, V., Frishman, D., Kramer, S., (Eds.) *Proceedings of the German Conference on Bioinformatics. GCB 2003*, Vol. 1, Belleville Verlag Michael Farin, München, Germany pp. 135–140.

Sousa, C., Johansson, C., Charon, C., Manyani, H., Sautter, C., Kondorosi, A., Crespi, M. (2001) Translational and structural requirements of the early nodulin gene enod40, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex, *Mol. Cell. Biol.* **21**, 354–366.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes, *Nucleic Acids Res.* **26**, 148–153.

Szymanski, M., Barciszewska, M., Barciszewski, J., Erdmann, V. (2000) 5S ribosomal RNA database Y2K, *Nucleic Acids Res.* **28**, 166–167.

Tacker, M., Stadler, P.F., Bornberg-Bauer, E.G., Hofacker, I.L., Schuster, P. (1996) Algorithm independent properties of RNA structure prediction, *Eur. Biophy. J.* **25**, 115–130.

Tang, T., Bachellerie, J., Rozhdestvensky, T., Bortolin, M., Huber, H., Drungowski, M., Elge, T., Brosius, J., Hüttenhofer, A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archeon Archeoglobus fulgidus, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7536–7541.

Van de Peer, Y., De Rijk, P., Wuyts, J., Winkelmans, T., De Wachter, R. (2000) The European small subunit ribosomal RNA database, *Nucleic Acids Res.* **28**, 175–176.

van Zon, A., Mossink, M., Schoester, M., Scheffer, G., Scheper, R., Sonneveld, P., Wiemer, E. (2001) Multiple human vault RNAs. Expression and association with the vault complex, *J. Biol. Chem.* **276**, 37715–37721.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.* **20**(1), 44–50.

Walter, A., Turner, D., Kim, J., Lyttle, M., Müller, P., Mathews, D., Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 9218–9222.

Wang, X.J., Reyes, J.L., Chua, N.H., Gaasterland, T. (2004) Prediction and identification of *Arabidopsis thaliana* microrna and their mRNA targets, *Genome Biol.* **5**, R65, [Epub].

Washietl, S., Hofacker, I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics, *J. Mol. Biol.* **342**, 19–39.

Washietl, S., Hofacker, I.L., Stadler, P.F. (2005) Fast and reliable prediction of noncoding rnas, *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2454–2459.

Wassarman, K., Repoila, F., Rosenow, C., Storz, G., Gottesman, S. (2001) Identification of novel small RNAs using comparative genomics and microarrays, *Genes Dev.* **15**, 1637–1651.

Waterman, M.S. (1978) Secondary structure of single-stranded nucleic acids, *Adv. Math. Suppl. Studies* **1**, 167–212.

Waterman, M.S., Smith, T.F. (1978) Combinatorics of RNA hairpins and cloverleaves, *Stud. Appl. Math.* **60**, 91–96.

Westhof, E., Jaeger, L. (1992) RNA pseudoknots, *Curr. Opin. Struct. Biol.* **2**, 327–333.

Witwer, C., Hofacker, I.L., Stadler, P.F. (2004) Prediction of consensus RNA secondary structures including pseudoknots, *IEEE/ACM Trans. Comp. Biol. Bioinf.* **1**, 65–77.

Witwer, C., Rauscher, S., Hofacker, I., Stadler, P. (2001) Conserved RNA secondary structures in picornaviridae genomes, *Nucleic Acids Res.* **29**, 5079–5089.

Wolfinger, M.T., Svrcek-Seiler, W.A., Flamm, C., Hofacker, I.L., Stadler, P.F. (2004) Exact folding dynamics of RNA secondary structures, *J. Phys. A: Math. Gen.* **37**, 4731–4741.

Wuchty, S., Fontana, W., Hofacker, I., Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures, *Biopolymers* **49**, 145–165.

Wuyts, J., De Rijk, P., Van de Peer, Y., Winkelmans, T., De Wachter, R. (2001) The European large subunit ribosomal RNA database, *Nucleic Acids Res.* **29**, 175–177.

Xia, T., SanatLucia J. Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H. (1998) Parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs, *Biochemistry* **37**, 14719–14735.

Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule, *Science* **244**, 48–52.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res.* **31**, 3406–3415.

Zuker, M., Sankoff, D. (1984) RNA secondary structures and their prediction, *Bull. Math. Biol.* **46**, 591–621.

Zuker, M., Jacobson, A.B. (1998) Using reliability information to annotate RNA secondary structures, *RNA* **4**, 669–679.

Zwieb, C. (1996) The uRNA database, *Nucleic Acids Res.* **24**, 76–79.

Zwieb, C., Wower, J. (2000) tmRDB (tmRNA database), *Nucleic Acids Res.* **28**, 169–170.

Zwieb, C., Wower, I., Wower, J. (1999) Comparative sequence analysis of tmRNA, *Nucleic Acids Res.* **27**(10), 2063–2071.

RNA Translation to Protein: *see* Translation of RNA to Protein