Review

From Structure Prediction to Genomic Screens for Novel Non-Coding RNAs

Jan Gorodkin[1,2]*, Ivo L. Hofacker[1,3]*

1 Center for non-coding RNA in Technology and Health, IBHV University of Copenhagen, Frederiksberg, Denmark, 2 Division of Genetics and Bioinformatics, IBHV University of Copenhagen, Frederiksberg, Denmark, 3 Department of Theoretical Chemistry, University of Vienna, Wien, Austria

Abstract: Non-coding RNAs (ncRNAs) are receiving more and more attention not only as an abundant class of genes, but also as regulatory structural elements (some located in mRNAs). A key feature of RNA function is its structure. Computational methods were developed early for folding and prediction of RNA structure with the aim of assisting in functional analysis. With the discovery of more and more ncRNAs, it has become clear that a large fraction of these are highly structured. Interestingly, a large part of the structure is comprised of regular Watson-Crick and GU wobble base pairs. This and the increased amount of available genomes have made it possible to employ structure-based methods for genomic screens. The field has moved from folding prediction of single sequences to computational screens for ncRNAs in genomic sequence using the RNA structure as the main characteristic feature. Whereas early methods focused on energy-directed folding of single sequences, comparative analysis based on structure preserving changes of base pairs has been efficient in improving accuracy, and today this constitutes a key component in genomic screens. Here, we cover the basic principles of RNA folding and touch upon some of the concepts in current methods that have been applied in genomic screens for *de novo* RNA structures in searches for novel ncRNA genes and regulatory RNA structure on mRNAs. We discuss the strengths and weaknesses of the different strategies and how they can complement each other.

Introduction

Non-coding RNA genes (ncRNAs) have emerged as major players in the cell and are involved in both housekeeping functions as well as regulation. They are characterized as functional transcripts that do not code for proteins and can be processed in numerous ways, see e.g., [1,2]. An abundant class of ncRNA genes are the micro RNAs (miRNAs), which have received considerable attention e.g., [3–5]. This can be observed through the rapid growth in the literature, not only for miRNAs [6], but also for ncRNAs in general [7]. Furthermore, regulatory RNA structure in UTR regions of protein-coding genes is also an exciting, emerging field.

The roles of ncRNAs are diverse and not only include regulation of protein coding genes [8], but also inactivation of other gene classes (e.g., imprinting [9,10]), alternative splicing [11], and modifying other ncRNAs [12], to mention just a few examples. Thus the miRNAs are but one among several other classes of ncRNAs. Novel classes of small ncRNA genes such as piRNAs [13,14] and hpRNAs [15] have also been reported. Recently, long intervening ncRNAs (lincRNAs) have been found. These are mRNA-like transcripts that lack protein-coding potential, contain exon intron structure, and are apparently largely unstructured [16]. The repertoire of ncRNAs is rapidly expanding and RNA-seq sequencing techniques, in combination with computer methods, are expected to give rise to a general expansion of the RNA universe. These RNA families are collected in the Rfam database [17] in the form of structural alignments and consensus structures. In a number of cases, such as SRP RNAs and tmRNAs [18], Rfam is based on pre-existing curated RNA structural alignments from specialized databases. This important resource is also often used to construct and test RNA structure prediction tools [7].

The size variation of ncRNAs is extreme, ranging from ~ 20 nucleotides (nt) for small interfering RNAs and miRNAs to $\sim 100,000$ nt for the *air* RNA [10]. ncRNAs are not only located in intergenic regions, that is outside of protein coding genic regions, but they are also found in introns. In the latter case they are either processed out during splicing, or they represent independent transcripts that come with their own promoters, as e.g., in *Caenorhabditis elegans* [19]. There are also examples of ncRNAs overlapping coding regions [20]. In addition, mRNAs may contain functional *cis*-acting RNA structures, such as the iron-responsive element [21] in vertebrates or riboswitches in bacteria [22].

As can be seen by inspection of Rfam, a solid volume of ncRNAs and regulatory RNAs come with a characteristic and functional RNA structure, which often is more conserved in evolution than its primary sequence. In order to find ncRNA genes, it therefore makes sense to search for RNA (secondary) structure rather than primary sequence. Computationally, this is a much more challenging and demanding problem than searching protein coding space, as there are no regular signals in RNA structured sequence such as open reading frames.

However, searching for RNA secondary structure is likely not to provide us with all functional non-protein-coding transcripts, since the emerging compilation of long ncRNAs seems to indicate that these in general are not densely structured, even though they

might contain structural regions. This is exemplified by a mouse transcriptomic analysis that revealed thousands of such transcripts based upon full-length cDNA sequencing [23]. It remains to be systematically investigated whether these RNAs harbor characteristic structures that carry out specific functions, and thus if searching for RNA structure in general is a sufficient starting point to search for ncRNAs. It is worth noting that the only functionally well-characterized lncRNA, hotair, does have functional RNA structures [24–26].

Here, we focus on describing the principles of searching for *de novo* RNA structures in genomic sequences, being aware that the gene (and functional transcript) itself can be (much) larger than the (predicted) structure and that overlapping predicted structures can be in the same functional transcript.

Parameters of the Search Space

Searching for novel RNA secondary structures requires that functional structures can be distinguished (e.g., by their folding energy) from those generated on shuffled sequences of the same composition, that is, the background. In general, programs like `mfold` and `RNAfold` will fold *any* RNA sequence you feed into them. Whether the structure (or parts thereof) is actually trustworthy is of course determined by the user.

It turns out that for most known ncRNAs (with miRNAs as a notable exception), it is generally not sufficient to screen individual genomes using minimum free energy folding (of a sequence in some fixed size window), since neither folding energies nor the resulting structures provide a reliable signal. Although ncRNAs tend to have somewhat more stable structures than expected by chance, the difference in folding energies between random sequences created by shuffling and native ncRNAs is in general too small to distinguish real ncRNAs from decoys [27,28]. However, the often stronger conservation of the (secondary) structure compared to the primary sequence can be used to enhance the discrepancy to the background. For example, for a human RNA sequence CCCCCCCAGUUGGGGGG that forms a simple hairpin, the mouse version could be CACCCCCA-GUUGGGGUG such that a GC base pair in human corresponds to an AU base pair in mouse. Not only do such features destroy conservation of primary sequence, but the base pairs can also be separated essentially by the full length of the sequence. Hence, meaningful *in silico* screens can be carried out on comparative genomic data, but using complex algorithms that take long-range base pairs into account.

The ideal search scenario is illustrated by a toy example in Figure 1, where we have randomized some sequences (shuffling the order of the nucleotides) and implanted a small hairpin conserved only in structure. These can be considered as a set of corresponding (but poorly conserved) sequences that do not necessarily have much in common except for common RNA structure. This sequence set can be searched (sequences on the left side) and a joint structure extracted (on the right side) where base pairs are represented by matching parentheses. In real examples, the "background part" of sequences is never so strongly divergent and neither are the sequences of the contained motifs (while the structure is convergent). This, of course, creates challenges for the prediction scheme.

Overall, *in silico* searches for ncRNAs can in essence be carried out in the following three ways: (i) by sequence/structure similarity to already known ncRNAs, (ii) by searching for specific ncRNA classes, e.g., miRNAs and snoRNAs, and (iii) *de novo* searches. Here, we focus on *de novo* searches, but briefly touch upon the others below.

Sequence Similarity Search for ncRNAs

The basic form of similarity search is purely sequence based using `BLAST` [29], and this approach has apparently not been reported in the literature for anything other than finding near identical sequences, e.g., genome and EST annotation projects [30,31]. The more advanced approach is to include the RNA secondary structure as done for covariance models such as `INFERNAL` and `RaveNnA` [32–34]. These constitute a probabilistic model of the RNA structure together with the corresponding sequence variation (e.g., compensating base pairs). More specifically, they employ *stochastic context-free grammars* (SCFGs), an extension of hidden Markov models (HMMs), that can cope with the long-range base pair interactions. An alternative (which is faster) is to extract patterns for RNA motif search, e.g., `RNAmotif` [35].

To obtain good models, well-curated data (structural RNA alignments) are needed, which can be obtained either from specialized databases, as in the case of RNAseP RNA and SRP RNAs [18,36], or from the meta database `Rfam`. Curating these and conducting homology-based searches comes with its own set of issues, which is described elsewhere [7].

Class-specific searches use distinctive features of an RNA class to search for novel, but not necessarily homologous, members of that class. miRNAs are such an example that can be identified on the basis of the characteristic stem-loop shape of the precursor either encoded as explicit rules or combined with machine learning techniques [6]. Another example is the well-known tRNA-scanSE program to search for tRNAs [37]. Similar types of searches have also been employed for other families, and incorporating this information is generally expected to help span greater distances in the evolutionary tree than what can be done solely from (present) covariance models. The principal reason is RNA structure itself changes, so that models made for one family cannot readily be applied to another. Well-known examples are RNAse P RNA [12] and telomerase RNA [38]. A recent advance in the `INFERNAL` package is that it can search for local structural matches.

As previously mentioned, *in silico* screens currently involve searching for *de novo* RNA structure, but there have also been a few cases employing GC content as an indicator of RNA sequence structure in certain organisms (extremophiles with biased AU content) [39–41]. Here, we focus on describing the principles and the concepts of *de novo* searches. When there is overlap with similarity search methods, this will be mentioned. We will concentrate, however, on the concepts and not on reporting what one actually can expect to find and what to do with these sequences. The latter aspects are reviewed e.g., in [42].

To summarize, in Figure 2 there are two basic flow charts of current similarity searches to provide mapping of homologous ncRNAs and regions of synteny for related genomes. The latter can be used as an extra layer of confirmation for the raw similarity search, but also to investigate if genomic rearrangements have taken place. Clearly, synteny can yield further support for the outcome of an *in silico* screen.

RNA Structure and Folding

As mentioned above, folding of single sequences is in general not sufficient to reliably detect RNA structure. Still, the principle of folding single sequences is fundamental in basically all computational approaches constructed to search for RNA structure in genomic sequence. The structured RNA molecules by nature take a characteristic three-dimensional (3D) structure. As depicted in Figure 3, even though it is still difficult to predict 3D from 2D structure, most contacts between bases are already part of the secondary structure. Moreover, the canonical base pairs

Figure 1. Searching for common RNA secondary structure in unaligned sequences. The scenario of searching for common RNA structure in sequences (left) that are otherwise unrelated (here generated by shuffling the order of the nucleotides in real sequences). This structure can either represent portions of an ncRNA gene or a structural RNA element in an mRNA. The search result in a multiple structural alignment (right) is typically based on the pattern of obtained compensating changes.
doi:10.1371/journal.pcbi.1002100.g001

making up the secondary structure can be reasonably well predicted without any knowledge of tertiary structure. This makes the minimum free energy secondary structure a useful abstraction of the full 3D structure. Current methods do generally focus on the RNA secondary structure, even though the awareness and feasibility of taking the 3D structure into account is improving. It is beyond the scope of this text to go deeper into this. Unless mentioned otherwise, we will from now on write RNA structure as a shorthand for RNA secondary structure. The RNA secondary structure can be represented in numerous ways, as depicted in Figure 4.

Concepts of Folding RNA Sequences

The basic folding algorithm goes back to the early work of Ruth Nussinov [43], who proposed a simple dynamic programming algorithm to find the maximum number of base pairs for an RNA

sequence. The idea is to keep track of the number of base pairs of any sub-sequence starting at some position, say i, and ending at position j. Given that the sequence is L nucleotides long, the recursion requires that $1 \leq i < j \leq L$. Additionally, pseudoknots are ignored as a first approximation. Pseudoknots can be considered as higher-order base pairing interactions and would correspond to having lines crossing in the outer left part of diagram shown in Figure 4. Including pseudoknots results in much more complex algorithms with higher time and memory consumption.

Thus, starting with (unpaired) sub-sequences of length one and extending (and meeting the first base pair at some point), one can consider a structure on the sub-sequence $x[i..j]$. Such structure can be formed in only two distinct ways from shorter structures: Either the starting nucleotide i is unpaired, in which case it is followed by an arbitrary structure on the shorter sequence $x[i+1..j]$, or the first nucleotide is paired with some partner base, say k. In the
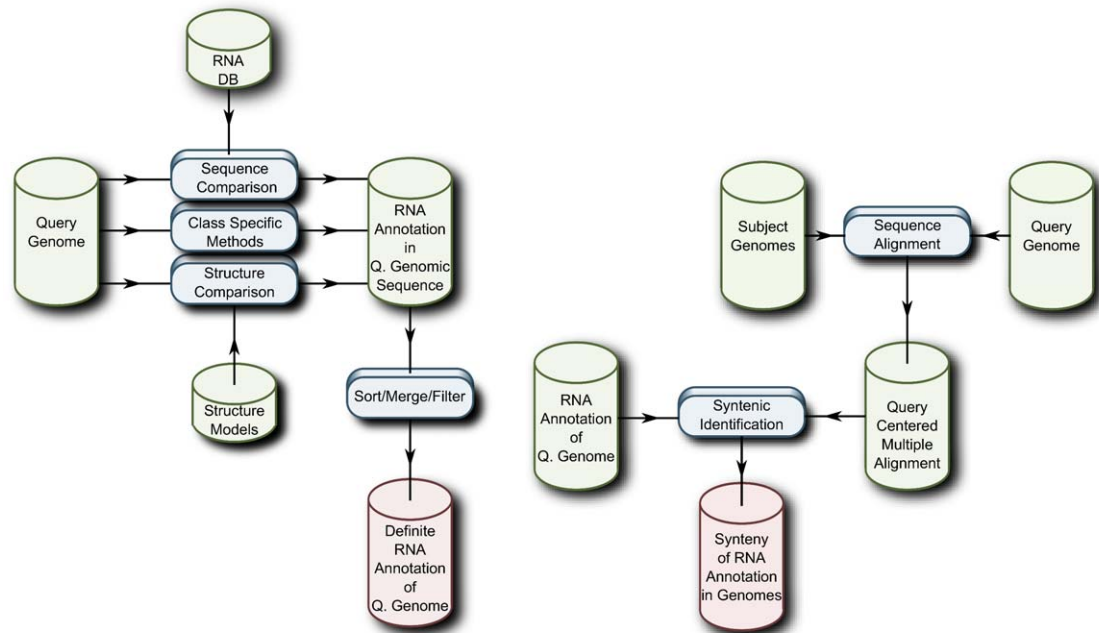

Figure 2. Basic flow homology (left) search in combination with identification of syntenic regions (right) of related genomes. (Figure courtesy of Christian Anthon.)
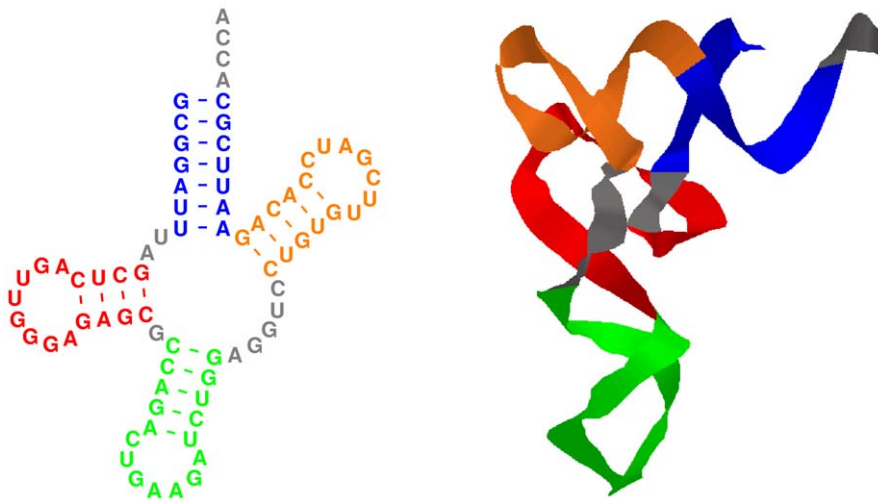doi:10.1371/journal.pcbi.1002100.g002

Figure 3. An example of 2D (left) and 3D (right) representations of RNA structures, here illustrated for a tRNA. The RNA secondary structure is an important step towards the full 3D structure. (Figure from [116].)
doi:10.1371/journal.pcbi.1002100.g003

latter case the rule that base pairs must not cross implies that we have independent secondary structures on the sub-intervals $x[i+1..k-1]$ and $x[k+1..j]$. Graphically, we can write this decomposition of the set of structures as shown in Figure 5.

Denoting E_{ij} as the maximum number of base pairs (or optimal energy) for a secondary structure on $x[i..j]$ corresponding to the left side of the equation, we see that E_{ij} is the optimal choice among each of the alternatives. In this context, independence of two substructures in the paired cases implies that we have to optimize these substructures independently. Using β_{ij} as 1 if x_i and x_j base pair and zero otherwise, we arrive at the recursion:

$$E_{ij} = \max \begin{cases} E_{i+1,j} \\ \max_{k,(i,k)\text{ pairs}}\{E_{i+1,k-1}+E_{k+1,j}+\beta_{ik}\} \end{cases}, \quad (1)$$

where the maximum runs over $i+1 < k \leq j$. Rather than having the parameter β_{ij} one or zero and rather than counting the maximum number of base pairs, we can let β_{ij} take negative values depending on the type of base pair, that is, by replacing β_{ij} with $\beta_{x[i]x[j]}$ to take the individual base pairs into account, and then replace the max in the recursion by min. An example of filling out the dynamical programming matrix is shown in Figure 6. The recursion in Equation 1 is a simplification (and less ambiguous) of

a more general form of the Nussinov algorithm. A good introduction is given in [44].

Towards a Full Folding Algorithm

This simple model is still too inaccurate, since it does not capture energetically important structure motifs, such as stacked pairs, bulges, and various types of loops (hairpin, multi, interior, and exterior). The more realistic "nearest-neighbor" energy model is therefore based on loops, rather than base pairs. A complete set of loop energies is available from the group of Doug Turner [45]. Stacked pairs, for example, consist of two consecutive base pairs and are the major source of stabilizing energy. Each possible stacking comes with its own free energy as listed in Figure 7. It can be observed that GCs have lower binding values and therefore form more stable stacks and thereby structures. This relates to the issue of searching for RNA structures in GC-rich regions in the genomes. In general, loop energies depend on the loop type and its size, and sequence dependence is conferred only through the base pairs closing the loop and the unpaired bases directly adjacent to the pair (the *terminal mismatches*). The general form of loop energy is therefore

$$E_{loop} = E_{mismatch} + E_{size} + E_{special}, \quad (2)$$
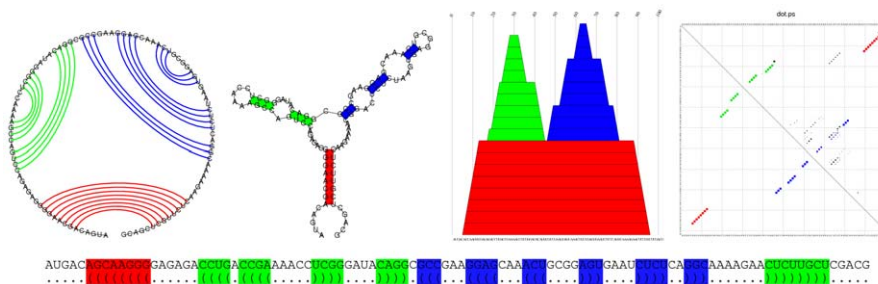
Figure 4. Representations of RNA (secondary) structure. From top left: A circle plot, a conventional secondary structure diagram, a mountain plot, and a dot plot. The bottom diagram shows the secondary structure in dot-bracket notation, where a base pair is represented by matching parentheses. The respective colors in each diagram represent the same base pairs. The structure shown is a glycine riboswitch from *B. subtilis*, Rfam family RF00504.
doi:10.1371/journal.pcbi.1002100.g004

Figure 5. Decomposition of RNA secondary structures for the Nussinov algorithm. The decomposition is unambiguous in the sense that each structure can only be decomposed in a single way.
doi:10.1371/journal.pcbi.1002100.g005

where the last term is used for special cases, e.g., to assign bonus energies to unusually stable *tetra loops*. While the model allows only Watson-Crick (AU, UA, CG, and GC) and wobble pairs (GU, UG), non-standard base pairs in helices are treated as special types of interior loops. Therefore, an extended dynamic programming algorithm is needed and replaces the one shown above.

Using the loop-based energy model is essential in order to achieve reasonable prediction accuracies. On average, current energy models achieve accuracies of $\sim 50-70\%$ in terms of the percentage of correct base pairs [46]. Prediction accuracy tends to fall somewhat with sequence length [47]. This effect could be simply due to combinatorics (long RNAs have more *wrong* structures), or because long sequences are kinetically trapped in structures other than the ground state. Recent approaches combine structure-probing experiments and use the following information for single/double-stranded positions as constraints to the folding algorithms to obtain higher accuracy [48,49].

The more standard energy model results in somewhat more complicated recursions and requires additional tables. However, memory and CPU requirements remain $\mathcal{O}(L^3)$ and $\mathcal{O}(L^2)$ as in the Nussinov algorithm. The factor L^3 comes from the time it takes to fill out the upper half of the matrix of size L^2 and then check for adding sub-structures (the k index in Equation 1). The crucial quantity in the loop-based version is the optimal free energy for a sub-sequence $x[i..j]$ enclosed by a base pair (i,j). In order to compute that, we now have to distinguish between the different types of loops that can be closed by i and j. For a complete set of corresponding recursions see e.g., [50].

Folding of Randomized Sequences

While it seems natural to detect ncRNA genes on the basis of structure prediction, the task is far from straightforward. The problem is that almost any RNA sequence will form some kind of secondary structure. The real challenge is therefore to distinguish whether a structure is spurious or may constitute a functional structure. Unfortunately, structures formed by functional ncRNAs do not look significantly different from structures formed by random sequences [51], as illustrated in Figure 8. By random sequences we denote sequences for which the order of the nucleotides has been shuffled. Often this is done by preserving the

di-nucleotide order, as that has an impact on the stacking of base pairs.

In fact, when Rivas and Eddy set out to build a general RNA gene finder based on this principle, they had to conclude that secondary structure alone is generally not significant enough for the detection of ncRNAs [27]. Subsequent studies [52] focused on folding energies and showed that (i) functional RNAs tend to be slightly more stable than randomized RNAs, (ii) the difference is statistically significant, but too small to be of much use without additional criteria, and (iii) that for a fair comparison randomized sequences should be generated such that the di-nucleotide content (not just nucleotide composition) is conserved.

A notable exception are microRNAs [53] which form unusually stable structures.

Extracting Structure from Multiple RNA Sequences

As single sequences are not sufficient to extract a clear signature of RNA structure, and since RNA structure can be more conserved than sequence, multiple (orthologous/syntenic) sequences can be searched to find a common structure. It is particularly of interest to detect or exploit *compensating* base changes, as these indicate conserved structure in spite of varying sequence as exemplified in the toy example in Figure 1. Below, we conceptually describe approaches to predicting consensus RNA structure from either aligned or unaligned sequences, an essential step towards searching for RNA structure in genomic sequence.

Mutual Information

Given a multiple sequence alignment (typically made without knowledge of the structure), the most common way to quantify covariation for the purpose of RNA secondary determination is by measuring the *mutual information* content [54,55]:

$$M_{ij} = \sum_{X,Y} f_{ij}(XY) \log \frac{f_{ij}(XY)}{f_i(X)f_j(Y)}, \qquad (3)$$

where i and j are two columns of a multiple sequence alignment, $f_i X$ denotes the frequency of nucleotide X in column i, and
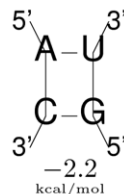

Figure 6. Free energies for stacked pairs and loops in kcal/mol. Note that both base pairs have to be read in 5′-3′ direction.
doi:10.1371/journal.pcbi.1002100.g006

Folding using Nussinov's Algorithm

$$\beta(C,G) = \beta(G,C) = 3 \quad \beta(A,U) = \beta(U,A) = 2$$
$$\beta(G,U) = \beta(U,G) = 1$$

$$E_{ij} = E_{i+1,j} \mid E_{i+1,k-1} + E_{k+1,j} + \beta(x_i, x_k)$$

.	(	(	.	(	.	.	.	)	)	)	
A	G	C	A	C	A	C	A	G	G	C	
0	0	0	0	0	0	3	3	3	6	9	A
	0	0	0	0	0	3	3	3	6	9	G
		0	0	0	0	0	3	6	6	C
			0	0	0	0	3	3	3	A
				0	0	0	3	3	3	C
					0	0	0	0	0	0	A
						0	0	0	0	C
							0	0	0	A
								0	0	G
									0	G
										C

Figure 7. Filled dynamic programming matrix E_{ij} for the toy sequence AGCACACAGGC. Values giving rise to the optimal folding energy of -9 are shown in red.
doi:10.1371/journal.pcbi.1002100.g007

$f_{ij}(XY)$ denotes the frequency of co-occurrence of the nucleotides X and Y.

Mutual information makes no use of pairing rules and can therefore be used to detect tertiary interactions as well. However, the number of sequences needed to reliably deduce secondary structures from mutual information alone is prohibitive for most classes of RNA. Nonetheless, alternative versions of the mutual information content have been shown to drastically lower the required number of sequences [56–58]. In any case, however, it makes good sense to combine co-variance analysis with structure prediction techniques. A manual approach to optimize the alignment is to revise the alignment based upon computation of the mutual information content, a process which recently has been automated in several projects, e.g., [59–61]. In a prediction screen, the consensus structure predictions are often based on a fixed pre-computed sequence alignment.

Folding Multiple Alignments of RNA Sequences

Consider a multiple alignment for which the mutual information content has been computed, then one simple way to extract the information about base pairs would be to employ a Nussinov-style algorithm to maximize the amount of mutual information between paired columns. In general, such an approach is insufficient, as a number of structural features cannot be taken into account, for example base pair stacking. An alternative is to combine the information from covarying base changes with a standard dynamic programming folding algorithm. In the `RNAalifold` program this is done simply by averaging the folding energy over all sequences, thus, e.g., the energy contribution of a stacked pair in the consensus structure is taken as the average of the stacking energy over all sequences in the alignment. To make best use of the covariation information, this average folding energy is augmented by a covariance term that is added as a pseduo-energy. Instead of mutual information (Equation 3), the following covariation term is employed:

$$C_{ij} = \sum_{XY,X'Y'} f_{ij}(XY)\mathbf{D}_{XY,X'Y'}f_{ij}(X'Y'), \qquad (4)$$

where the 16×16 matrix $\mathbf{D}$ is chosen such that compensatory mutations receive a bonus of -2 kcal/mol, consistent mutations (such as G-C $\rightarrow$ C-U) receive -1 kcal/mol, conserved pairs get a score of 0, and non-canonical pairs incur a penalty of 1 kcal/mol. In contrast to mutual information, this covariance term explicitly favors consistent mutation and tends to be less noisy for alignments with few sequences.

A widely used alternative, but similar approach, is to compute probabilities for alignment columns (based on 4×4 substitution rates) to be single stranded (unpaired) and probabilities for columns to be base paired (based on 16×16 substitution rates) and search for the structure that leads to the highest alignment probability. This approach is taken in the SCFG program `Pfold`, which aims to maximize the joint probability of consensus structure and alignment [62]. More precisely, it computes the probability $P(D|\sigma,T,M)$ of an alignment D given a consensus structure σ, a phylogenetic tree T, and a model of substitution rates M. This uses a Felsenstein model [63], as is usual in maximum likelihood tree estimation, for single-stranded and base-paired columns, respectively. In addition, it uses an SCFG to compute the prior probability of a structure $P(\sigma|M)$, and thereby the joint probability $P(D,\sigma|T,M) = P(D|\sigma,T,M) \cdot P(\sigma|M)$. Recently, the concepts of `Pfold` were extended to a maximum
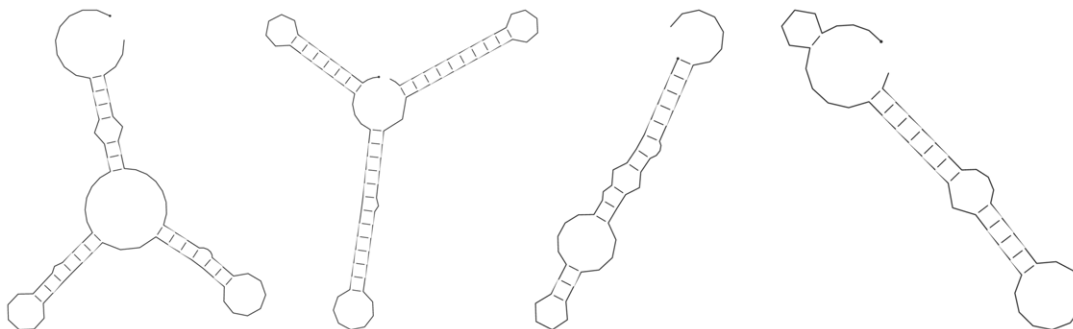

Figure 8. Structure prediction for two non-coding RNA sequences (DsrA and DicF) and respective (shuffled) sequences with the same length and nucleotide composition. Most readers will not be able to distinguish between the real and randomized scenarios.
doi:10.1371/journal.pcbi.1002100.g008

expected accuracy framework, `PETfold`, to simultaneously optimize phylogenetic and energetic information [64].

Under ideal conditions, i.e., well-conserved structure, many compensatory mutations, and error-free alignments, all these algorithms produce near-perfect predictions. For realistic datasets, the challenges lie in dealing with (small) structural variations between the sequences, while being not too sensitive to alignment errors, and dealing gracefully with the lack of covariation.

Simultaneously Folding and Aligning RNA Sequences

Consensus structure prediction exploits the co-variation signal in an alignment, and this signal should increase as sequences become more diverged. A potential problem in applying sequence-based alignments for RNA structure prediction is, however, that with lower sequence similarity, alignments become more inaccurate, eventually leading to a breakdown of structure prediction. Empirically, this limit has been found to lie at about 60% pairwise sequence identity, both for `RNAalifoldZ` [65] and in a study by Gardner et al. [66], who showed for tRNAs that around this similarity sequence-based alignment methods drastically lose the ability to reproduce the alignment, whereas structure-based methods are still providing fairly good results. A toy example in Figure 9 illustrates how sequence similarity can be insufficient for comparing structured RNA sequences.

In 1985, Sankoff [67] published the first method for simultaneously folding and aligning N sequences of length L, a method that has time and memory complexities of $O(L^{3N})$ and $O(L^{2N})$, respectively. This basically makes the algorithm intractable for more than two sequences as well as for long sequences. Intuitively, for two sequences all folds in the one sequence are to be compared with all folds in the other, leading to twice as high an exponent, e.g., $O(L^6)$ instead of $O(L^3)$. This intractable high complexity has prompted several creative attempts at simplified versions of the Sankoff algorithm, as well as completely different types of approaches, e.g., [68]. Complementary to folding alignments, approaches folding the individual sequences and aligning the structures have been proposed, e.g., [69].

Some of the first implementations for RNA structure alignments are based on SCFGs [70,71] and avoid the high cost of the Sankoff algorithm by using an iterative approach that alternates between aligning sequences to a covariance model and deducing a refined covariance model from the alignment ([70]).

The first simplified implementation of the Sankoff algorithm was the first version of `FOLDALIGN` [72], which was restricted to stem-loop structures only. Later, more complete versions were published and the first full-scale implementation for two sequences

was `dynalign` [73,74]. A nice SCFG framework was also introduced in `stemloc` and later `consan` methods [75–77]. Later, PMcomp [78] and `LocARNA` [79] introduced the use of pre-computed base pair probability matrices to reduce computational cost (`PMcomp`) and memory (`LocARNA`). Common for these methods is that when structurally aligning two sequences, the recursion involves a *four dimensional* dynamical programming matrix. Essentially, Equation 1 can be extended to a $E_{ij,kl}$ where the sub-sequences $x[i..j]$ and $y[k..l]$ are simultaneously folded and aligned. The scoring scheme (energy model) thus has to be able to score (mis)matches between unpaired nucleotides as well as between base pairs. For the latter, one often uses the so-called ribosum matrices [80], derived from substitution frequencies in ribosomal RNAs, but also pair probabilities or even the energies of base pair stacking.

Recently, basic conceptual improvements to the Sankoff-style approach as introduced in `FOLDALIGN` [81] have been implemented. The first improvement was introduction of *sparsification*, in which not all computations of what correspond to the k index in the Equation 1 need to be carried out, as a number of configurations are the same, but obtained in different ways from composition of various sub-structures. The other improvement was a heuristic approach that basically *prunes* away cells in the dynamical programming matrix that never exceed a length-dependent threshold. This could be accomplished by filling out the dynamical programming matrix "ahead of time" (see Figure 10 for details).

Additional methods (not explicitly employed for ncRNA gene finding) have been published since and we refer to [42] for further details.

Whereas most methods perform *global* alignments, a few do local structural alignments. These include `FOLDALIGN` and `LocaRNA`, which conduct pairwise local structural alignments, as well as `CMfinder` [82].

RNA Structure-Based ncRNA *In Silico* Screens

Here, we describe the basic principles applied for the search of structured RNAs in genomic sequence and we refer to [42] for a detailed overview and discussion of the outcome. There are two main directions that have been applied for the *de novo* search for RNA structure, which is, as indicated above, a trade off between computational resources and the ability to explore the size of the search space. The two directions are, one that employs sequence-based alignments and one that also exploits synteny/orthology, but allows for structural (re-)alignment of the sequences. This is also sketched in Figure 11.
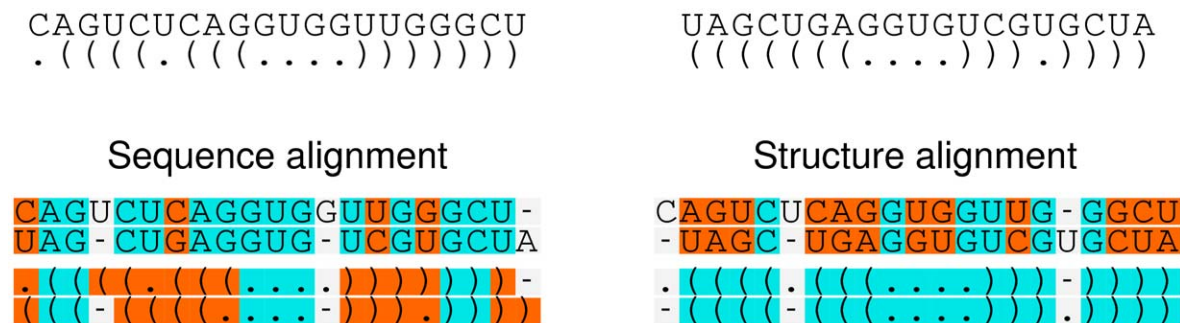

Figure 9. Two toy sequences that, if aligned only by their sequence, do not match in secondary structure. If correctly aligned, low sequence similarity between the two sequences does not hinder the revelation of structure.
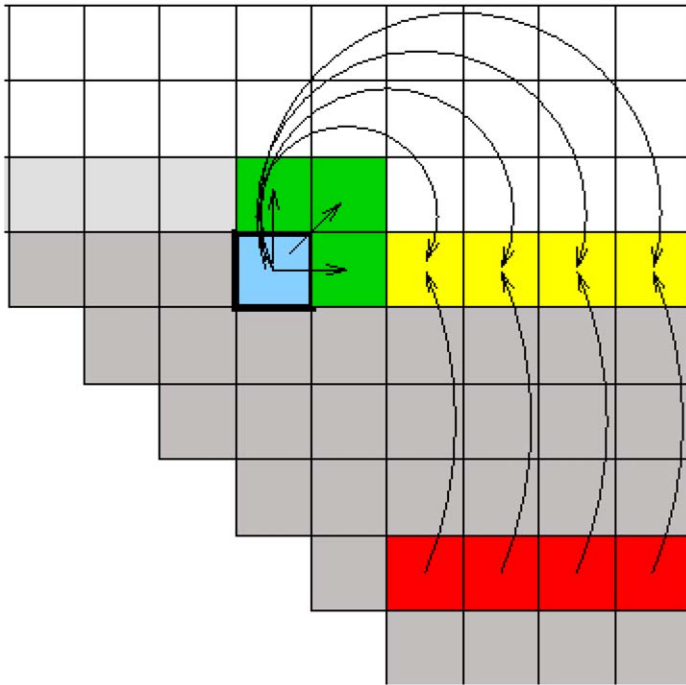doi:10.1371/journal.pcbi.1002100.g009

Figure 10. Filling out the dynamical programming matrix "ahead of time". That is, for the current position in the sequence just partially filling out future cells, either for the first time, or by updating the maximum score in the particular cell. All grey cells, including the blue cell and the current cell (i,j of a single sequence), have been completely computed. The green and yellow cells are partially filled out, making part use of the red cells (previously computed). (The figure is from the supplemental material of [81].)
doi:10.1371/journal.pcbi.1002100.g010

Screens on Sequence-Based Alignments

These screens are typically carried out by using a sliding window, that is, a pre-defined window of some size is moved along a set of multiple aligned genomes (typically MAF [Multiple Alignment Format] blocks from the UCSC browser [83]). The alignment is based on sequence similarity and the window slides a

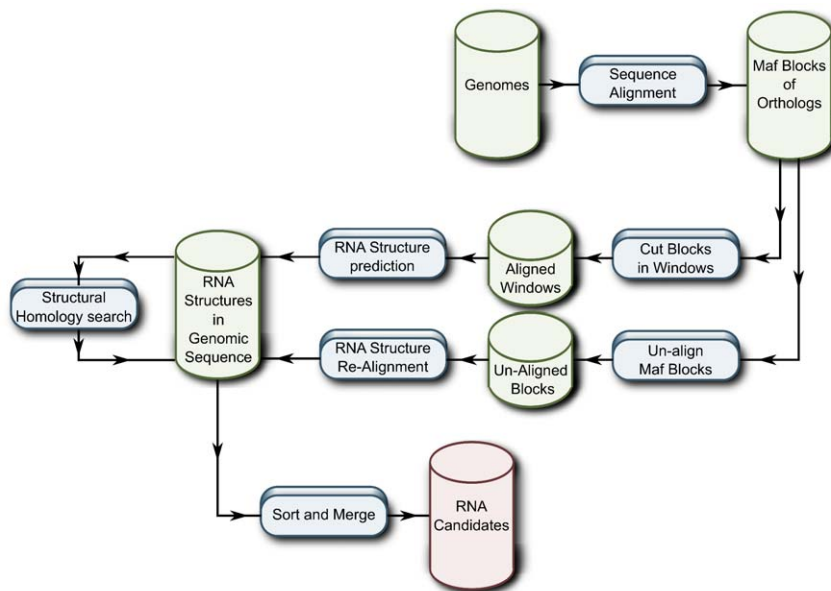**Figure 11. The basic flow of strategies for *de novo* prediction of RNA structures in genomic sequences.** Given the strategy of applying multiple organism sequences, orthologs are already obtained. For the homology search using the obtained *de novo* candidates, these can be compared in syntenic regions as for obtained homology candidates. (Figure courtesy of Christian Anthon.)
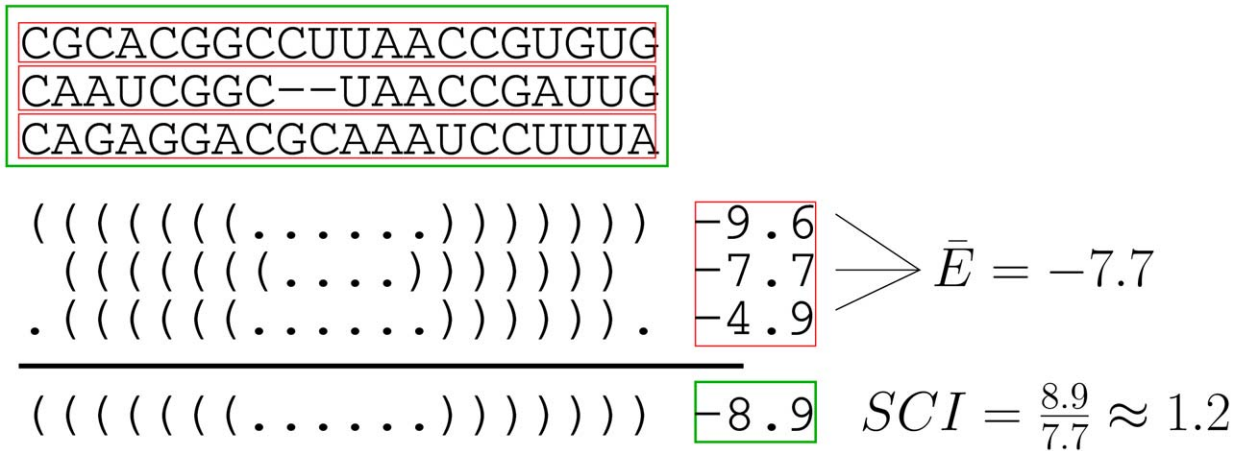doi:10.1371/journal.pcbi.1002100.g011

Figure 12. Computation of the SCI from a multiple alignment.
doi:10.1371/journal.pcbi.1002100.g012

number of nucleotides (e.g., half or quarter of the window size) in each step. In each window a consensus structure prediction is performed. By the end of the screen, various types of post processing are carried out, such as ranking the findings, estimating a false positive rate, determining strand specificity, and finding overlapping regions.

A potential drawback of the procedure is that results depend not only on the quality of the input alignments, but also on the windowing procedure. Windows should be large enough to fully cover ncRNAs (or at least a complete substructure), but should not be much larger than the smallest ncRNAs one wants to detect. A window size of, e.g., 120 nt, as has been used in RNAz screens (see below), is large enough to ensure that almost all miRNAs precursors will be detected. However, for maximum sensitivity, it can make sense to repeat screens using different window sizes.

An early reasonably successful attempt to predict structured RNAs from sequence alignments was qrna [84], which employed three different models of sequence evolution: a pair of HMMs describes the null model of sequences evolving without position dependent constraints, a second HMM that produces pairs of codons and models the evolution of protein coding sequences, and finally a pair of SCFGs is responsible for determining the evolution of sequence pairs with a common secondary structure. qrna computes the likelihood of the input alignment for each model, and identifies the model that yields the highest likelihood for the input alignment. qrna was successfully used to predict ncRNAs candidates in *E. coli* and *S. cerevisiae* [85,86], some of which were verified experimentally. A limitation of qrna is that it only works on pairwise alignments. With the more recent method, Evofold [87] tries to extend the qrna approach of model comparison to multiple alignments. It adopts the pfold approach of modelling the joint probability of consensus structure and alignment by combining a phylogenetic model (substitution process along the branches of a tree) with a simple SCFG to compute the *a priori* probability of a structure.

In contrast to the SCFG-based approaches, the AlifoldZ and RNAz programs are based on energy-directed folding. In [65] it was shown that (in contrast to single-sequence folding) the joint folding energy of real ncRNAs can be distinguished from the folding energies of randomized alignments. A natural measure to assess whether an RNA is unusually stable is to compute a *z*-score over folding energies $z = \frac{E - \mu}{\sigma}$ where μ and σ are the mean and standard deviation of randomized sequences obtained by shuffling.

The idea in AlifoldZ is simply to compute the *z*-score using the energies of consensus structures as returned by RNAalifold. This is straightforward except that it requires a method to randomize alignments. Simply shuffling columns would result in alignments with unusual gap and conservation patterns (e.g., many short gaps instead of a few longer gaps). AlifoldZ therefore uses a conservative shuffling where only columns that display the same gap pattern and similar conservation can be swapped.

The shuffling procedure, however, results in a somewhat slow procedure. RNAz [88] therefore aims to avoid shuffling altogether. It uses energy *z*-scores for single sequences only and combines it with a separate measure of structure conservation. Importantly, the *z*-scores for single sequences can be estimated, as it turns out that the mean energy μ and standard deviation σ are simply functions of the sequence length and composition. RNAz therefore uses a support vector machine (SVM) (for a tutorial, see e.g., [89]) to train regression models for μ and σ, which allows computation of *z*-scores with only a single call to the folding algorithm. The latest version of RNAz [90] improves detection accuracy by using a regression model based on di-nucleotide content rather than nucleotide frequencies. To quantify structural conservation, RNAz uses a *structure conservation index* (SCI), defined as the ratio of the energy returned from consensus structure prediction E_A divided by the average folding energy of the individual sequences $SCI = E_A/\bar{E}$, see Figure 12. Finally, a SVM takes the *z*-score and SCI as input and classifies the alignment (of the given window) as containing a significant RNA structure or not.

The Sankoff-based method Dynalign was applied in a screening approach using a fixed size window, but allowing for realignment (by Dynalign) and training of an SVM on such alignments. For low sequence similarity candidates (with identity less than 50%), it (not surprisingly) performs better than RNAz [91]. Subsequently, Dynalign has been optimized to lower its computational resources by employing an HMM for pre-processing the input and applying the HHM-based alignment as a constraint [74].

Local Searches

A local search for RNA structure deviates from that of sequence-based alignments in two main ways. Firstly, even though the alignment is used to indicate orthology or synteny, the alignment itself is ignored and the combined sequence structure approach is applied to the sequences. Secondly, the approach is

not bound by any window, so does not suffer from limitations such as adding too much flanking region and/or partial overlap to a real RNA structure, both of which can result in erroneous detection of RNA structures. In contrast, the local search approaches do not suffer from these limitations, but come with a set of their own to lower the computational overhead and make the methods practical. These limitations include a limited motif size, typically $\sim 200 - 500$ nt, though this might change in the future.

In the Sankoff-based approach `FOLDALIGN`, constraints other than those mentioned above made genome-wide screens possible. Two corresponding genomic sequences of lengths L_1 and L_2 were screened, but since the motif size was limited to size $\lambda \leq L_1, L_2$, it was only necessary to store a 4D matrix constrained by λ (typically ~ 200 nt) rather than the full (large) sequence lengths. Essentially, the dynamical programming matrix slides along the two genomes and for each position throws away elements corresponding to positions no longer included by the motif range while adding new ones. To screen (genomic) sequences, one of the sequences is chopped into pieces of size l_s, where a default value is $l_s = 2\lambda - 1$ and where two consecutive pieces overlap $\lambda - 1$ nucleotides. Without employing pruning, this doubles the running speed as compared to storing the entire 4D programming matrix in memory. This approach was applied to screen corresponding but unaligned sequences between human and mouse [92].

While the current local alignment version of `FOLDALIGN` is limited to two sequences, it is also of interest to conduct a screen involving multiple sequences. The program `CMfinder` [82] searches a set of unaligned sequences using seed structures found from energy folding. It aims exactly to do what is outlined in Figure 1. The principle is summarized in Figure 13 and holds significant overlap to the early SCFGs [70]. The candidates are used to construct an initial alignment from which a covariance model is constructed and used to make further searches. Additional findings are incorporated into the model and a new search is made until convergence is reached. As in the work of Eddy and Durbin, an expectation maximization (EM) algorithm was employed to find the optimal local structure. `CMfinder` was also recently applied to screen for ncRNAs in prokaryotes [93,94] and has been a main tool in riboswitch discovery, e.g., [94]. An additional strength is that if some of the sequences do not contain the RNA structure, they will simply be ignored, whereas the sequence alignment–based methods discussed above try to predict an RNA structure in all sequences.

An overview of the methods applied in *in silico* screens along with a short description of what they have been applied on can be found in [42].

False Discovery Rates

A main issue that comes with all the methods for *de novo* RNA structure searches is they have high false positive rates, around 50% [42]. Furthermore, a comparison of the ENCODE regions [95] that comprise one percent of the human genome show little overlap between `RNAz`, `Evofold`, and `CMfinder`. Even though the methods work in quite different ways, they all aim to fulfill the same task. This clearly shows that the area still needs to mature. A future direction is to improve the background model for the screens, e.g., by using di-nucleotide shuffling [90]. A major challenge lies in providing good background models for shuffling multiple alignments. Recent advances in that area include methods like `SISSIz` [96] and `Multiperm` [97].

The `Multiperm` program shuffles the multiple alignments, while preserving gap and local patterns of conservation, while also preserving the approximate di-nucleotide frequencies, which is a main concern. The `SISSIz` program simulates (using a phylogenetic substitution model) a multiple alignment with a given dinucleotide content and does preserve, on average, local conservation patterns and gap structure. To our knowledge, the two programs have not been systematically benchmarked, but in our experience they are of approximately the same quality (unpublished observations).

Performance Evaluation

Evaluating the performance of both RNA structure prediction and RNA gene finding is a subtle task. In both cases, a comparison to known (blinded to the experiment) data is required. RNA structure prediction is typically evaluated by comparison to curated structure data, e.g., [61]. From the number of (in)correctly predicted base pairs one computes accuracy measures, such as the positive predictive value (PPV) [98] and specificity, or Matthews correlation coefficient [99]. The latter is for RNA structure prediction well approximated by the geometric mean of the sensitivity (SEN) and PPV [100]. Note that the SCI measure is *not* suitable for performance evaluation, since it does not compare predictions to a blind dataset. SCI is a measure of divergence of the structures in the multiple alignment, and a high SCI does not necessarily imply correct performance, but merely states that the consensus structure is in good agreement with the structure of the individual sequences. Still, the entire structure prediction can be wrong.

For RNA gene finding, the genomic locations of predicted structures are compared to the locations of known RNAs (in blind dataset). Overlap of prediction and known gene (by some threshold) are used to state that a known RNA gene has been correctly predicted, see e.g., [81]. A major problem, however, is to measure the false positives, because a prediction in a given genomic location might indicate a so far unannotated ncRNA gene. What can be measured, however, is how many of the known ncRNA genes are missed in some benchmark dataset.

Discussion

Approaches for *de novo* and *in silico* searching for structured RNAs is a highly difficult task that exceeds "regular" finding of protein coding genes in complexity due to the lack of regular patterns (such as codon bias). Algorithms have to take long-range interactions (secondary structure) into account, and typically work in a comparative manner requiring several homologous sequences.

The current algorithmic approaches using sequence-based alignments are much faster than using structural alignments [42]; however, structural alignments can take regions with weak sequence conservation into account more accurately. An observation from the `CMfinder` screen on the ENCODE regions was that the `CMfinder` alignment was similar to the original alignment for MAF blocks with high sequence similarity, but showed significant rearrangement for low similarity blocks [101].

A major challenge is the quality of currently available genomic MAF alignments. Especially if the number of species is large, alignment blocks are quite short such that an ncRNA may be broken up into multiple blocks. In this case one can try to extend or merge MAF blocks without losing too many species. Often, MAF blocks appear to be broken by gaps in one organism (unpublished observations).

When screening for RNA structures in genomic sequence, the respective methods optimizes a scoring function and within that function seeks an optimal structure. However, the structure predicted might well be suboptimal for a number of reasons. These include inaccuracies of the energy model, kinetic folding
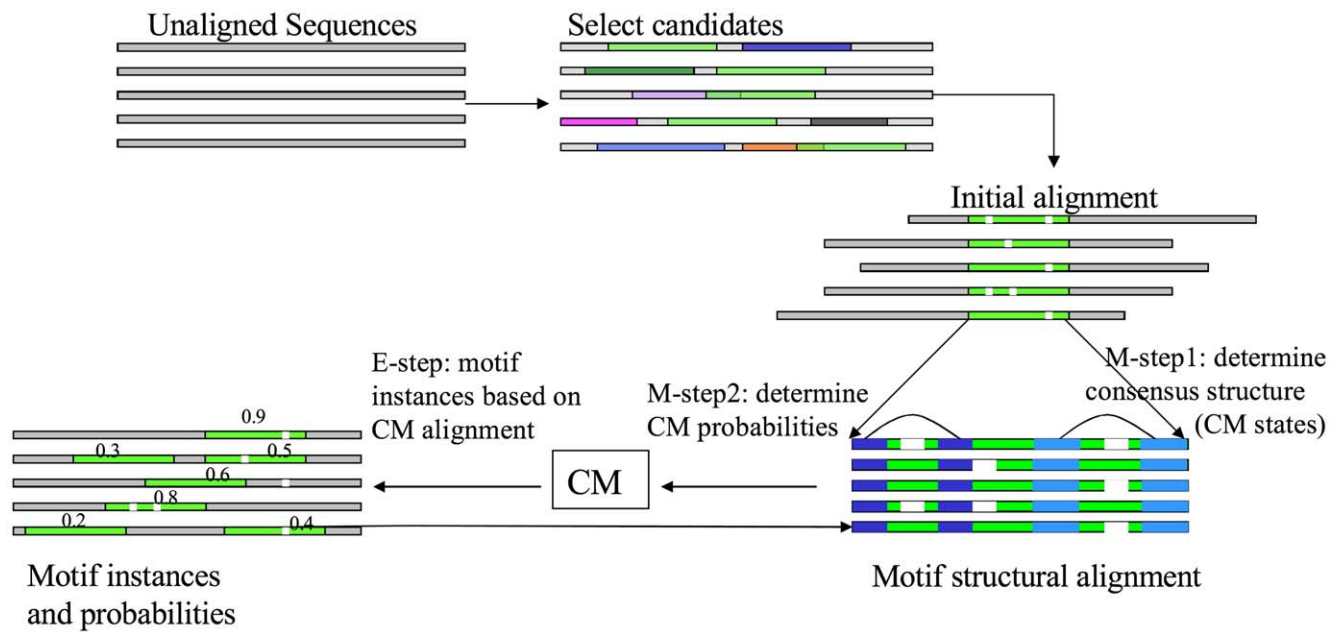
Figure 13. Searching unaligned sequences using `CMfinder`. After construction of an initial alignment (based on energy folded seeds), a covariance model is constructed and used to make further searches. Additional findings are incorporated into the model and novel searches are made until convergence was reached. (The figure was kindly provided by Zizhen Yao.)
doi:10.1371/journal.pcbi.1002100.g013

effects, as well as neglecting tertiary structure. Adding covariance information can dramatically improve the quality of structure prediction, but is dependent on the quality of the alignment. This is an issue in particular due to the limited quality of genome-wide alignments.

As the number of species grows, alignment blocks tend to become smaller. This imposes length constraints on the length of ncRNAs that can be detected, and may in the future require more sophisticated pre-processing of input alignments. In addition, the different search strategies have their own constraints on the length of their motifs (due to fixed window size, computational complexity, etc.) and thus have the same issues as for limited size MAF blocks. Currently, genomic screens typically result in a number of overlapping predictions, and the entire region is then merged into a candidate region for which there sometimes is not an entire structure prediction, e.g., [101,102]. In fact, an open challenge is to make a good strand discriminator, as a prediction on one strand can imply an almost equally good prediction on the other strand. Some work has been initiated in this area [103].

Compensating base pairs are clearly important, even though systematic analyses to study the impact have not been carried out. As discussed in [42], the overlap between `RNAz`, `EVOfold`, and `CMfinder` on the ENCODE regions [95] was poor, and a main difference was that a substantial amount of `CMfinder` candidates had more and more of the MAF blocks re-aligned as sequence similarity dropped, suggesting that compensating base changes are important in lesser regions. In a study of known RNAs from Rfam, it was concluded the that `MULTIz` alignments were relatively accurate, but with room for better alignments in a number of regions [104]. Thus, a factor contributing to the lack of detection of novel RNAs could be the ability to include compensating changes into the alignments.

Suboptimal structures have not yet been taken systematically into account in ncRNA gene finding methods. However, at least in some cases, they might be essential for the detection of functional RNAs.

Riboswitches, for example, are known to change conformation, and it is therefore expected that such types of information can add value to a genomic screen in general. To our knowledge, there have not been any systematic studies to compare predicted RNA structures from *in silico* screens with experimental data.

Simultaneously with the potential for RNA structure in the genome, a number of recent studies have shown the existence of long non-coding RNAs (lncRNAs), which are long transcripts. Presumably, these lncRNAs are largely unstructured [105]. However, recently, one of these lncRNAs was shown to have an enhancer-like function [106] that was coupled to the presence of short RNA structures in the lincRNA. These lncRNAs have been revealed to cover a variety functions [107], including epigenetic gene silencing [108], antisense regulation [109], and possibly chromatin organization, to promote long-range gene activation [110], to mention just a couple of examples. Over time there have been some attempts to distinguish coding from non-coding sequence on transcript. For a recent approach (post the lncRNA awareness), see [111].

Whether all lincRNAs contain local structured domains remains an open question. In [101] a functional RNA structure (67 nt) was predicted within a 2.8-kb ncRNA expressed in the brain, and subsequent studies revealed that this ncRNA also has overlap to `RNAz` predictions. Scenarios like this add to the challenge of arriving at full-length and/or functional transcripts from the RNA structure predictions, and it appears that RNA structure predictions cannot stand alone and will need to be accompanied by other types of data and possibly follow-up experiments to assign functional information.

Recently, exciting experimental developments have opened the arena for high-throughput structure probing on a transcriptome scale [112,113]. These methods promise to provide useful data that can complement the computational screens, but are still in their early phase, each with their own challenges. For example, none yet work *in vivo*. Other sources for probing data are also

promising to provide information applicable to a transcriptome-wide scale [114]. Incorporating such data in folding algorithms, including those used for genomic screens, will therefore be highly relevant. Emerging work in that area has recently been initiated [115].

References

1. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al. (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17: 669–681.
2. Mattick JS, Taft RJ, Faulkner GJ (2010) A global view of genomic information–moving beyond the gene and the master regulator. Trends Genet 26: 21–28.
3. Filipowicz W, Bhattacharyya S, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9: 102–114.
4. Guarnieri D, DiLeone R (2008) MicroRNAs: a new class of gene regulators. Ann Med 40: 197–208.
5. Georges M, Coppieters W, Charlier C (2007) Polymorphic miRNA-mediated gene regulation: contribution to phenotypic variation and disease. Curr Opin Genet Dev 17: 166–176.
6. Lindow M, Gorodkin J (2007) Principles and limitations of computational microRNA gene and target finding. DNA Cell Biol 26: 339–351.
7. Menzel P, Gorodkin J, Stadler PF (2009) The tedious task of finding homologous non-coding RNA genes. RNA 15: 2075–2082.
8. Lewis B, Burge C, Bartel D (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.
9. Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, et al. (2009) The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. Development 136: 525–530.
10. Sleutels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415: 810–813.
11. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 39: 925–938.
12. Ellis JC, Brown JW (2010) The evolution of rnase p and its rna. In: Liu F, Altman S, eds. The RNA world Springer Science + Business Media, LLC. pp 17–40.
13. Lau N, Seto A, Kim J, Kuramochi-Miyagawa S, Nakano T, et al. (2006) Characterization of the piRNA complex from rat testes. Science 313: 363–367.
14. Kim V (2006) Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. Genes Dev 20: 1993–1997.
15. Okamura K, Chung W, Ruby J, Guo H, Bartel D, et al. (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature 453: 803–806.
16. Park Y, Oh H, Meller VH, Kuroda MI (2005) Variable splicing of non-coding roX2 RNAs influences targeting of MSL dosage compensation complexes in Drosophila. RNA Biol 2: 157–164.
17. Gardner JPP. Daub, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. (2009) Rfam: updates to the RNA families database. Nucl Acids Res 37: D136–D140.
18. Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, et al. (2006) The tmRDB and SRPDB resources. Nucleic Acids Res 34: D163–D168.
19. Deng W, Zhu X, Skogerb? G, Zhao Y, Fu Z, et al. (2006) Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res 16: 20–29.
20. Warden C, Kim S, Yi S (2008) Predicted functional RNAs within coding regions constrain evolutionary rates of yeast proteins. PLoS ONE 3: e1559. doi:10.1371/journal.pone.0001559.
21. Hentze MW, Kühn LC (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. Proc Natl Acad Sci U S A 93: 8175–8182.
22. Nudler E, Mironov AS (2004) The riboswitch control of bacterial metabolism. Trends Biochem Sci 29: 11–17.
23. Consortium TF, the RIKEN Genome Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 66,770 full-length cDNAs. Nature 420: 563–573.
24. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329: 689–693.
25. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464: 1071–1076.
26. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106: 11667–11672.
27. Rivas E, Eddy S (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 7: 583–605.
28. Workman C, Krogh A (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucl Acids Res 27: 4816–4822.
29. Altschul SF, Madden TL, Sch?ffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucl Acids Res 25: 3389–3402.
30. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
31. Gorodkin J, Cirera S, Hedegaard J, Gilchrist MJ, Panitz F, et al. (2007) Porcine transcriptome analysis based on 97 non-normalized cDNA libraries and assembly of 1,021,891 expressed sequence tags. Genome Biology 8: R45.
32. Weinberg Z, Ruzzo WL (2004) Faster genome annotation of non-coding RNA families without loss of accuray. In: Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB), ACM Press. pp 243–251.
33. Nawrocki EP, Eddy SR (2007) Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput Biol 3: e56. doi:10.1371/journal.pcbi.0030056.
34. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.
35. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, et al. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucl Acids Res 29: 4724–4735.
36. Brown JW (1999) The Ribonuclease P Database. Nucleic Acids Res 27: 314.
37. Lowe T, Eddy S (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 25: 955–964.
38. Xie M, Mosig A, Qi X, Li Y, Stadler PF, et al. (2008) Structure and function of the smallest vertebrate telomerase RNA from teleost fish. J Biol Chem 283: 2049–2059.
39. Klein RJ, Misulovin Z, Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA 99: 7542–7547.
40. Schattner P (2002) Searching for RNA genes using base-composition statistics. Nucl Acids Res 30: 2076–2082.
41. Larsson P, Hinas A, Ardell DH, Kirsebom LA, Virtanen A, et al. (2008) De novo search for non-coding RNA genes in the AT-rich genome of Dictyostelium discoideum: performance of Markov-dependent genome feature scoring. Genome Research 18: 888–899.
42. Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, et al. (2010) De novo prediction of structured RNAs from genomic sequences. Trends Biotechnol 28: 9–19.
43. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ (1978) Algorithms for loop matching. SIAM J Appl Math 35: 68–82.
44. Eddy SR (2004) How do RNA folding algorithms work? Nat Biotechnol 22: 1457–1458.
45. Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38: D280–D282.
46. Mathews DH, Sabina J, Zuker M, Turner H (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Biol 288: 911–940.
47. 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: 105.
48. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., et al. (2009) Architecture and secondary structure of an entire hiv-1 rna genome. Nature 460: 711–716.
49. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. (2010) Genome-wide measurement of rna secondary structure in yeast. Nature 467: 103–107.
50. Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, et al. (2008) Variations on RNA folding and alignment: Lessons from benasque. J Math Biol 56: 129–144.
51. Fontana W, Konings DA, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33: 1389–1404.
52. Clote P, Ferré F, Kranakis E, Krizanc D (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11: 578–591.
53. Bonnet E, Wuyts J, Rouze P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20: 2911–2917.
54. Chiu DK, Kolodziejczak T (1991) Inferring consensus structure from nucleic acid sequences. CABIOS 7: 347–352.
55. Gutell RR, Woese CR (1990) Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. Proc Natl Acad Sci U S A 87: 663–667.

56. Gutell RR, Power A, Hertz GZ, Putz E, Stormo GD (1992) Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. Nucl Acids Res 20: 5785–5795.

57. Gorodkin J, Heyer LJ, Brunak S, Stormo GD (1997) Displaying the information contents of structural RNA alignments: the structure logos. CABIOS 13: 583–586.

58. Gorodkin J, Stærfeldt HH, Lund O, Brunak S (1999) Matrixplot: visualizing sequence constraints. Bioinformatics 15: 769–770.

59. Griffiths-Jones S (2005) RALEE–RNA ALignment editor in Emacs. Bioinformatics 21: 257–259.

60. Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M (2006) 4SALE–a tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics 7: 498.

61. Andersen ES, Lind-Thomsen A, Knudsen B, Kristensen SE, Havgaard JH, et al. (2007) Semiautomated improvement of RNA alignments. RNA 13: 1850–1859.

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

63. Felsenstein J (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. J Mol Evol 17: 368–376.

64. Seemann SE, Gorodkin J, Backofen R (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res 36: 6355–6362.

65. Washietl S, Hofacker IL (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.

66. Gardner P, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33: 2433–2439.

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

68. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5: 140.

69. Siebert S, Backofen R (2007) Methods for multiple alignment and consensus structure prediction of RNAs implemented in MARNA. Methods Mol Biol 395: 489–502.

70. Eddy S, Durbin R (1994) RNA sequence analysis using covariance models. Nucl Acids Res 22: 2079–2088.

71. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, et al. (1994) Stochastic context-free grammars for tRNA modeling. Nucl Acids Res 22: 5112–5120.

72. Gorodkin J, Heyer LJ, Stormo GD (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucl Acids Res 25: 3724–3732.

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

74. Harmanci AO, Sharma G, Mathews DH (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign. BMC Bioinformatics 8: 130.

75. Holmes I, Rubin GM (2002) Pairwise RNA structure comparison with stochastic context-free grammars. In: Altman RB, Dunker AK, Hunter L, Klein TE, eds. Proceedings of the 7th Pacific Symposium on Biocomputing. Available: http://psb.stanford.edu/psb-online/. Accessed. pp 163–174.

76. Holmes I (2005) Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics 6: 73.

77. Dowell RD, Eddy SR (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics 7: 400.

78. Hofacker IL, Bernhart SH, Stadler PF (2004) Alignment of RNA base pairing probability matrices. Bioinformatics 20: 2222–2227.

79. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 3: e65. doi:10.1371/journal.pcbi.0030065.

80. Klein RJ, Eddy SR (2003) RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4: 44.

81. Havgaard JH, Torarinsson E, Gorodkin J (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Computational Biology 3: e193. doi:10.1371/journal.pcbi.0030193.

82. Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder–a covariance model based RNA motif finding algorithm. Bioinformatics 22: 445–452.

83. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37: D755–D761.

84. Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2: 8.

85. Rivas E, Klein RJ, Jones TA, Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 11: 1369–1373.

86. McCutcheon JP, Eddy SR (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. Nucl Acids Res 31: 4119–4128.

87. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2: e33. doi:10.1371/journal.pcbi.0020033.

88. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A 102: 2454–2459.

89. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4: e1000173. doi:10.1371/journal.pcbi.1000173.

90. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF (2010) RNAZ 2.0: Improved noncoding RNA detection. Pac Symp Biocomput 15: 69–79.

91. Uzilov AV, Keegan JM, Mathews DH (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 7: 173.

92. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Research 16: 885–889.

93. Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, et al. (2007) A computational pipeline for high- throughput discovery of cis-regulatory noncoding RNA in prokaryotes. PLoS Comput Biol 3: e126. doi:10.1371/journal.pcbi.0010126.

94. Weinberg Z, Barrick J, Yao Z, Roth A, Kim J, et al. (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res 35: 4809–4819.

95. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1 Nature 447: 799–816.

96. Gesell T, Washietl S (2008) Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics 9: 248.

97. Anandam P, Torarinsson E, Ruzzo WL (2009) Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. Bioinformatics 25: 668–669.

98. Dowell RD, Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics 5: 71.

99. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochem Biophys Acta 405: 442–451.

100. Gorodkin J, Stricklin SL, Stormo GD (2001) Discovering common stem-loop motifs in unaligned RNA sequences. Nucl Acids Res 29: 2135–2144.

101. Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, et al. (2008) Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. Genome research 18: 242–251.

102. Washietl S, Hofacker IL (2007) Identifying structural non-coding RNAs using RNAz. In: Baxevanis A, Davison D, eds. Current protocols in bioinformatics Volume 1. John Wiley & Sons. doi:10.1002/0471250953.bi1207s19.

103. Reiche K, Stadler PF (2007) RNAstrand: reading direction of structured RNAs in multiple sequence alignments. Algorithms Mol Biol 2: 6.

104. Wang AX, Ruzzo WL, Tompa M (2007) How accurately is ncRNA aligned within whole-genome multiple alignments? BMC Bioinformatics 8: 417.

105. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.

106. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46–58.

107. Hung T, Chang HY (2010) Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol 7: 582–585.

108. Nagano T, Fraser P (2009) Emerging similarities in epigenetic gene silencing by long noncoding RNAs. Mamm Genome 20: 557–562.

109. Hawkins PG, Morris KV (2010) Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. Transcription 1: 165–175.

110. Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, et al. (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472: 120–124.

111. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, et al. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16: 1478–1487.

112. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, et al. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods 7: 995–1001.

113. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. Nature 467: 103–107.

114. Weeks KM (2010) Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol 20: 295–304.

115. Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A 106: 97–102.

116. Gorodkin J, Knudsen B (2000) RNA informatik. Naturens Verden 11–12: 2–9.