From consensus structure prediction to RNA gene finding

Stephan H Bernhart[a], Ivo L Hofacker[a,1]

[a] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

Abstract

Reliable structure prediction is a prerequisite for most types of bioinformatical analysis of RNA. Since the accuracy of structure prediction from single sequences is limited, one often resorts to computing the *consensus structure* for a set of related RNA sequences. Since functionally important RNA structures are expected to evolve much more slowly than the underlying sequences, the pattern of sequence (co-)variation can be exploited to dramatically improve structure prediction. Since a conserved common structure is only expected when the RNA structure is under selective pressure, consensus structure prediction also provides an ideal starting point for the *de novo* detection of structured non-coding RNAs.

Here we review different strategies for the prediction of consensus secondary structures, and show how these approaches can be used to predict non-coding RNA genes.

Key words: RNA, RNA secondary structure, non-coding RNA genes, RNA structure prediction, ncRNA gene finding

Email addresses: berni@tbi.univie.ac.at (Stephan H Bernhart), ivo@tbi.univie.ac.at (Ivo L Hofacker)

[1] Ivo Hofacker is best known as the maintainer of the widely used Vienna RNA package. Currently, he holds a position as associate professor at the University of Vienna.

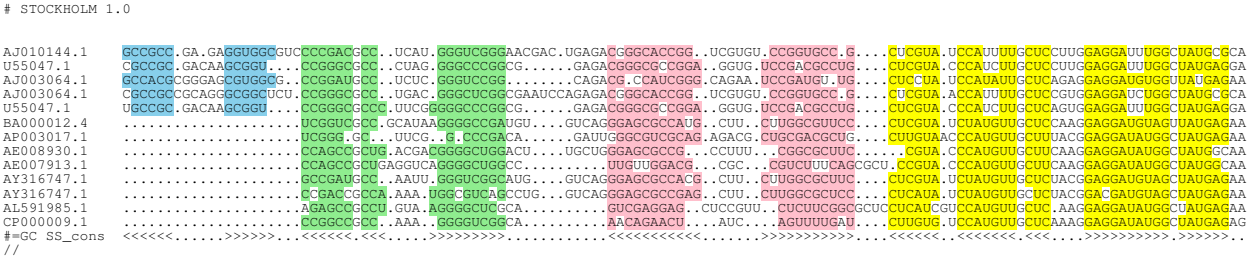

Figure 1: Stockholm format seed alignment of the ROSE – Repression of heat shock gene expression element (taken from Rfam 9.1 [7], id RF00435). The first stem is present in only 5/13 sequences, but still part of the Rfam consensus structure.

1. Introduction

In the last decade or so, the important roles that RNA has in all known organisms has been brought to the attention of life scientists. The identification of new classes of RNA molecules, such as miRNAs or piRNAs, has lead to the question of how to identify other non-coding RNAs that may be not as abundant as these RNA families. Basically, there are two possibilities to tackle that problem. The experimental approach, aided by the advent of next generation sequencing [1–3], aims to sequence the whole RNome of an organism. A limitation of this approach is however, that RNAs that are expressed only in specific tissues, developmental stages, or stress conditions, are likely to be missed. The bioinformatical approach usually depends on comparative genomics. Here, evolutionary conservation of secondary structure (or its absence) is used to decide whether a part of the genome is a non-coding RNA or not. When candidates for RNA genes have been found either experimentally or with computational means, the next task is to characterize them. A comparatively fast way to get information on a RNA molecule is its secondary structure. If one can find related sequences (mostly in related species), a consensus structure prediction is the best way to predict a secondary structure with computational means. In this contribution, we want to show the extensive parallels between in silico non coding RNA gene finding and RNA consensus structure prediction.

2. Consensus structure prediction

Correctly predicting the secondary (pseudo-knot free) structure of a single RNA molecule is a difficult task. While the basic algorithms used for these predictions have been around for a long time [4, 5], they suffer both from the simplifications of the underlying models as well as from the uncertainty concerning the exact values of the energy parameters. These problems lead to typical prediction accuracies (measured as the fraction of correctly predicted base pairs) of between 45% and 70% [6]. Luckily, in these days of next generation sequencing and ubiquitous genome projects, we can use another source of information besides the nucleotide sequence of a single RNA molecule. Evolutionary constraints on the function, and thus on the form of a molecule, can be used to find out more about its functional structure. The reasoning is simple: related RNA molecules with identical functions are thought to have identical or related structures. With this information, we try to find the best structure that the whole set of related molecules share. As an example, many tRNA sequences will not show the ”traditional” clover leaf shape when folded as a single molecule. A consensus structure prediction from a few sequences suffices to correctly identify the clover leaf structure. One should note however, that the notion of consensus structure is not entirely well defined. It is unclear if, e.g., structural insertions in some molecules should be considered to be part of the consensus structure or not, see Fig. 1. In particular, this makes it difficult to compare the quality of the different programs.

A benchmark comparing several prediction methods was published in 2004 [8], and the accompanying BRAliBase database has been widely used since. In fact, the availability of good reference data may have facilitated recent improvement of tools. An objective comparison of prediction accuracy is, however, complicated by the fact that in many cases program parameters have been optimized using the same data.

2.0.1. Covariance and covariance scores

Structural conservation means that while the sequence is changed, the ability to form a base pair is retained. For example, a conserved base pair may be realized as an AU pair in one sequence, and as a GC in another. This is known

as a *compensatory* mutation, and regarded as one of the hallmarks of conserved structures. Mutations where only one side of a base pair changes (such as GU to GC) are called *consistent* mutations and provide a much weaker signal. Many of the "classical" RNA structures, e.g. for tRNAs and ribosomal RNAs, were derived manually and purely on the basis of such co-variation signals.

To find out whether a substitution pattern is due to structural conservation, diverse measures can be used. The "classical" measure is the mutual information between two columns i and j of an alignment:

$$M_{ij} = \sum_{a,b \in \mathcal{A}} f_{i,j}(ab) \ln_2 \frac{f_{i,j}(ab)}{f_i(a) f_j(b)},$$

Here, $f_{i,j}(ab)$ is the frequency of co-occurrence of bases a and b in the two alignment columns, and $f_i(a)$ (resp. $f_j(b)$) the frequency of occurrence of base a (b) in the column i (j), and $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$. Because mutual information makes no use of RNA base pairing rules, it can be used to detect tertiary interactions just as well as canonical base pairs. For the same reason, the measure tends to be very noisy unless a large number (tens or hundreds) of sequences is available. To some extend this can be mediated by performing the sum over canonical base pairs only.

Another popular measure, that is more suitable for small sequence sets, is the *ad hoc* score of RNAalifold [9]. The alifold score simply counts the number of compensatory and consistent mutations between pairs of sequences, as well as the number of sequences that cannot form a base pair. More precisely it computes:

$$\begin{aligned}
\gamma(i,j) &= \gamma'(i,j) - \delta \sum_{\alpha \in \mathbb{A}} \begin{cases} 0 \text{ if } (\alpha_i \alpha_j) \in \mathcal{B} \\ 0.25 \text{ if } \alpha_i \wedge \alpha_j \text{ are gaps} \\ 1 \text{ otherwise} \end{cases} \\
\gamma'(i,j) &= \frac{1}{2} \sum_{\alpha,\beta \in \mathbb{A}, \alpha \neq \beta} \begin{cases} h(\alpha_i,\beta_i) + h(\alpha_j,\beta_j) & \text{if } (\alpha_i,\alpha_j) \in \mathcal{B} \wedge (\beta_i,\beta_j) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases},
\end{aligned}$$

where α, β are lines of the alignment $\mathbb{A}$, the hamming distance $h(a,b)$ equals 0 if $a = b$ and 1 otherwise, and $\mathcal{B} = \{\mathsf{AU}, \mathsf{UA}, \mathsf{CG}, \mathsf{GC}, \mathsf{GU}, \mathsf{UG}\}$ contains the canonical base pairs. The parameter δ weights the importance of counter examples, RNAalifold sets the default value to $\delta = 1$. The alifold score, as well as extensions that includes stacking of consecutive base pairs performed especially well in a survey by Lindgreen et al. [10].

In their RNA homology search tool RSEARCH [11], Klein and Eddy introduced the so called RIBOSUM pair substitution matrices, derived from structural alignments of SSU ribosomal RNAs. The substitution of nucleotides a and b in one sequences by c and d in another is scored by

$$R(ab,cd) = \log\left(f(ab;cd)/g(a)g(c)g(b)g(d)\right).$$

Thus, the RIBOSUM score is a log-odds ratio $\log(P/Q)$ for pairs of columns. The numerator $f(ab;cd)$ is the frequency with which ab pairs aligned with cd pairs are observed in paired columns of the SSU alignments, and the denominator Q is the probability of seeing the four nucleotides $\{a,b,c,d\}$ independently anywhere in the SSU sequences. It thus represents a statistical comparison between two models, conserved base pairs and independently evolving sites. Note that there are several other possibilities to define Q, such as probabilities to see unaligned base pairs (ab,cd) or alignment columns $acbd$. As with standard substitution scores, different matrices are needed for closely related and highly divergent sequences. Alternatively, a continuum of score matrices can be derived from a single instantaneous rate matrix, which can be estimated e.g. using XRATE [12]. The latter approach is in particular favoured by methods that make use of phylogenetic trees.

2.0.2. Approaches on consensus structure prediction

In general the problem of (structural) alignment and consensus structure prediction are closely related. Gardner and Giegerich [8] therefore defined three approaches to the problem of consensus structure prediction.

i Align first, then predict structure for the alignment

ii predict the structure for single sequences, then align these structures

iii align and predict the structure at the same time

However, subsequent work from the same group introduced a consensus shape prediction that does not fit into this scheme [13]. We therefore have grouped methods in the following into:

1 Stucture prediction from a fixed alignment

2 Methods that simultaneously compute alignment and structure

3 Alignment free methods

Furthermore, programs can be split into those that work globally and those that can also predict local structure.

Note that almost all programs mentioned below consider only pseudo-knot free structures, i.e. structure without crossing base pairs. Standard standard dynamic programming algorithms for RNA secondary structure prediction cannot deal with such structures, making pseudo-knot prediction a computationally hard problem. For the few programs that do support pseudo-knots, this is explicitly mentioned below.

2.1. Stucture prediction from a fixed alignment

Because multiple sequence alignments tend to be readily available, and because of the speed and convenience of the methods, structure prediction from a fixed alignment is still the most commonly used approach. The obvious drawback is that accuracy of the predicted consensus structure will be dependent on the quality of the alignment. In practice, popular sequence alignment methods like ClustalW [14] yield sufficiently good alignments for sequences with a similarity of 70% and above [15]. And for high quality, hand crafted alignments, such as Rfam alignments, most of the programs below give very good results.

In the following we have picked two programs, RNAalifold and Pfold, to be discussed in more detail, since they are widely used and form the basis for prominent ncRNA gene finders. The two programs also serve as representatives for the two prominent approaches to RNA structure prediction, namely energy directed folding and SCFGs. As this field is growing rapidly, we can not give a full list of all programs for consensus structure prediction, but present examples covering the different techniques.

2.1.1. Pfold

The use of stochastic context free grammars (SCFGs) for structure prediction is exemplified by the Pfold [16] program. Given an alignment A, Pfold first computes a phylogenetic tree T from the alignment. In conjunction with a sequence evolution model consisting of two rate matrices, a 4x4 matrix for the evolution of unpaired sites, and a 16x16 matrix for paired sites, the maximum likelihood approach of Felsenstein [17] allows to compute the probability $P(A|T,\sigma,M)$ of observing the alignment given the tree T, the secondary structure σ, and the rate model M. In order to obtain an optimal structure one also needs the prior probability of a secondary structure $P(\sigma)$, which is computed using a small SCFG. For reference the grammar with its production probabilities is given below:

$$\begin{aligned} S &\rightarrow LS\ (86.9\%) \,|\, L\ (13.1\%) \\ F &\rightarrow dFd\ (78.8\%) \,|\, LS\ (21.2\%) \\ L &\rightarrow s\ (89.5\%) \,|\, dFd\ (10.5\%) \end{aligned}$$

Here, the $L \rightarrow s$ rule produces an unpaired position, while the right hand side dFd produces two paired positions. Note that the Pfold SCFG is one of the smallest SCFGs that takes stacking of base pairs into account: A helix is initiated via $L \rightarrow dFd$ with a low probability of about 10%, while the probability to extend an existing helix (via $F \rightarrow dFd$) is close to 80%. By multiplying production probabilities of the SCFG with column probabilities, one obtains an extended SCFG that emits alignment columns, sometimes called a phylo-SCFG. The standard CYK algorithm can now be used to obtain the structure σ that maximizes $P(A|T,\sigma,M)P(\sigma)$, and therefore $P(\sigma|A,T,M)$.

2.1.2. RNAalifold

RNAalifold [9] faithfully implements the idea of "folding an alignment". In other words it is a generalization of the standard dynamic programming (DP) RNA folding algorithm as introduced by Zuker et al. [4] to alignments. To score the energy of a structural motif, the energy contributions of the single sequences are averaged, and a co-variation score is added to every base pair. This generalization to an alignment can also be done for McCaskill's [18] partition function variant of DP RNA folding. In a recent contribution [19], the prediction accuracy of RNAalifold has been improved by introducing a better treatment of gaps and RIBOSUM based co-variation scores.

Both Pfold and RNAalifold can be used to predict not only a single optimal structure, but also pair probabilities which provide a measure of confidence.

2.1.3. Other programs

ILM [20], iterative loop based matching, uses a Nussinov style loop matching algorithm. For alignments, scores for the base pairs within the algorithm are computed using mutual information scores. The best helix of the secondary structure is found, cut out of the alignment and the remaining parts of the alignment are iteratively folded again. This procedure makes it possible to predict pseudo knots in a comparatively fast way.

KNetFold [21] represents a machine learning approach to consensus structure prediction. It uses a k-nearest neighbor net to classify pairs of alignment columns as either paired or unpaired, based on three descriptors: mutual information, the fraction of compatible sequences, and average base pair probability. Several filters are employed to get to the final prediction, including a minimum helix length, as well as discarding all but the highest ranked pair for each base. The program can be used to predict structures with pseudo-knots.

BayesFold [22] uses Bayesian reasoning to combine information from different sources, in order select the best structure from a list of candidate structures, as provided, e.g., by RNAsubopt [23]. An interesting aspect is that it allows inclusion of chemical probing information in addition to thermodynamics and co-variation.

The McCaskillMEA approach [24] is closely related to RNAalifold, but does not use co-variation explicitly. Instead, it first computes base pair probabilities for each sequence, and from this the average pair probability for each pair of columns. A modified Nussinov algorithm is then used to compute the structure of maximum expected accuracy (in the simplest case the structure maximizing the sum of pair probabilities).

The same idea of superimposing base pair probability matrices has been used already early on in Alidot [25, 26] and Construct [27, 28]. Newer versions of Construct offer a sophisticated GUI that allows the user not only to predict and visualize consensus structures from a fixed alignment, but also to interactively edit and optimize the alignment.

All programs above have time and memory requirements that scale as $O(n^3)$ and $O(n^2)$, respectively, for alignments of length n. In practice, most have comparable runtimes, with KNetFold typically being the slowest.

2.2. Simultaneously computing alignment and structure

From the point of view of structure prediction, sequences that are homologous but highly diverged are ideal, since they should contain a maximum of covariance information. However, accurately aligning sequences becomes harder, and sequence alignments diverge more and more from the structurally correct alignment. In practice, pure sequence alignments become unsuitable for structure prediction at a pairwise sequence similarity of about 50%, at the latest. There is, however, a number of structure based alignment programs that can improve the performance of these alignments.

Today, the most popular approaches without a fixed alignment are variants of the Sankoff algorithm [29]. Some of these have been used as ncRNA gene finders (see below), they are however quite slow compared to methods on fixed alignemtns and therefore usually restricted to two sequences.

2.2.1. Sankoff based approaches

When David Sankoff introduced his algorithm for simultaneous folding and aligning in 1985, it was seen as a purely theoretical exercise since the cost in terms of CPU $O(n^6)$ and memory $O(n^4)$, for two sequences of length n, would make it impractical. Today, quite a number of implementations are available and becoming more widely used. At the heart of these implementations are heuristics that reduce the search space by restricting possible consensus structures, possible alignments, or both.

The earliest such attempt was Foldalign [30], which originally allowed only unbranched stem-loop structures (newer versions lift this restriction). Since Foldalign focuses on *local* alignments it also restricts the maximum length of the final alignment λ, as well as the length difference δ between the aligned sequence pieces, leading to a time complexity of $O(n^2\lambda^2\delta^2)$. The latest version [31] additionally introduces a *pruning* technique which discards sub-alignments whose score does not exceed a length-dependent threshold, removing them from the dynamic programming matrix.

Another early program, Dynalign [32], is noteworthy for implementing the full Turner energy model, as used for single sequence structure prediction. Thus, it tries to find the alignment and consensus structure that yields the best free energy averaged over the two sequences. Dynalign originally restricted alignments only by demanding that the length of two aligned subsequences differ by no more than a constant M. This limits the alignment path to a band close to the diagonal of the dynamic programming matrix.

In the SCFG based StemLoc program, Ian Holmes [33] replaces this fixed band by an *alignment envelope* and a *fold envelope*, restricting possible alignments and structures, respectively. The envelopes are computed by performing standard sequence alignments, and by folding the individual sequences. The envelopes are then constructed from the n_a best sequence alignments and the n_f best secondary structures, with n_a and n_f user settable parameters.

A different approach to restricting possible alignments is taken by the SCFG using Consan [34]. Here, high scoring local sequence alignments are pre-computed and used as "pins", i.e. matches that the resulting structural alignments must contain.

Similar to the fold envelopes above, PMcomp [35] pre-computes a matrix of pair probabilities for each sequence, and allows only pairs with a probability exceeding some threshold to be formed as part of the consensus structure. For a fixed probability threshold this already reduces runtime from $O(n^6)$ to $O(n^4)$, and the technique has since been adopted by many other tools, such as the latest version of Dynalign [36]. PMcomp/PMmulti was also one of the first tools to produce multiple alignments via progressive pairwise alignment. More efficient implementations of these ideas are nowadays available in FoldalignM [37] and LocARNA [38]. LocARNA in particular uses the restriction of possible pairs to not only speed up the algorithm, but also to reduce memory requirements from $O(n^4)$ to $O(n^2)$. In addition, LocARNA allows global and local alignments, as well as *structure local* alignments which allow for insertion/deletion of whole substructures.

While LocARNA restricts only the fold space, the recent RAF program [39] adds a simultaneous restriction of possible alignments: A pre-processing step uses sequence alignment algorithms to compute match probabilities between any positions in the two sequences. The Sankoff phase then excludes low probability matches from the search space leading to an expected runtime of $O(n^2)$.

Sankoff based methods have improved rapidly both in terms of speed and accuracy, sometimes making benchmarks obsolete within a year. An important distinction for the enduser is, however, that not all programs offer local alignment modes and multiple alignments, see Table 1 below. With the exception of Dynalign, most programs focus more on alignment quality than structure prediction quality. Even though they usually provide both alignment and consensus structure as output, it can often be beneficial to recompute the consensus structure based on this structural alignment using conventional tools such as RNAalifold.

2.2.2. Non-Sankoff approaches to structural alignment

Given the high computational cost of the Sankoff based methods, it is natural to consider heuristics that avoid this type of algorithm altogether. A very early approach is using a genetic algorithm to optimize the alignment and the structure [40].

CARNAC [41] is a stem-based consensus structure prediction tool that is divided into three steps. In the first step, a dynamic programming secondary structure prediction is used to identify all potential stems of all sequences. The second step consists of pairwise comparisons of the sequences to select the best stems. Highly conserved anchor points for the pairwise alignments restrict the number of possible stem to stem alignments. Furthermore, co-variations are required for stems to be aligned, which further reduces the number of possible stem matches. Using a DP algorithm, the optimal secondary structure for every pair of sequences is found. A final consensus structure is then computed by greedily combining stems that are weighted according to a stem graph.

RNA Sampler [42] first computes the probabilities that two bases are aligned and the base pair probability of the single sequences. It then aligns single stems, which are defined as at least three consecutive base pairs. After aligning all pairs of stems, the best aligned stems are kept and compatible aligned stems (called blocks) are combined to create

a secondary structure. After that, alignment probabilities and base pairing probabilities are updated, and the procedure is iteratively repeated.

MASTR [43] iteratively improves the structure prediction as well as the sequence alignment of a set of RNA sequences using simulated annealing with moves that change either sequence alignment or consensus structure. The cost function being minimized is composed of sequence conservation, co-variation and base-pairing probabilities.

Simulfold [44] is similar in spirit, but more ambitious. Given sequence data D it simultaneously optimizes consensus structure S, alignment A, and the phylogenetic tree T using a Markov chain Monte Carlo method, thus sampling S, A, T from the posterior probability $P(S, A, T|D)$. The move set consists of changes to branch lengths, tree topology, or compound changes of structure and alignment. While the time for a single simulation step is linear in sequence length and number of sequences, a large number of steps is needed until the simulation converges. To speed up convergence of the procedure can be quite slow, it can be helpful to include data like a known alignment, evolutionary tree or secondary structure and restrict the algorithm to run on the unknown parts. Since the method is not based on dynamic programming, it can include pseudo-knots in the prediction.

CMfinder [45] is a tool for finding RNA motifs. Thus, rather than performing a global alignment of the sequences it identifies structurally similar subsequences. It first predicts a list of local structures (motifs) by folding all substrings of each sequence, and weights them by the energy divided by motif length. It then performs pairwise structure comparisons between the motifs from all sequences, and picks the most "central" motif, i.e. the one with highest similarity to motifs in other sequences. An initial alignment is built by aligning the central motif to its closest match from each other sequence. The algorithm then iteratively improves the alignment by (i) computing a consensus structure (ii) translating the alignment with consensus structure into a covariance model, CM [46] and (iii) aligning each sequence against the CM to obtain a new alignment. The procedure can optionally be repeated with another initial motif.

2.3. Alignment free methods

RNAcast [13] is special in the sense that it does not require or build an alignment. The approach is based on predicting coarse grained structures, so called *abstract shapes*. An example abstract shape would be the "cloverleaf" shape, encompassing all structures with three hairpins and an enclosing multi-loop. For each sequence the RNAshapes program is used to compute all shapes with energies within some interval of the optimum. Since the number of shapes is so much smaller than the number of full structures, the same shapes are expected to occur in many of the predictions. The consensus shape is then simply the highest ranked shape common to all sequences. Once a consensus shape has been determined one can ask for the best full structure of this shape for each sequence (the so-called shape representative or *shrep*). If an alignment is desired, this can be computed *post factum* by aligning the structures.

The web server WAR [47] provides an easy to use platform to simultaneously use many of the structural alignment methods mentioned above. It makes it possible to compare many predictions and also to use a "majority vote" approach. Thus, it is very well suited for non experts to quickly generate alignments and consensus secondary structures.

3. RNA gene finding

As pointed out by Rivas & Eddy [48], secondary structure prediction on a single sequence is insufficient to reliably predict ncRNA genes. Therefore, the reverse of the reasoning applied to the consensus structure prediction problem is used for non-coding gene prediction in silico. If a structure is evolutionary conserved in spite of sequence variation, then the structure must be subject to selection and thus be functional. Consensus structure prediction is therefore an ideal starting point for ncRNA prediction, but has to be augmented by a suitable measure of significance.

For SCFG based methods, the natural approach is to do model comparison between a model for structured RNA (which also yields the consensus structure prediction), and a null model describing the genomic background. Alternatively, one can extract signals indicating functional structures from the prediction, which can then be used as descriptors for a machine learning approach. A recent comparison of such signals is given in Gruber et al. [49] (see also below for SCI).

Strictly speaking, all methods discussed below predict functional RNA structures, not ncRNA *genes*. This means that unstructured ncRNAs are generally undetectable for these methods. Moreover, the ends of the detected structures need not coincide with transcript boundaries. Cis-regulatory structures, for example, can be identified even though they are not independent transcripts.

Furthermore, we discuss only the *ab initio* prediction of novel ncRNA genes. In order to find new members of already known ncRNA families, one would rather resort to (structure based) homology search methods, see the review of Mosig et al. in this issue.

3.1. Alignment based methods

The first practical tool for *de novo* ncRNA gene finding was QRNA [50]. The program takes as input a pairwise sequence alignment that is then analyzed by three probabilistic models: A pair SCFG, i.e. an SCFG that emits columns of a pairwise alignment, is used to compute the probability that the input data are due to an underlying secondary structure, a hidden Markov model (HMM) emitting aligned codon pairs checks whether the input alignment might represent a protein coding region, and another HMM represents the null model of independently evolving columns. The model that yields the highest likelihood of the input data is then declared the winner. In practice, the biggest shortcoming of QRNA is that it can be used only on pairwise alignments.

The limitation to pairwise alignments is lifted in EvoFold [51], which also serves as an illustrative example for the close relation between the non-coding gene finding and consensus structure prediction. EvoFold implements an SCFG based on the one in Pfold, but poses a slightly different question: Structure prediction asks for the most likely structure given the alignment, while ncRNA detection asks for the likelihood of the alignment given a structural evolution model. As in QRNA this is then compared to the likelihood of the alignment in a null model for uncorrelated evolution, the final score being the log likelihood ratio between the two models. Besides the original scan in human, EvoFold was, e.g., used to scan the genome of 12 drosophilids for ncRNA genes [52].

The RNA-decoder tool [53] is similar to EvoFold, but is intended to detect regions with functional structure within a longer alignment. Rather than scoring the alignment using two different models, it employs a high-level grammar that switches between two sub-models for structured and un-structured parts of the alignment. It is noteworthy for being the only tool that explicitly models RNA structures that overlap protein-coding regions, as are frequently observed in RNA viruses.

AlifoldZ [54] and RNAz [55] are directly based on consensus structure prediction from RNAalifold. In the case of AlifoldZ the RNAalifold energy E_{Alif} for the input alignment is compared to the energies of randomized alignments produced by shuffling. This is done by computing a z-score $z = (E_{\mathrm{Alif}} - < E >)/\sigma$, where $< E >$ and σ are the mean and standard deviation over the randomized alignments.

RNAz uses a machine learning technique, a support vector machine (SVM), for the final decision whether or not the input alignment harbors a structural RNA. From folding the individual sequences as well as consensus structure prediction, it extracts two important descriptors: The "structure conservation index" (SCI) as a measure of structural conservation, and the average energy z-score of the individual sequences as a measure of thermodynamic stability. The SCI is computed as

$$\mathrm{SCI} = \frac{E_{\mathrm{Alif}}}{\frac{1}{N}\sum_i E_i}$$

Thus, if all sequences will fold into the same structure anyway, the SCI equals 1 (or slightly above if there is covariation), while it approaches zero if no common structure can be formed. Rather than computing the z-score via explicit shuffling (as in the case of AlifoldZ), RNAz uses a second SVM to estimate mean and standard deviation from sequence length and composition. This makes RNAz much faster and better suited for large genome wide screens. Small RNAz screens, e.g., on bacterial genomes can even be performed on-line at the RNAz web server [56]. RNAz has been used in a number of ncRNA screens, including an initial screen in humans [57], but also nematodes [58], plasmodium [59] and arabidopsis [60].

The z-score computations above crucially depend on a method to generate randomized sequences and alignments, and most ncRNA screens have used randomized alignments to estimate their false discovery rate. For single sequences, it is generally recommended to use di-nucleotide shuffling, i.e. using randomized sequences with the same di-nucleotide content. Alignments, however, can generally not be shuffled, while simultaneously preserving di-nucleotide content, gap structure, and the local degree of conservation. An alternative to shuffling, is to generate randomized alignments by simulating sequence evolution along a phylogenetic tree. Two such frameworks were recently introduced in [61], where it was shown that di-nucleotide corrected null data can improve AlifoldZ predictions, and [62] who observe that less realistic null models generally lead to underestimating the false positive rates of gene finders.

Program	web server	source code	model	features/limitations
Pfold	y	on request	SCFG	
RNAalifold	y	y	Turner	
ILM	y	-	Turner	pseudoknots
KNetFold	y	register	Turner	pseudoknots
BayesFold	IE only	-	Turner	can include probing data
McCasklill MEA	-	y	Turner	
ConStruct	-	y	Turner	interactive, GUI
Foldalign	y	y	Turner	pairwise only, local
Dynalign	-	register	Turner	pairwise only
StemLoc	-	y	SCFG	local
Consan	-	y	SCFG	pairwise only
FoldalignM	-	y	Turner	
LocaRNA	y	y	Turner	local
RAF	-	y	CLLM	
CARNAC	y	y	Turner	
RNASampler	-	register	Turner/CLLM	pseudoknots
MASTR	y	y	Turner	
Simulfold	-	y	Entropy based	pseudoknots
CMfinder	y	y	Turner	local only
RNAcast	y	y	Turner	
Gene finders				
QRNA	y	y	SCFG	pairwise only
EvoFold	-	y	SCFG	
RNA-Decoder	-	-	SCFG	for coding regions
RNAz	y	y	Turner	

Table 1: List of the programs, the availability of source code and web service, and the type pf model underlying structure prediction. CLLM:conditional log-linear models (trained parameters), SCFG: Stochastic context free grammar, Turner: free energy.

3.2. Predictions based on structural alignments

Non-coding RNAs are often characterized by very fast rates of sequence evolution, making it difficult to obtain reliable sequence based alignments. It is therefore tempting to base ncRNA screens on some type of structural alignment. Note, however that an initial (sequence based) alignment is needed to define syntenic regions that can then be re-aligned. The first such attempt was a Foldalign based screen of human versus mouse [63]. Since Foldalign performs local alignments in a scanning fashion, ncRNA candidates were identified by simply selecting those local alignments with an exceptionally good score. At the time, the project was a tour de force, requiring about 5 months on a cluster of 70 CPUs. Current implementations should be able to repeat the experiment in less than 1/10th of the time. Nevertheless, it demonstrated that ncRNAs can be found in regions that are not alignable on the sequence level.

Similarly, Dynalign has been used for ncRNA screens in prokaryotes [64]. As in RNAz, a support vector machine was trained to identify ncRNAs based on the Dynalign output. The SVM descriptors were the Dynalign folding energy, the nucleotide composition, and length of the two sequences. As expected, the approach yields significantly better predictions than RNAz for sequences with less than 50% sequence identity, albeit at significantly higher computational cost.

More recently, CMfinder was used in a screen of the ENCODE regions of the human genome [65]. While Foldalign and Dynalign are restricted to pairwise comparisons, CMfinder could make full use of the large number of sequenced species for these regions. Candidates were selected primarily on the basis of an ad hoc score combining global and local sequence similarity, fraction of paired bases in the motif, and number of sequences that can realize the motif.

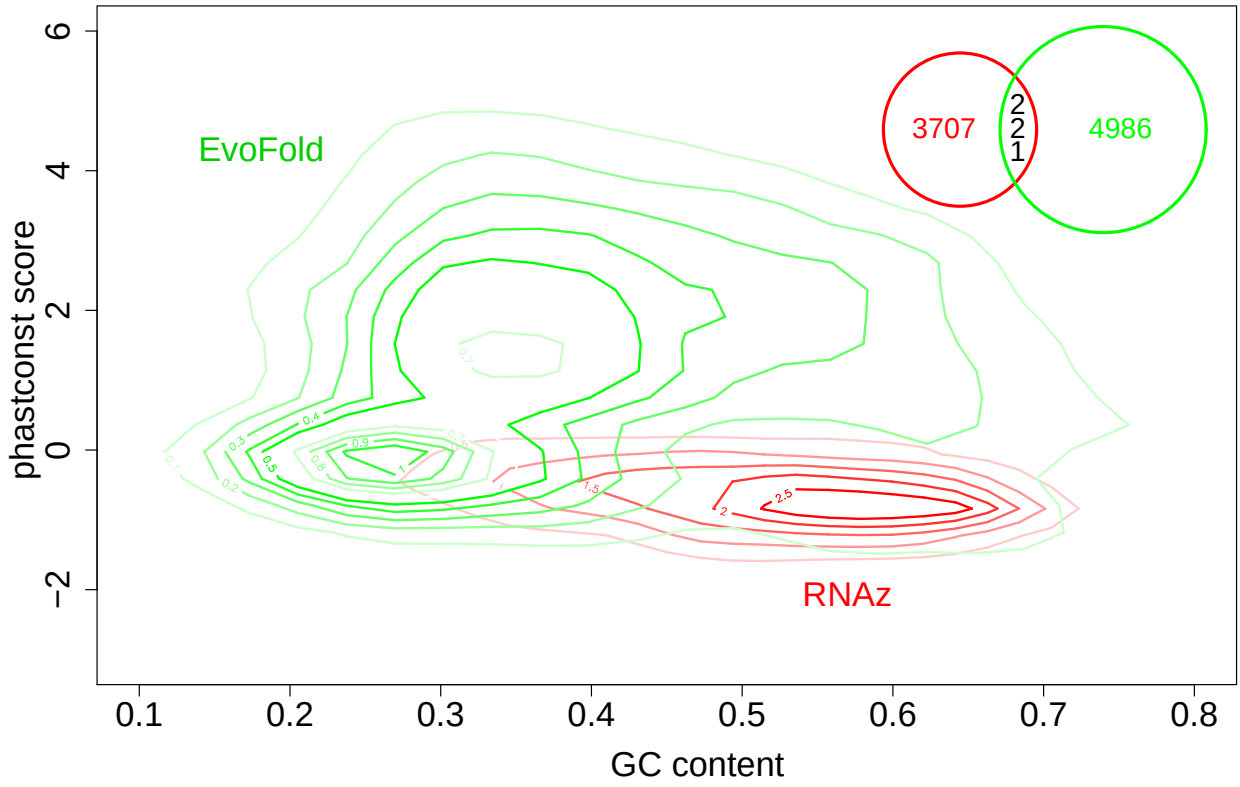

Figure 2: Comparison of EvoFold and RNAz predictions for the ENCODE regions of the human genome. The densities of the predictions with high significance is shown as a function of GC content and conservation measured by the phastCons program. RNAz is best at detecting ncRNAs with slightly elevated GC content and significant sequence variation. Evofold is most sensitive at low GC contents and high sequence conservation. The intersections of the prediction results is therefore necessarily small (Venn diagram in upper right corner). Figure modified from [66].

4. Discussion

A variety of different approaches has been brought to bear on the problem of consensus structure prediction. Given a high quality input alignment several of the available methods achieve prediction accuracies around 90%. Benchmarks should however be taken with a grain of salt, since (i) different programs employ slightly different notions about what constitutes a consensus structure, and (ii) reference structures have often been computed using tools very similar to those being benchmarked. Since ease of use and availability can be important factors in the choice of tool, a summary is provided in table 1.

Structural alignment methods have made remarkable progress over the last years. The Sankoff algorithm, long thought to be computationally unfeasible, is now available in several implementations that are fast enough for large scale use.

All these encouraging results are however obtained for relatively homogeneous families of RNAs with little to no structural variation within the family. For heterogeneous families with significant structural plasticity manual construction of alignments and consensus structures is still the rule. None of the programs discussed here would, for example, be much help in aligning 7SK sequences from mammals and insects [67].

The situation is also different in the field of ncRNA gene finders. While there exist several approaches with broadly similar performance, the overlap between predictions obtained in the same screens from different tools is surprisingly small [65, 66]. In part this may be due to the significant false positive rate, however a closer comparison of RNAz and Evofold predictions reveals that they are maximally sensitive in different and largely disjoint genomic regions (see fig. 2). This suggests that the true complement of ncRNAs in the human genome might still be underestimated by current gene-finders.

5. Key points

- Non-coding RNA gene finding and RNA consensus structure prediction are key problems in modern bioinformatics.
- For consensus structures, several different approaches lead to high accuracy predictions.
- Structural alignment tools improved a lot over the last few years.
- Today, gene finders are based on consensus structure prediction tools.
- Unstructured non-coding RNA genes can thus not be predicted.

References

[1] Margulies, M, Egholm, M, Altman, WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature Sep 2005:437(7057):376–380.

[2] Bennett, ST, Barnes, C, Cox, A, et al. Toward the 1,000 dollars human genome. Pharmacogenomics Jun 2005: 6(4):373–382.

[3] Schuster, SC. Next-generation sequencing transforms today's biology. Nat Methods Jan 2008:5(1):16–18.

[4] Zuker, M, Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res Jan 1981:9(1):133–148.

[5] Nussinov, R, Jacobson, AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A Nov 1980:77(11):6309–6313.

[6] Doshi, KJ, Cannone, JJ, Cobaugh, CW, et al. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics Aug 2004: 5:105–105.

[7] Gardner, PP, Daub, J, Tate, JG, et al. Rfam: updates to the RNA families database. Nucleic Acids Res Jan 2009: 37(Database issue):136–140.

[8] Gardner, PP, Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics Sep 2004:5:140–140.

[9] Hofacker, IL, Fekete, M, Stadler, PF. Secondary structure prediction for aligned RNA sequences. J Mol Biol Jun 2002:319(5):1059–1066.

[10] Lindgreen, S, Gardner, PP, Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics Dec 2006:22(24):2988–2995.

[11] Klein, RJ, Eddy, SR. Rsearch: finding homologs of single structured RNA sequences. BMC Bioinformatics Sep 2003:4:44–44.

[12] Klosterman, PS, Uzilov, AV, Bendaña, YR, et al. Xrate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics 2006:7:428.

[13] Reeder, J, Giegerich, R. Consensus shapes: an alternative to the sankoff algorithm for RNA consensus structure prediction. Bioinformatics Sep 2005:21(17):3516–3523.

[14] Thompson, JD, Higgins, DG, Gibson, TJ. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res Nov 1994:22(22):4673–4680.

[15] Wilm, A, Mainz, I, Steger, G. An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol Biol 2006:1:19–19.

[16] Knudsen, B, Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res Jul 2003:31(13):3423–3428.

[17] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981: 17(6):368–376.

[18] McCaskill, JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers May-Jun 1990:29(6-7):1105–1119.

[19] Bernhart, SH, Hofacker, IL, Will, S, et al. Rnaalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 2008:9:474–474.
[20] Ruan, J, Stormo, GD, Zhang, W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics Jan 2004:20(1):58–66.
[21] Bindewald, E, Shapiro, BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA Mar 2006:12(3):342–352.
[22] Knight, R, Birmingham, A, Yarus, M. Bayesfold: rational 2 degrees folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. RNA Sep 2004:10(9):1323–1336.
[23] Wuchty, S, Fontana, W, Hofacker, IL, et al. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 1999:49:145–165.
[24] Kiryu, H, Kin, T, Asai, K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. Bioinformatics Feb 2007:23(4):434–441.
[25] Hofacker, IL, Fekete, M, Flamm, C, et al. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. Nucl Acids Res 1998:26:3825–3836.
[26] Hofacker, IL, Stadler, PF. Automatic detection of conserved base pairing patterns in RNA virus genomes. Comp & Chem 1999:23:401–414.
[27] Lück, R, Gräf, S, Steger, G. Construct: a tool for thermodynamic controlled prediction of conserved secondary structure. Nucleic Acids Res Nov 1999:27(21):4208–4217.
[28] Wilm, A, Linnenbrink, K, Steger, G. Construct: Improved construction of RNA consensus structures. BMC Bioinformatics 2008:9:219.
[29] Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math 1985:45:810–825.
[30] Gorodkin, J, Heyer, LJ, Stormo, GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res Sep 1997:25(18):3724–3732.
[31] Havgaard, JH, Torarinsson, E, Gorodkin, J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol Oct 2007:3(10):1896–908.
[32] Mathews, DH, Turner, DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol Mar 2002:317(2):191–203.
[33] Holmes, I. Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics 2005:6:73 [epub].
[34] Dowell, RD, Eddy, SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics 2006:7:400.
[35] Hofacker, IL, Bernhart, SH, Stadler, PF. Alignment of RNA base pairing probability matrices. Bioinformatics Sep 2004:20(14):2222–2227.
[36] Harmanci, AO, Sharma, G, Mathews, DH. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign. BMC Bioinformatics 2007:8:130.
[37] Torarinsson, E, Havgaard, JH, Gorodkin, J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics Apr 2007:23(8):926–32.
[38] Will, S, Reiche, K, Hofacker, IL, et al. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comp Biol 2007:3:e65.
[39] Do, CB, Foo, CS, Batzoglou, S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics Jul 2008:24(13):68–76.
[40] Chen, JH, Le, SY, Maizel, JV. Prediction of common secondary structures of RNAs: a genetic algorithm approach. Nucleic Acids Res Feb 2000:28(4):991–999.
[41] Perriquet, O, Touzet, H, Dauchet, M. Finding the common structure shared by two homologous RNAs. Bioinformatics Jan 2003:19(1):108–116.
[42] Xu, X, Ji, Y, Stormo, GD. RNA sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. Bioinformatics Aug 2007:23(15):1883–1891.
[43] Lindgreen, S, Gardner, PP, Krogh, A. Mastr: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics Dec 2007:23(24):3304–3311.
[44] Meyer, IM, Miklós, I. Simulfold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a bayesian mcmc framework. PLoS Comput Biol Aug 2007:3(8).

[45] Yao, Z, Weinberg, Z, Ruzzo, WL. Cmfinder–a covariance model based RNA motif finding algorithm. Bioinformatics Feb 2006:22(4):445–452.
[46] Eddy, SR, Durbin, R. RNA sequence analysis using covariance models. Nucleic Acids Res Jun 1994: 22(11):2079–2088.
[47] Torarinsson, E, Lindgreen, S. War: Webserver for aligning structural RNAs. Nucleic Acids Res Jul 2008:36(Web Server issue):W79–84.
[48] Rivas, E, Eddy, SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics Jul 2000:16(7):583–605.
[49] Gruber, AR, Bernhart, SH, Hofacker, IL, et al. Strategies for measuring evolutionary conservation of RNA secondary structures. BMC Bioinformatics 2008:9:122.
[50] Rivas, E, Eddy, SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2001:2:8–8.
[51] Pedersen, JS, Bejerano, G, Siepel, A, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol Apr 2006:2(4).
[52] Stark, A, Lin, MF, Kheradpour, P, et al. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. Nature Nov 2007:450(7167):219–232.
[53] Pedersen, JS, Meyer, IM, Forsberg, R, et al. A comparative method for finding and folding rna secondary structures within protein-coding regions. Nucleic Acids Res 2004:32(16):4925–36.
[54] Washietl, S, Hofacker, IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 2004:342:19–39.
[55] Washietl, S, Hofacker, IL, Stadler, PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A Feb 2005:102(7):2454–2459.
[56] Gruber, AR, Neuböck, R, Hofacker, IL, et al. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Res Jul 2007:35(Web Server issue):W335–8.
[57] Washietl, S, Hofacker, IL, Lukasser, M, et al. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol Nov 2005:23(11):1383–1390.
[58] Missal, K, Zhu, X, Rose, D, et al. Prediction of structured non-coding RNAs in the genomes of the nematodes caenorhabditis elegans and caenorhabditis briggsae. J Exp Zoolog B Mol Dev Evol Jul 2006:306(4):379–392.
[59] Mourier, T, Carret, C, Kyes, S, et al. Genome-wide discovery and verification of novel structured RNAs in plasmodium falciparum. Genome Res Feb 2008:18(2):281–292.
[60] Song, D, Yang, Y, Yu, B, et al. Computational prediction of novel non-coding RNAs in arabidopsis thaliana. BMC Bioinformatics 2009:10 Suppl 1.
[61] Gesell, T, Washietl, S. Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics 2008:9:248.
[62] Varadarajan, A, Bradley, RK, Holmes, IH. Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. Genome Biol Oct 2008:9(10):R147.
[63] Torarinsson, E, Sawera, M, Havgaard, JH, et al. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res Jul 2006:16(7):885–889.
[64] Uzilov, AV, Keegan, JM, Mathews, DH. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 2006:7:173–173.
[65] Torarinsson, E, Yao, Z, Wiklund, ED, et al. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. Genome Res Feb 2008:18(2):242–251.
[66] Washietl, S, Pedersen, JS, Korbel, JO, et al. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 2007:17:852–864.
[67] Gruber, AR, Kilgus, C, Mosig, A, et al. Arthropod 7sk RNA. Mol Biol Evol Sep 2008:25(9):1923–30.