THERMODYNAMIC MATCHERS: STRENGTHENING THE SIGNIFICANCE OF RNA FOLDING ENERGIES

T. Höchsmann*, M. Höchsmann and R. Giegerich
Faculty of Technology, University Bielefeld,
Bielefeld, Germany
Email: {thoechsm,mhoechsm,robert}@techfak.uni-bielefeld.de*

Thermodynamic RNA secondary structure prediction is an important recipe for the latest generation of functional non-coding RNA finding tools. However, the predicted energy is not strong enough by itself to distinguish a single functional non-coding RNA from other RNA. Here, we analyze how well an RNA molecule folds into a particular structural class with a restricted folding algorithm called Thermodynamic Matcher (TDM). We compare this energy value to that of randomized sequences. We construct and apply TDMs for the non-coding RNA families RNA I and hammerhead ribozyme type III and our results show that using TDMs rather than universal minimum free energy folding allows for highly significant predictions.

1. INTRODUCTION

In this section, we discuss shortly the state of the art in RNA gene prediction and classification, and give an outline of the new approach presented here.

1.1. RNA gene prediction and classification

The term "RNA genes" is defined, for the purpose of this article, as those RNA transcripts that are not translated to protein, but carry out some cellular function by themselves. Recent increased interest in the manifold regulatory functions of RNA have led to the characterization of close to 100 classes of functional RNA[1, 2]. These RNA regulators mostly exert their function via their tertiary structure.

RNA genes are more difficult to predict than protein coding genes for two reasons: (1) There is no signal such as an open reading frame, which would be a first necessary indicator of a coding region. (2) Comparative gene prediction approaches are difficult to apply, because sequence need not be preserved in order to preserve a functional structure. In fact, structure preservation in the presence of sequence variation is the best indicator of a potentially interesting piece of RNA[3, 4]. This means that, in one way or another, structure must play an essential role in RNA gene prediction.

Whereas the full 3D structure of an RNA molecule currently cannot be computed, its 2D structure, the particular pattern of base pairs that form helices, bulges, hairpins etc., can be determined by dynamic programming algorithms based on an elaborate thermodynamic model[5–7]. Unfortunately, the minimum free energy (MFE) structure as defined and computed by this model is often weakly determined, and does not necessarily correspond to the functional structure *in vivo*. And of course, *every* single stranded RNA molecule, be it functional or not, attains *some* structure.

However, if there is a functional structure, preserved by evolution, it should be well-defined, according to two criteria:

- Energy Criterion: The energy level of the MFE structure should be relatively low, to ensure that the structure is stable enough to execute a specific function.
- Uniqueness Criterion: The determined MFE structure should not be challenged by alternative foldings with similar free energy.

Much work has been invested in the Energy Criterion: Can we move a window along an RNA sequence, determine the MFE of the best local folding, and where it is significantly lower than for a random sequence, may we hope for an RNA gene, because evolution has selected for a well-defined structure? Surprising first results were reported by Seffens &

*Corresponding author.

Digby[8], indicating that mRNAs (where one would not even expect such an effect) had lower energies than random sequences of the same nucleotide composition. However, this finding was refuted by Workman & Krogh[9], who showed that this effect goes away when considering randomized sequences with conserved dinucleotide composition. Rivas & Eddy[10] studied the significance of local folding energies in detail, reporting two further caveats: First, local inhomogeneity of CG content can produce a seemingly strong signal. Second, variance in MFE values is high, and a value of at least 4 standard deviations from the mean (a Z-score of 4) should be required before a particular value is considered an indicator of structural conservation. In most recent work, Clote et al.[11] studied several functional RNA families, comparing their MFE values against sequences of the same dinucleotide composition. They found that, on the one hand, there is a signal of smaller-than-random free energy, but on the other hand, it is not significant enough to be used for RNA gene prediction.

A weak signal can be amplified by using a comparative approach. Washietl et al.[12] suggest that, by scanning several well-aligned sequences, significant Z-scores can be obtained. The tool *RNAz*[3] is based on this idea. Of course, a good sequence alignment is not always available.

All in all, it has been determined that the Energy Criterion is not useless, but also not strong enough by itself to distinguish functional RNA genes from other RNA.

A first move to incorporate the Uniqueness Criterion has been suggested by Le et al.[13]. They compute scores based on energy differences: They compare the MFE value to the folding energy of a "restrained structure", which is defined by forbidding all base pairs observed in the MFE structure. This essentially partitions the folding space into two parts, taking the MFE structure *within* each part as the representative structure. This can be seen as a binary version of the *shape representative structures* defined by Giegerich et al.[14]. Just recently, the complete probabilistic analysis of abstract shapes of RNA has become possible[15], which would allow us to base the Le et al. approach on probabilities derived from Boltzmann statistics. This appears to be a promising route to follow. Here, however, we take yet another road in the same direction.

1.2. Outline of the new approach

After gene prediction via the Energy Criterion, the next step is to analyze the candidate structure, in order to decide whether it is a potential member of a known functional class. The structural family models provided in Rfam[16, 17] are typically used for this purpose. We suggest to combine the second step with the first one: We ask how well the molecule folds into a particular structural class, and compare this energy value to that of randomized sequences. We shall show that in this way we can obtain significant Z-scores. Note that this approach contains the earlier one as a special case: If the "particular class" holds all feasible structures, we are back with simple MFE folding. The Le et al. approach, by contrast, is not subsumed by this idea, as their partitioning is derived from the sequence at hand, while ours is set *a priori*.

The term *Thermodynamic Matcher* (TDM) has been suggested by Reeder et al.[18] for an algorithm that folds an RNA sequence into a particular type of structure in the energetically most favorable way. This is similar to using covariance models based on stochastic context free grammars, but uses thermodynamics rather than statistics. A first example of a TDM was the program *pknotsRG-enf*, which folds an RNA sequence into the energetically best structure containing *at least one* pseudoknot somewhere.

Although the idea of specialized thermodynamic folding appears to be an attractive supplement to covariance models[16], to our knowledge, no other TDMs have been reported. This is most likely due to the substantial programming effort incurred when implementing such specialized folding algorithms under the full energy model. However, these efforts are reduced by the technique of *algebraic dynamic programming*[19, 20], which allows to produce such a folding program – at least an executable draft – in one afternoon of work. Subsequent experimentation may be required to make the draft more specific, as explicated below. By this technique, we have been able to produce TDMs for nine RNA families so far, and our results show that using TDMs rather than universal MFE folding allows for highly significant

predictions.

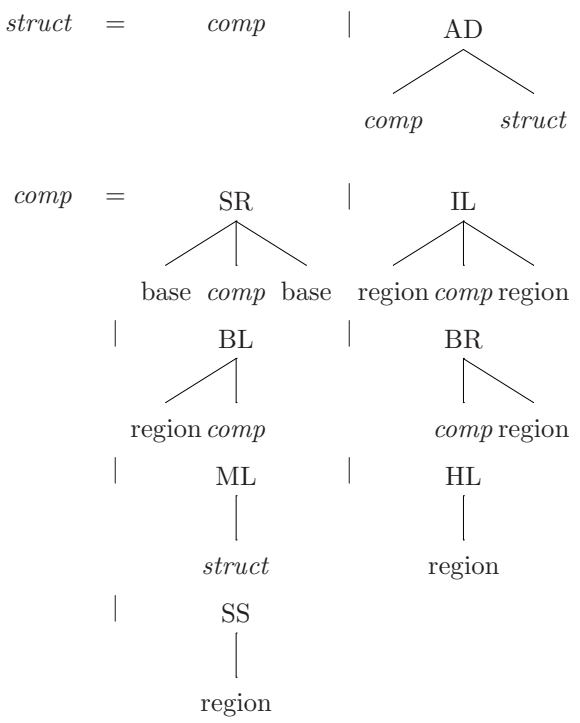


Fig. 1. General folding grammar $\mathcal{G}_{GF}$: The terminal symbol "base" denotes one of the nucleotides A,C,G,U, and "region" is a sequence of nucleotides. *struct* and *comp* are non-terminal symbols, and the corresponding productions are shown above. These productions can be read as follows: An RNA secondary structure can be a single component or a component next to some other *struct*. A component is either a single stranded region (SS), or it is composed (AD) from stacking regions (SR) and loops (BR,BL,IL,ML), which can be arbitrarily nested and terminated by a hairpin loop (HL).

The same results as with our TDMs in this paper can be computed using *RNAmotif*[21], by using the free energy as score function. However our motifs will result in exponential many structures for a input sequence. For every structure the energy has to be separately computed resulting in exponential runtime.

1.3. Tree grammars

RNA secondary structure, excluding pseudoknots, can be formally defined with regular tree grammars[15]. Similar to context free string grammars, a set of rules, called productions, transforms non terminal symbols into trees labeled with terminal and non terminal symbols. Formally, a tree grammar $\mathcal{G}$ is a tuple (Σ, V, P, A) where Σ is a set of terminal symbols, V is a set of variables with $\Sigma \cap V = \emptyset$, P is a production set, and A is a designated variable called *axiom*. The language $\mathcal{L}(\mathcal{G})$ of a tree grammar $\mathcal{G}$ is the set of trees that do not contain variables, which can be derived by iteratively applying productions starting with the axiom.

Figure 1 shows tree grammar $\mathcal{G}_{GF}$ for RNA secondary structures. $\mathcal{G}_{GF}$ is a simplified version of the base grammar our TDMs are derived from, which is more complex and takes into account the latest energy model for RNA folding. We use $\mathcal{G}_{GF}$ to illustrate the basic concepts of TDMs. Note that the sequence of leaf nodes (in left-to-right order) for a tree $T \in \mathcal{L}(\mathcal{G})$ is the primary sequence for T. RNA structure prediction and stochastic context free grammar approaches to align RNA structures, are problems of computing an optimal derivation for a primary sequence.

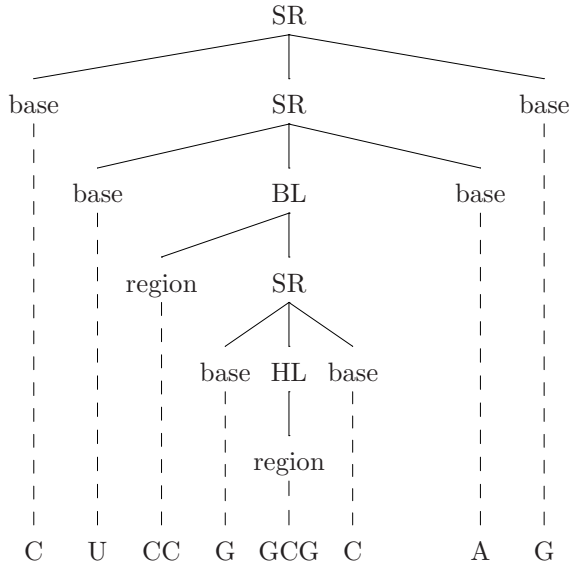


Fig. 2. This is one possible derivation of the grammar $\mathcal{G}_{GF}$ for the sequence "CUCCGGCGCAG". Note that this is just one of many possible trees/structures.

2. THERMODYNAMIC MATCHERS

The RNA folding problems means finding the energetically best folding for a given sequence under a certain model. Throughout this article, we consider the Zuker & Stiegler model, which describes the structure space and energy contributions for RNA

secondary structures and is used in a wide range of folding routines[7, 15, 6]. As indicated above, the structure space for an RNA molecule can be defined with a tree grammar and the folding problem becomes a parsing problem[19, 20]. We use this view and express (or restrict) folding spaces in terms of tree grammars, thereby obtaining thermodynamic matchers. The informal notion of a structural *motif* is formally modeled by a specialized tree grammar.

Let $\mathcal{G}$ be a grammar that describes the folding space for some structural motif, e.g. only those structures that have a tRNA-like hairpin structure. $\mathcal{G}$ typically differs from $\mathcal{G}_{GF}$ by absence of some rules, while other rules may be duplicated and specialized. $F_{\mathcal{G}}$ denotes the structure space for the grammar $\mathcal{G}$, in other words: all possible trees that can be derived from the grammar's axiom. A thermodynamic matcher $TDM_{\mathcal{G}}(s)$ is an algorithm that calculates the minimum free energy and the corresponding structure from the structure space $F_{\mathcal{G}}$ for some nucleotide sequence s. $MFE_{\mathcal{G}}(s)$ is the minimum free energy calculated by $TDM_{\mathcal{G}}(s)$. Since the same energy model is used, the minimal free energy of the restricted folding can not be lower than the minimal free energy of the general folding, we always have $MFE_{\mathcal{G}}(s) \geq MFE_{GF}(s)$. Note that it is not always possible to fold a sequence into a particular motif. In this case, a TDM returns an empty result.

2.1. Z-scores

A Z-score is the distance from the mean of a distribution normalized by the standard deviation. Mathematically: $Z(x) = (x - \mu)/\delta$, with μ being the mean and δ the standard deviation. Z-scores are useful for quantifying how different from normal a recorded value is. This concept has been applied to eliminate an effect that is well known for minimum free energy folding: The energy distribution is biased by the G/C content of a sequence as well as its length and dinucleotide composition.

To calculate the Z-score for a particular sequence, the distribution of MFE values for random sequences with the same dinucleotide composition must be known. The lower the Z-score, the lower is the energy compared to energies from random sequences. Clote et. al.[11] observed that Z-score distributions for RNA genes are lower than Z-score distribution for random RNA. However, this difference is fairly small and only significant if the whole distribution is considered. It is not sufficient to distinguish an individual RNA gene from random RNA[10]. The reason for the insufficient significance of Z-scores are the combinatorics of RNA folding. There is often *some* structure in the complete search space that obtains a low energy.

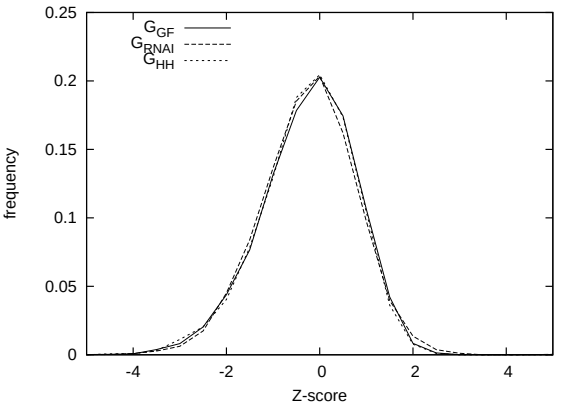


Fig. 3. Z-score histogram for 10000 random sequences with a length of 100 nucleotides, for two TDMs and the general folding.

Here, our aim is not the general prediction of non-coding RNA, but the detection of new members of a known, or at least defined, RNA family. By restricting the folding space, we can, as we demonstrate in Section 3, shift Z-scores for family members into a significant zone. Structures with $MFE_{GF} = MFE_{\mathcal{G}}$ for a grammar $\mathcal{G}$ get a lower Z-score, since the distribution $MFE_{\mathcal{G}}$ for random RNA is shifted to higher energies. Even if this seems to be right for the grammars used in this paper, the effect of a folding space restriction on the energy distribution is not obvious. Clearly, the mean is shifted to more positive values, but the effect on the variance is not yet understood mathematically. Therefore, our applications must provide evidence that the Z-scores are affected in the desired way.

Let $D_{\mathcal{G}}(s)$ be the frequency distribution of MFE values for random sequences with the same dinucleotide frequency as s, i.e. the minimum free energy versus the fraction of structures s' obtaining that energy with $TDM_{\mathcal{G}}(s')$. $Z_{\mathcal{G}}(s)$ is the Z-score for a sequence s with respect to the distribution $D_{\mathcal{G}}(s)$.

The value-mean and the standard deviation can be determined by a sampling procedure. For our experiments, we generate 1000 random sequences preserving the dinucleotide frequencies of s.

The distribution of Z-scores for random RNA sequences is shown in Figure 3. Interestingly, a restriction of the folding space does not affect the Z-score distribution. At least this holds for the TDMs shown in this paper. For a reliable detection of RNA genes, a Z-score of lower than -4 is needed[10]. Our experiments showed that over 99.98% of random RNAs have Z-scores greater then -4. To distinguish RNA genes from other RNA on a genomic scale, a threshold should be set to a Z-value such that the number of false predictions is trackable.

2.2. Design and implementation

Designing a thermodynamic matcher means defining its structure space. On the one hand it must be large enough to support good sensitivity, and on the other hand it must be small enough to provide good specificity. A systematic analysis of the relation between structure space restriction and its effect on specificity and sensitivity of MFE based Z-scores is subject of our current research.

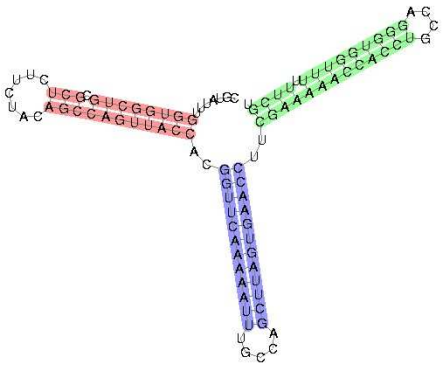


Fig. 4. Consensus structure for RNAI genes taken from the Rfam database.

The design of a TDM for an RNA gene requires a consensus structure. If an RNA family is listed in the Rfam database, the consensus shown there is a good starting point; at least the structural part of it. Alternatively, the consensus of known sequences can be obtained with programs that predict a common structure, like *PMmulti*[22] and *RNAcast*[23].

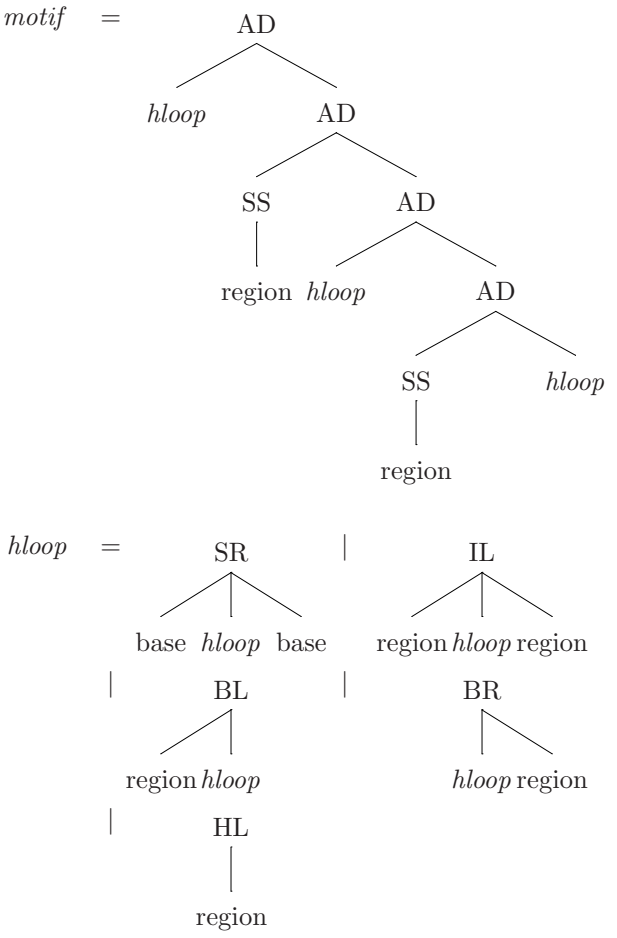


Fig. 5. Simplified version of the grammar $\mathcal{G}_{RNAI}$. Reconsider the grammar in Figure 1. Instead of an axiom that derives arbitrary RNA structures, the axiom *motif* derives three hairpin loops (*hloop*) connected by single stranded regions.

We now exemplify the design of a TDM. For instance, we are interested in stable secondary structures that consist of three hairpin loops separated by single stranded regions, like the structures of RNAI genes as shown in Figure 4. A specialized grammar for RNAI must only allow structures compatible with this motif. A simplified version of the grammar $\mathcal{G}_{RNAI}$, which abstracts from length constraints for stems and loops, is given in Figure 5.

Since we want to demonstrate that with a search

space reduction new members of an RNA family can be detected by their energy based Z-score, we do not incorporate explicit sequence constraints in a thermodynamic matcher other than those necessary to form the required base-pairs. However, this could be easily incorporated in our framework.

We use the algebraic dynamic programming (ADP) framework[19] to turn RNA secondary structure space grammars into thermodynamic matchers. In the context of ADP, writing a grammar in a text based notation is equivalent to writing a dynamic programming structure prediction program. This approach is similar to using an engine for searching with regular expressions. There is no need to implement the search routines, it is only a matter of specifying the search results. A grammar, which constitutes the control structure of an unrestricted folding algorithm, is augmented by an *evaluation algebra* incorporating the established energy rules[5]. All TDMs share these rules, only the grammar changes.

The time complexity of a TDM depends on the motif complexity. If multiloops are included the runtime is $O(n^3)$ where n is the length of the sequence that is folded. Without multiloops the time complexity is $O(n^2)$, if the size of bulges and loops is bounded by a constant. In both cases the memory consumption scales with $O(n^2)$.

3. RESULTS

We constructed TDMs for the non-coding RNA families RNAI and hammerhead type III ribozyme (hammerheadIII) taken from the Rfam database Version 7.0[16, 17]. All TDMs used in this section utilize the complete energy model for RNA folding[6] and therefore have more complex grammars than the grammars presented to explain our method.

To assess if TDMs can be used to find candidates for an RNA family, we searched for known members in genomic data. The known members are those from Rfam seeds, which are experimental validated. We apply our TDMs to genomes containing the seed sequences and measure the relation between Z-score threshold, sensitivity, and specificity. We define sensitivity as TP/(TP+FN) and specificity as TN/(TN+FP), where TP is the number of true positives, TN is the number true negatives, FP is the number of false positives, and FN is the number of false negatives.

3.1. RNA I

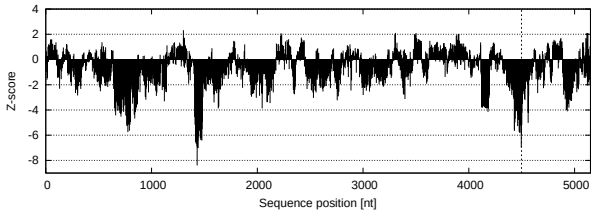
Replication of ColE1 and related bacterial plasmids is initiated by a primer, the plasmid encoded RNAII transcript, which forms a hybrid with its template DNA. RNAI is a shorter plasmid-encoded RNA that acts as a kinetically controlled suppressor of replication and thus controls the plasmid copy number[24]. Sequences coding for RNAI fold into stable secondary structures with Z-scores reaching from -3.6 to -6.7 (Table 1).

Table 1. Z-score for the RNAI seed sequences computed with $TDM_{\mathcal{G}_{GF}}$ and $TDM_{\mathcal{G}_{RNAI}}$.

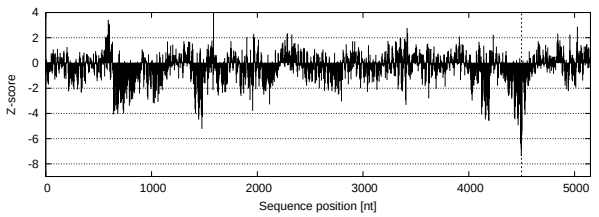
EMBL Accession number	$Z_{\mathcal{G}_{GF}}$	$Z_{\mathcal{G}_{RNAI}}$
AF156893.2	-6.61	-7.31
X80302.1	-4.88	-6.20
Y17716.1	-5.74	-6.29
Y17846.1	-5.06	-6.16
U80803.1	-6.33	-6.84
D21263.1	-3.96	-5.33
S42973.1	-4.53	-5.82
U65460.1	-6.73	-7.41
X63534.1	-3.63	-5.41
AJ132618.1	-5.93	-6.71

The Rfam consensus structure consists of three adjacent hairpin loops connected by single stranded regions (Figure 4). Structures for this consensus are described by the grammar $\mathcal{G}_{RNAI}$ (Figure 5). If we allow for arbitrary stem lengths in our motif, all structures that consist of three adjoined hairpins would be favored by $TDM_{\mathcal{G}_{RNAI}}$. This has an undesired effect: It would be possible to fold a sequence, that folds (with general folding) into a single hairpin with low energy, into a structure with one long and two very short hairpins. Although the energy of the restricted folding is higher than the energy of the unrestricted folding, it would still obtain a good energy resulting in a low Z-score. Clearly, these structures do not really resemble the structures of RNAI genes. In refinement, each stem loop is restricted to a minimal length of 25 nucleotides and the length of the complete structure is restricted to up to 100 nucleotides. These restrictions are compatible with the consensus of RNAI and increase the sensitivity

and specificity of $TDM_{\mathcal{G}_{RNAI}}$. Sequences from the seed obtain $Z_{\mathcal{G}_{RNAI}}$ values between -5.33 and -7.41 (Table 1). For random RNA the frequency distribution of $Z_{\mathcal{G}_{RNAI}}$ is similar to $Z_{\mathcal{G}_{GF}}$ (see Figure 3). The $Z_{\mathcal{G}_{RNAI}}$ score difference is large enough to distinguish RNAI genes from random RNA.



(a) General folding ($TDM_{\mathcal{G}_{GF}}$)



(b) Restricted Folding ($TDM_{\mathcal{G}_{RNAI}}$)

Fig. 6. TDM scan for RNAI in a plasmid of *Klebsiella pneumoniae* (EMBL Accession number AF156893). The known RNAI gene is located at position 4498 indicated by the dotted vertical line. (a) In steps of 5 nucleotides, the score $Z_{\mathcal{G}_{GF}}$ is shown for the following 100 nucleotides and for their reverse complement. The Z-scores for both directions are drawn versus the same sequence position. The position where the known RNAI gene starts achieves a low Z-score, but there is another position with a lower Z-score (position ~ 1450) and positions with nearly as low scores (around position 750). (b) shows corresponding values for $Z_{\mathcal{G}_{RNAI}}$. The RNAI gene now clearly separates from all other positions. Sequences that fold into some unrelated stable structure are penalized because they cannot fold into a stable RNAI structure.

To verify whether RNAI genes can also be distinguished from genomic RNA, we applied our matcher to 10 plasmids that contain the seed sequences (one in each of them). The Plasmid length ranges from 108 to 8193 nucleotides in this experiment. All plasmids together have a length of ~ 27500 nucleotides. For each plasmid, a 100 nucleotides long window was slid from 5' to 3' with a successive offset of 5. $Z_{\mathcal{G}_{RNAI}}$ was computed for every window. RNA I can be located on both strands of the plasmid. Therefore, $TDM_{\mathcal{G}_{RNAI}}$ was also applied to the reverse complement. Overall, this results in ~ 11000 $Z_{\mathcal{G}_{RNAI}}$ scores. An RNAI sequence was counted as positive hit if a Z-score in the range of 5 nucleotides to the left or right of the starting position of an RNAI gene has a Z-score equal or lower than the current threshold. In this region, no negative hits are counted. Figure 6 shows the result for a plasmid of *Klebsiella pneumoniae*.

It is also possible to use a complete sequence as input for a TDM. However, this will return the best substructure (or substructures) in terms of energy, which not always corresponds to the substructure with the lowest Z-score.

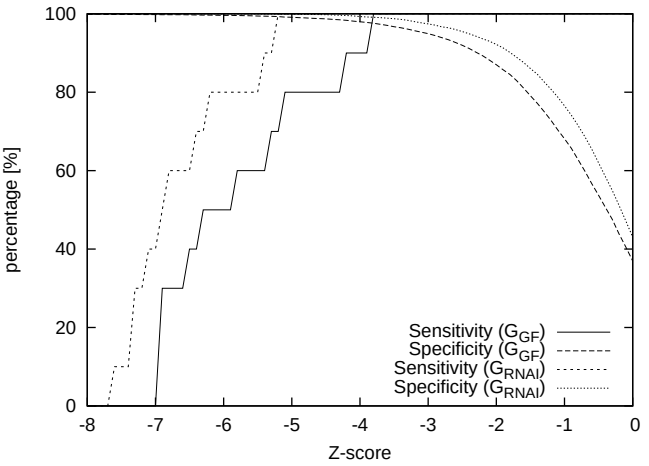


Fig. 7. Sensitivity and specificity versus the Z-value threshold. $TDM_{\mathcal{G}_{RNAI}}$ improves sensitivity and specificity compared to $TDM_{\mathcal{G}_{GF}}$.

If we set the Z-score threshold to -5, we obtain for $TDM_{\mathcal{G}_{RNAI}}$ a sensitivity of 100% and a specificity of 99.89%, which means 10 true positives and 12 false positives (for all plasmids). For $TDM_{\mathcal{G}_{GF}}$, we obtain only a sensitivity of 80% and a specificity of 99.10%, which means 8 true positives and 99 false positives. A threshold of -3.5 is required to find all RNAI genes of the seed. The specificity in this case is 96.71% resulting in 362 false positives. (Figure 7). Although the specificity is fairly low, it makes a big difference to the number of false positives for genome wide applications.

3.2. Hammerhead ribozyme (type III)

The hammerhead ribozyme was originally discovered as a self-cleaving motif in viroids and satellite RNAs. These RNAs replicate using the rolling circle mech-

anism, which generates long multimeric replication intermediates. They use the cleavage reaction to resolve the multimeric intermediates into monomeric forms. The region able to self-cleave has three base paired helices connected by two conserved single stranded regions and a bulged nucleotide. Hammerhead type III ribozymes (HammerheadIII) form stable secondary structures with Z-scores varying from -6 to -2 for general folding.

The seed sequences from the Rfam database vary in their length. 6 sequences have a length of around 80 nucleotides. All other seed sequences are around 55 nucleotides long. To be able to use length constraints, which are not too vague, we removed the 6 long sequences for our experiment. Thus, $TDM_{\mathcal{G}_{HH}}$ is not designed to search for HammerheadIII candidates with a sequence length larger than 60 nucleotides.

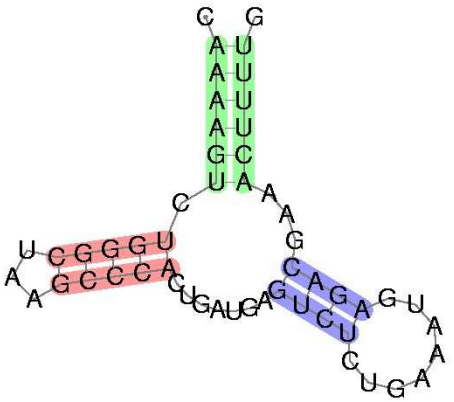


Fig. 8. Consensus structure for hammerhead ribozyme type III genes taken from the Rfam database.

Grammar $\mathcal{G}_{HH}$ describes the folding space for the consensus structure shown in Figure 8. The maximal length of our motif is 60 nucleotides. The single stranded region between the two stem loops in the multiloop has to be between 5 and 6 nucleotides long. The stem lengths are not explicitly restricted. $TDM_{\mathcal{G}_{HH}}$ improves the distribution of Z-scores for the seed sequences (Figure 9).

Most sequences now obtain a Z-score smaller than -4, but some obtain a higher score. These sequences are only about 45 nucleotides long. They fold into two adjacent hairpin loops and do not form a multiloop with $TDM_{\mathcal{G}_{GF}}$ They are forced into our HammerheadIII motif with considerable higher free energy. If a family has many members, it might be necessary to separately consider subfamilies.

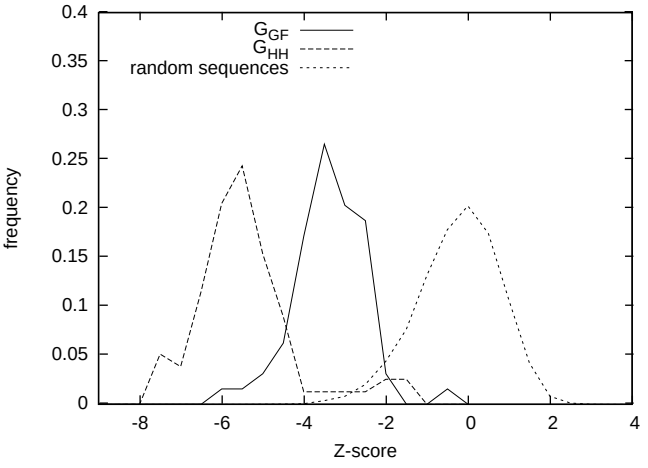


Fig. 9. Z-scores distribution for 68 hammerhead ribozyme type III sequences.

We applied $TDM_{\mathcal{G}_{HH}}$ to 59 viroid sequences with length of 290 to 475 nucleotides. HammerheadIII can be located on both strands of the DNA. Each sequence contains one or two HammerheadIII genes. A 60 nucleotides long window was slid from 5' to 3' with a successive offset of 2. For the sequence (and for its reverse complement), of each window $Z_{\mathcal{G}_{HH}}$ was computed. Overall, this resulted in $\sim$ 19500 scores. An HammerheadIII sequence was counted as positive hit if a Z-score in the range of 3 nucleotides to the left or right of the starting position of an HammerheadIII gene has a Z-score equal or lower than the current threshold. In this region, no negative hits are counted. The sensitivity and specificity depending on the Z-score threshold is shown in Figure 10. The sensitivity is improved significantly compared to $TDM_{\mathcal{G}_{GF}}$. However, the specificity is lower for Z-scores thresholds smaller than -3, which is the relevant region. It turned out that many false positives with Z-values of smaller -4 maybe true positives, which are not part of the Rfam seed, but are predicted as new RNAI candidate genes in Rfam. Figure 11 shows sensitivity and specificity if false nega-

tives, that are candidate genes in Rfam, are counted as true positives. All RNA candidate genes that are provided in Rfam achieve low Z-scores as shown in Figure 12. Unlike *Infernal*[16], which is used for the prediction of candidate family members in Rfam, we use pure thermodynamics rather than a covariance based optimization. This gives further and independent evidence for the correctness of both predictions.

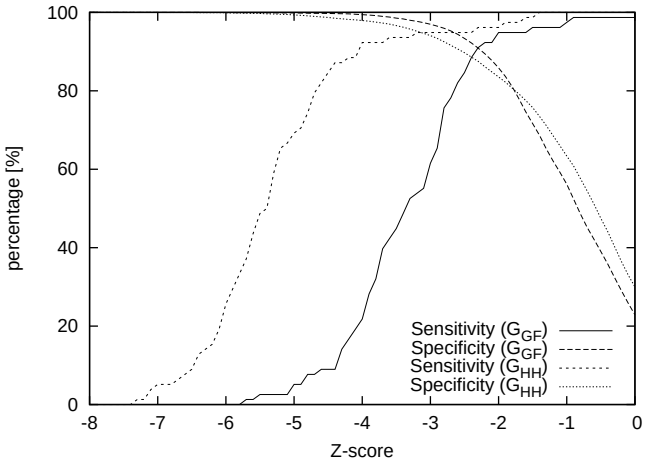


Fig. 10. Selectivity and specificity versus the Z-value threshold. $TDM_{\mathcal{G}_{HH}}$ improves sensitivity and specificity compared to $TDM_{\mathcal{G}_{GF}}$.

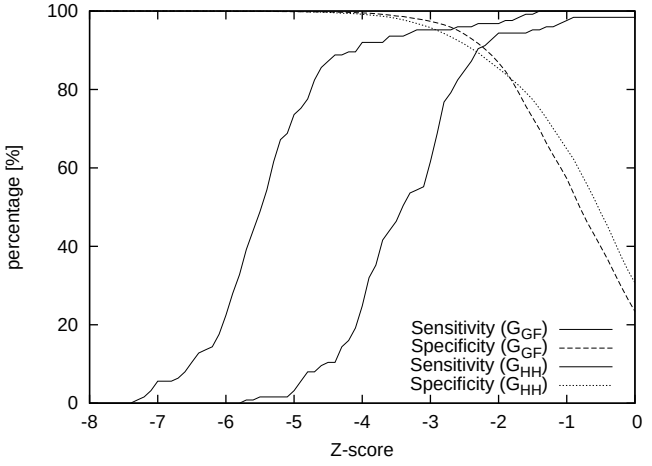


Fig. 11. Selectivity and specificity versus the Z-value threshold. $TDM_{\mathcal{G}_{HH}}$ improves sensitivity and specificity compared to $TDM_{\mathcal{G}_{GF}}$. Candidates predicted by Rfam are treated as positive hits.

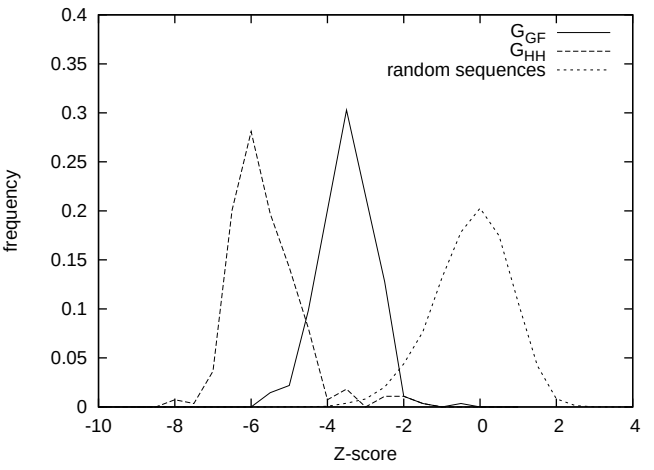


Fig. 12. Distribution of Z-scores for all 274 HammerheadIII gene and gene candidate sequences taken from the Rfam database.

4. DISCUSSION

The current debate about the quality of thermodynamic prediction of RNA secondary structures is extended by our observations regarding specialized folding spaces. It is well known that the MFE structure from predictions in most cases only shares a small number of base-pairs that can be detected by more reliable sources than MFE such as compensational base mutations. This is a consequence of the combinatorics of the RNA folding space, which provides many "good" foldings. Thus, MFE on its own can not be used to discriminate non-coding RNAs. We demonstrated that, given a consensus structure for a family of non-coding RNA, a restriction of the folding space to this family prunes low energy foldings for non-coding RNA that do not belong to this family. The overlap of Z-score distributions for MFE values for family members and non-family members can be reduced by our technique resulting in a search technique with high sensitivity and specificity, called *thermodynamic matching*.

In our experiments for RNA I and the hammerhead type III riboyzme, we did not include other restrictions than size restrictions for parts of the structure. These matchers can be fine tuned and can also include sequence restrictions, which could further increase their sensitivity and specificity. It is also possible to include H-type pseudoknots in the motif using techniques presented in Ref. 18.

We demonstrated that a TDM can detect members of RNA families by scanning single sequences. It seems promising to extend the TDM approach to scan aligned sequences using a combined energy and covariance scoring in spirit of RNAalifold[12]. This should further increase selectivity, or, if this is not necessary, allow "looser" motif definitions.

A question that arises from our observations is: Can our TDM approach be incorporated in a gene prediction strategy? If we would guess a certain motif and find stable structures with significant Z-scores, they might be somehow biologically relevant.

In a current research project, we focus on a systematic generation of TDMs for known RNA families from the Rfam database. We are also working on a graphical user interface to facilitate biologists to create their own TDMs, without requiring the knowledge of the underlying algebraic dynamic programming technique.

Beside the two RNA families shown here we have implemented TDMs for 7 other non-coding RNA families, including transfer RNA, micro RNA percursor and the Nanos 3' UTR translation control element. The results were consistent with our observations for RNAI and the hammerhead ribozyme given here, and will be used to analyze further the predictive power of thermodynamic matchers.

ACKNOWLEDGEMENTS

We thank Marc Rehmsmeier for helpful discussions and Michael Beckstette for comments on the manuscript.

References

1. A. F. Bompfünewerer, C. Flamm, C. Fried, G. Fritzsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S. J. Prohaska, B. M. R. Stadler, P. F. Stadler, A. Tanzer, S. Washietl, and C. Witwer, "Evolutionary patterns of non-coding RNAs," *Theor. Biosci*, vol. 123, pp. 301–369, 2005.
2. S. R. Eddy, "Non-coding RNA Genes and the Modern RNA World," *Nature Reviews Genetics*, vol. 2, pp. 919–929, 2001.
3. S. Washietl, I. L. Hofacker, and P. F. Stadler, "From The Cover: Fast and reliable prediction of noncoding RNAs," *PNAS*, vol. 102, no. 7, pp. 2454–2459, 2005.
4. E. Rivas and S. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," *BMC Bioinformatics*, vol. 2, no. 1, p. 8, 2001.
5. D. H. Turner, N. Sugimoto, and S. M. Freier, "RNA Structure Prediction," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 17, no. 1, pp. 167–192, 1988.
6. M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucl. Acids Res.*, vol. 31, no. 13, pp. 3406–3415, 2003.
7. I. L. Hofacker, "Vienna RNA secondary structure server," *Nucl. Acids Res.*, vol. 31, no. 13, pp. 3429–3431, 2003.
8. W. Seffens and D. Digby, "mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences," *Nucl. Acids Res.*, vol. 27, no. 7, pp. 1578–1584, 1999.
9. C. Workman and A. Krogh, "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution," *Nucl. Acids Res.*, vol. 27, no. 24, pp. 4816–4822, 1999.
10. E. Rivas and S. R. Eddy, "Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs," *Bioinformatics*, vol. 16, no. 7, pp. 583–605, 2000.
11. P. Clote, F. Ferre, E. Kranakis, and D. Krizac, "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency," *RNA*, vol. 11, no. 5, pp. 578–591, 2005.
12. S. Washietl and I. L. Hofacker, "Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics," *J Mol Biol*, vol. 342, pp. 19–30, 2004.
13. S.-Y. Le, J.-H. Chen, D. Konings, and J. Maizel, Jacob V., "Discovering well-ordered folding patterns in nucleotide sequences," *Bioinformatics*, vol. 19, no. 3, pp. 354–361, 2003.
14. R. Giegerich, B. Voss, and M. Rehmsmeier, "Abstract Shapes of RNA," *Nucl. Acids Res.*, vol. 32, no. 16, pp. 4843–4851, 2004.
15. B. Voss, R. Giegerich, and M. Rehmsmeier, "Complete probabilistic analysis of RNA shapes," *BMC Biology*, vol. 4, no. 5, 2006.
16. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database," *Nucl. Acids Res.*, vol. 31, no. 1, pp. 439–441, 2003.
17. S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucl. Acids Res.*, vol. 33, no. suppl 1, pp. D121–124, 2005.
18. J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinformatics*, vol. 5, no. 104, 2004.
19. R. Giegerich, C. Meyer, and P. Steffen, "A discipline of dynamic programming over sequence data,"

Science of Computer Programming, vol. 51, no. 3, pp. 215–263, 2004.

20. P. Steffen and R. Giegerich, "Versatile and declarative dynamic programming using pair algebras," *BMC Bioinformatics*, vol. 6, no. 224, 2005.
21. T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath, "RNAMotif, an RNA secondary structure definition and search algorithm," *Nucl. Acids Res.*, vol. 29, no. 22, pp. 4724–4735, 2001.
22. I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler, "Alignment of RNA Base Pairing Probability Matrices," *Bioinformatics*, vol. 20, pp. 2222–2227, 2004.
23. J. Reeder and R. Giegerich, "Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction," *Bioinformatics*, vol. 21, no. 17, pp. 3516–3523, 2005.
24. Y. Eguchi and J. Itoh, T Tomizawa, "Antisense RNA," *Annu. Rev. Biochem.*, vol. 60, pp. 631–652, 1991.