RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures

Andreas R. Gruber[1], Stephan H. Bernhart[1],
You Zhou[1,2], and Ivo L. Hofacker[1]
[1] *Institute for Theoretical Chemistry*
University of Vienna, Währingerstraße 17, 1090 Wien, Austria
[2] College of Computer Science and Technology
Jilin University, Changchun 130012, China
{agruber, berni, ivo}@tbi.univie.ac.at, zyou@jlu.edu.cn

Abstract: The search for local RNA secondary structures and the annotation of unusually stable folding regions in genomic sequences are two well motivated bioinformatic problems. In this contribution we introduce `RNALfoldz` an efficient solution two tackle both tasks. It is an extension of the `RNALfold` algorithm augmented by support vector regression for efficient calculation of a structure's thermodynamic stability. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

1 Introduction

Over the past decade noncoding RNAs (ncRNAs) have risen from a shadowy existence to one of the primary research topics in modern molecular biology. Today computational RNA biology faces challenges in the ever growing amount of sequencing data. Efficient computational tools are needed to turn these data into information. In this context, the search for locally stable RNA secondary structures in large sequences is a well motivated bioinformatic problem that has drawn considerable attention in the community. `RNALfold` [HPS04] has been the first in a series of tools that offered an efficient solution to this task. Instead of a straight-forward, but costly sliding window approach a dynamic programming recursion has been formulated that predicts all stable, local RNA structures in $\mathcal{O}(N \times L^2)$, where L is the maximum base-pair span and N the length of the sequence. Since its publication, the `RNALfold` algorithm has inspired a lot of work in this field, see e.g. `Rnall` by Wan *et al.* [WLX06] or `RNAslider` by Horesh *et al.* [HWL+09]. All contributions so far in this field focused on improving the computational complexity of the algorithm, but none of the approaches has ever been used to unravel results of biological significance. In particular, *de novo* detection of functional RNA structures has been addressed, but application on a genome-wide scale with a low false discovery rate seems still out of reach. Even on the moderately sized genome of *E. coli* (4.6 Mb) one is drowning in hundreds of thousands of local structures. Unlike in the well established field of protein coding gene detection where one can exploit signals like codon usage, functional

RNA secondary structures, in general, do not show strong characteristics that make them easily distinguishable from random decoys. Successful approaches for ncRNA detection operating solely on a single sequence [HHS08, JWW^{+}07] are limited to specific RNA classes, where some outstanding characteristics can be harnessed. There is no master plan for the detection of functional RNA structures, but one would certainly want to limit the `RNALfold` output to a reasonable amount. So far, only the minimum free energy (MFE) of the locally stable secondary structures, which is intrinsically computed by the algorithm, has been considered as potential discriminator to limit the number of secondary structures. As demonstrated clearly by Freyhult and colleagues [FGM05] the MFE is roughly a function of the length of the sequence and the G+C content. Even normalizing the MFE by length of the sequence does not serve as a good discriminator between shuffled or coding sequences and functional RNA structures. A strategy that does work, however, is to compare the native MFE E of the RNA molecule to the MFEs of a set of shuffled sequences of same length and base composition [LM89]. This way we can evaluate the thermodynamic stability of the secondary structure. A common statistical quantity in this context is the z-score, which is calculated as follows

$$z = \frac{E - \mu}{\sigma}$$

where μ and σ are the average and the standard deviation of the energies of the set of shuffled sequences. The more negative the z-score the more thermodynamically stable is the structure. Efficient estimation of a sequence's z-score has been a profound problem already addressed in the very beginnings of computational RNA biology. A first strategy to avoid explicit shuffling and folding was based on table look-ups of linear regression coefficients [CLS^{+}90]. Clote and colleagues [CFKK05] introduced the concept of the asymptotic z-score, where the efficient calculation is also solved via table look-ups. The current state-of-the art approach for fast and efficient estimation of the z-score is to use support vector regression [WHS05].

The study by Clote and colleagues and a follow up to Chen *et al.* (1990) [LLM02] also report on the effort to predict thermodynamically stable structures using a sliding window approach. In this contribution we present `RNALfoldz` an algorithm that combines local RNA secondary structure prediction and the efficient search for thermodynamically stable structures. `RNALfoldz` is an extension of the `RNALfold` algorithm augmented by support vector regression for efficient calculation of a sequence's z-score. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

2 Methods

2.1 Fast estimation of the z-score using support vector regression

For the efficient estimation of the z-score we follow the strategy first introduced by Washietl *et al.* [WHS05]. Instead of explicit generation and folding of shuffled sequences in order to determine the average free energy and the corresponding standard deviation support vector regression (SVR) models are trained to estimate both values. As described in detail in the previous work, we used a regularly spaced grid to sample sequences for the training set. Synthetic sequences ranged from 50 to 400 nt in steps of 50 nt. The G+C content, A/(A+T) ratio and C/(C+G) ratio were, however, extended to a broader spectrum, now ranging from 0.20 to 0.80 in steps of 0.05. A total of 17,576 sequences were used for training. For each sequence of the training set 1,000 randomized sequences were generated using the Fisher-Yates shuffle algorithm, and subsequently folded with `RNAfold` with dangling ends option `-d2` [HFS+94]. SVR models for the average free energy and standard deviation were trained using the `LIBSVM` package (`www.csie.ntu.edu.tw/~cjlin/libsvm`). While in the previous work input features and the dependent variables were normalized to a mean of zero and a standard deviation of one, we apply here a different normalization strategy that leads to a significantly lower number of support vectors for the final models. For the regression of the average free energy model the dependent variable is normalized by the length of the sequence, while for the standard deviation it is the square root of the sequence length. The length still remains in the set of input features and is scaled from 0 to 1. Other features remain unchanged. We used a RBF kernel, and optimized values for the SVM parameters were determined using standard protocols for this purpose. Final regression models were selected by balancing two criteria: (i) mean absolute error (MAE) on a test set of 5,000 randomly drawn sequences of arbitrary length (50-400) from the human genome, and (ii) complexity of the model (number of support vectors) , which translates to following procedure: from the top 10% of regression models in terms of MAE we selected the one that had the lowest number of support vectors. For the average free energy regression we selected a model with a MAE of 0.453 and 1,088 support vectors, and for the standard deviation regression a model with a MAE of 0.027 and 2,252 support vectors. To validate our approach we finally compared z-scores derived from the SVR to traditionally sampled z-scores on a set of 1,000 randomly drawn sequences from the human genome. The correlation coefficient (R) is 0.9981 and the MAE is 0.072. This is in fair agreement to results obtained when comparing two sets of sampled z-scores (R: 0.9986, MAE: 0.054, number of shuffled sequences = 1,000).

2.2 Adaption of the `RNALfold` algorithm

RNALfold computes locally stable structures of long RNA molecules. It uses a Zuker type secondary structure prediction algorithm [ZS81] and restricts the maximum base pair span to L bases to keep the structures local. The sequence is processed from the 3' (the sequence length n) to the 5' end. In order to keep the number of back trace operations low and the output at moderate size, we want to avoid backtracing structures that differ only by unpaired regions. Furthermore, only the longest helices possible are of interest. To achieve this, a structure starting at base i is only traced back if the total energy $F(i,n)$ is smaller than that of its 3' neighbor $F(i+1,n)$ while its 5' neighbor has the same energy: $F(i-1,n) = F(i,n) < F(i+1,n)$. The local minimum structure is found by identifying the pairing partner j of i so that $C(i,j) + F(j+1,n) = F(i,n)$, i.e. the minimum energy from i to n is decomposed into the local minimum part i, j and the rest of the molecule. Here, $C(i, j)$ stands for the energy of a structural feature enclosed by the base pair i, j. As a result of this, the output of `RNALfold` contains components, i.e. structures that are enclosed by a base pair, only. Before we actually start the trace back, we evaluate two new criteria: (1) the sequence of the structure traced back has to be within the training parameters of the SVR, and (2) the z-score of the energy of this structure has to be lower than a predefined bound. Criterion (1) is simply imposed by the training boundaries of the SVMs. Boundaries have, however, been chosen carefully to cover a broad range of today's known spectrum of functional RNA structures. 99.79% of the sequences in the `Rfam` v. 10 full data set match the base composition requirements of the SVR and 90% of `Rfam` RNA families are in within the sequence length restrictions.

In order to get the exact sequence composition that is needed for the SVR evaluations, the 3' end of the structure (j) has to be computed first. This is done by a first, short backtracing step, where the decomposition $F(i, n) = C(i, j) + F(j + 1, n)$ is used to find j. Subsequently, the average free energy given the base composition of the sequence $s(i, j)$ is computed by calling the corresponding SVR model. The SVR model for the standard deviation has approximately twice the number of support vectors as the average free energy model. To minimize calls of this model, first the minimal standard deviation for the particular sequence length is looked up. We can then, using the free energy of $C(i, j)$, calculate a lower bound of the z-score. Only if this lower bound is below the minimal required z-score, the support vector regression for the standard deviation is called to calculate the actual z-score. If the z-score then still meets the minimal z-score criterion, the structure is fully traced back and printed out.

3 Results

The concept of fast and efficient estimation of the z-score by support vector regression was first introduced by Washietl *et al.* [WHS05], and implemented in the noncoding RNA gene finder `RNAz`. The speed up of this approach compared to explicit shuffling and folding, which is usually done on 1,000 replicas, is tremendous, at minimum a factor of 1,000. Moreover, computing time is invariant to the length of the sequence, while RNA folding is of complexity of $\mathcal{O}(N^3)$. When considering the z-score as evaluation criterion in the `RNALfold` algorithm, calculation of the z-score becomes a time consuming factor, as in a worst case scenario it has to be done almost for every nucleotide of the sequence. It is therefore a crucial concern to use support vector models that do not only have good accuracy, but also a moderate number of support vectors (SVs). In this work we extended the z-score support vector regression to cover a broader range of the sequence spectrum, but managed at the same time to build models that have significantly less SVs than the models used by `RNAz`. This was accomplished by normalizing the dependent variables of the regression, i. e. the average free energy and the standard deviation, by the sequence length. The dependent variables do not strictly linearly correlate with the sequence length and so we have to keep the sequence length as an input feature. Nevertheless, redundant points are created in the training set, which eventually leads to a smaller space to be trained. For the average free energy model and the standard deviation model we were able to achieve a 6.3 and a 2.7 fold reduction, respectively, in the number of SVs compared to the `RNAz` equivalents.

3.1 Evaluation of **`RNALfoldz`** predicition accuracy

For the task of predicting local RNA secondary structures one would particularly be interested in following criteria: (i) to which extent can functional ncRNAs be discovered, (ii) how well do the molecule's predicted boundaries match to the real coordinates, and (iii) is there any significant difference between native, biological sequences and random decoys. To address these questions, we applied `RNALfoldz` to the genome of *E. coli* (Accession number: CP000948). A maximum base-pair span L of 120 nt and a z-score cutoff of -2 was used. Setting the cutoff at -2 is for sure restrictive, but it should still cover a large fraction of the ncRNA repertoire. Both strands were considered. A total of 202,126 structures have been obtained. In comparison, the regular `RNALfold` returned a total of 1,387,136 structures, 824, 000 of which have a length $\geq$ 50 nt. The `RNALfoldz` output (a true subset of the `RNALfold` output) is only a forth of the regular `RNALfold` output.

The *E. coli* genome Genbank file lists 119 ncRNAs with a maximum length of 120 nt in its current annotation. To investigate the extent annotated ncRNAs are covered in the `RNALfoldz` output, we define for a `RNALfold`/`RNALfoldz` prediction to be counted as hit a minimal coverage of 90% of the ncRNA sequence. Giving this setup a total of 106 and 89 ncRNAs can be found in the `RNALfold` and `RNALfoldz` output, respectively. Detailed results for each RNA gene are shown in an online supplementary table. With a z-score cutoff of -2, 17 ncRNAs that were found by `RNALfold` are not in output set of `RNALfoldz`. The detection success is directly proportional to the reduction rate of the `RNALfold` output. Modulating the z-score cutoff affects both quantities (Fig. 1). The failure to detect the 13 ncRNAs that were missed by both `RNALfold` and `RNALfoldz` results from the fact that the `RNALfold` algorithm predicts only self-contained RNA structures. For example, the two ncRNA genes *rprA* and *ryeE* that have only low covering `RNALfoldz` hits, have indeed multi-component structures at the MFE level (abstract shape notation [GVR04]: `[][][][]`, `[][][]`). In these cases `RNALfoldz` will rather produce multiple hits than one single hit covering the whole ncRNA. Overall, our findings confirm that most *E. coli* small ncRNAs are indeed more thermodynamically stable than expected by chance and that the `RNALfoldz` algorithm is able to detect these structures efficiently.

We further investigated how precisely the `RNALfoldz` predictions map to the coordinates of the annotated ncRNAs. This is a legitimate issue, but one has to keep in mind that functional RNAs adopt their structure at the transcription level, while in this experiment we used the genomic sequence to detect these structures. So it might easily happen that the RNA is predicted in a bigger structural context than its actual size. The underlying dynamic programming algorithm is the same for `RNALfold` and `RNALfoldz`, and hence results discussed here do hold for both versions. In this work we define *noise* as the fraction of the `RNALfoldz` hit that does not overlap with the annotated ncRNA. In total, 34 out of

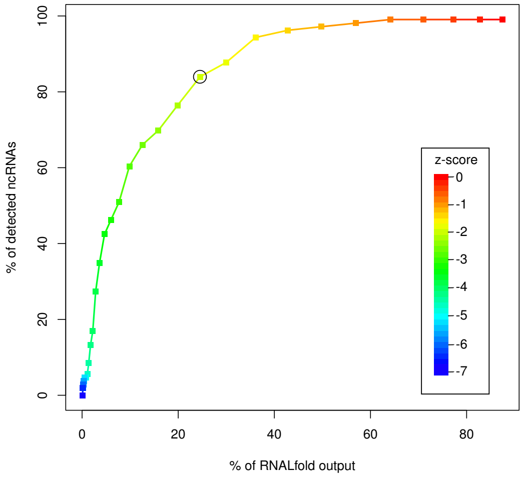

Figure 1: Non-coding RNA detection success vs. reduction of the `RNALfold` output. Given a z-score cutoff of 0 only one structure prediction is missed in the `RNALfoldz` output. With a z-score cutoff of -2 (circle) we see a four-fold reduction of the output, while at the same time covering 84% of the correct `RNALfold` ncRNA predictions.

the 89 predictions have less than 10% noise. Averaged over all hits ($\geq$ 90% coverage) we see *noise* of 18%. Using a classic sliding window approach with a length of 120 nt, one would expect more than 33% noise for a window containing a tRNA sequence (average length of tRNAs in E. coli: 78 nt). In the `RNALfoldz` output we find that 29 out of 73 tRNA predictions have less than 10% noise.

Finally, we address the significance of the predictions when compared to randomized controls. Therefore, we performed the same experiment on randomized sequences generated by (i) mononucleotide shuffling, (ii) simulation with an order-0 Markov model (mononucleotide frequencies) , and (iii) simulation with an order-1 Markov model (dinucleotide frequencies). Shuffling and simulations were done with `shuffle` from Sean Eddy's squid library using default parameters. A detailed comparison of the results of these four experiments is shown in Fig. 2. In general, we observe a tendency to more stable structures in the native sequence than in any randomized sequence. Structures with a z-score $\leq$ -8 are profoundly enriched in the native sequence, which might point to biological relevance of these structures. These are, however, extremes and most ncRNAs will have z-score values in a much higher range.

The value -2 for the z-score cutoff in this experiment was chosen arbitrarily. Moving to an even lower value than -2 will reduce the false discovery rate, but at the same time limit the number of ncRNAs that show such high thermodynamic stability. Using the `RNALfoldz` output from the experiment with randomized sequences (order-1 Markov model), we can calculate an empirical precision or positive predictive value (PPV), which is simply the

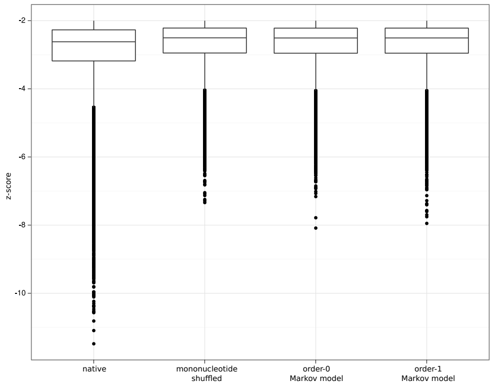

Figure 2: Comparison of the distribution of stable secondary structures from the native *E. coli* genome and randomized controls. The native *E. coli* sequence has a strong tendency to more stable local secondary structures. `RNALfoldz` predictions with a z-score below -8 are exclusively found in the native sequence.

proportion of true positives against all positive results. Assuming that thermodynamic stability is inherently linked to biologically function, we declare any `RNALfoldz` prediction with a z-score below a certain threshold from the native sequence and the randomized sequence as true positive and as false positive, respectively. Using then a PPV of 0.75, which corresponds to 25% estimated false positives, and, hence, a deduced z-score cutoff of -3.86 we can find 25 of the 106 annotated ncRNAs that are detectable with the `RNALfold` algorithm, while reducing the `RNALfoldz` to 21,715 predictions (3% of the `RNALfold` output). We further investigated if we can determine more specific z-score cutoffs when the `RNALfoldz` output is partitioned into different structural classes. This is motivated by the reasonable assumption that, for example, a short stable hairpin is more likely formed by chance than a stable, structurally more complex, multi-branched molecule. Hence, one would set different z-score cutoffs for different structural classes. To investigate this claim we map the MFE structures to the corresponding abstract RNA shape at the highest abstraction level. At this abstraction level only the helix nesting pattern is considered. As an example, the well-known cloverleaf structure of tRNA molecules is then simply represented as `[[][][]]`. The six major structural classes are shown in Tab. 1. We further partition structures according to their length into two classes *short* ($\leq$ 85 nt) and *long*.

Fig. 3 shows structure class specific precision values in dependency of the z-score, for those three classes that show the most deviation from the population precision. Using now class-specific z-score values when filtering the `RNALfoldz` output we can raise our prediction count from 25 to 38 ncRNAs, while keeping the expected false-positive rate fixed at 25%. The total number of `RNALfoldz` predictions increases slightly to 23,225.

Table 1: Major structural classes in the *E. coli* genome

frequency	abstract shape	length class	figure code	class specific z-score cutoff (PPV 0.75)
27%	[[][]]	long		-3.60
26%	[[][]]	short	SC2	-4.14
21%	[]	short	SC3	-4.16
7%	[[][[][]]]	long		-3.80
7%	[[[][]][]]	long		-3.74
4%	[]	long	SC6	-3.35
8%	rest			-3.35

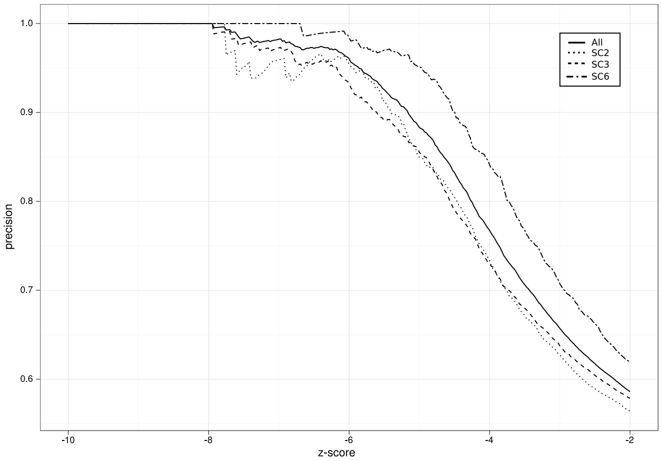

Figure 3: Precision values of different structural classes by the z-score. The solid line represents the whole `RNALfoldz` output.

3.2 Timing

The overall complexity $\mathcal{O}(N \times L^2)$ of the core algorithm does not change, the z-score calculation just adds a constant factor. We benchmarked both implementations on an Intel Quad Core2 CPU with 2.40 GHz. Detailed results are shown in Tab. 2.

At a maximal base-pair span of 120 nt `RNALfold` is able to scan at a speed of approx. 250 kb/min. At the same settings and with a minimal z-score cutoff of -2 scanning speed drops to 153 kb/min for `RNALfoldz`. The performance clearly depends on the number of times the support vector regression has to be called. When moving to a lower z-score cutoff of -4 the scanning speed increases to 207 kb/min.

Table 2: Timing results [sec] for RNALfold and RNALfoldz.

L	RNALfold		RNALfoldz	
		z-score $\leq$ -2	z-score $\leq$ -3	z-score $\leq$ -4
120	1,123	1,842	1,477	1,359
240	2,629	3,922	3,321	3,105

4 Discussion

In this work we have presented an extension of the RNALfold algorithm to predict thermodynamically stable, local RNA secondary structures. Using fast support vector regression models to calculate the z-score this approach comes with only a minor overhead in execution time compared to the former version, while yielding at the same time a much lower number of relevant structures. We have demonstrated that already with a z-score cutoff of -2, approx. 80% of the annotated *E. coli* small ncRNAs can be found in the RNALfoldz output. Comparison to randomized genome sequences showed that the native *E. coli* genome sequence has a strong bias to more stable secondary structures. This shift is, however, not significant enough to qualify RNALfoldz as a stand-alone RNA gene finder with an acceptable false discovery rate. We see the role of RNALfoldz mainly as a first filtering step in a cascade of computational ncRNA detection steps. In particular, the intersection of data from high throughput sequencing, promoter and transcription termination signals (see e.g. [SNS^{+}10]) or G+C content on AT rich genomes with RNALfoldz hits could be of value.

In this contribution, we assume that thermodynamic stability is inherently coupled to biological function. This is certainly true to some extent, but there are also a lot of RNA classes where stability is not a major issue for function, e.g. C/D box snoRNAs or ncRNAs that form interaction with other RNAs. It is therefore highly unlikely that these RNA classes will show up in the RNALfoldz output. In this context, RNALfoldz can, however, be used to define a set of highly stable loci which can then be excluded from further analysis.

It has been noted early on that thermodynamic stability alone is not a sufficient discriminator to distinguish ncRNAs from random sequences [RE00]. This is the main reason why most ncRNA gene finders rely solely on signals from evolutionary conservation of RNA secondary structures, or use thermodynamic stability only as an additional feature. These methods are clearly limited by the comparative genomics data available. Investigation of species that are distantly related to any species sequenced so far, or species specific RNA genes are, hence, out of scope for these methods. The RNALfoldz algorithm presented in this work will not be a magic tool suddenly shedding light on these dark areas. The search for extraordinarily stable structures, however, can help to give first clues to putatively functional RNA secondary structure elements, where other methods fail.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects “bioinformatics integration network III” and “regulatory non-coding RNAs”. `RNALfoldz` is part of the Vienna RNA package and can be obtained free of charge at: `http://www.tbi.univie.ac.at/~ivo/RNA/`. An electronic supplement located at `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/RNALfoldz` compiles training and test data.

References

[CFKK05] P Clote, F Ferré, E Kranakis, and D Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, 2005.

[CLS+90] J H Chen, S Y Le, B Shapiro, K M Currey, and J V Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Comput Appl Biosci*, 6(1):7–18, 1990.

[FGM05] E Freyhult, P P Gardner, and V Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241–241, 2005.

[GVR04] R Giegerich, B Voss, and M Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res*, 32(16):4843–4851, 2004.

[HFS+94] I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[HHS08] J Hertel, I L Hofacker, and P F Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2008.

[HPS04] I L Hofacker, B Priwitzer, and P F Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190, 2004.

[HWL+09] Y Horesh, Y Wexler, I Lebenthal, M Ziv-Ukelson, and R Unger. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics*, 10:76–76, 2009.

[JWW+07] P Jiang, H Wu, W Wang, W Ma, X Sun, and Z Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*, 35(Web Server issue):339–344, 2007.

[LLM02] S Y Le, W M Liu, and J V Maizel. A data mining approach to discover unusual folding regions in genome sequences. *Knowledge-Based Systems*, 15(4):243 – 250, 2002.

[LM89] S Y Le and J V Maizel. A method for assessing the statistical significance of RNA folding. *J Theor Biol*, 138(4):495–510, 1989.

[RE00] E Rivas and S R Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

[SNS+10] J Sridhar, S R Narmada, R Sabarinathan, H Y Ou, Z Deng, K Sekar, Z A Rafi, and K Rajakumar. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*, 5(8), 2010.

[WHS05] S Washietl, I L Hofacker, and P F Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, 2005.

[WLX06] X F Wan, G Lin, and D Xu. Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. *J Bioinform Comput Biol*, 4(5):1015–1031, 2006.

[ZS81] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.