Pacific Symposium on Biocomputing 15:69-79(2010)

RNAZ 2.0: IMPROVED NONCODING RNA DETECTION

ANDREAS R. GRUBER[1,2], SVEN FINDEIß[1], STEFAN WASHIETL[2,3],
IVO L. HOFACKER[2] AND PETER F. STADLER[1,2]

[1] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig*
Härtelstrasse 16-18, D-04107 Leipzig, Germany

[2] *Institute for Theoretical Chemistry, University of Vienna*
Währingerstrasse 17, A-1090 Wien, Austria.

[3] *European Molecular Biology Laboratory – European Bioinformatics Institute*
Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

`RNAz` is a widely used software package for *de novo* detection of structured noncoding RNAs in comparative genomics data. Four years of experience have not only demonstrated the applicability of the approach, but also helped us to identify limitations of the current implementation. `RNAz 2.0` provides significant improvements in two respects: (1) The accuracy is increased by the systematic use of dinucleotide models. (2) Technical limitations of the previous version, such as the inability to handle alignments with more than six sequences, are overcome by increased training data and the usage of an entropy measure to represent sequence similarities. `RNAz 2.0` shows a significantly lower false discovery rate on a dinucleotide background model than the previous version. Separate models for structural alignments provide an additional way to increase the predictive power. `RNAz` is open source software and can be obtained free of charge at: `http://www.tbi.univie.ac.at/~wash/RNAz/`

Keywords: RNA structure; noncoding RNA; structure conservation; comparative genomics; gene prediction

1. Introduction

Noncoding RNAs (ncRNAs) are transcripts that are not translated to proteins but function directly on the RNA level. During the past few years it has become evident that such "RNA genes" are more common than previously thought. MicroRNAs, for instance, have profoundly changed our view of gene regulation, and several completely new classes of ncRNAs were discovered recently.[1] They have been found to be involved in such diverse processes as transcriptional regulation,[2–4] post-transcriptional regulation,[5] chromatin modification and epigenetics,[6,7] and development.[8] Non-coding RNAs thus are key players in cellular regulation, a realization that has also moved the computational analysis and the annotation of ncRNAs at genome-wide scales into the focus of attention.

With the rapidly increasing availability of genomic sequence data, the *de novo* prediction of ncRNAs is of particular interest. While protein gene prediction is a classical problem in computational biology and has been studied for more than 15 years, RNA gene prediction is still in its infancy. Nevertheless, significant progress has been made regarding the prediction of "structured ncRNAs". This class of ncRNAs is characterized by evolutionary conserved secondary structures which appear to be important for their function. Most of the well-characterized ncRNAs belong to this class. Leading software tools developed for *de novo* RNA gene finding therefore use evolutionary conservation of functional secondary structures as the main signal to detect these ncRNAs.[9–13]

`RNAz` also detects structural ncRNAs by means of a comparative approach. In addition to measuring evolutionary conservation, however, it also explicitly evaluates the thermodynamic stability of the secondary structure.[14] A support vector machine (SVM) is then used to evaluate both criteria. `RNAz 1.0` has been used successfully to map structural ncRNAs in a wide variety of genomes.[15–20] A large number of these predictions have also been verified experimentally.[21–23] Moreover, the generic approach and many algorithmic details developed for `RNAz 1.0` have been re-used, extended, and adapted to other problems in the field of RNA gene-finding.[11,24–30]

The wide-spread use of `RNAz 1.0` also helped to identify some of its limitations and to point our directions

for improvements. In this contribution, we describe a major update of the RNAz program. It is based on the results of two follow-up studies,[31,32] on our experiences gained during many real-life applications, in particular the ENCODE pilot project,[33,34] and last but not least, on the user feedback we received over the past four years.

One major improvement is that RNAz 2.0 now allows to calculate thermodynamic stability scores based on a dinucleotide background model. It has been noted early-on that folding algorithms utilizing stacking energies of adjacent base-pairs in their energy model are sensitive to the dinucleotide content.[35] In the context of genome-wide ncRNA predictions, this effect can lead to an increased number of false positive calls as pointed out several times.[32,33,36] The new dinucleotide model in RNAz 2.0 now avoids this source of potential false positives and increases the accuracy of the program.

Another major limitation of RNAz 1.0 was the fact that only alignments with at most six sequences could be scored. This rather arbitrary restriction was the result of the limited amount of comparative data sets that were available at the time. During the past few years, however, comparative data sets have grown massively and therefore we adapted the algorithm to allow flexible analysis of alignments of any size.

2. Methods

2.1. *Overview of the RNAz algorithm*

RNAz predicts functional RNA structures on two independent criteria: (i) thermodynamic stability and (ii) structural conservation.

A common way to express thermodynamic stability is in terms of a z-score. This is simply the number of standard deviations by which the minimum free energy (MFE) deviates from the mean MFE of a set of randomized sequences with the same length and base composition. A negative z-score thus indicates that a sequence is more stable than expected by chance. As this procedure involves energy evaluation of a large set of random sequences it is not applicable for large-scale genomic screens. RNAz instead uses support vector regression (SVR) to estimate the mean and the standard deviation based on the nucleotide composition of a sequence.

RNAz evaluates evolutionary conservation of RNA structures in terms of the structure conservation index (SCI). A consensus secondary structure is predicted using the RNAalifold algorithm,[42] which is an extension of standard minimum free energy folding algorithms with the constraint that all sequences have to fold into a common structure. Compensatory mutations, i.e. mutations that preserve a certain base pair, yield bonus energies, while inconsistent mutations add penalty energies. RNAz measures structural conservation by calculating the ratio of the consensus folding energy to the unconstrained folding energies of the single sequences.

Both criteria are combined by another support vector machine model that classifies the input alignment as "structural RNA" or "other". A graphical overview of the RNAz algorithm is depicted in Fig. 1. In the following, we describe independent refinements of these steps that improve the overall prediction accuracy of the RNAz approach.

2.2. *z-score regression for dinucleotide shuffled sequences*

As in RNAz 1.0, we use support vector regression to compute z-scores for folding energies because the direct approach via repeated shuffling and folding is too costly for genome-wide applications.

In order to efficiently train the regression engine of RNAz 2.0, we used the following grid-like procedure: We first generated synthetic sequences of length 50 with G+C content, A/(A+U) ratio, and C/(C+G) ratio ranging from 0.20 to 0.80 in steps of 0.05. For each of these start sequences we then generated 500,000 mononucleotide shuffled sequences and discarded those sequences where the relative difference between the observed dinucleotide frequency and the expected frequency exceeded the threshold of 1.5. Evaluation on human ENCODE sequences showed that only a small fraction of approximately 1% of the sequences have a higher value and it was hence considered to be a reasonable threshold. Sequences of length 100, 150 and

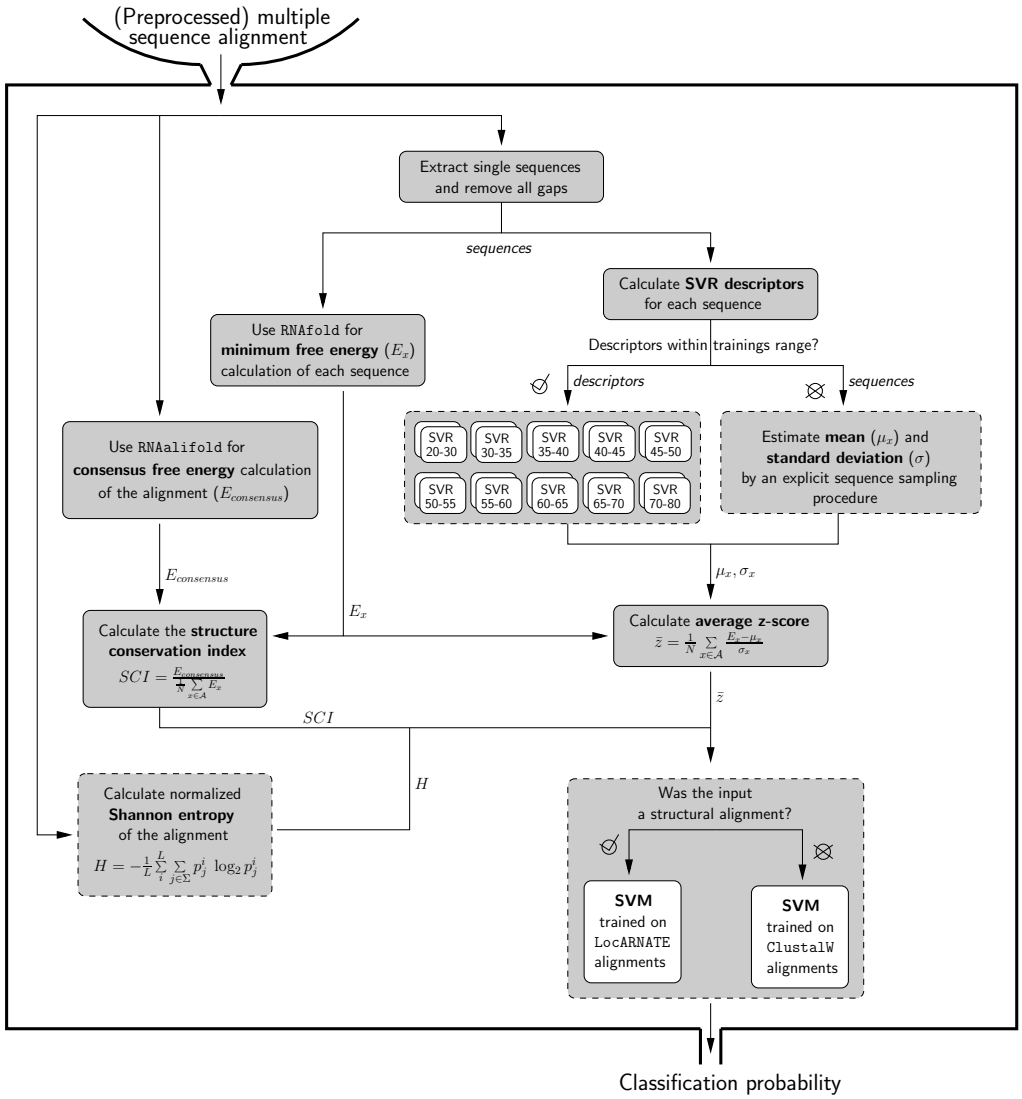

Fig. 1. Outline of the RNAz 2.0 work-flow and algorithm. In a first step large genomic multiple alignments are processed using rnazWindow.pl into smaller alignments. This filtering procedure involves several steps: (i) overlapping windows given a fixed window and step size are created, (ii) sequences that contain too many gaps are removed and (iii) from the remaining sequences only those sequences are kept that meet a predefined average pairwise identity threshold. The resulting alignments are then separately processed by RNAz. First, structure and energy predictions are performed for both the single sequences and the alignment. These results can be immediately combined to calculate the SCI as the measure of the evolutionary conservation of the RNA sequences in the alignment. In a second step, the mean free energy and the standard deviation used for the calculation of the z-score are estimated. For this purpose descriptors based on the nucleotide composition (G+C content, A/(A+U) ratio, C/(C+G) ratio, all 16 dinucleotide frequencies and the length of the sequence) are calculated for each sequence. If descriptors are within the training boundaries they are passed to the corresponding support vector regression (SVR) based on the G+C content. Otherwise, the mean and the standard deviation are evaluated explicitly by folding of 1,000 randomized sequences with the same dinucleotide composition. In a final step the average z-score of the sequences, the SCI and the normalized Shannon entropy of the alignment are passed to the classification SVM, which returns a probability estimate that the given alignment harbours thermodynamically stable and/or evolutionary conserved RNA secondary structures. Parts that are highlighted in dashed boxes are new or modified components of RNAz algorithm. RNAfold and RNAalifold are part of the Vienna RNA Package. Numbers in the SVR boxes indicate the G+C content the particular SVR is trained on. For a detailed explanation of the formulas we refer to section 2.3.

200 where then generated by concatenating the initial set of sequences 2 to 4 times. This initial set can be generated very quickly and served as the basis for the selection of a much smaller, approximately evenly

spaced, training set with representative dinucleotide frequencies. A sequence from the initial set was only added to the representative training set if the Euclidean distance of the dinucleotide frequencies to any sequence already present in the representative set was above a certain threshold (0.075 for a G+C content of 0.20 and 0.80, 0.100 for a G+C content of 0.25, 0.30, 0.70 and 0.75, and 0.125 for the remaining range). For the final training set we also added sequences of length 75, 125 and 175, which were generated as described above, resulting in a total of 1,155,737 training instances.

For each of these instances, we generated 1,000 randomized sequences by the Altschul-Erikson algorithm[37] with the same dinucleotide composition and used `RNAfold`[38] with parameter `-d2` to evaluate their folding free energy. More than 1 million training instances are by far too many to be used in SVM training procedures in reasonable time. For this reason we split the training instances into smaller subsets according to their G+C content. In total we have 10 subsets with at most 150,000 training instances. We used the SVM library `LIBSVM` to train regression models for the mean and the standard deviation for each of the ten subsets. As input features we used the G+C content, the A/(A+U) ratio, the C/(C+G) ratio, all 16 dinucleotide frequencies and the length of the sequence scaled to the interval [0,1]. The regression for estimating the mean free energy was trained to learn energy per nucleotide, while the standard deviation was not scaled. We chose the ν variant of regression and a radial basis function kernel. The standard grid search approach was used to find optimal combinations for SVM parameters. Regression accuracy was monitored on an independent test set compiled from randomly selected sequences of variable length from 50 to 200 nt from the human ENCODE regions. The average number of support vectors for the mean and the standard deviation regression models are 8,763 and 8,607, respectively.

2.3. *Training data generation and training of the SVM classifier*

Training and test sets are based on the data available in the `Rfam` 9.1 database.[39] 93 RNA families were selected based on their signals for thermodynamic stability and structural conservation. The `RNAz 2.0` training set covers a broad range of different RNA families including major classes such as tRNAs, snoRNAs, microRNAs, riboswitches, and bacterial regulatory RNAs.

For each RNA family, a set of alignments with varying numbers of sequences and average pairwise identities was generated using the following strategy: `Rfam` full alignments were used if they contained less than 300 sequences, otherwise we used the seed alignments. For our purpose the use of at most 300 sequences proofed well to generate a set of alignments over the desired range of average pairwise identities. `Rfam` alignments were utilized only as a source to retrieve family members of a particular ncRNA class and only extracted, ungapped RNA sequences were used for subsequent analyses.

First, `Rfam` alignments were filtered to remove nearly identical sequences, so that the training alignments contained sequences with at most 98% identity. The sequences were then re-aligned using `ClustalW`. For each of these ncRNA family alignments we then proceeded as follows: for each number of sequences from 2 to 15 we generated at most 10 alignments with a randomly chosen average pairwise identity between 50 and 98% and with a maximum relative difference in sequence lengths of 65% using `rnazWindow.pl` which is part of the `RNAz` analysis pipeline.[44]

To ensure that this set of positive training examples contained only instances with good structural conservation signals we filtered alignments by using tree editing distances between the structures of the sequences in the alignment as a quality measure of structural conservation. Ordered, rooted trees can be deduced from the dot-bracket notation of RNA secondary structures. Tree editing defines a metric in the space of trees by a set of operations (deletions, insertion and relabeling of nodes) and hence can be used to calculate distances between RNA secondary structures.[31] For each alignment we extracted sequences, removed gaps and calculated the averaged pairwise tree editing distance using `RNAdistance` with options `-d2 -Dh` to enable dangling ends and to use the HIT representation for RNA secondary structures. We repeated this for a set of 100 randomized alignments and calculated an empirical p-value as a measure of structural conservation. Alignments with a p-value higher than 0.05 were removed from the training set. Alignments retained after this filtering procedure were realigned with `ClustalW` with standard options for

application to sequence-based alignments.

For the generation of structural alignments for the training set we chose to use `LocARNATE`,[40] which is a structural alignment program based on the Sankoff algorithm for the simultaneous solution of the RNA folding and the alignment problem. `LocARNATE` uses `RNAfold` for structure predictions and hence the same energy parameters as `RNAz` does. `LocARNATE` was called with options `--no-seq --no-struc` to generate global, structural alignments.

Negative instances of the training set were generated by shuffling using `multiperm`[41] v. 0.9.3 if the normalized Shannon entropy of the alignment[31] was less than 0.50. Otherwise, alignments were simulated using `SISSIz`[32] to ensure full randomization for the more diverse alignments where shuffling can become inefficient. The final training set was composed of 10,538 alignments for each the positive and the negative class.

The `RNAz 2.0` SVM classifier uses three features to detect structured noncoding RNAs: (i) the average minimum free energy z-score $\bar{z}$ estimated from a dinucleotide shuffled background, (ii) the SCI and (iii) the normalized Shannon entropy H of the alignment as a measure for the content of evolutionary information.

Consider an alignment $\mathcal{A}$ consisting of N sequences. Let E_x denote the minimum free energy of sequence x, and let μ_x and σ_x be the mean and standard deviation, respectively, of the folding energies of a large number of random sequences of the same length and same dinucleotide composition as x. The averaged z-score of the alignment $\mathcal{A}$ is defined as

$$\bar{z} = \frac{1}{N} \sum_{x \in \mathcal{A}} \frac{E_x - \mu_x}{\sigma_x}$$

The SCI of an alignment is given as the fraction of the consensus folding free energy ($E_{consensus}$) to the average of the folding free energies of the single sequences:

$$SCI = \frac{E_{consensus}}{\frac{1}{N} \sum\limits_{x \in \mathcal{A}} E_x}$$

The normalized Shannon entropy H of an alignment $\mathcal{A}$ of RNA sequences over the alphabet $\Sigma = \{$A, C, G, U, -$\}$ is defined as the sum of the Shannon entropies of the individual columns divided by the length of the alignment denoted by L:

$$H = -\frac{1}{L} \sum_{i}^{L} \sum_{\alpha \in \Sigma} p_\alpha^i \ \log_2 \ p_\alpha^i$$

The probability p_α^i is approximated by the observed frequency of character α in alignment column i (normalized by the number N of sequences in the alignment). All features were scaled to a range of [-1,1]. Standard grid search combined with a 10-fold cross validation was applied to find optimized SVM parameters. Among the models with the best cross-validation accuracy (top 20) we chose the model that showed best performance on an independent test set created the same way as the training set. The output of the final classification SVM is a probability estimate that the input alignment contains thermodynamically stable and/or structurally conserved RNA sequences.

A second, independent, SVM classifier was trained on sequence/structure-based alignments generated by `LocARNATE` using the same procedure.

3. Results

3.1. *Dinucleotide based z-scores*

To estimate the mean and standard deviation of folding energies for mononucleotide shuffled sequences it is feasible to sample uniformly simply by varying variables describing the four mononucleotide frequencies and the length of the sequence on a grid. This approach cannot, however, be extended that easily for dinucleotide shuffled sequences. One has to consider the much larger space of dinucleotide compositions that is occupied by

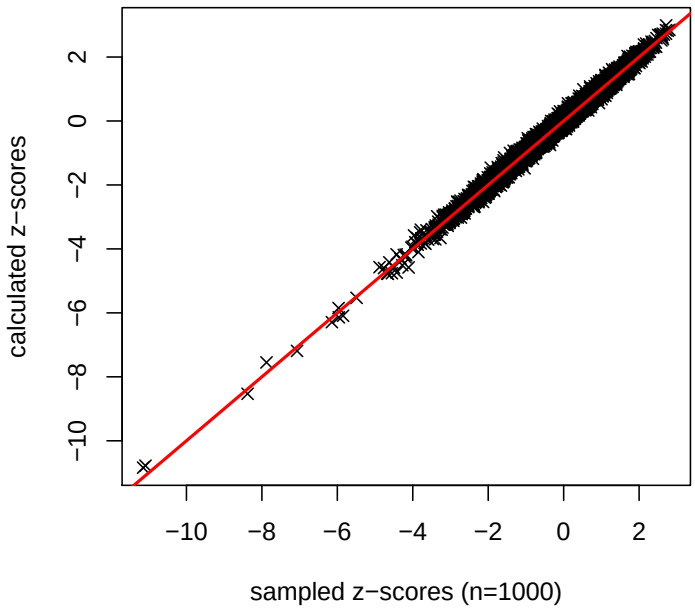

Fig. 2. z-scores calculated by support vector regression in comparison with z-scores determined from 1,000 random samples preserving dinucleotide frequencies for 10,000 randomly drawn sequences from the human ENCODE regions. Correlation of z-scores is 0.996 and the mean absolute error is 0.076.

sequences of practical interest. In this work we use a grid-like approach, where we first apply uniform sampling to cover the mononucleotide space and then choose, for each data point in the grid, a representative set of sequences that covers the dinucleotide space for that particular base composition. However, this procedure still gave more than one million training instances. The training data was split into different ranges of the G+C content to guarantee efficient training and fast prediction. This comes at the price of increased memory consumption but keeps the number of support vectors comparable to the approach used in `RNAz 1.0`. Accuracy of the z-score regression for dinucleotide shuffled sequences was evaluated on 10,000 randomly chosen sequences of variable length from 50 to 200 nt from the human ENCODE regions[34] (Fig. 2) and genomic sequences of *D. melanogaster* and *E. coli*. The mean absolute error (MAE) and the correlation (R) of z-scores calculated by SVM regression compared to z-scores determined from 1,000 random samples is 0.0748 and 0.996, respectively ($n = 30,000$; genomic sequence from ENCODE regions, *D. melanogaster*, and *E. coli*). Comparisons of z-scores determined from 1,000 dinucleotide shuffled sequences to 100 dinucleotide shuffled sequences (MAE= 0.107, $R = 0.992$) and to 1,000 mononucleotide shuffled samples (MAE= 0.420, $R = 0.916$) clearly demonstrate that our method is a suitable approach for fast and efficient estimation of dinucleotide controlled z-scores. `RNAz 1.0` also showed restrictions on the base composition because of the training range of the SVR. This limitation is now overcome by explicit generation of shuffled sequences once the base composition of a sequence is out of the training range. Since boundaries have been chosen broadly (e.g. G+C content from 20 to 80%) this will only apply in a small minority of cases.

3.2. *New training sets and improved classification model*

Since the postulation of the SCI, it has been a major point of criticism that the SCI evaluates structural conservation on the energy level rather than on the RNA structures themselves. However, in previous study[31] it has been shown that the SCI is on average the most powerful method and that it is only outperformed by

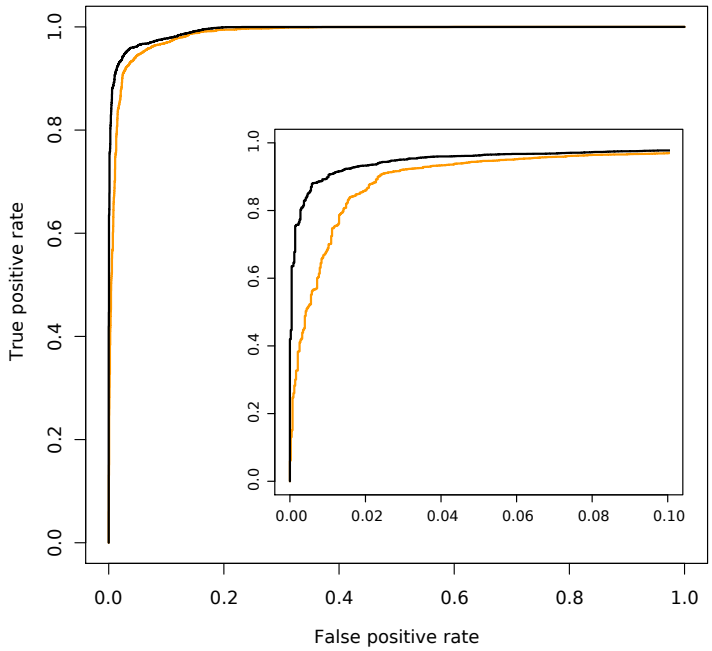

Fig. 3. Accuracy of `RNAz 2.0` classification (black) vs. `RNAz 1.0` classification (orange) on a previously published data set for the evaluation of noncoding RNA gene finders.[32] The positive instance data set consists of 4,303 alignments of structural RNA families (5S ribosomal RNA, U2 spliceosomal RNA, tRNA, Hammerhead ribozyme, U3 snoRNA, U5 spliceosomal RNA, Group II catalytic intron, and Mir-10 microRNA) with two to six sequences per alignment. The negative instance data set consists of 4,303 alignments taken from random genomic location, which resemble approximately the same dinucleotide composition and conservation degree as the positive set. The inset shows the region of high specificity were `RNAz 2.0` clearly outperforms the old version.

other approaches in the high sequence identity range. Attempts to use other conservation measure methods than the SCI, however, failed to give results of comparable quality (data not shown).

To use the SCI for efficient classification one has to take into account the average pairwise identity and the number of sequences as well. Due to the lack of comparative data at the time of training of the initial `RNAz` algorithm limits on these two descriptors were rather arbitrarily chosen. In this work we generated a new training set covering a broader range of RNA families and evaluate sequence variation in terms of the normalized Shannon entropy which has been shown to combine both sequence variation and the number of sequences into one measure.[31] This does not only result in dimensionality reduction of the final classification model, but also overcomes the need to set an upper boundary to the number of sequences in an alignment.

The new `RNAz 2.0` algorithm now uses the average z-score of the sequences in the alignment based on a dinucleotide background model, the SCI and the normalized Shannon entropy as features in the final classification model. To evaluate the predictive power of `RNAz 2.0` we chose a test set used in a previous study.[32] This test set is especially well suited as it contains randomly chosen genomic regions from vertebrate alignments as negative controls. The background dinucleotide content in vertebrate genomes is known to be the main reason for false positive calls in `RNAz 1.0`.[33] Although both versions perform well on this test set, `RNAz 2.0` clearly outperforms version `1.0` in the high specificity range (Fig. 3). For example, at the generally used 0.01 false-positive cutoff, `RNAz 2.0` shows 0.899 sensitivity compared to 0.688 in the old version.

It is a well known fact that sequence-based alignment methods fail to give high quality alignments regarding RNA secondary structures in low average pairwise identity ranges. By using structural alignments one can expect an improvement in discrimination capability of the SCI for alignments with low sequence

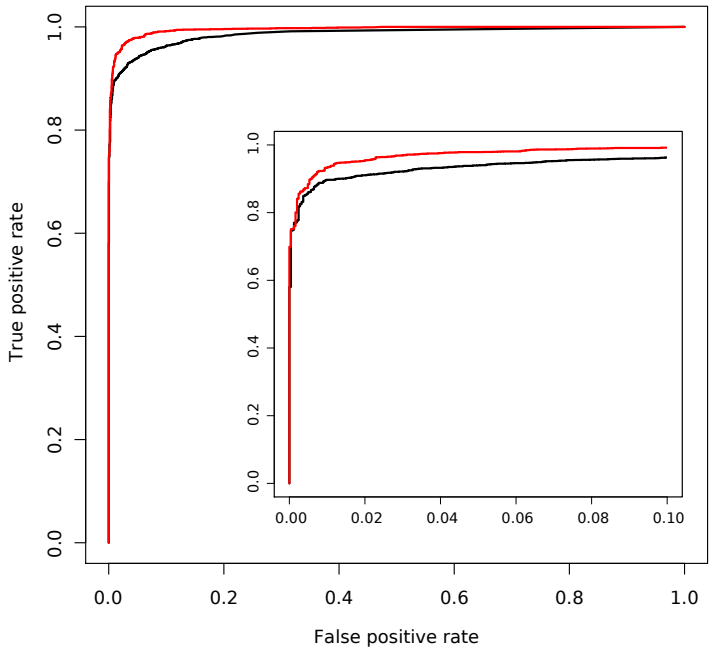

Fig. 4. ROC curves for the RNAz 2.0 prediction accuracy on sequence-based alignments (black) vs. structural alignments (red). A significant improve of the overall predictive power of RNAz 2.0 is achieved by use of structural alignments. The test set is composed of 2,455 alignments of various ncRNA families with an average pairwise identity between 30 and 70%, as well as a negative set consisting of 2,455 alignments derived by randomization of reference alignments with multiperm or SISSIz as described in section 2.2. Sequence-based alignments were generated with ClustalW, while structural alignments were generated with LocARNATE.

similarity.[11] Therefore, we trained a separate SVM decision model based on sequence/structure alignments, similar to the approach used in RSSVM.[30] Structural alignments were generated using LocARNATE, a multiple alignment variant of LocARNA.[46] As depicted in Fig. 4 structural alignments improve the overall predictive power of RNAz.

Recent studies (e.g. Washietl *et al.*[33]) have shown that RNAz suffers from a high false discovery rate (FDR). We therefore evaluated the performance of both versions for the human ENCODE regions. 17-way MAF alignments based on the human genome assembly hg.17 were downloaded from the UCSC genome browser. In total we screened 193,634 MAF alignments derived by pre-filtering with rnazWindow.pl with standard options (window length is 120 nt, step size is 40 nt, average pairwise identity the resulting alignment is optimized to is 80%, and at most six sequences are allowed). Both reading directions were considered in our analysis. A dinucleotide background model was generated with SISSIz[32] and all hits detected by RNAz on this data set were considered to be false positives. Results are summarized in Tab. 1. While RNAz 1.0 shows a very high FDR of around 80%, the FDR of RNAz 2.0 is much lower being around 54% for high confident hits (classification probability > 0.9). It seems noteworthy, that in a previous study[33] the FDR for RNAz 1.0 on ENCODE data was estimated to be around 50%. This estimate was based on a rather simplistic *ad hoc* method to correct for the dinucleotide bias. The new results are based on the more accurate SISSIz null model and demonstrate that RNAz 1.0 is even more affected by the dinucleotide bias than previously assumed. The new version, however, reduces this source of false positives significantly.

To investigate a potential G+C bias of RNAz that was observed for version 1.0,[33] we also trained a classification model that included the G+C content as fourth feature. This additional feature, however, had little impact on the predictions. In particular, the distribution of the G+C content of the positive predictions

Table 1. Comparison of the false discovery rate (FDR) based on ENCODE regions and a dinucleotide background model for low (P > 0.5) and high (P > 0.9) confidence hits. A hit corresponds to a single alignment derived from pre-filtering of ENCODE MAF alignments with `rnazWindow.pl`.

	`RNAz 1.0`		`RNAz 2.0`	
	# low conf.	# high conf.	# low conf.	# high conf.
ENCODE regions	17,814	6,854	6,880	2,259
background	14,489	5,596	4,090	1,219
estimated FDR	81%	82%	59%	54%

remained nearly unchanged (data not shown). This suggests that the elevated G+C content of `RNAz` hits is not an artificial bias, but rather reflects the G+C content of true functional RNAs. Consistent with this observation, the G+C bias of structured RNAs has been used successfully for *de novo* prediction of RNA genes.[43] Preliminary analysis of the ENCODE data showed that the effect is smaller for `RNAz 2.0` than in the earlier version.

3.3. *Computational speed*

The performance of `RNAz 2.0` in comparison to `RNAz 1.0` was benchmarked on 50,000 randomly chosen MAF alignments from the ENCODE data set. Alignment length was 120 nucleotides and alignments contained at most six sequences. Experiments were conducted on an Intel Xeon 2.40GHz CPU. For each alignment both reading directions were examined, resulting in a total of 100,000 alignments that had to be scored. The execution time required by `RNAz 1.0` was 202 min, `RNAz 2.0` with explicit shuffling switched off was 252 min and `RNAz 2.0` using explicit shuffling was 1,230 min. Although explicit shuffling had to be used for only 1% of the sequences (5,524 out of 549,210), it comes with an tremendous overhead increasing the run time of `RNAz 2.0` almost 5-fold. We extracted those alignments where explicit shuffling was used and compared the classification probability to the one derived from calling `RNAz` with option `--no-shuffle` to avoid explicit shuffling. For the vast majority of cases (96%) the change in classification probability was less than 1%. For this data set the maximal observed difference was 0.21. In general, we observed larger differences in the range from 0.2 to 0.8 than in the regions close to 0 or 1.

With option `--no-shuffle`, `RNAz 2.0` has an execution time that is increased by about 25% compared to `RNAz 1.0`.

4. Future directions

In this work we present a major update of the `RNAz` algorithm. Evaluation of thermodynamic stability has been improved by considering a dinucleotide background model. This directly translates into a significantly lower false discovery rate. In addition to the dinucleotide z-score, the overall prediction accuracy is improved by a combination of the use of a new training set and the normalized Shannon entropy as a measure of sequence variation. Furthermore, the updated version is not any more restricted to limitations concerning the base composition or number of sequences in the input alignment.

The generation of structural alignments is computationally expensive but we showed that they can improve the `RNAz` classification power. This is true in particular for alignments of low average pairwise identity. Given that the overall computational complexity of `LocARNATE` is $O(n^4)$, the routine use of structural alignments on a genome-wide scale is still out of reach, at least when off-the shelf hardware is used. In general, it has to be questioned if ncRNA gene finding would benefit from realigning genomic alignments available to date with a structural aligner. These alignments have been generated by means of sequence-only based methods and therefore are not likely to contain homologous RNA sequences that evolve fast on nucleotide level but retain structural conservation. A feasible strategy, however, is the pre-selection of

syntenic regions based on better-conserved flanking regions.[13] Such an approach could be employed for the detection of conserved local structures in the untranslated regions of protein-coding mRNAs, where orthology is established based on similarities of the much better conserved coding sequences. The re-scoring of positively scored hits of a sequence-based `RNAz` screen after re-aligning them with a structural aligner may help to increase the overall accuracy, in particular for relatively poorly conserved alignment slices. One could also use `RNALfold`[45] augmented with the z-score prediction engine of `RNAz` to screen for loci that show signature of increased thermodynamic stability then re-evaluate these loci using structural alignments with `RNAz 2.0` to also account for structural conservation.

An open question, not covered by this work, is how to address the growing number of species in genomic alignments. The use of the normalized Shannon entropy helped us to remove the upper limit on the number of sequences in the alignment. Preliminary analysis of `RNAz 2.0` on multiz 44-way, 28-way and 17-way alignments shows, however, that the simple use of more sequences does not necessarily correlated with improved classification power. To a large extent the increased conservation signal is counteracted by increasing levels of alignment errors. Structural variation of the ncRNAs themselves also poses technical challenges. To-date, an algorithm that addresses both possible misalignments and structural variation is still missing.

RNA secondary structure prediction is sensitive to the exact ends of the input sequence. The use of arbitrarily determined alignment windows of fixed width thus introduces noise. This issue will be alleviated in a forthcoming update of `RNAz` that addresses the pre-processing of long genomic alignments. Here, the sliding window approach will be replaced by the systematic use of `RNALalifold`,[47] an algorithm that computes locally stable consensus RNA secondary structures. These are then used to extract alignments of self-contained (sub)structures for `RNAz` scoring.

`RNAz 2.0` was trained on two particular alignment methods, `ClustalW` for sequence-based alignments and `LocARNATE` for structure-based alignments. As `RNAz` uses a machine learning approach, we have to expect some influence of the alignment algorithm since the features passed to the SVM implicitly also incorporate properties of the alignment algorithms themselves. It may thus become necessary to either re-align the input data or to train decision models for alternative alignment methods.

Supplementary material

An Electronic Supplement located at `www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-026/` compiles a supplemental figure and data sets used in this work.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network III" and "noncoding RNA II", the University of Vienna and the German Research Foundation (grants STA 850/7-1 under the auspices of SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes").

References

1. P. P. Amaral *et al.*, *Science* **319**, 5871 (2008).
2. T. Kuwabara *et al.*, *Cell* **116**, 6 (2004).
3. J. Feng *et al.*, *Genes Dev* **20**, 11 (2006).
4. C. A. Espinoza *et al.*, *RNA* **13**, 4 (2007).
5. A. Pagano *et al.*, *PLoS Genet* **3**, 2 (2007).
6. A. Wutz, *Trends Genet* **23**, 9 (2007).
7. J. L. Rinn *et al.*, *Cell* **129**, 7 (2007).
8. P. P. Amaral and J. S. Mattick, *Mamm Genome* **19**, 7-8 (2008).
9. E. Rivas and S. R. Eddy, *BMC Bioinformatics* **2**, (2001).
10. J. S. Pedersen *et al.*, *PLoS Comput Biol* **2**, 4 (2006).
11. A. V. Uzilov *et al.*, *BMC Bioinformatics* **7**, (2006).
12. Z. Yao *et al.*, *Bioinformatics* **22**, 4 (2006).

13. E. Torarinsson *et al.*, *Genome Res* **16**, 7 (2006).
14. S. Washietl *et al.*, *Proc Natl Acad Sci USA* **102**, 7 (2005).
15. S. Washietl *et al.*, *Nat Biotechnol* **23**, 11 (2005).
16. K. Missal *et al.*, *Bioinformatics* **21**, (2005).
17. K. Missal *et al.*, *J Exp Zoolog B Mol Dev Evol* **306**, 4 (2006).
18. D. Rose *et al.*, *BMC Genomics* **8**, (2007).
19. D. Rose *et al.*, *J Bioinform Comput Biol* **6**, 6 (2008).
20. A. M. McGuire and J. E. Galagan, *PLoS One* **7**, 3 (2008).
21. C. Weile *et al.*, *BMC Genomics* **8**, (2007).
22. T. Mourier *et al.*, *Genome Res* **18**, 2 (2008).
23. C. del Val *et al.*, *Mol Microbiol* **66**, 5 (2007).
24. J. Hertel *et al.*, *Bioinformatics* **24**, 2 (2008).
25. J. Hertel *et al.*, *Bioinformatics* **22**, 14 (2006).
26. K. Reiche *et al.*, *Algorithms Mol Biol* **2**, (2007).
27. P. P. Gardner *et al.*, *Nucleic Acids Res* **33**, 8 (2005).
28. P. W. Hsu *et al.*, *Nucleic Acids Res* **33**, (2006).
29. T. Sandmann and S. M. Cohen, *PLoS ONE* **2**, 11 (2007).
30. X. Xu *et al.*, *PLoS Comput. Biol.*, **5**, (2009).
31. A. R. Gruber *et al.*, *BMC Bioinformatics* **9**, (2008).
32. T. Gesell and S. Washietl, *BMC Bioinformatics* **9**, (2008).
33. S. Washietl *et al.*, *Genome Res* **17**, 6 (2007).
34. ENCODE Project Consortium, *Nature* **447**, 7146 (2007).
35. C. Workman and A. Krogh, *Nucleic Acids Res* **27**, 24 (1999).
36. T.Babak *et al.*, *BMC Bioinformatics* **8**, (2007).
37. S. F. Altschul and B. Erickson, *Mol Biol Evol.* **2**, 6 (1985).
38. I. L. Hofacker *et al.*, *Monatsh. Chem.* **125**, (1994).
39. P. P. Gardner *et al.*, *Nucleic Acids Res.* **37**, (2008).
40. W. Otto *et al.*, *Proceedings of the German Conference on Bioinformatics* **P-136**, (2008).
41. P. Anandam *et al.*, *Bioinformatics* **25**, (2009).
42. I. L. Hofacker *et al.*, *J.Mol.Biol.* **319**, (2002).
43. M. M. Meyer *et al.*, *BMC Genomics* **10**, (2009).
44. S. Washietl, *Methods Mol Biol.* **395**, (2007).
45. I. L. Hofacker *et al.*, *Bioinformatics* **20**, (2004).
46. S. Will *et al.*, *PLoS Comput. Biol.*, **3**, (2007).
47. A. F. Bompfünewerer Consortium, *J Exp Zoolog B Mol Dev Evol.* **308**, (2007).