Fast and reliable prediction of noncoding RNAs

Stefan Washietl[1], Ivo L. Hofacker[1] & Peter F. Stadler[1,2,*]

[1]Department of Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

[2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

[*]Corresponding author. studla@bioinf.uni-leipzig.de

Abstract

We report an efficient method to detect functional RNAs. The approach, which combines comparative sequence analysis and structure prediction, yields excellent results already for a small number of aligned sequences and is suitable for large scale-genomic screens. It consists of two basic components: (1) a novel measure for RNA secondary structure conservation based on computing a consensus secondary structure, and (2) a measure for thermodynamic stability, which — in the spirit of a z-score — is normalized w.r.t. both sequence length and base composition but can be calculated without sampling from shuffled sequences. Functional RNA secondary structures can be identified in multiple sequence alignments with high sensitivity and high specificity. We demonstrate that this approach is not only much more accurate than previous methods but also significantly faster. The method is implemented in the program RNAz, which can be downloaded from http://www.tbi.univie.ac.at/~wash/RNAz/.

As a first application we screened all alignments of length $N \geq 50$ in the CORG database, which compiles conserved non-coding elements in upstream regions of orthologous genes from human, mouse, rat fugu and zebrafish. We recovered *all* of the known non-coding RNAs and *cis*-acting elements with high significance and found compelling evidence for many other conserved RNA secondary structures not described so far.

1 Introduction

Traditionally, the role of RNA in the cell was considered mostly in the context to protein gene expression, limiting RNA to its function as mRNA, tRNA, and rRNA. The discovery of a diverse array of transcripts that are not translated to proteins but rather function as RNAs has changed this view profoundly [1, 2, 3]. Non-coding RNAs (ncRNAs) are involved in a large variety of processes, including gene regulation [4], maturation of mRNAs, rRNAs, and tRNAs, or X-chromosome inactivation in mammals [5]. In fact, a large fraction of the mouse transcriptome consists of non-coding RNAs [6], and about half of the transcripts from Human chromosomes 21 and 22 are non-coding [7, 8]. Structured RNA motifs furthermore function as *cis*-acting regulatory elements within protein coding genes. Also here, new intriguing mechanisms are being discovered [9].

Hence, a comprehensive understanding of cellular processes is impossible without considering RNAs as key players. Efficient identification of *functional* RNAs (ncRNAs as well as *cis*-acting elements) in genomic sequences is, therefore, one of the major goals of current bioinformatics. Notwithstanding its utmost biological relevance, *de novo* prediction is still a largely unsolved issue. Unlike protein coding genes, functional RNAs lack in their primary sequence common statistical signals that could be exploited for reliable detection algorithms. Many functional RNAs, however, depend on a defined secondary structure. In particular, evolutionary conservation of secondary structures serves as compelling evidence for biologically relevant RNA function. Comparative studies therefore seem to be most promising approach. To date, complete genomic sequences of related species have been sequenced for almost all genetic model organisms as for example bacteria [10, 11], yeasts [12], nematodes [13, 14] and even mammals [15, 16, 17]. Several studies identified a large collection of evolutionary conserved non-coding elements in mammalian (or, more generally, vertebrate) genomes and it must be expected that a significant fraction of them are functional RNAs [18, 19, 20, 21].

Possible candidates, however, were identified only sporadically so far [19, 21], simply because there are no reliable tools to scan multiple sequence alignments for functional RNAs. The most widely used program `QRNA` [22], which has been successfully used to identify ncRNAs in bacteria [23] and yeast [24], is not suitable for screens of large genomes: `QRNA` is limited to pairwise alignments and its reliability is low especially if the evolutionary distance of the two sequences lies outside the optimal range. An alternative approach, `ddbRNA` [25], suffers from similar problems and so far has not been used in a real-life application. `MSARi` [26], on the other hand, gains its drastically enhanced accuracy from the large amount of information contained in large multiple sequence alignments of 10–15 sequences with high sequence diversity. At present, however, data-sets of this kind are simple not available at a genome-wide scale at least for multicellular organisms.

In this contribution we address the problem using an alternative approach: We combine a measure for thermodynamic stability with a novel measure for structure conservation. Using a combination of both scores we are able to efficiently detect functional RNAs in multiple sequence alignments of only a few sequences. Our method is substantially more accurate than `QRNA` or `ddbRNA` and performs better on on pairwise alignments than `MSARi` does on alignments with 15 sequences. On the large diverse alignments used for testing `MSARi` in ref. [26], our `RNAz` program achieves 100% sensitivity at 100% specificity.

2 Results

2.1 The structure conservation index

In a recent contribution [27] we have demonstrated that the program `RNAalifold` (which was originally developed for *prediction* of secondary structure in aligned sequences [28]) can also be used for *detection* of evolutionarily conserved secondary structure. `RNAalifold` implements a consensus folding algorithm generalizing the standard dynamic programming algorithms for RNA secondary structure prediction algorithms [29] by adding sequence covariation terms to the folding energy model [30, 31]. More precisely, a consensus minimum free energy (MFE) is computed for an alignment that is composed of an energy term averaging the energy contributions of the single sequences and a covariance term rewarding compensatory and consistent mutations [28]. As this consensus MFE is difficult to interpret in absolute terms, we previously used a time-consuming random sampling method to assess its significance [27]. This approach would require massive computational effort even for small-sized genomes and it does not seem practicable for large genomes as for example the human genome.

A much more efficient normalization can be achieved, however, by comparing the consensus MFE with the MFEs of each individual sequence in the alignment. To this end, we fold the alignment and calculate the consensus MFE E_A of the alignment using `RNAalifold`. If the sequences in the alignment fold into a conserved common structure, the average $\bar{E}$ of the individual MFEs will be close to the MFE of the alignment, $E_A \approx \bar{E}$. Otherwise, the MFE of the alignment will be much higher (indicating a less stable structure) than the average of the individual sequences, $E_A \gg \bar{E}$. We therefore define the *structure conservation index* (SCI) as

$$\text{SCI} = E_A/\bar{E}$$

An SCI close to zero indicates that `RNAalifold` does not find a consensus structure while a set of perfectly conserved structures has SCI ≈ 1. A SCI larger than 1 indicates a perfectly conserved secondary structure which is in addition supported by compensatory and/or consistent mutations, which contribute a covariance score to E_A.

2.2 A normalized measure for thermodynamic stability

It is widely believed that MFE-predictions cannot be used for detection of functional RNAs after an in-depth study on the subject [32]. Although thermodynamic stability is not significant alone, it still can be used as valuable diagnostic feature since functional RNAs are indeed more stable than random sequences to some degree [32, 27]. This effect is particularly dramatic in the case of microRNA precursors [33].

The significance of a calculated MFE value m is assessed by comparison with a large sample of random sequences. This approach was introduced 16 years ago [34] and it is still widely used today [35, 18, 19]. Typically, the normalized z-score $z = \frac{m-\mu}{\sigma}$ is used, where μ and σ are the mean and standard deviations, resp., of a large number of random sequences of the same length and same base or dinucleotide composition.

The parameters μ and σ are, by construction, functions of length and base composition. In the case of RNA molecules we found that they can be computed very accurately from a relatively simple regression model which we obtained by means of

a standard implementation of a support vector machine (SVM) [36] regression algorithm (see Methods for details). In order to calibrate the model we used 1000 random sequences for each of approximately 10000 points evenly spaced in the variable space spanned by chain length and base composition. Independent SVMs were trained for μ and σ.

The accuracy of the SVM regression model is verified by comparing z-scores from the SVM approach with z-scores obtained by sampling, Fig. 1. We find that the correlation between sampled values and SVM values is as good as the correlation between two independently sampled z-scores for the same test sequence at a sample size of 1000. We can therefore replace the time-consuming sampling procedure by the SVM estimate without a significant loss of accuracy, while saving about a factor of 1000 in CPU time.

2.3 Classification based on both scores

In order to to classify alignments as a "functional RNA" or "other" we have to determine the separatrix between "functional RNAs" and "other sequences" in the SCI/z-score plane. Again, this is a typical application for SVMs; we therefore trained a binary classification SVM on test sets encompassing all major known classes of ncRNAs:

We generated test alignments using `ClustalW` of 12 well known ncRNA classes from Rfam [37] as well as random controls for which any native secondary structure is removed by shuffling the alignment positions (see Methods), and computed z-score and SCI. Fig. 2 illustrates the results for a test set of tRNAs and 5S-rRNAs. Supplementary Fig. 1 shows the results for the other ncRNA classes. We find that the combination of both scores reliably separate the native alignments from the randomized controls in two dimensions.

In order to improve the performance of the binary classification SVM we use not only z-score and SCI but also the mean pairwise identity and the number of sequences in the alignment as input parameters. In essence, this teaches the SVM to interpret the information contained in the numerical value of the SCI depending on the sequence variation in the alignment. This is necessary because the information content of a multiple alignment strongly depends on these parameters: in the extreme case, an alignment of identical sequences has SCI $= 1$ but does not contain any information about structural conservation at all. Since we use a randomized control which has the same number of sequences and the same pairwise sequence conservation together with each positive example, the calibration process is not biased by these additional variables.

The class probability p estimated by the SVM provides a convenient significance measure. Tab. 1 shows the sensitivity and specificity for detecting different ncRNA classes at different probability cutoffs. We used alignments with mean pairwise sequence identities between 60% and 100% and 2–4 sequences per alignment. At a cutoff of $p = 0.9$, we can detect on average 75.27% at a specificity of 98.93%.

The accuracy of the classification depends quite strongly on the type of the ncRNA. We can find most RNA classes with high sensitivities in the range of 80%–100%. Only two of the twelve classes in out test set (U70 snoRNA and tmRNA) are difficult to detect. The scatter-plots (Supplementary Fig. 1) show that the U70 is quite stable but not very well conserved, whereas the tmRNA has a conserved secondary structure that is obviously not very stable and moreover contains pseudo-knots. Alignments with more sequences are needed to detect also these two RNA classes quantitatively.

We emphasize that, although we use here a machine learning approach for classification, we do **not** train the SVM on specific sequences, sequence patterns, structure

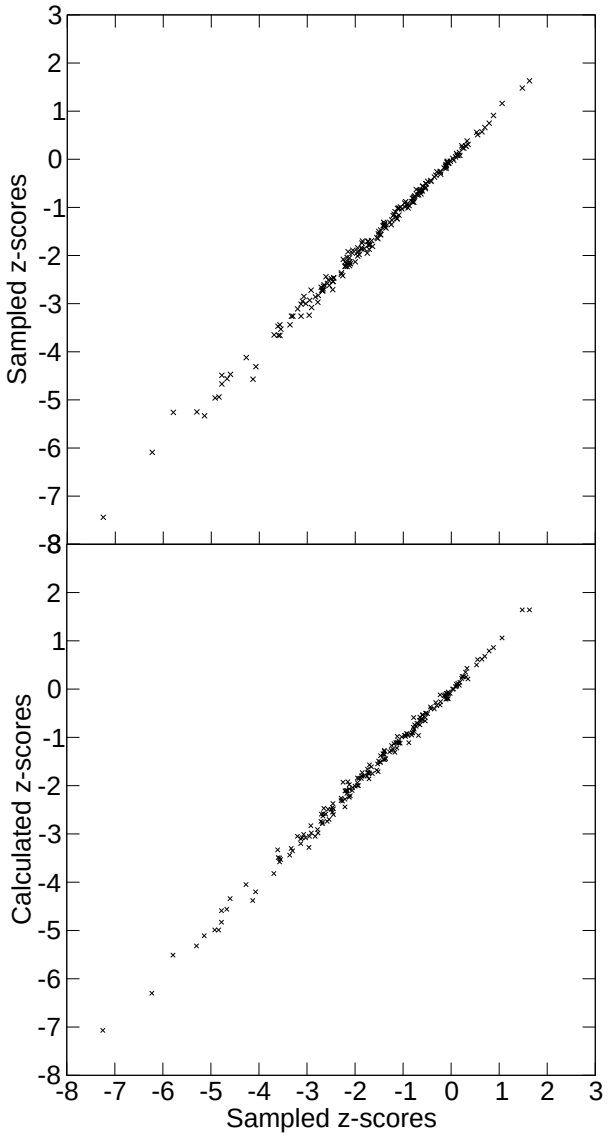


Figure 1: z-scores calculated by support vector machine regression in comparison to z-scores determined from 1000 random samples for each data point. As test sequences we chose 100 sequences from random locations in the human genome and 100 known ncRNAs from the `Rfam` database [37]. Upper panel: Correlation of z-scores from two independent samplings (mean squared error: 0.00990). Lower panel: Correlation of calculated z-scores and sampled z-scores (mean squared error: 0.00998)

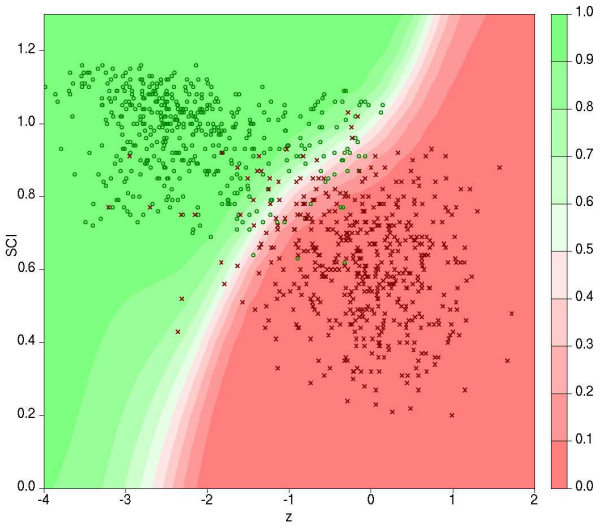


Figure 2: Classification based on z-scores and SCI using a support vector machine. Alignments of tRNAs and 5S-rRNAs with 2–4 sequences per alignment and mean pairwise identities between 60% and 90% are shown. Green circles represent native alignments, red crosses represent shuffled random controls. The background color ranging from red to green indicates the RNA-class probability for different regions of the z–SCI-plane.

motifs, conservation patterns, or base-compositions. We use the SVM solely as a guide to interpret the SCI and z-score which represent two diagnostic features that do not contain any information that is specific for a particular class of ncRNAs. In fact, it would be interesting to replace the SVM by a direct statistical model. In order to demonstrate that our classification procedure is generally applicable and not biased towards ncRNA classes of the training set, we trained the SVM excluding particular classes of ncRNAs and used those models to classify the excluded ncRNAs and their randomized controls. The sensitivities summarized in Tab. 1 can therefore also be expected for novel classes of structured ncRNAs.

2.4 Comparison to other methods

RNAseP and SRP RNAs have repeatedly been used for benchmarking ncRNA detection algorithms [22, 26]. We therefore use these datasets here as well. For the comparison to QRNA and ddbRNA we used pairwise and three-way alignments with mean pairwise identities between 60% and 90%, respectively. In contrast to the previous section we exclude alignments with identities higher than 90% since both QRNA and ddbRNA are known to perform poorly on such input data. We used a cut-off of $p = 0.9$ for RNAz and chose the cut-offs for the other programs in a way that the specificity is at least 90%. Results are summarized in Tab. 2. We find that RNAz is substantially more sensitive on both pairwise and three-way alignments than QRNA and ddbRNA and at the same time has a larger specificity.

Table 1: Detection performance for different classes of ncRNAs

		Cutoff					
		0.5		0.9		0.99	
ncRNA Type	N	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
5S ribosomal RNA	297	81.48 (242)	96.63 (10)	68.69 (204)	99.33 (2)	33.00 (98)	100.00 (0)
tRNA	329	94.83 (312)	93.62 (21)	90.27 (297)	97.87 (7)	75.68 (249)	99.70 (1)
SRP RNA	464	100.00 (464)	96.55 (16)	96.55 (448)	98.92 (5)	66.16 (307)	100.00 (0)
RNAse P	291	98.97 (288)	96.22 (11)	84.19 (245)	99.31 (2)	56.70 (165)	100.00 (0)
U2 spliceosomal RNA	351	98.58 (346)	97.72 (8)	95.44 (335)	99.15 (3)	66.67 (234)	99.72 (1)
U5 spliceosomal RNA	285	91.58 (261)	98.25 (5)	81.75 (233)	100.00 (0)	70.53 (201)	100.00 (0)
U3 snoRNA	277	83.75 (232)	98.56 (4)	62.82 (174)	99.28 (2)	44.40 (123)	99.64 (1)
U70 snoRNA	363	61.16 (222)	96.69 (12)	35.54 (129)	98.90 (4)	17.91 (65)	99.72 (1)
Hammerhead III ribozyme	271	100.00 (271)	95.20 (13)	98.15 (266)	98.89 (3)	89.67 (243)	99.26 (2)
Group II catalytic intron	407	78.62 (320)	96.31 (15)	76.90 (313)	98.53 (6)	25.31 (103)	100.00 (0)
tmRNA	386	24.87 (96)	96.37 (14)	18.65 (72)	98.19 (7)	8.55 (33)	99.48 (2)
micro RNA mir-10	380	100.00 (380)	95.26 (18)	97.63 (371)	99.21 (3)	62.37 (237)	100.00 (0)
Total	4101	84.17 (3452)	96.42 (147)	75.27 (3087)	98.93 (44)	50.18 (2058)	99.80 (8)

Results for alignments with 2–4 sequences and mean pairwise identities between 60% and 100% are shown. N is the number of alignments in the test set. For each native alignment, one randomized alignment was produced, and randomized alignments classifi fed as ncRNA were counted as false positives. Sensitivity and specifi city are shown in percent for three cutoffs of the RNA class probability predicted by the SVM. Absolute numbers of true positives and false negatives are shown in brackets.

Table 2: Detection performance (Sensitivity/Specificity) for SRP- and RNAseP-alignments with mean pairwise identities between 60% and 90%

	Number of sequences in alignment		
Program	2	3	10
QRNA	42.9/92.9	—	—
ddbRNA	45.4/98.5	58.0/94.5	—
MSARi	—	—	appr. 56/100
RNAz	87.8/99.5	94.1/99.6	100/100

Table 3: CPU-time in seconds for 1000 alignments on an Intel 2.4 GHz Pentium 4

	Alignment length		
Program	100	200	300
QRNA	485	4044	14777
ddbRNA	741	921	1522
RNAz	163	375	754

We also tested our method on larger alignments with 10 sequences as used for benchmarking `MSARi`. We generated 150 alignments which had mean pairwise identities between 50% and 70%. Our SVM classification model is currently trained only for up to six sequences so we did not use it for the classification of this test set. It turns out, however, that the simple rule SCI ≥ 0.3 and $z \leq -1.5$ perfectly separates the native alignments from the controls with 100% sensitivity and 100% specificity using *either* of the two scores without help of a SVM. Although the alignments produced by `ClustalW` are, at this level of sequence similarity, structurally not perfectly correct, our consensus folding algorithm still finds the correct common structure and the SCI is still significant, albeit at lower levels.

As of writing this paper, no executable version of `MSARi` was available so we can only compare `RNAz` with the published results: according to [26] `MSARi` achieves at best a sensitivity of 56% at 100% specificity for `ClustalW` alignments of $N = 10$ RNAseP or SRP sequences.

2.5 Implementation and run-time

The method described above was implemented in `RNAz` using the `C`-programming language. The time complexity of our method is $\mathcal{O}(N \times n^3)$, where N is the number of sequences and n is the length of the alignment. Tab. 3 compares the runtime for pairwise alignments of different lengths between `RNAz` and the alternative methods: `RNAz` is not only more accurate but also significantly faster than the other methods. (A comparison with `MSARi` was not possible since no implementation is publicly available. It should have similar run times as `RNAz`, however, since it also uses the RNA folding routines of the `Vienna RNA Package` as the rate limiting step.)

2.6 Screening the CORG database for functional RNA structures

The CORG (**CO**mparative **R**egulatory **G**enomics) database is a collection of conserved sequence elements in non-coding, genomic DNA [38]. The release 2.0 contains multiple sequence alignments of conserved elements in the upstream regions (up to 15 kb from the translation start) of orthologous protein-coding genes from human, mouse, rat, fugu and zebrafish. We focus here on the 4263 conserved non-coding blocks (CNBs) that are longer than 50 nucleotides.

We scanned the alignments using RNAz; after clustering overlapping and redundant CNBs we found 89 distinct regions are predicted as structural RNA with $P > 0.5$. Of these, 28 score with $P > 0.9$, see Tab.4. Among the predicted RNAs we can find all known ncRNAs from Rfam [37] and the miRNA registry [39] that are located in the upstream regions of known protein-coding genes. We identified six micro RNAs with $P > 0.99$ and the snoRNA U93 with $P = 0.72$. Furthermore, we also could reliably ($P > 0.98$) detect known structural *cis*-acting elements [40], in particular we encountered 4 internal ribosome entry sites (IRES) [41] and one iron response element (IRE) [42].

Thus only 11 of the 89 RNAz-hits are known ncRNAs or *cis*-acting structures. This leaves us with 78 candidates, 17 of which have RNAz probabilities above $P = 0.9$. We estimated the specificity in this screen by scoring random controls and found that the $p = 0.5$ and $p = 0.9$ cut-offs have associated specificities of 99.2% and 99.9%, respectively. This is even higher than in the test examples; we are therefore confident that most of these hits are true positives.

Tab. 4 lists the top hits and their genomic context. We found several hits in 5'-untranslated regions of protein coding genes, as for example in NFAT5, the only known transcription factor involved in the osmoregulation in mammalian cells. NFAT5 has a spliced 5'-UTR and in one exon we found a stable and conserved stem-loop structure (CNB-405712). Interestingly, several splice variants of this mRNA exist, some of which have this exon while others do not. We suspect that CNB-405712 is an important regulatory module of the NFAT5 mRNA.

Significant hits were also found in introns, even though introns are not systematically covered in the current release of CORG. For example, CNB-284325 is a structurally highly conserved element supported by many compensatory mutations in the intron of a muscle specific *LIM* domain protein. This structure is probably part of a ncRNA.

Some other hits are not directly related to any known protein coding genes. CNB-134297 is an exceptionally large (appr. 1800 nucleotides) conserved region without any annotation or predicted coding capacity. We scanned alignments $>$300 in sliding windows of size 300 and slide 50. In this special case, significant RNA structures were predicted in several independent windows. Also this region is thus a strong candidate for a novel ncRNA.

The CORG database sporadically contains alignments of coding regions and we also found significant secondary structures in some of them (e.g. CNB-453969: $P = 0.999$). In some instances we could only detect a signal in the reverse complement strand compared to the mRNA, possibly indicating structured antisense transcripts. For some hits, this prediction is additionally supported by EST data. We routinely scanned the reverse complement for all alignments, because RNAz scores are generally higher if the RNA in question is provided in the correct orientation. The snoRNA U93 found in CNB-470004 is a good example demonstrating the remarkable sensitivity of RNAz. It is predicted as RNA with $P = 0.72$ in its correct orientation while there is no

Table 4: Top-scoring alignments in the CORG database

CORG ID	P	Genomic context	Function
110355	1.0	5'-UTR of "Di George syndrome critical region gene 8"	IRES
194820	1.0		micro RNA: mir-196b
226470	1.0		microRNA: mir-10a
288188	1.0		micro RNA: mir-10b
393758	1.0	5'-UTR of "Solute carrier family 40" (iron-regulated transporter)	IRE
119596	0.999		micro RNA: mir-34b
159932	0.999		micro RNA: mir-138-2
373196	0.999	Not annotated	unknown
453969	0.999	Coding exon of "Retinoic acid induced 17"	unknown
461749	0.999	Coding exon of "CIN85-associated multi-domain containing RhoGAP"	unknown
264053	0.997	5'-UTR of "Brain chitinase like protein 2"	IRES
376858	0.997	Not annotated	unknown
405712	0.997	5'-UTR exon of "nuclear factor of activated T-cells 5, tonicity-responsive (NFAT5)"	unknown
391315	0.996		micro RNA: mir196a-2
386451	0.985	5'-UTR of a hypothetical protein	unknown
260572	0.984	Upstream of a hypothetical protein	unknown
430443	0.983	Upstream/5'-UTR of "Hairy and enhancer of split 1"	unknown
57635	0.980	5'-UTR of a hypothetical protein	IRES
238772	0.980	5'-UTR of a hypothetical protein	IRES
284325	0.964	Intron of "Skeletal muscle LIM-protein 2"	unknown
134297	0.963	Not annotated	unknown
501416	0.961	Coding region of hypothetical protein	unknown
363131	0.950	Upstream of "Eyes absent 1"	unknown
386639	0.950	5'-UTR of "Ribosomal protein L12"	unknown
143688	0.938	Upstream of "Zinc finger protein 503"	unknown
456164	0.921	Intron of the spliced 5'-UTR of "Checkpoint suppressor 1"	unknown
154812	0.918	Upstream/5'-UTR of "Basic helix-loop-helix domain containing, class B 5"	unknown
406119	0.902	Upstream of "Zinc finger protein of the cerebellum 3"	unknown

IRE, iron response element. IRES, internal ribosome entry site.

significant signal in the reverse complement strand ($P = 0.06$).

A detailed description of all 89 hits can be found online under `www.tbi.univie.ac.at/papers/SUPPLEMENTS/RNAz/`, where we provide links to the UCSC genome browser [43] allowing a detailed study of the genomic context for all hits (annotation, mRNA structure, ESTs etc.). Unlike other methods, `RNAz` does not only predict the existence of a functional RNA element it also predicts an accurate model of the consensus structure. These can also be found in the on-line material together with the annotation of compensatory mutations.

3 Discussion

We have described here a novel, versatile method for detecting functional RNAs in genomic screens. This approach can reliably detect a surprisingly wide variety of different ncRNAs and *cis*-acting RNA elements using only evolutionary conservation and thermodynamic stability as characteristic signal. Although conceptually simple, the structure conservation index proved to be a convenient and effective measure of structural conservation. Our stability measure, on the other hand, shows that contrary to common belief thermodynamic stability can be useful for ncRNA detection. As a consensus of several independent sequences in an alignment, stability can be a significant measure. Furthermore, we have demonstrated that a properly normalized stability measure can be directly calculated without the need for time consuming sampling of shuffled sequences or alignments. Our results show that `RNAz` is suitable for large scale genomic annotation whenever alignments can be obtained.

Since a wealth of genomic data together with new methods for generating high quality alignments [44] are already available, we are currently preparing genome-wide screens for various organisms including human. Aided by visualization tools [43], we aim to draw genome wide-maps of significant RNA structures. This approach of "computational RNomics" opens a completely new perspective which, as we hope, will result in the discovery of new terrain in the expanding RNA world of cellular mechanisms.

4 Methods

4.1 Calculation of the SCI

For minimum free energy RNA folding we used the `C`-libraries of the Vienna RNA package version 1.5 [45]. We used `RNAfold` for folding single sequence and `RNAalifold` [28] for consensus folding of aligned sequences. The same folding parameters were used for both algorithms to ensure that the obtained MFE values are comparable. For the covariation part of `RNAalifold` we used default parameters. Gaps were removed for single sequence folding.

4.2 Calculation of z-scores using support vector machine regression

To calculate z-scores by regression analysis we used the following procedure: We generated synthetic sequences of different length and base composition. The length of the test sequences ranged from 50 to 400 nucleotides in steps of 50. To quantify

base composition, we used the GC/AT, A/T and G/C ratios of the sequences and chose values for all ratios ranging from 0.25 to 0.75 in steps of 0.05. This resulted in 10648 points in a 4 dimensional space of the independent variables. For each of the points we calculated the mean and standard deviation of the MFE of 1000 random sequences, representing the dependent variables in our regression.

We used the support vector machine library `LIBSVM` [46] to train two regression models for mean and standard deviation. Input data for the SVM were scaled to mean of 0 and standard deviation of 1.

We chose the ν variant of regression and a radial basis function (RBF) kernel. We optimized the parameters and found $\nu = 0.5$, $\gamma = 1$ and $C = 5$ to yield the best results. Finally, we obtained two models for the mean and standard deviation we used for z-score calculation described in the main text.

The traditional sampling of z-scores depends on the randomization of the native sequence by shuffling the positions. In this context it was pointed out by Workman & Krogh [47] that a correct randomization procedure should conserve dinucleotide content because of the energy contributions of stacked base pairs in the energy model. In principle, the regression model could be extended to use dinucleotide frequencies. The good results with the simple model, however, allow us to neglect this effect.

4.3 Generation of the test alignments

Sequences for the test alignments were taken from the Rfam database [37] with the exception of the SRP and RNaseP test sets which were taken from other sources [48, 49] in order to use the same data as previous studies [22, 26]. We used the procedure as previously described [27] to generate test sets consisting of a reasonable number of non-redundant alignments of different size, with a defined range of mean pairwise identities and in which all are sequences approximately equally represented.

4.4 Randomization of the test alignments

The program `shuffle-aln.pl` [27] was used to generate the randomized controls for alignments with up to $N = 6$ sequences. In brief, this program implements a randomization algorithm that takes care not to introduce randomization artifacts and produces random alignments of the same length, the same base composition, the same overall conservation, the same local conservation pattern, and the same gap pattern at the input alignment. For the large alignments ($N = 10$) we employed the same procedure as in ref. [26]: We completely shuffled all columns and re-aligned the alignment afterwards using `ClustalW`.

4.5 Support vector machine classification

A binary classification support vector machine again using `LIBSVM` was trained to classify alignments as "RNA" or "other sequence". Input parameters are the MFE z-scores of the individual sequences in the alignments (without gaps), the SCI of the alignment, the mean pairwise identity and the number of sequences in the alignment. For the final calibration of the SVM in the current implementation of `RNAz` we used all classes of ncRNA with the exception of tmRNAs and U70 snoRNAs. For the tests on known families presented in this paper, we generated models from all classes, leaving out one class at a time. In all cases, we used alignments with mean pairwise identities between appr. 50% and 100% and 2–6 sequences per alignment. For each native

alignment we included one randomized version in the training set. All parameters were scaled linearly from -1 to 1. We used an RBF-kernel and the parameters $\gamma = 2$ and $C = 32$ to train the models. The probability estimation option was used to obtain a model with probability information.

4.6 Test of other programs

We used QRNA version 1.2b and ddbRNA as available from the author[1] (version of July 2004). For the tests shown in Tab. 2, we chose the cutoffs log-odd$= 15$ for QRNA and $K = 1.5$ ddbRNA, respectively. For RNAz we used a cutoff of $p = 0.9$ and customized models for the SVM that excluded both SRP and RNAseP from the training set.

Acknowledgments

We thank Christoph Dieterich and Martin Vingron for permission to use release 2.0 of their CORG databases prior to publication. Discussion with Paul Gardener and Andrea Tanzer are gratefully acknowledged. This work was supported in part by the Austrian *Gen-AU bioinformatics integration network* sponsored by BM-BWK and BMWA, the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. P-15893, and the Bioinformatics Initiative of the *Deutsche Forschungsgemeinschaft*, BIZ-6/1-2.

References

[1] Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919–929 (2001).

[2] Storz, G. An expanding universe of noncoding RNAs. *Science* **296**, 1260–3 (2002).

[3] Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930–939 (2003).

[4] He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**, 522–31 (2004).

[5] Avner, P. & Heard, E. X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet* **2**, 59–67 (2001).

[6] Suzuki, M. & Hayashizaki, Y. Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *BioEssays* **26**, 833–843 (2004).

[7] Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).

[8] Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).

[9] Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29**, 11–17 (2004).

[1] http://www.tigem.it/Research/DiBernardoPersonalWebPage.htm

[10] McClelland, M. *et al.* Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res.* **28**, 4974–4986 (2000).

[11] Florea, L., McClelland, M., Riemer, C., Schwartz, S. & Miller, W. EnteriX 2003: Visualization tools for genome alignments of Enterobacteriaceae. *Nucleic Acids Res.* **31**, 3527–3532 (2003).

[12] Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).

[13] C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).

[14] Stein, L. D. *et al.* The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).

[15] The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

[16] International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

[17] Consortium, R. G. S. P. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).

[18] Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–5 (2004).

[19] Bejerano, G., Haussler, D. & Blanchette, M. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* **20 Suppl 1**, I40–I48 (2004).

[20] Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–93 (2003).

[21] Margulies, E. H., Blanchette, M., Haussler, D. & Green, E. D. Identification and characterization of multi-species conserved sequences. *Genome Res* **13**, 2507–18 (2003).

[22] Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).

[23] Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373 (2001).

[24] McCutcheon, J. P. & Eddy, S. R. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.* **31**, 4119–4128 (2003).

[25] di Bernardo, D., Down, T. & Hubbard, T. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics* **19**, 1606–11 (2003).

[26] Coventry, A., Kleitman, D. J. & Berger, B. MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* (2004).

[27] Washietl, S. & Hofacker, I. L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342**, 19–30 (2004).

[28] Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066 (2002).

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

[30] Walter, A. E. *et al.* Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA* **91**, 9218–9222 (1994).

[31] Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).

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

[33] Bonnet, E., Wuyts, J., Rouzé, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding rnas, have lower folding free energies than random sequences. *Bioinformatics* (2004). In press.

[34] Le, S. V., Chen, J. H., Currey, K. M. & Maizel Jr., J. V. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.* **4**, 153–159 (1988).

[35] Bonnet, E., Wuyts, J., Rouze, P. & Van De Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **in press** (2004).

[36] Cristianini, N. & Shawe-Taylor, J. *An Introduction to Support Vector Machines* (Cambridge University Press, 2000).

[37] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).

[38] Dieterich, C., Wang, H., Rateitschak, K., Luz, H. & Vingron, M. Corg: a database for comparative regulatory genomics. *Nucleic Acids Res* **31**, 55–7 (2003).

[39] Griffiths-Jones, S. The microrna registry. *Nucleic Acids Res* **32 Database issue**, D109–11 (2004).

[40] Pesole, G. *et al.* Utrdb and utrsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mrnas. update 2002. *Nucleic Acids Res* **30**, 335–40 (2002).

[41] Le, S. Y. & Maizel Jr., J. V. A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. *Nucl. Acids Res.* **25**, 362–369 (1997).

[42] Hentze, M. W. & Kuhn, L. C. Molecular control of vertebrate iron metabolism: mrna-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc Natl Acad Sci U S A* **93**, 8175–82 (1996).

[43] Kent, W. J. *et al.* The human genome browser at ucsc. *Genome Res* **12**, 996–1006 (2002).

[44] Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708–15 (2004).

[45] Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie* **125**, 167–188 (1994).

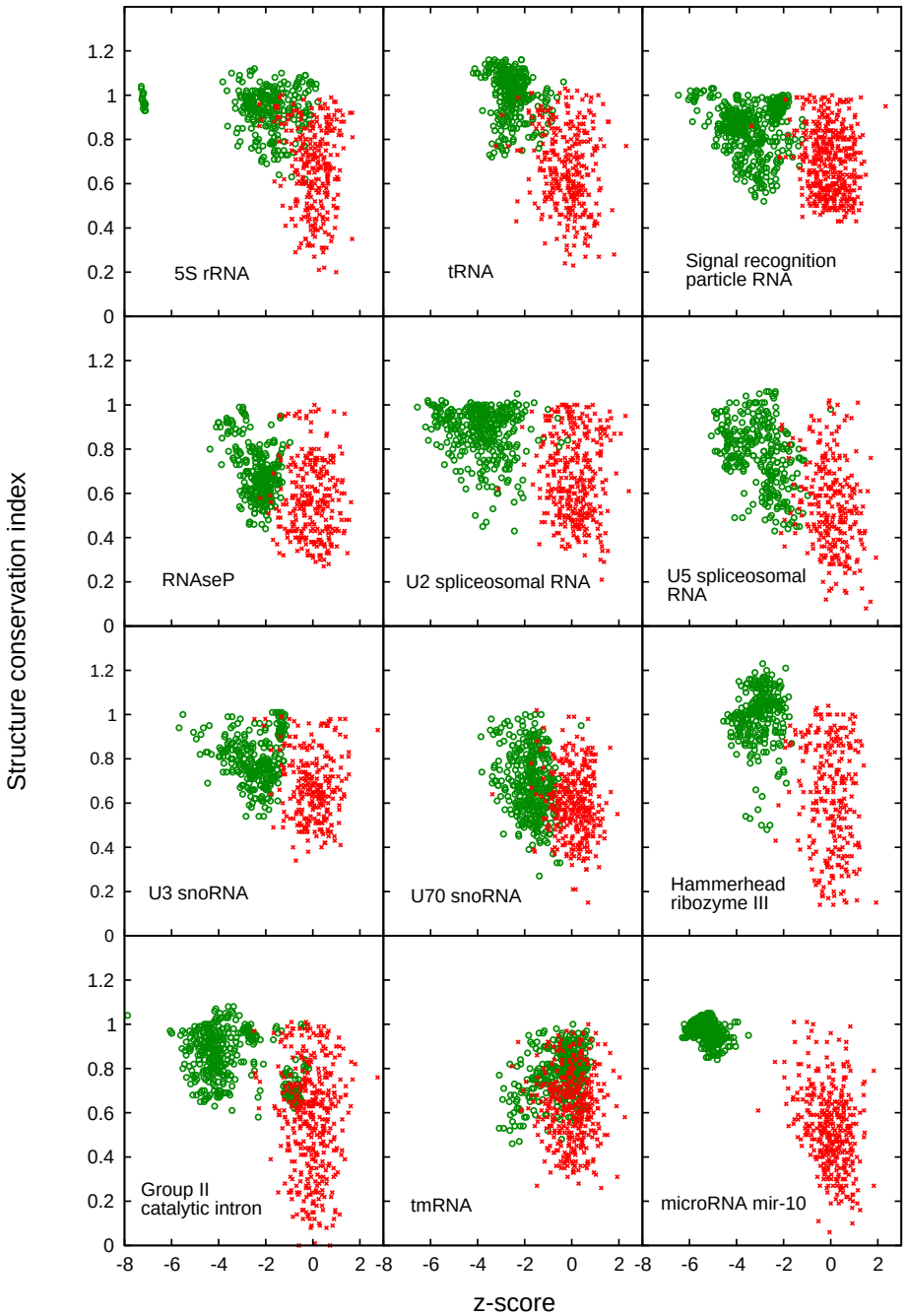
[46] Chang, C.-C. & Lin, C.-J. *LIBSVM: a library for support vector machines* (2001). Software available at `http://www.csie.ntu.edu.tw/čjlin/libsvm`.

[47] Workman, C. & Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27**, 4816–4822 (1999).

[48] Rosenblad, M. A., Gorodkin, J., Knudsen, B., Zwieb, C. & Samuelsson, T. SRPDB: Signal recognition particle database. *Nucleic Acids Res.* **31**, 363–364 (2003).

[49] Brown, J. W. The ribonuclease P database. *Nucleic Acids Res* **27**, 314 (1999).

Supplementary Figure 1



Separation of native alignments (green) from random controls (red) for various classes of ncRNAs. The test sets are the same as used in Table 1 with mean pairwise identities between 60% and 100% and 2–4 sequences per alignment.