Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics

Stefan Washietl [a] and Ivo L. Hofacker [a,b]

[a] *Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*

[b] *To whom correspondence should be addressed.* ivo@tbi.univie.ac.at

Abstract

Facing the ever-growing list of newly discovered classes of functional RNAs, it can be expected that further types of functional RNAs are still hidden in recently completed genomes. The computational identification of such RNA genes is, therefore, of major importance. While most known functional RNAs have characteristic secondary structures, their free energies are generally not statistically significant enough to distinguish RNA genes from the genomic background. Additional information is required. Considering the wide availability of new genomic data of closely related species, comparative studies seem to be the most promising approach. Here we show that prediction of consensus structures of aligned sequences can be a significant measure to detect functional RNAs. We report a new method how to test multiple sequence alignments for the existence of an unusually structured and conserved fold. We show for alignments of six types of well known functional RNA that an energy score consisting of free energy and a covariation term significantly improves sensitivity compared to single sequence predictions. We further test our method on a number of non coding RNAs from *C. elegans*/*C. briggsae* and seven Saccharomyces species. Most RNAs can be detected with high significance. We provide a Perl implementation which can be readily used to score single alignments and discuss how the methods described here can be extended to allow for efficient genome-wide screens.

Key words: Minimum free energy folding, consensus secondary structure prediction, non coding RNAs, comparative genomics, randomizing multiple sequence alignments

Manuscript *April 19, 2004*

1 Introduction

In the past few years our knowledge on the molecular and cellular functions of RNA has increased dramatically. In particular the identification of numerous RNA transcripts that function directly as RNA without ever being translated to protein (non coding RNAs) has made clear that the traditional view of RNA must be extended profoundly. To mention just one example, the discovery of micro RNAs[1–3] has led to an new paradigm of RNA-directed gene expression regulation. There are many other examples of such new 'RNA-genes'.[4,5]

Another aspect of RNA function are *cis*-acting regulatory elements within protein coding genes. A recent example is the regulation of metabolic pathways in bacteria through 'riboswitches' . These riboswitches occur in leader sequences of operons and interact directly with small metabolites[6] in order to control protein expression.

These findings not only force experimental biologists to reconsider their strategies and methods, but also pose new challenges to bioinformatics. In particular, the computational identification of functional RNAs in is a major, yet largely unsolved, issue.

Current methods mostly are based on similarity searches and are successful in the identification of functional RNAs that are members of already known families.[7–10] A more general approach that detects new classes of functional RNAs without relying on any *a priori* knowledge would be helpful. This, however, proved to be difficult. In contrast to protein coding genes, which show strong statistical signals like open reading frames and codon bias, the primary sequences of functional RNAs seem to lack comparable signals completely.

Since most known functional RNAs depend on a defined secondary structure, it was suggested by Maizel and co-workers that functional RNAs have a more stable secondary structure than expected by chance.[11–13] However, efforts to build a general RNA gene finder based on secondary structure prediction failed. Rivas & Eddy had to conclude in an in-depth study on the subject that secondary structure alone is generally not significant enough for the detection of non-coding RNAs.[14] Some other statistical measures, partly derived from secondary structure predictions, have been proposed.[15–17] Still, additional information seems to be required for reliable predictions on a genome-wide scale.

The most promising source of information comes from comparative studies. Already, a number of complete genomes from closely related species are available. Some of them have been sequenced solely for the purpose of genome comparisons. Readily available sets for comparison are: more than 15 enteric bacteria,[18,19] seven yeast species,[20,21] two nematodes[22,23] and the two mammalian genomes from human[24] and mouse.[25] Facing the ever-growing pace of

genome projects, even more can be expected in the near future.

`QRNA` is a program that makes use of this comparative information and scans pairwise alignments for conserved secondary structures using probabilistic models based on stochastic context free grammars.[26] This approach has been applied successfully to predict candidates for non coding RNAs in *E.coli* and *S. cerevisiae*. Some of which could be verified experimentally.[27,28]

In this contribution, we propose an alternative method to assess a multiple sequence alignment for the existence of a conserved secondary structure. We compute an averaged folding energy of aligned sequences, that also takes into account sequence covariations. Following the ideas of the Maizel group, we compare this to a set of random alignments in order to estimate if there is an unusually stable and conserved fold. We address the question, if this can be a significant measure to detect functional RNAs in genome-wide screens.

2 Results and Discussion

2.1 MFE predictions for single sequences are of limited statistical significance.

Secondary structure is a useful level on which to understand RNA function. Fairly reliable models can be predicted with computational methods. Since many known functional RNAs are tied to a defined secondary structure, such predictions appear a straightforward measure for their detection. However, prediction programs readily calculate minimum free energy (MFE) structures also for arbitrary random sequences. The question arises, if natural RNAs are more stable (have lower MFE) than random sequences. This question was partly addressed previously.[14] Here, we test it again for sequences from a set of six structural RNA families (tRNA, 5S rRNA, Hammerhead ribozyme type III, Group II catalytic intron, Signal recognition particle RNA, U5 spliceosomal RNA). We used `RNAfold` for the prediction and calculated z-scores from a sample of 100 random sequences (see Material and Methods). The results are shown in Table 1. On average, the structural RNAs have all z-scores clearly below zero, meaning they have lower folding energy than the random samples. Is this significant enough to reliably distinguish single sequences from the random background? Fig. 2 illustrates this for the tRNA test set. The topmost panel shows the distribution of z-scores for 579 tRNAs together with the z-scores of 579 random sequences (one shuffled version for each tRNA). If we use a conservative limit of -4 to define a significant z-score, we can only detect 2% of the tRNAs. To detect half of all tRNAs we would have to lower the cutoff to -1.8. Then, however, we would encounter 4% of false positives. For

genome-wide screens where a huge number of candidates has to be scored, this selectivity is too low (especially for a corresponding sensitivity of only 50%). Some of the tested families form more stable structures (e.g. Group II catalytic intron: average z=−3.88, Hammerhead ribozyme III: z=−3.08) but generally the native sequences are not efficiently separated from the bulk of random sequences.

An additional point seems noteworthy regarding these experiments. Workman & Krogh[29] pointed out that dinucleotide content influences secondary structure predictions, because of the energy contributions of stacked base-pairs. A correct randomization procedure should, therefore, generate random sequences of the same dinucleotide content. It is impossible to consider this in the randomization of multiple sequence alignments (see next section). For single sequences, however, we performed the z-score calculations with both mono- and dinucleotide shuffled random sequences. The results (Table 1) show that a systematic bias is not recognizable for our test sets. The values differ only minimally and the mononucleotide-shuffled z-scores. Thus, while dinucleotide composition was important in the study of Workman & Krogh where long ($>$ 500 nucleotides) mRNAs are tested for an (obviously non-existent) subtle bias towards lower folding energies, it can be neglected in our case.

2.2 Additional information from aligned sequences shifts MFE predictions towards significant levels.

The results so far show that folding energy is indeed a characteristic signal of (structural) ncRNAs, but is in itself not sufficient for a reliable detection. Given the availability of comparative data mentioned in the introduction, we wondered how to efficiently make use of this information. We use the program `RNAalifold`, which was originally developed to predict consensus secondary structures of aligned sequences.[30] `RNAalifold` calculates an averaged minimum free energy for the alignment, incorporating covariance information into the energy model. We consider `RNAalifold`-MFEs to be a good measure for the existence of a conserved fold and a good alternative for the probabilistic approach implemented in `QRNA`. `RNAalifold` makes use the standard energy model for RNA secondary structures, and thus reduces to simple MFE structure prediction in the case of single sequences. For an alignment of several sequences the energy model is augmented through covariance information. Furthermore, `RNAalifold` is not limited in the number of input sequences.

To test if the consensus folding of homologous sequences is more significant than the folding of single sequences, we generated test sets of multiple sequence alignments from the same RNA families as before and subsequently calculated z-scores based on `RNAalifold`-MFEs. For this purpose we had to

develop a reliable randomization procedure for multiple sequence alignments. Our algorithm takes care not to introduce randomization artifacts (see section Material and Methods and Fig. 5) and generates random alignments of the same length, the same base composition, the same overall conservation, the same local conservation and the same gap pattern. This is the most conservative randomization procedure possible but it is effective enough to remove correlations arising from secondary structures.

The results for the z-score calculations are summarized in Table 1 and Fig. 1. If we compare the average z-score from the single sequences to the average z-scores of the pairwise alignments (N=2), we observe in all cases that the average z-score drops by almost 2. It further drops for the alignments consisting of three and four sequences. We want to recall that the units of z-scores are standard deviations, so that even small changes shift the sensitivity significantly (for fixed z-score threshold). In Table 1 we calculated the sensitivities for a threshold of -4. In Fig. 2 the z-score distribution is shown for the tRNA alignments with varying N. Folding of pairwise alignments instead of single sequences improves sensitivity from 2.1% to 71.1%. For $N = 4$, the native alignments are completely separated from the random alignments and almost all score below -4 (98.4%).

2.3 z-scores of random alignments are well approximated by a standard normal distribution

Sensitivity and selectivity depend on a predefined z-score threshold. Assuming a normal distribution, one can readily estimate the false-positive rate of a given z-score threshold. For -4 it is below 0.003%, meaning one false positive can be expected in approximately 31500 alignments. Since one should not expect the MFEs of randomized alignments to follow the normal distribution exactly, it makes sense to empirically estimate the significance of z-scores. The distribution of 11633 random z-scores is shown in Fig. 3. It is, indeed, well approximated by a standard normal distribution (mean $\mu = 0.01$, standard deviation $\sigma = 0.99$). The distribution is slightly skewed with a fat tail: There are apparently more z-scores below -3 than z-scores above $+3$. This tail is not due to our shuffling algorithm. Single sequences (whether mono- and dinucleotide shuffled) show the same skew in the distribution (not shown), as noted also in other studies.[14]

The expected and observed frequencies for several thresholds can be found in Fig. 3. The observed frequencies are slightly above the theoretically expected ones. For the experiments shown here, all thresholds below -3 have a false-positive rate below 1% and can be regarded as significant.

The threshold of -4 used so far has a false-positive rate of 0.06% and thus represents a rather conservative definition of significance. It must be noted, however, that all these values are based on the 11633 `ClustalW` alignments made from Rfam entries in this study. For genome-wide studies it cannot be assumed that the genomic background behaves exactly like random alignments and it might be possible that various inhomogeneities cause more false positives than experienced here. The false-positive rate will depend on preparation of the data (e.g. masking of repeats and low complexity regions) and the quality of the alignments.

2.4 Sensitivity depends on sequence divergence and alignment method.

`RNAalifold` takes a multiple sequence alignment as input. It can predict an existing consensus structure only if the sequence alignment reflects common structural properties. Ideally, one would like to feed `RNAalifold` with structurally aligned sequences. However, existing algorithms,[31] are much too slow to make this a feasible alternative for a large number of alignments, so that typically alignments based on sequence similarity alone will be used. To test to which extent the performance of our method depends on the alignment method, we did the following experiment: We took 73 eukaryotic SRP-RNAs and generated 2083 pairwise alignments with a wide variety of pairwise identities. For this test set, manually curated structural alignments exist.[32] We calculated z-scores for structurally aligned pairs and for `ClustalW` aligned pairs (Fig. 4). The detection performance for the structural alignments constantly increases with increasing sequence divergence over the full range of pairwise identities. This is exactly what could have been expected, since higher sequence divergence means more information-rich covariances. From appr. 60% to 100% pairwise identity, the z-scores of the sequence based alignments are essentially the same. Below 60%, the detection performance drops remarkably. Extrapolating from this example, we can conclude that there is obviously no need for structural alignments above 65% pairwise identity and that our method scores best somewhere between 60% and 70%.

2.5 Most structural ncRNAs from S. cerevisiae and C. elegans can be significantly detected.

The results so far show that detection sensitivity highly depends on the quality of the available data. A large number of homologous sequences with high divergence (but still alignable) is desirable. In real-life applications, such ideal data sets will hardly be found. To test our method not only for the rather artificial data sets taken from Rfam, we created tests sets of known non-

coding RNAs from *C. elegans* and *S. cerevisiae* (see Material and Methods). The genome of a second nematode *C. briggsae* was finished recently.[22] For *S. cerevisiae*, unassembled draft sequences of six related yeast species exist.[20,21] In the case of the yeast sequences, we tried an automatic alignment procedure. We chose `MultiPipMaker`,[33] which is currently the only program available which can align a reference sequence to unassembled contigs on a genome wide level off-the-shelf. Sometimes it worked well, but in many cases manual refinement was necessary.

Table 2 and 3 show the results for the genomic examples. For scanning whole genomes it will not be feasible to predict structures longer than appr. 200 nucleotides. We therefore scored alignments longer than 150 columns using a sliding window (size 150, slide 20) and report the lowest z-score obtained. To estimate the contribution of secondary structure stability alone, we also scored single sequences from *C. elegans* and *S. cerevisiae* using `RNAfold`.

We found that ncRNA sequences are highly conserved between *C. elegans* and *C. briggsae*. Pairwise identities are above 90% in most cases. Still, most genes score well below -4. Some of them (e.g. SRP RNA or let-7 pre-miRNA) form exceptionally stable structures that can also be detected by single sequence predictions without problems. However, the alignment scores are more significant in all cases with values below the single scores in the order of appr. one standard deviation. Only the spliceosome RNAs U4 and U6 cannot be detected. This shows the inherent limitation of this method. U6 for example is known to form extensive *inter*molecular interactions with U4 rather than forming a stable *intra*molecular secondary structure. U6 only features a short 5'-stem loop. Although predicted by `RNAalifold` in the native alignment, this loop is too short to be significantly different from the random background.

For the yeast genes we encounter similar results. Here we have more sequences (up to seven homologs) and sometimes also higher divergence with mean pairwise identities below 90%. Therefore, the alignment scores differ significantly (up to 4 standard deviations) from the single sequence scores. Many genes which would have been missed with a -4 threshold with the single score can be reliably detected by the alignment score. As seen before, RNA genes lacking a stable secondary structure are missed. This is the case for all C/D snoRNAs. The H/ACA snoRNAs show z-scores around -3, but the typical two stem loops of H/ACA snoRNAs are not stable enough to be detected at a threshold of -4. The covariance information could help in this case, but the H/ACA snoRNAs in yeast are too conserved (appr. 90% mean pairwise identity).

Our method is readily available to analyze a given multiple sequence alignment. For example, if a new gene has been cloned and found to have an evolutionary conserved untranslated region, it can be tested for the existence of an unusually stable and/or conserved secondary structure.

Our results show that the sensitivity and selectivity are suitable even for genome wide scans. Some important issues have to be considered regarding such large scale applications.

A straightforward approach to fold large genomic regions is to apply a sliding window. As already mentioned, the maximum length is practically limited to appr. 200 nts. Although many known functional RNA structures are longer than 200 nucleotides, this seems to be long enough to detect local substructures. However, a sliding window has several other drawbacks. Only for a step-size of one, all possible regions are covered. In practice, the use of a much larger step-size is inevitable which leaves us with a 'blind spot' and many relevant local structures are ignored. Another problem arises, if for example a small structured motif of 50 nts should be detected within a much longer window of 200 nts. This will result in a low signal to noise ratio which probably hinders detection. Again for performance reasons, the use of different sized windows is not an alternative. To avoid problems of that kind, a local prediction algorithm is desirable. Such an algorithm is for example implemented in the probabilistic model of `QRNA`. Similarly, energy based dynamic programming algorithms can be modified to allow for the efficient prediction of all locally stable structures of a given maximum size, as shown recently in our group.[34] In principle, the idea can also be applied to `RNAalifold` for local consensus structure prediction.

Generally, the `RNAalifold` algorithm is fast for moderate window sizes. Middle-sized genomes like *S. cerevisiae* or *C. elegans* could be analyzed within hours on a modern desktop computer. However, the Monte Carlo procedure to estimate the significance imposes a serious performance problem. A direct measure for the significance of a calculated MFE would have to consider alignment properties like the GC-content, the degree of conservation, the gap pattern and of course the length of the structure. It appears difficult to put all this together into a meaningful *ad hoc* score.

Shuffling is the only remedy but, theoretically, a genome must be folded 200 times if both forward and reverse strand are analyzed with a sample number of 100. In practice, the number of calculations can be reduced drastically. First, only conserved regions have to be analyzed and even in closely related species only a fraction of the genome can be reliably aligned on the nucleotide

level. Second, `RNAalifold` will not predict stable consensus structures in all regions. There is no sense in extensively shuffling and folding a structure which is not even stable in its native conformation. Third, we only want to test if a structure has a z-score below a certain threshold, we are not interested in the exact z-score if it is above the threshold. This means that we can roughly estimate the z-score based on a small sample and then decide if it is worth to do a precise evaluation. E.g., if the estimated z-score from a sample of 10 is above -1 it is unlikely the real z-score will fall below -3.5. For these reasons number of computations can be reduced remarkably. Finally, the folding of random samples can be performed independently and is, therefore, an easily parallelizable task. To conclude, our method is computationally demanding but feasible if reduced to the essential.

3 Conclusions

In this work we have introduced z-scores of `RNAalifold` MFEs as a measure for the detection of functional RNA structures. The combination of free energy and covariance used by `RNAalifold` provides a reliable measure to distinguish functional from random RNAs. We have shown for several test cases, that this method can detect known structural RNAs with high sensitivity *and* selectivity. This is not only true for ideal data sets featuring high sequence divergence, even for datasets with few and closely related sequences as in the case of *C. elegans*/*C. briggsae*, it shows good detection performance and clearly outperforms single sequence predictions. Encouraged by these results, we are currently working on a general (structural) RNA gene finding program based on the ideas discussed here. We hope that this will be a useful addition to the arsenal of today's sequence analysis tools.

4 Supplementary Material and Programs

Supplementary material including all data sets is available on our website `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/Alifoldz/`.

A Perl 5 script `alifoldz.pl` which implements the procedures shown here can be downloaded from the same location. It depends on the `RNAalifold` program which can be downloaded as part of the `Vienna RNA Package` from `http://www.tbi.univie.ac.at/RNA/`. Another Perl script `shuffle-aln.pl` is provided, which implements the shuffling algorithm described in this work. It might be useful also for other purposes.

5 Methods

5.1 Co-folding of aligned sequences

We use the program `RNAalifold`[30] from the `Vienna RNA package`[35] version 1.5 to perform consensus secondary structure predictions of multiple sequence alignments. `RNAalifold` essentially uses the same algorithms[36] and energy parameters[37,38] as standard programs for minimum free energy (MFE) prediction. The energy contributions of the single sequences in the alignment are averaged. Covariance information is incorporated into the energy model by rewarding compensatory and consistent mutations, while non compatible base-pairs are penalized. `RNAalifold` thus calculates a combined MFE composed of an energy term and a covariance term.[30] We simply call this MFE of the alignment, although it is of course not an energy in a strict physical sense. `RNAalifold` depends on some predefined parameters. We used standard parameters throughout this work to ensure consistency.

Secondary structure prediction for single sequences was performed using `RNAfold` with standard parameters.

5.2 Estimation of statistical significance

In analogy to previous work[11–13] we use a Monte Carlo approach to estimate statistical significance of a MFE. For each alignment we generate 100 random alignments (see next section). We then calculate the MFE m of the native alignment and the mean μ together with the standard deviation σ of the random samples. The significance of m is expressed in units of standard deviations from the mean as a z-score $z = \frac{m-\mu}{\sigma}$. Negative z-scores indicate that the MFE of the native alignment is lower than those of the randomized alignments.

5.3 Randomization of multiple sequence alignments

The randomization procedure is of crucial importance for the calculation of meaningful z-scores. A straightforward algorithm would simply shuffle the columns of the alignment. This would result in an alignment of the same length, the same base composition and the same overall conservation. However, the gap structure and the local conservation pattern would be different. Possible consequences for consensus folding and z-score calculations are illustrated in Fig. 5. If there is for example a gap of length 10 in the alignment, the shuffling probably would produce 10 gaps of length 1. This can result in

artefactual low z-scores since many gaps spread over the complete alignment can remarkably impair the consensus folding, while one long gap probably does not. The same is true for local conservation patterns, meaning that a well conserved column `AAAAAGG` should not be shuffled with a less conserved column `AGUACUA`, but rather with a column `CCCCCAA` of the same pattern. We considered this in our shuffling algorithm: First we collect all columns which have the same gap structure and local conservation pattern into individual groups of columns. We memorize which column of the initial alignment has which pattern. Subsequently, we shuffle the groups individually using a standard procedure.[39] Finally, we reassemble the alignment. Since the shuffling procedure of the individual sets is provably random and independent from each other, all possible alignments are sampled with the same probability.

It must be pointed out that we only shuffle columns with exactly the same pattern of nucleotide succession (i.e. we shuffle `AAAAAGG` with `CCCCCAA` but no with `CCAAAAA`). Alternatively, one might shuffle columns of the same *degree* of conservation but different pattern. While we cannot think of a possible scenario where this could introduce randomization artifacts, we decided to use the more restrictive version here.

As the conservative shuffling procedure restricts the possible number of permutations, the question arises if it is effective enough to destroy a secondary structure. It is known that if only a small fraction (around 10%) of a sequence is randomly mutated this leads almost certainly to unrelated structures.[40] These theoretical considerations, as well as our computational results, suggest that the shuffling procedure is effective enough to destroy any native secondary structures.

5.4 Randomization of single sequences

Single sequences were randomized both by mono- and dinucleotide shuffling (see Results and Discussion for further explanation). Mononucleotide shuffling was performed simply by shuffling the single nucleotides of the sequences. For dinucleotide shuffling, we used a recent implementation by Clote *et al.* (`http://clavius.bc.edu/~clotelab/`) of an algorithm developed by Altschul & Erickson.[41]

5.5 Creation of test sets

Most of the RNA sequences used in this work were taken from the Rfam database release 5.0.[42] We took the sequences from the *full* alignments of Hammerhead ribozyme III (RF00008), Group II catalytic intron (RF00029)

and U5 spliceosomal RNA (RF00020). For tRNA (RF00005) and 5S rRNA (RF00001) we used the sequences from the *seed* alignment. In the case of tRNA, the number of the sequences in the seed alignment was reduced to 579 (we removed every second of the 1161 sequences). The signal recognition particle RNA test set was taken from the SRP database.[32] We used the 73 eukaryotic sequences that could be found in the database as of January 2004.

To get a reasonable number of non-redundant alignments of different size N (2 to 4 sequences) within a defined range of mean pairwise identity (65% to 85%) and ideally with all sequences of the test set equally represented, we used the following procedure: First, we roughly clustered the sequences using `BlastClust` (available from NCBI, `http://www.ncbi.nlm.nih.gov/`) and created clusters with approximate pairwise identities between 60% and 95%. Within those clusters we computed all possible combinations for a given N. From each cluster we randomly chose a varying number of combinations taking into account the size of the cluster. This should avoid that the resulting alignments are made up just by a fraction of the sequences of the initial test set (which can easily happen because the number of possible combinations can get very large). In the next step, the collected sequence combinations were realigned using `ClustalW`[43] and the mean pairwise identities were calculated. For most of the experiments presented in this paper, we eventually used alignments with mean pairwise identities between 65% and 85%.

5.6 Genomic examples

For the *C.elegans*/*C.briggsae* alignments, we tried to take one example of each ncRNA family (excluding tRNAs and rRNAs) reported in.[22] If available, sequences were simply taken from the respective Rfam family. *C. elegans* RNA genes which could not be found in Rfam were taken from Wormbase release 117 (`www.wormbase.org`) and the corresponding *C. briggsae* homologs were searched using `BLASTN`. We could not find annotated sequences of RNAse P and U3 snoRNA although they have been reported to exist.[22]

The alignments of the yeast examples were created in a semi-automatic way: We downloaded the (draft-)sequences of seven yeast species from the Saccharomyces Genome Database (`ftp://ftp.yeastgenome.org/yeast`): *S. cerevisae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *S. kluyveri*. Next, we created chromosome wide multiple sequence alignments using `MultiPipmaker`.[33] Then we extracted the regions of annotated RNA genes known for *S. cerevisiae*. Most of the resulting alignments needed manual refinement. We removed single sequences which obviously did not align well in the automatic alignment and occasionally performed a re-alignment using `ClustalW`. We included all non-tRNA, non-rRNA genes which produced

a reasonable multiple alignment after this procedure. In the case of the numerous snoRNAs, however, we took only five examples of the H/ACA-type and five of the C/D-type each (picked out arbitrarily, i.e. the first five in the alphabetical annotation list).

Acknowledgments

Useful comments from Peter F. Stadler are gratefully acknowledged. This project has been partly funded by the Austrian Gen-AU bioinformatics integration network sponsored by BM-BWK and BMWA, as well as the *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893.

References

1. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

2. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.

3. Lee, R. C. & Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.

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

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

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

7. Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

8. Lowe, T. M. & Eddy, S. R. (1999). A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.

9. Edvardsson, S., Gardner, P. P., Poole, A. M., Hendy, M. D., Penny, D. & Moulton, V. (2003). A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics*, **19**, 865–873.

10. Klein, R. J. & Eddy, S. R. (2003). RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.

11. Le, S. V., Chen, J. H. & and, K. M. C. (1988). A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, **4**, 153–159.

12. Le, S. Y., Chen, J. H. & Maizel, J. V. (1989). Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic Acids Res*, **17**, 6143–6152.

13. Chen, J. H., Le, S. Y., Shapiro, B., Currey, K. M. & Maizel, J. V. (1990). A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, **6**, 7–18.

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

15. Schultes, E. A., Hraber, P. T. & LaBean, T. H. (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76–83.

16. Le, S. Y. & and, K. Z. (2002). RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res*, **30**, 3574–3582.

17. Le, S. Y., Chen, J. H. & and, D. K. (2003). Discovering well-ordered folding patterns in nucleotide sequences. *Bioinformatics*, **19**, 354–361.

18. McClelland, M., Florea, L., Sanderson, K., Clifton, S. W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R. K. & Miller, W. (2000). 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.

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

20. Cliften, P. F., Hillier, L. W., Fulton, L., Graves, T., Miner, T., Gish, W. R., Waterston, R. H. & Johnston, M. (2001). Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.

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

22. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D. H., Fulton, L. A., Fulton, R. E., Griffiths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R. & Waterston, R. H. (2003). The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.*, **1**, E45.

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

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

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

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

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

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

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

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

31. Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.

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

33. Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E. D., Hardison, R. C. & Miller, W. (2003). MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, **31**, 3518–3524.

34. Hofacker, I. L., Priwitzer, B. & Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

35. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, **125**, 167–188.

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

37. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Muller, P., Mathews, D. H. & Zuker, M. (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA*, **91**, 9218–9222.

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

39. Knuth, D. E. (1973). *Fundamental Algorithms*, vol. 3 of *The Art of Computer Programming*, 237. Addison-Wesley, Reading, Massachusetts.

40. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. (1994). From sequences to shapes and back: a case study in rna secondary structures. *Proc. R. Soc. Lond. B Biol. Sci.*, **255**, 279–284.

41. Altschul, S. F. & Erickson, B. W. (1985). Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, **2**, 526–538.

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

43. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Table 1. z-scores and detection sensitivities for single and aligned sequences of various functional RNAs

	Single sequence				Number of sequences in alignment											
					2				3				4			
ncRNA Type	n	Z_{mono}	Z_{di}	S	n	ID	Z	S	n	ID	Z	S	n	ID	Z	S
tRNA	579	-1.84	-1.71	2.24	329	76.60	-5.15	71.12	479	73.29	-6.13	84.47	244	75.65	-6.76	98.36
5S rRNA	606	-1.62	-1.71	5.11	87	77.34	-3.89	40.23	81	80.03	-5.26	70.37	102	79.24	-5.12	69.61
Hammerh. III	251	-3.08	-3.17	8.80	94	76.07	-5.50	80.85	120	78.44	-6.10	93.33	130	79.74	-6.11	98.46
Gr. II Intron	116	-3.88	-3.77	44.82	109	75.98	-5.79	89.91	138	76.26	-7.00	94.20	134	76.06	-7.03	96.27
SRP RNA	73	-3.37	-3.09	34.24	135	77.29	-6.52	89.63	55	78.42	-7.09	90.91	50	78.75	-7.59	92.00
U5	199	-2.73	-2.38	17.58	110	74.32	-4.36	49.09	125	74.88	-5.14	64.80	127	74.57	-5.43	71.65

n . . . number of sequences/alignments scored, ID . . . average mean pairwise identity, Z . . . average z-score, S . . . sensitivity (% below -4).

Table 2
z-scores of ncRNAs in *C. elegans* aligned to homologs of *C. briggsae*

				z-score	
ncRNA Type	No. of Seqs.	Identity (%)	Length	Single	Alignment
SRP RNA	2	83.8	296	−5.5	−7.9
U1 spliceosome RNA	2	91.5	165	−4.6	−5.0
U2 spliceosome RNA	2	94.5	193	−5.0	−5.9
U4 spliceosome RNA	2	99.3	139	+0.7	+0.2
U5 spliceosome RNA	2	92.7	123	−2.3	−5.0
U6 spliceosome RNA	2	98.0	102	−0.8	−0.4
let-7 pre-miRNA	2	89.0	73	−7.5	−8.4
lin-4 pre-miRNA	2	90.0	70	−4.1	−4.8
SL2 RNA	2	91.3	103	−2.5	−3.6

Table 3
z-scores of ncRNAs in *S. cerevisiae* aligned to homologs of six related yeast species

					z-score	
ncRNA Type	Gene Name	No. of Seqs.	Identity (%)	Length	Single	Alignment
SRP RNA	SCR1	5	78.5	709	−2.2	−5.0
MRP RNA	NME1	7	81.5	355	−4.6	−8.9
RNAse P RNA	RPR1	7	72.3	402	−3.8	−6.7
U1 spliceosome RNA	snR19	5	82.9	683	−3.2	−6.7
U4 spliceosome RNA	snR14	7	88.0	165	−2.4	−4.2
U5 spliceosome RNA	snR7-L	5	88.0	218	−3.6	−4.5
	snR7-S	5	91.2	181	−3.3	−4.5
U6 spliceosome RNA	snR6	7	92.8	122	−1.9	−0.3
H/ACA snoRNA	snR3	4	89.5	196	−2.3	−2.8
	snR5	5	91.8	213	−1.5	−2.6
	snR8	5	94.4	197	−1.7	−1.9
	snR9	5	88.5	191	−1.3	−3.2
	snR10	7	83.4	280	−2.1	−3.8
C/D snoRNA	snR4	5	77.3	190	−1.3	−1.6
	snR13	6	89.9	127	−2.7	−2.9
	snR38	7	84.0	100	−0.1	0.0
	snR39	7	83.2	97	−0.4	−0.2
	snR40	6	80.7	99	+0.7	0.0

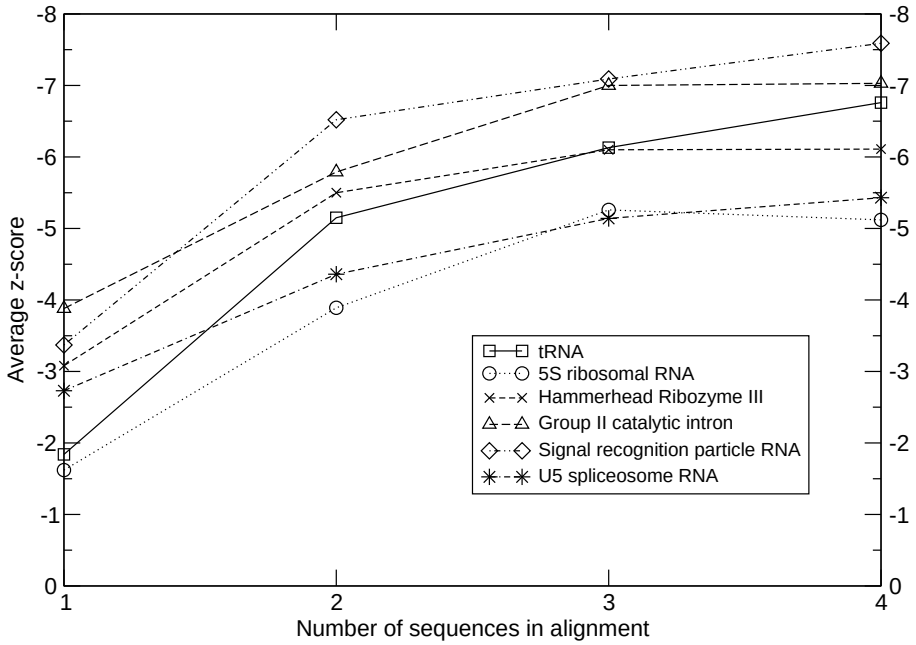

Fig. 1. Mean z-scores of various RNA types dependent on the number of sequences in alignment. $N = 1$ means `RNAfold` predictions for single sequences. Mean pairwise identities of the alignments are between 65% and 85%. See Table 1 for more details.

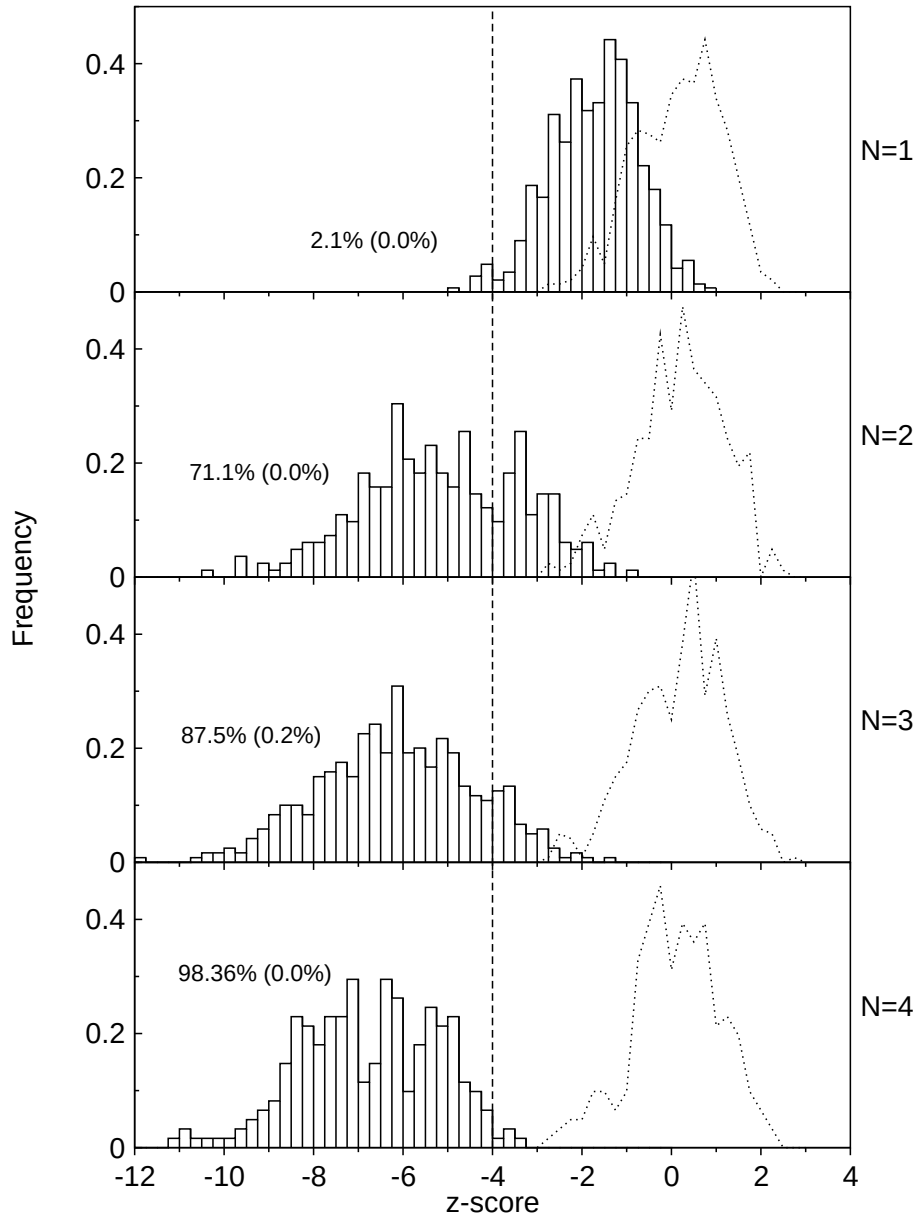

Fig. 2. Distribution of z-scores for the tRNA test sets. The distribution of native z-scores are shown as bars. The distribution of z-scores of the corresponding random sequences are shown as dashed line. N is the number of sequences in the alignment. $N = 1$ means `RNAfold` predictions for single sequences. The sensitivity (percent of native alignments with a z-score below a threshold of -4) and the selectivity (percent of random alignments with z-scores below -4) are shown for each set.

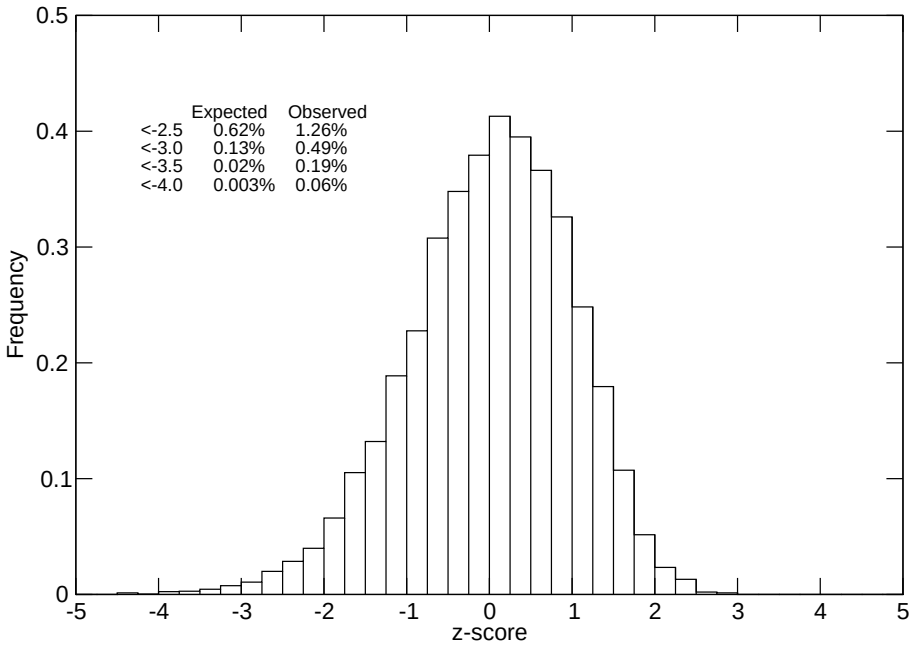

Fig. 3. Frequency distribution of 11633 random z-scores. The theoretically expected (for a standard normal distribution) and empirically observed frequencies below certain thresholds are shown.

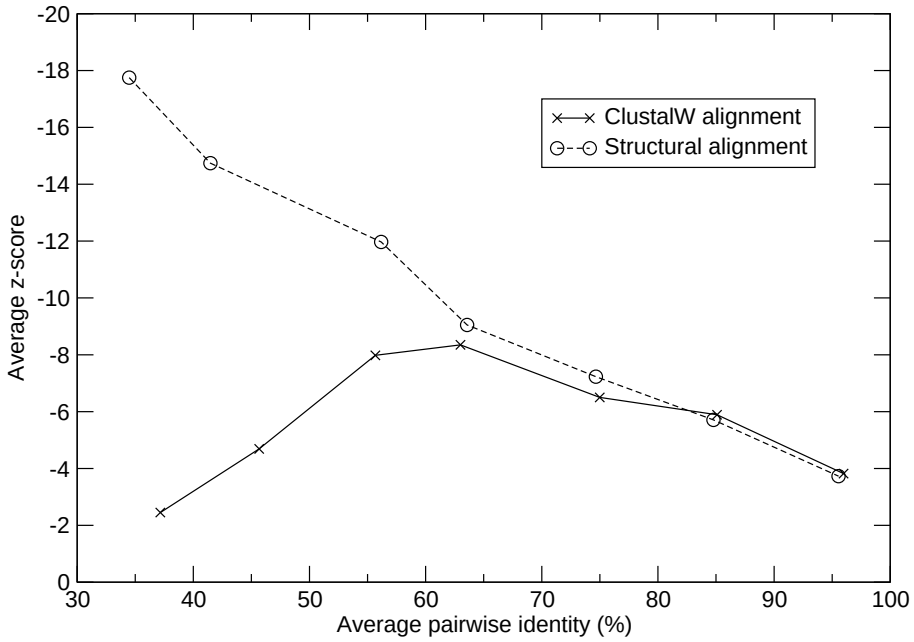

Fig. 4. Average z-scores of structural and sequence-based pairwise alignments of SRP RNAs versus pairwise identity. 2083 alignments were scored and average z-scores where calculated for seven intervals of pairwise identities between 30% an 100%. The average z-scores are plotted against the average pairwise identities calculated for each interval.

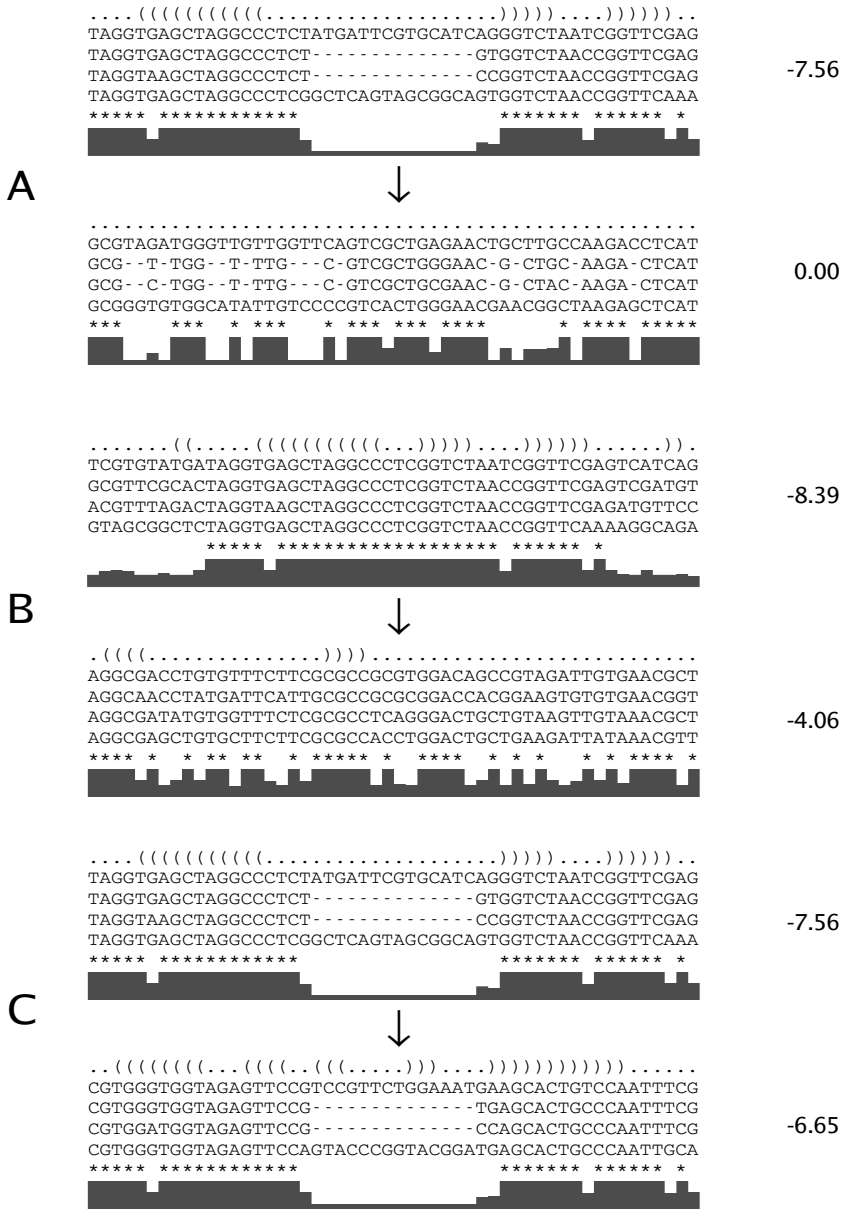

Fig. 5. Randomization of multiple sequence alignments. Three examples of shuffled alignments are shown. In A and B, the alignments are randomized by simply shuffling the columns. In C, only columns of the same gap pattern and local conservation pattern are shuffled. The degree of conservation is illustrated by black bars of varying size and asterisk for perfectly conserved columns. Each alignment was folded using `RNAalifold`. The consensus secondary structure prediction is shown in dot/bracket-notation in the first line. The `RNAalifold`-MFE is shown next to the alignment. (A) The alignment has one long gap in the middle which is spread over the whole length of the alignment after shuffling. In the resulting random alignment, `RNAalifold` cannot predict a consensus secondary structure (MFE=0.0). This results in significant low z-scores (-4.1 in this special case) although there is no unusually stable structure in the initial alignment (see C). (B) A highly conserved block is embedded in a less conserved region. Shuffling destroys this block and the consensus structure of the resulting random alignment is thus more unstable. Artifacts of this kind can lead to low z-scores and thus false positives. (C) The same alignment as in A is shuffled using our conservative algorithm. The randomized alignment retains the gap pattern and local conservation pattern of the initial alignment. It has a comparable MFE although the consensus structure is completely different (they do not have a single base pair in common). Using this shuffling procedure, we obtain a meaningful z-score of -0.8.