

Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics

Stefan Washietl and Ivo L. Hofacker*

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

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 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 *Caenorhabditis elegans*/*Caenorhabditis briggsae* and seven *Saccharomyces* species. Most RNAs can be detected with high significance. We provide a Perl implementation that can be used readily to score single alignments and discuss how the methods described here can be extended to allow for efficient genome-wide screens.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: conserved secondary structure; consensus structure prediction; non-coding RNAs; comparative genomics; randomizing multiple sequence alignments

*Corresponding author

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; ncRNAs) 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 a 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 concerns *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 metabolites⁶ 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 genomes 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.^{7–11} A more general approach that

Abbreviations used: MFE, minimum free energy; RUF, RNAs of unknown function; SGD, *Saccharomyces* Genome Database; ncRNA, non-coding RNA.

E-mail address of the corresponding author:
ivo@tbi.univie.ac.at

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.^{12–14} 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 ncRNAs.¹⁵ Some other statistical measures, partly derived from secondary structure predictions, have been proposed.^{16–18} 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,^{19,20} seven yeast species,^{21,22} two nematodes^{23,24} and the two mammalian genomes from human²⁵ and mouse.²⁶ 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.²⁷ This approach has been applied successfully to predict candidates for non-coding RNAs in *Escherichia coli* and *Saccharomyces cerevisiae*, some of which could be verified experimentally.^{28,29}

Here, 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 of whether this can be a significant measure to detect functional RNAs in genome-wide screens.

Results and Discussion

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 of whether natural RNAs are more stable (have lower MFE) than random sequences. This question has been partly addressed.¹⁵ Here, we test it again for sequences from a set of six structural RNA families (tRNA, 5 S 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 Methods). The results are shown in Table 1. On average, the structural RNAs all have 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? Figure 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 detect only 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 & Krogh³⁰ 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 the 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 are not necessarily below the dinucleotide-shuffled scores. Thus, while dinucleotide composition was important in the study by Workman & Krogh, where long (> 500 nt) mRNAs are tested for an (obviously non-existent) subtle bias towards lower folding energies, it can be neglected in our case.

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)

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

ncRNA type	Single sequence		Number of sequences in alignment											
			2				3				4			
	n	Z_{mono}	n	ID	Z	S	n	ID	Z	S	n	ID	Z	S
tRNA	579	-1.84	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	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	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	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	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	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).

ncRNAs, but is in itself not sufficient for a reliable detection. Given the availability of comparative data mentioned in Introduction, we wondered how to make use of this information efficiently. We use the program RNAalifold, which was originally developed to predict consensus secondary structures of aligned sequences.³¹ 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 of 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 Methods and Figure 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 Figure 1. If we compare the average z-score from the single sequences to the average z-scores of the pairwise alignments ($N=2$),

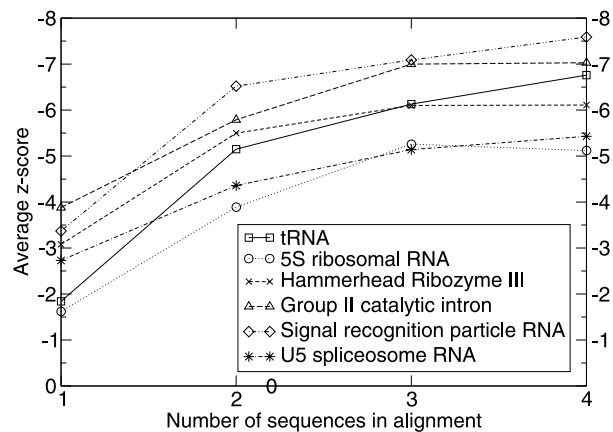


Figure 1. Mean z-scores of various RNA types dependent on the number of sequences in alignment. $N=1$ means RNAfold predictions for single sequences. Mean pairwise identities of the alignments are between 65% and 85%. See Table 1 for more details.

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 Figure 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%).

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

Sensitivity and selectivity depend on a predefined z-score threshold. To estimate the false positive rate for our test set, we also scored a shuffled random control for each alignment in the set. The distribution of 5930 random z-scores is shown in Figure 3. Three alignments had z-scores below -4 . This means that the sensitivities shown in Table 1 have a corresponding false-positive rate of 0.05%. The form of the distribution is of

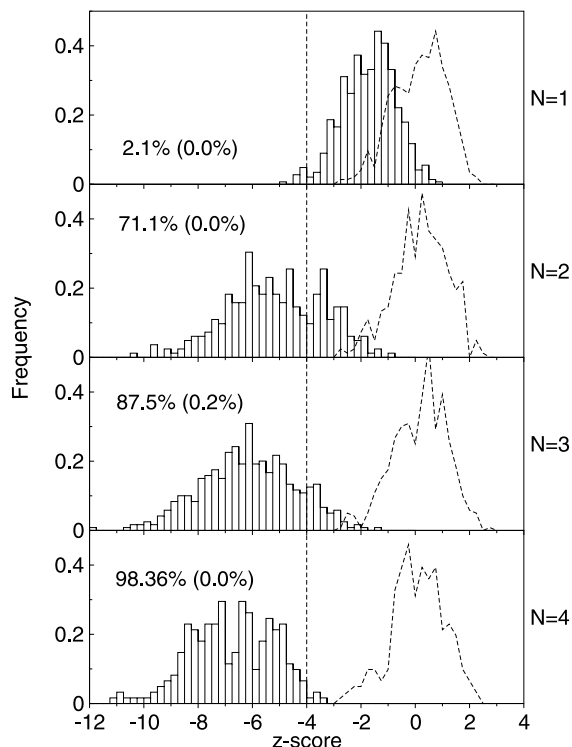


Figure 2. Distribution of z-scores for the tRNA test sets. The distribution of native z-scores are shown as bars. The distribution of z-scores of the corresponding random sequences are shown as broken line. N is the number of sequences in the alignment. $N=1$ means RNAfold predictions for single sequences.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

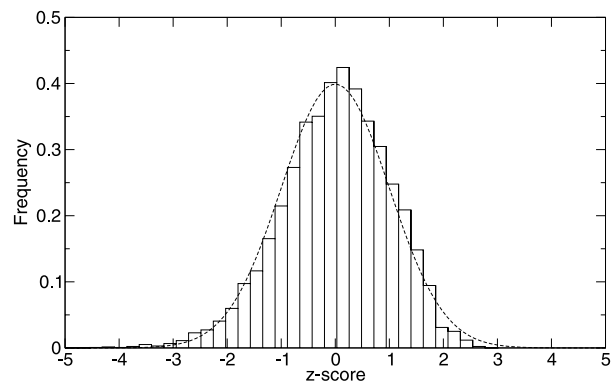


Figure 3. Frequency distribution of random z-scores. The bars show the distribution of z-scores of 5930 random alignments (mean 0.003 and standard deviation 0.989). The broken line shows a standard normal distribution.

particular interest. It can be fairly well approximated by a standard normal distribution. However, the distribution is slightly skewed with a negative 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- or dinucleotide shuffled) show the same skew in the distribution (not shown), as noted also in other studies.¹⁵ A possible explanation might be that we select the minimum free energy from random sequences and one could therefore expect behavior similar to an extreme value distribution. As it turns out, the extreme value distribution overestimates the tail significantly, giving a slightly worse fit than the normal distribution.

In any case, the significance of a given cutoff has to be estimated empirically. Especially 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. This is exactly what we find for automatically generated yeast alignments (see below) where the false-positive rate is significantly higher than on our test set of Clustal W alignments of Rfam sequences.

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 algorithms³² 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

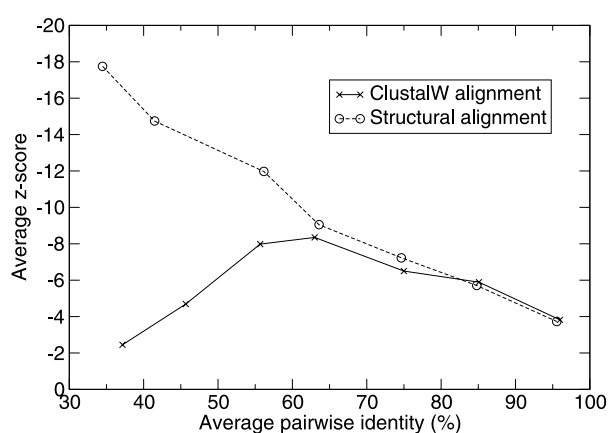


Figure 4. Average z-scores of structural and sequence-based pairwise alignments of SRP RNAs *versus* pairwise identity. The 2083 alignments were scored and average z-scores were calculated for seven intervals of pairwise identities between 30% and 100%. The average z-scores are plotted against the average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

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.³³ We calculated z-scores for structurally aligned pairs and for ClustalW aligned pairs (Figure 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 approximately 60–100% 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%.

Although the sensitivity will vary at different degrees of conservation, the practicability of our method is not limited to a specific interval of

pairwise identities. Since the z-score combines both energy contribution and the covariance contribution, we can detect stable structures even at 100% conservation. On the other hand, structures which are not exceptionally stable can be detected on the basis of covariance information if there is enough variation in the sequences. It must be pointed out that the selectivity is constant in any case and for all pairwise identities. In contrast, QRNA shows good performance at around 85% pairwise identity, but produces increasingly many false positives above this value.

Tests on ncRNAs from *C. elegans*

The results so far show that detection sensitivity depends highly on the quality of the available data. A large number of homologous sequences with high divergence (but still alignable) is desirable. However, in real-life applications, such ideal data sets will rarely be found. To test our method on more realistic data we created pairwise alignments of known ncRNAs from *C. elegans*²⁴ and *C. briggsae*²³ and calculated z-scores (Table 2). For scanning whole genomes it will not be feasible to predict structures longer than about 200 nt. 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* alone.

We found that the ncRNA sequences are highly conserved between *C. elegans* and *C. briggsae*. Pairwise identities are above 90% in most cases. Still, most RNA 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 of the order of about 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 intermolecular interactions with U4 rather than forming a stable intramolecular secondary structure. U6 only features a short 5'-stem loop. Although predicted by RNAalifold in the native alignment, this stem-loop is too short to be

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

ncRNA type	No. of seqs	Identity (%)	Length	z-Score	
				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. Sensitivity on known ncRNAs in *S. cerevisiae*

ncRNA type	Annotated genes	Detected genes ($z < -4$)	Sensitivity (%)
tRNA	275	28	10.2
rRNA	11	6	55.5
snRNA	6	4	66.7
C/D snoRNA	46	5	10.9
H/ACA snoRNA	20	14	70.0
Other ncRNAs of known function	4	4	100.0
ncRNAs of unknown function (RUF)	5	5	100.0

significantly different from the random background.

Tests on ncRNAs from *S. cerevisiae*

Pairwise alignments can easily be obtained by BLAST. However, if more than two genomes are available, multiple sequence alignments have to be generated. The generation of high-quality multiple sequence alignments on a genome-wide scale is a difficult task and still the subject of heavy research. We evaluated the performance of our method on automatically generated alignments on the genome of *S. cerevisiae* to draft sequences of six related yeast species.^{21,22} We chose MultiPipMaker,³⁴ which is currently the only program available which can align a reference sequence to unassembled contigs on a genome-wide level off-the-shelf.

To estimate the sensitivity for screening the yeast genome we used the following procedure: we generated multiple sequence alignments of all 16 chromosomes. We then extracted the regions of annotated ncRNAs. Since MultiPipMaker could not find homologous sequences in all species for all ncRNAs and sometimes only dubious fragments could be found, the alignments needed refinement. As before, we scanned the rough alignments in windows of size 150 and slide 20. Before a window was scored, poorly aligned sequences were discarded and the windows were realigned using CLUSTALW (see Methods). We then evaluated the MFE of the native alignment. Only if it was below -15 we eventually calculated the z -score. This could be a realistic scenario for a genome-wide scan. If we encountered a window having a z -score below -4 we regarded the ncRNA as detected.

Table 3 summarizes the results for the different ncRNA classes. Table 4 shows detailed predictions for selected ncRNAs. The alignment characteristics and z -scores are shown for the best-scoring 150 column window in each of the genes.

Only a small fraction of the transfer RNAs can be found. This is due to the high conservation ($>95\%$) of this class of RNAs. As expected, also the ribosomal RNAs are highly conserved between the closely related yeast species. Still, the large 18 S and 25 S subunits contain extremely stable local secondary structures that can be detected even at 100% conservation. As seen for *C. elegans*, RNA genes lacking a stable secondary structure are missed. This is true for some small nuclear RNAs and all C/D-type small nucleolar RNAs. The

H/ACA type snoRNAs on the other hand have a typical two stem-loop secondary structure and therefore 14 of 20 can be detected. The six that are missed score around -3 . Again, the high conservation (around 90%) hinders a more efficient detection.

All other known RNAs (SRP, RNaseP, RNase MRP and telomerase-RNA) can be detected. In addition to these, a previous screen using QRNA²⁹ had identified eight RNAs of unknown function (RUF). Of these RUF4, RUF6 and RUF7 could not be verified in additional experiments, and RUF8 has been found to be a coding mRNA (a correction has been issued for the original paper)[†]. This is consistent with our predictions: only RUF1, RUF2, RUF3 and the two copies of RUF5 have z -scores below -4 . We do not find significantly conserved secondary structures in RUF4, RUF6, RUF7 or RUF8.

To conclude, our method has good sensitivity in this test screen. Most of the structured RNAs which show some variation in sequence (i.e. which are not too conserved) could be detected.

To assess the false-positive rate for the cutoff of -4 in this experiment, we repeated the screen in exactly the same way, but shuffled the windows before calculating the z -score. In the 313 genes, we scored 807 randomized windows and encountered two windows scoring below -4 . This corresponds to a false-positive rate of 0.25% per alignment.

To estimate the false-positive rate not only on known ncRNAs but also on coding genes and other conserved regions, we randomized the complete chromosome 5 by shuffling the alignment in non-overlapping windows of length 150. We then scanned the random chromosome in non-overlapping windows of length 150 in the same way as before. We scanned both the forward direction and the reverse complement. This resulted in 2217 conserved blocks with RNAalifold MFE below -15 after the re-alignment step. Out of these, five had a z -score below -4 , which is a false-positive rate of 0.23% per alignment. This is approximately the same as we found for the randomized ncRNAs. The chromosome 5 is 574,860 base-pairs long. So we can expect around eight to ten false positives per megabase of the yeast genome in such a screen. However, this number of statistical false positives will depend on how many overlapping windows

[†] ftp://ftp.genetics.wustl.edu/pub/eddy/papers/2003-mccutcheon-yeast/correction_long.pdf

Table 4. The z-scores of selected ncRNAs in *S. cerevisiae*

ncRNA type	Gene name	No. of seqs	Identity (%)	z-Score	
				Single	Alignment
Signal recognition particle RNA	SCR1	4	85.6	-2.2	-4.2
RNAase P RNA	RPR1	4	85.0	-3.7	-6.5
RNAase MRP RNA	NME1	6	69.1	-4.8	-11.1
Telomerase RNA	TLC1	3	71.1	-4.5	-7.4
U1 spliceosomal RNA	snR19	6	74.3	-3.4	-8.5
U2 spliceosomal RNA	LSR1	3	74.9	-6.3	-6.5
U4 spliceosomal RNA	snR14	6	87.0	-1.8	-3.0
U5 spliceosomal RNA	snR7-L	5	80.6	-3.6	-5.5
	snR7-S	5	79.4	-3.4	-4.4
U6 spliceosomal RNA	snR6	6	90.9	-1.9	-2.3
RNAs of unknown function	RUF1	4	75.8	-4.4	-7.3
	RUF2	4	80.9	-4.0	-8.9
	RUF3	4	77.3	-3.9	-6.7
	RUF5-1	4	66.8	-3.0	-4.5
	RUF5-1	4	66.7	-2.4	-4.4

we score. This is yet another problem of using a sliding window (see the next section).

This number also does not include biological false positives as for example inverted repeats which could be interpreted as stable hairpins. Also pseudogenes could be a problem here. However, we expect our method to be quite robust to distinguish real ncRNAs from pseudogenes. Unlike other methods which search for sequence patterns, our method only relies on the conservation of a secondary structure. It is known that only a small number of random mutations destroy secondary structures³⁵ and it is thus unlikely that pseudogenes retain a conserved structure without evolutionary pressure.

Towards genome-wide scans

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 evolutionarily 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 about 200 nt. Although many known functional RNA structures are longer than 200 nt, this seems to be long enough to detect local substructures. However, a sliding window has several other drawbacks. Only for a step-size of 1, 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 nt should be detected within a much longer window of 200 nt. 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 by our group.³⁶ In principle, the idea can 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 at 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 doing a precise evaluation, e.g. if the estimated z-score from a sample of ten is above -1 it is unlikely the real z-score will fall below -3.5. For these reasons, the 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 without doubt computationally demanding but feasible at least for middle-sized genomes. Using our current Perl implementation, a scan of the yeast genome (about 13 MB) would take about 30 CPU days on a 2.8 GHz Pentium 4 processor (including the ClustalW realignment step). Considering the points mentioned above and using a fast implementation in C this procedure could be accelerated by an order of magnitude.

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 true only 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.

Supplementary Material and Programs

Supplementary material including all data sets is available on our website†.

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‡. Another Perl script `shufflealn.pl` is provided, which implements the shuffling algorithm described here. It might be useful also for other purposes.

Methods

Consensus folding of aligned sequences

We use the program `RNAalifold`³¹ from the Vienna RNA package³⁷ version 1.5 to perform consensus secondary structure predictions of multiple sequence alignments. `RNAalifold` essentially uses the same algorithms³⁸ and energy parameters^{39,40} 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.³¹ 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.

Estimation of statistical significance

In analogy to previous work¹²⁻¹⁴ we use a Monte Carlo approach to estimate statistical significance of a MFE. For each alignment we generate 100 random alignments (see the 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 $z = (m - \mu)/\sigma$. Negative z-scores indicate that the MFE of the native alignment is lower than those of the randomized alignments.

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 Figure 5. If there is, for example, a gap of length 10 in the alignment, the shuffling probably would produce ten gaps of length 1. This can result in artifactual 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 that 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.⁴¹ Finally, we reassemble the alignment. Since the shuffling procedure of the individual sets is probably 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 not 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

† <http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/Alifoldz/>

‡ <http://www.tbi.univie.ac.at/RNA/>

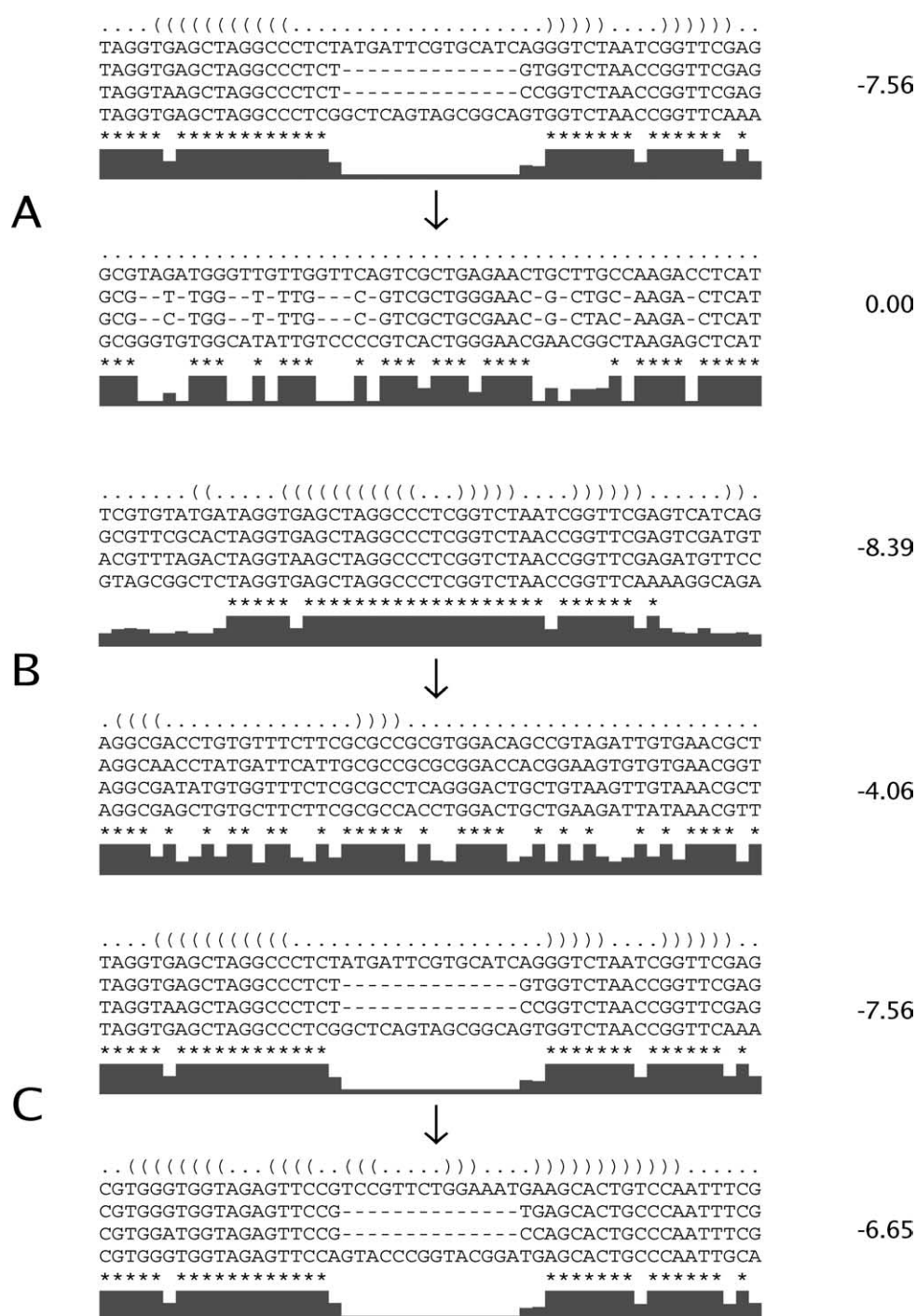


Figure 5. Randomization of multiple sequence alignments. Three examples of shuffled alignments are shown. In A and B, the alignments are randomized by simply shuffling the columns. In C, only columns of the same gap pattern and local conservation pattern are shuffled. The degree of conservation is illustrated by black bars of varying size and asterisks 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.

possible number of permutations, the question arises of whether 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.³⁵ These theoretical considerations, as well as our computational results, suggest that the shuffling procedure is effective enough to destroy any native secondary structures.

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.*† of an algorithm developed by Altschul & Erickson.⁴²

Creation of test sets

Most of the RNA sequences used in this work were taken from the Rfam database release 5.0.⁴³ 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.³³ 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 (2–4 sequences) within a defined range of mean pairwise identity (65–85%) 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‡) 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 the possibility 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 CLUSTALW⁴⁴ and the mean pairwise identities were calculated. For the experiments shown in Table 1 and Figures 1 and 2, we eventually used alignments with mean pairwise identities between 65% and 85%. To estimate the false positive rate, we generated a shuffled version of each of the alignments. Here we used all alignments generated by the procedure above. This set consisted of 5930 alignments with mean pairwise identities between 30% and 100%, GC-content between 30% and 70% and length between 50 and 350 columns. The test set included 3280 pairwise alignments, 1701 alignments with $N=3$ and 949 alignments with $N=4$.

ncRNAs from *C. elegans* and *C. briggsae*

For the *C. elegans*/*C. briggsae* alignments, we tried to take one example of each ncRNA family (excluding tRNAs and rRNAs)²³. 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§ 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.²³

Automatic alignments of yeast ncRNAs

The alignments of the yeast examples were created by the following procedure: we downloaded the (draft) sequences of seven yeast species from the Saccharomyces Genome Database (SGD||): *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *S. kluyveri*. Next, we created chromosome-wide multiple sequence alignments using MultiPipmaker³⁴ with the options “search both strands”, “single coverage” and “high sensitivity and low time limit”. From these raw alignments, we extracted the regions of ncRNA genes known for *S. cerevisiae* (according to the annotation table downloaded from SGD July 2003). The alignments were scanned in windows of 150 and slide 20. To get reasonable alignments, the following editing steps were performed for each window: the pairwise identity of the reference sequence from *S. cerevisiae* to all other sequences in the alignment was calculated. If it was below 60% the sequence was dropped. We included the gap character in the calculation of the similarity and thus excluded sequences that did not have a match in the region (i.e. only gaps) and also sequences that had only some short fragments aligned by MultiPipmaker. After this selection, we removed the gaps in the remaining sequences and re-aligned them using CLUSTALW. Finally, we calculated the z-scores for the re-aligned window. Since there is no sense in calculating a z-score if there is no stable secondary structure even in the native alignment, we only considered alignments which had a RNAali-fold MFE below -15 . For the random controls, we shuffled the window after the editing steps but before the MFE and the z-score was determined.

Acknowledgements

Useful comments from Peter F. Stadler are gratefully acknowledged. This project has been funded, in part, 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.

† <http://www.clavius.bc.edu/~clotelab/>

‡ <http://www.ncbi.nlm.nih.gov/>

§ <http://www.wormbase.org>

|| <ftp://ftp.yeastgenome.org/yeast>

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. *Nature 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. *Nucl. 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. Regalia, M., Rosenblad, M. A. & Samuelsson, T. (2002). Prediction of signal recognition particle RNA genes. *Nucl. Acids Res.* **30**, 3368–3377.
10. 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.
11. Klein, R. J. & Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinform.* **4**, 44.
12. Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V., Jr (1988). A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.* **4**, 153–159.
13. Le, S. Y., Chen, J. H. & Maizel, J. V., Jr (1989). Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucl. Acids Res.* **17**, 6143–6152.
14. Chen, J. H., Le, S. Y., Shapiro, B., Currey, K. M. & Maizel, J. V., Jr (1990). A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.* **6**, 7–18.
15. Rivas, E. & Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
16. Schultes, E. A., Hrabec, P. T. & LaBean, T. H. (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* **49**, 76–83.
17. Le, S. Y., Zhang, K. & Maizel, J. V., Jr (2002). RNA molecules with structure dependent functions are uniquely folded. *Nucl. Acids Res.* **30**, 3574–3582.
18. Le, S. Y., Chen, J. H., Konings, D. & Maizel, J. V., Jr (2003). Discovering well-ordered folding patterns in nucleotide sequences. *Bioinformatics*, **19**, 354–361.
19. McClelland, M., Florea, L., Sanderson, K., Clifton, S. W., Parkhill, J., Churcher, C. *et al.* (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. *Nucl. Acids Res.* **28**, 4974–4986.
20. Florea, L., McClelland, M., Riemer, C., Schwartz, S. & Miller, W. (2003). EnteriX 2003: visualization tools for genome alignments of Enterobacteriaceae. *Nucl. Acids Res.* **31**, 3527–3532.
21. Cliften, P. F., Hillier, L. W., Fulton, L., Graves, T., Miner, T., Gish, W. R. *et al.* (2001). Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11**, 1175–1186.
22. 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.
23. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N. *et al.* (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45.
24. *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
25. The Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
26. International Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
27. Rivas, E. & Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform.* **2**, 8.
28. 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.
29. McCutcheon, J. P. & Eddy, S. R. (2003). Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucl. Acids Res.* **31**, 4119–4128.
30. Workman, C. & Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.* **27**, 4816–4822.
31. Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066.
32. Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* **45**, 810–825.
33. Rosenblad, M. A., Gorodkin, J., Knudsen, B., Zwieb, C. & Samuelsson, T. (2003). SRPDB: signal recognition particle database. *Nucl. Acids Res.* **31**, 363–364.
34. Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A. *et al.* (2003). MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucl. Acids Res.* **31**, 3518–3524.
35. 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. Roy. Soc. ser. B*, **255**, 279–284.
36. Hofacker, I. L., Priwitzer, B. & Stadler, P. F. (2004). Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
37. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188.
38. Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9**, 133–148.
39. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H.,

- Muller, P., Mathews, D. H. & Zuker, M. (1994). Coaxial stacking of helices enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
40. 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.
41. Knuth, D. E. (1973). *The Art of Computer Programming*, vol. 3. Addison-Wesley, Reading, MA.
42. 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.
43. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2003). Rfam: an RNA family database. *Nucl. Acids Res.* **31**, 439–441.
44. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Edited by J. Doudna

(Received 21 April 2004; received in revised form 5 July 2004; accepted 9 July 2004)