Towards a general approach for the detection of non-coding RNAs by comparative genomics

Stefan Washietl

Institute for Theoretical Chemistry and Structural Biology, University Vienna

Bled, Slovenia 2004

Non coding RNAs everywhere. . .

Non coding RNAs (“RNA genes”) are transcripts that exert their function as RNA whithout being translated to protein.

- Well known examples directly or indirectly involved in protein gene expression:
  - Protein expression: **transfer RNA, ribosomal RNA**
  - Pre-mRNA splicing: **spliceosomal RNAs (U1,U2,U4,U5,U6,. . .)**
  - (r)RNA modification: **small nucleolar RNAs**
  - tRNA maturation: **Ribonuclease P**
  - Protein export: **Signal recognition particle RNA**
- Most prominent new class of non-coding RNAs: **microRNAs**
- Many other examples are currently emerging.

. . . and even more

- In complex organisms like human 97-98% of transcripts are ncRNAs.
- In few cases single ncRNAs have been described with interesting implications for physiology and phathology
  - **roX1/2 Xist/Tsix** are involved in X chromosome dosage compensation in mammals and drosphila, resp.
  - Y-chromosome specific **TTY2** family is expressed in testis and kidney
  - **Bic** is strongly upregulated in certain B-cell lymphomas
  - **SCA** is involved in the neurodegenerative disorder spinocerebellar ataxia type 8
  - **DISC2** is implicated in the molecular etiology of schizophrenia
  - Mutations in **RMRP** cause the development disorder cartilage-hair hypoplasia (CHH)
  - One of the known loci associated with autism encodes a ncRNA.

Computational identification of ncRNAs

- Based on *a priori* knowledge: find members of known families
  - Sequence similiarity alone: `BLASTN`
  - Sequence and additional motif information: specialized programs for e.g. tRNA or snoRNAs
- *De novo* prediction: find new genes and families
  - Unlike protein coding genes (ORFs, codon bias,. . .) ncRNAs lack statistical signals in primary sequence
  - Many known ncRNA have a characteristic secondary structure.

Is secondary structure prediction a reliable measure for the detection of ncRNAs?

z-score statistics

Has a natural occuring RNA sequence a lower minimum free energy (MFE) than random sequences of the same size and base composition?

1. Calculate native MFE m.
2. Calculate mean μ and standard deviation σ of MFEs of 100 shuffled random sequences.
3. Express significance in standard deviations from the mean as z-score

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

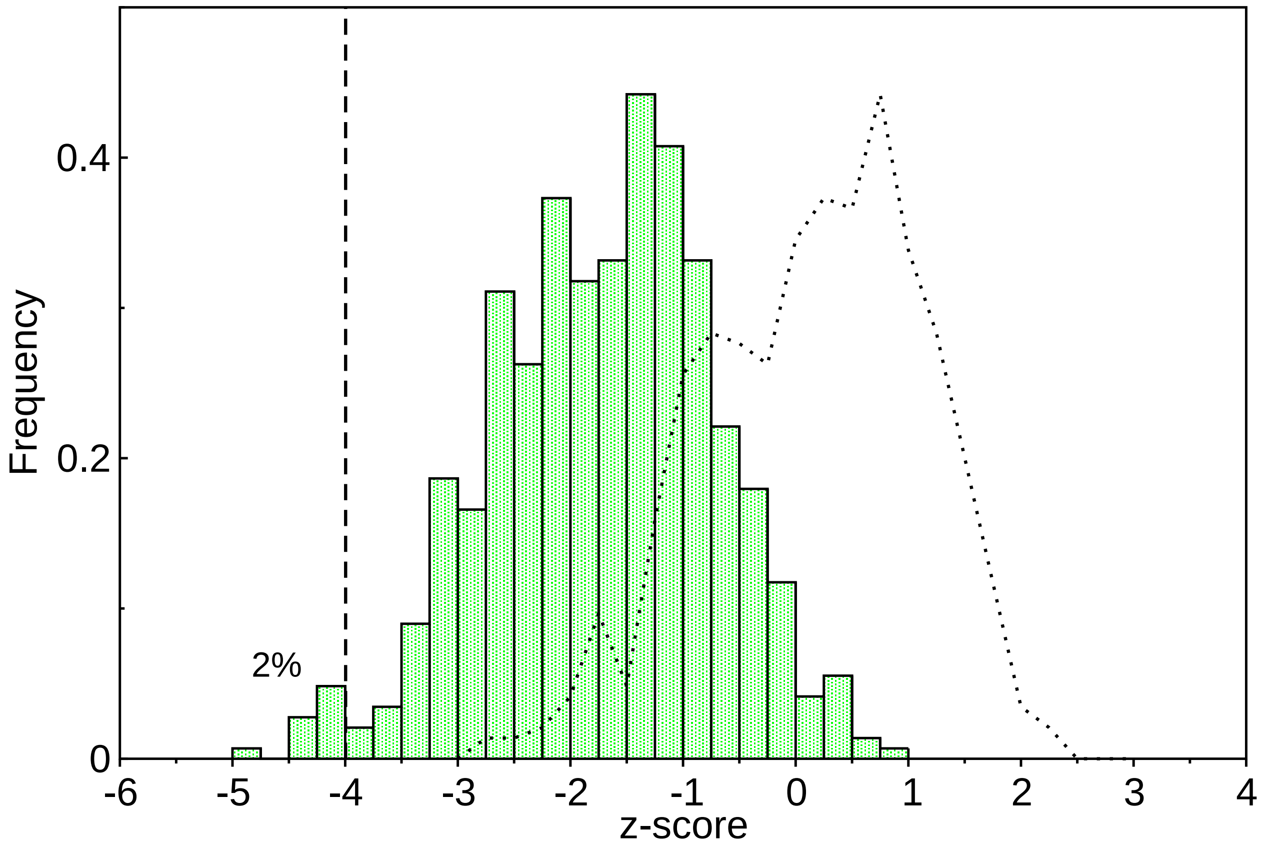
Negative z-scores indicate that the native RNA is more stable than the random RNAs.

MFE z-scores of known functional RNAs

ncRNA Type	No. of Seqs.	Mean z-score
tRNA	579	−1.84
5S rRNA	606	−1.62
Hammerhead ribozyme III	251	−3.08
Group II catalytic intron	116	−3.88
SRP RNA	73	−3.37
U5 spliceosomal RNA	199	−2.73

- Functional RNAs are clearly more stable than random sequences.
- Is this significant enough for genome wide screens?

z-score distribution for 579 tRNAs

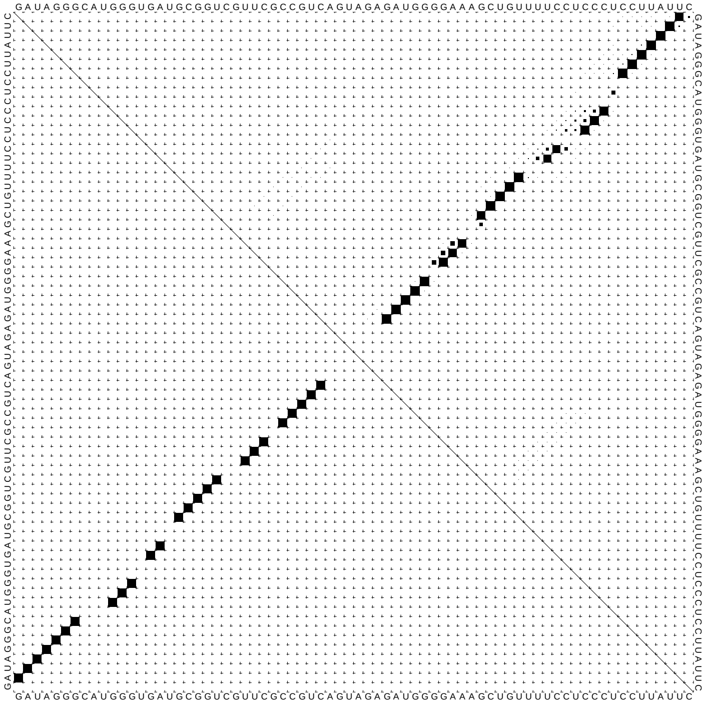


- Only 2% below a z-score threshold of -4.
- Native sequences are not clearly separated from the random bulk.

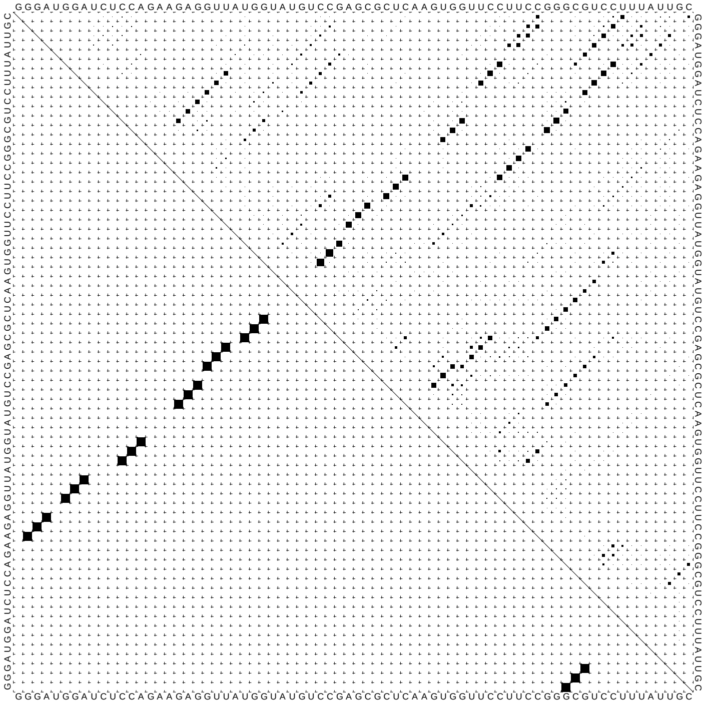
Well-definedness of RNA secondary structure

- At a given temperature RNA molecules form an ensemble of structures which is described by the Boltzmann distribution.
- If this ensemble is dominated by the ground state (MFE structure) we call the structure well-defined.

well-defined (mean distance = 3.27)



not well-defined (mean distance = 15.7)



A measure for well-definedness

- As measure for well-definedness we can use the mean distance between structures in the ensemble.
- For the so-called “base-pair distance” metric the mean distance can be calculated from the base-pair probability matrix as

$$\langle D \rangle = \sum_{i<j} p_{ij} - p_{ij}^2$$

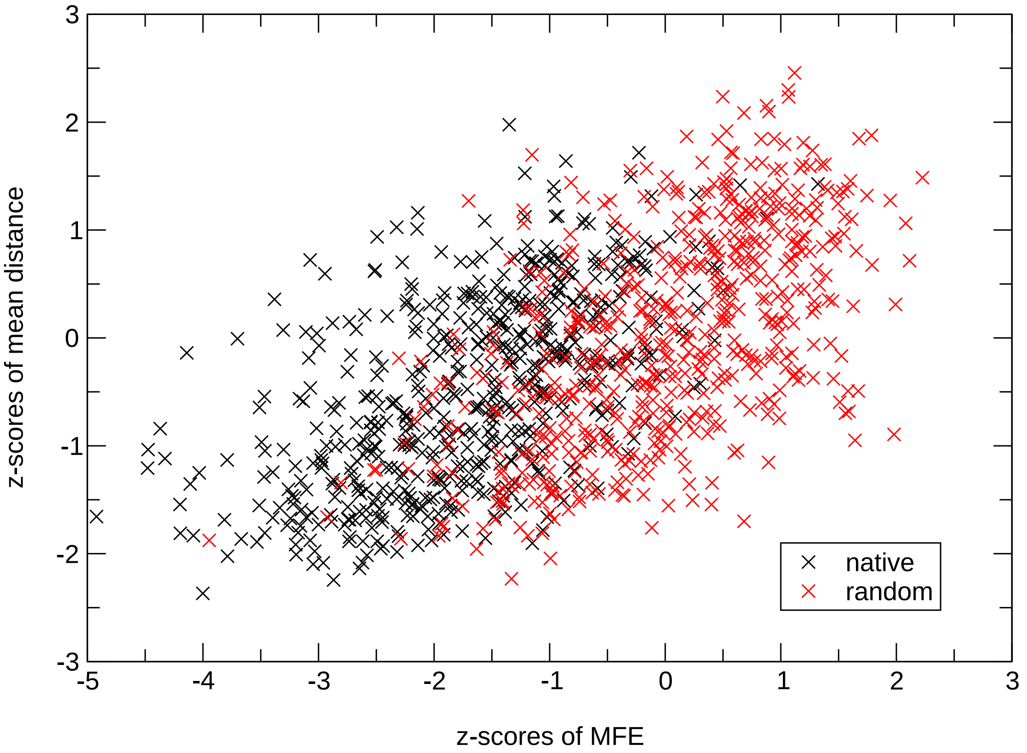
Are functional RNAs better defined than random seqeunces?

Well-definedness of functional RNAs

ncRNA Type	Mean z MFE	Mean z well-definedness
tRNA	-1.84	-0.5
5S rRNA	-1.62	-0.7
Hammerhead ribozyme III	-3.08	-1.5
Group II catalytic intron	-3.88	-1.2
U5 spliceosomal RNA	-2.73	-1.1

- z-scores for mean-distances are less significant than z-scores based on MFEs.
- Can a combination of both help?

Well-definedness and MFE are not independent



- Well-definedness and MFE are (to some degree) linear dependent.
- Well-definedness holds no additional information for our purpose.

Measures for single sequence predictions are not significant enough for detecting ncRNAs.

Comparative genomics at our hands

- Prokaryotes: **15 enteric bacteria**
- Yeast: **7 Sacharomyces species**
- Nematode: **C. elegans + C. briggsae** (C. remanei, C. japonica and CB5161 planned)
- Mammals: **Mouse, rat, human**

How can we make use of homologous sequences for ncRNA finding?

QRNA (Rivas & Eddy)

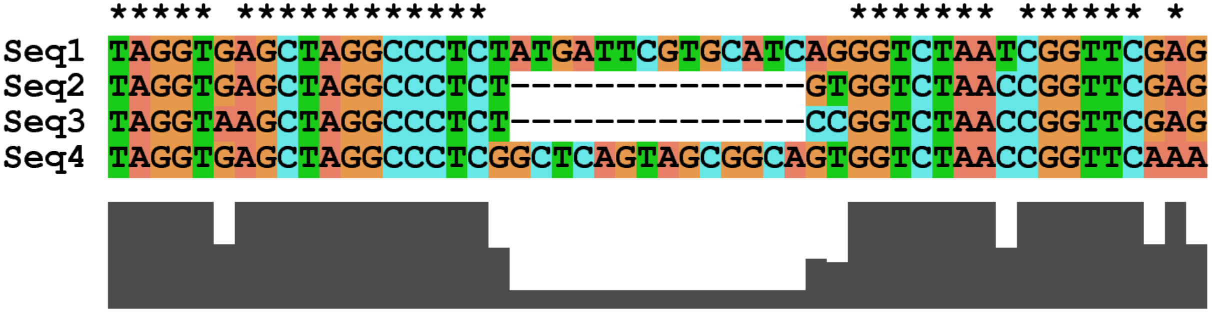
- For a given pairwise alignment decide if it is coding, structural RNA or neither.
- There is one probabilistic model for each case which evaluates the mutation pattern. The RNA model implements a probabilistic folding algorithm.
- `QRNA` can be useful to some degree but has several disadvantages:
  - The model parameters depend on many *ad hoc* assumptions and extrapolations.
  - Performance depends strongly on GC content and pairwise identity.
  - Sensitivity and selectivity is generally low for non-optimal data sets.
  - `QRNA` is relatively slow.
  - The probabilistic folding model is not optimal (e.g. trained with rRNAs and tRNAs and thus strongly biased).
  - `QRNA` is limited to pairwise alignments.

An alternative approach: RNAalifold

- `RNAalifold` performs MFE folding of a multiple sequence alignment
- It essentially uses the same algorithms and energy parameters as `RNAfold`.
- Energy contributions of the single sequences are averaged.
- Covariance information is incorporated into the energy model:
  - Consistent and compensatory mutations are rewarded.
  - Non compatible base pairs are penalized.
- It calculates a (pseudo-)MFE consisting of an energy term and a covariance term.

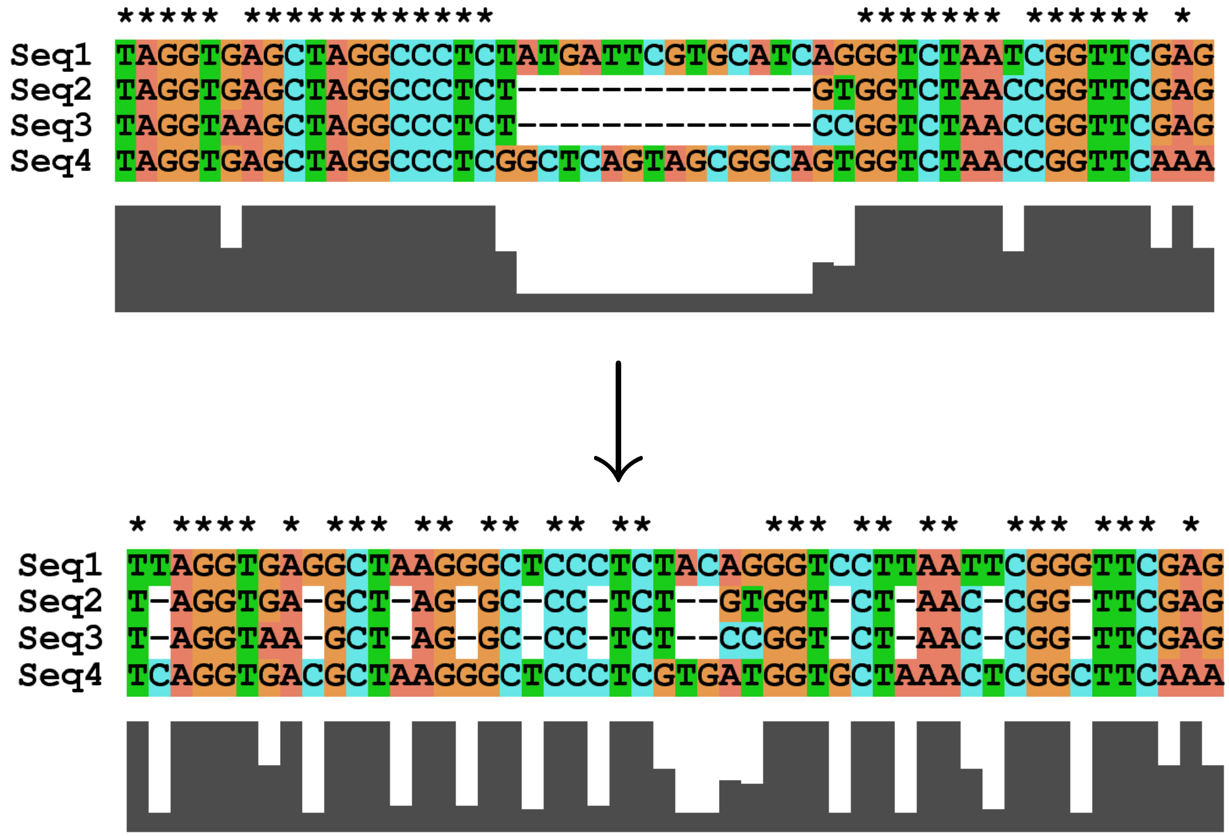
Can we use this MFE to assess an alignment for the existance of an unusually stable and/or conserved secondary structure?

How not to shuffle a MSA



```
      ***** ************                 ******* ****** * 
Seq1  TAGGTGAGCTAGGCCCTCTATGATTCGTGCATCAGGGTCTAATCGGTTCGAG
Seq2  TAGGTGAGCTAGGCCCTCT--------------GTGGTCTAACCGGTTCGAG
Seq3  TAGGTAAGCTAGGCCCTCT--------------CCGGTCTAACCGGTTCGAG
Seq4  TAGGTGAGCTAGGCCCTCGGCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
```

How not to shuffle a MSA

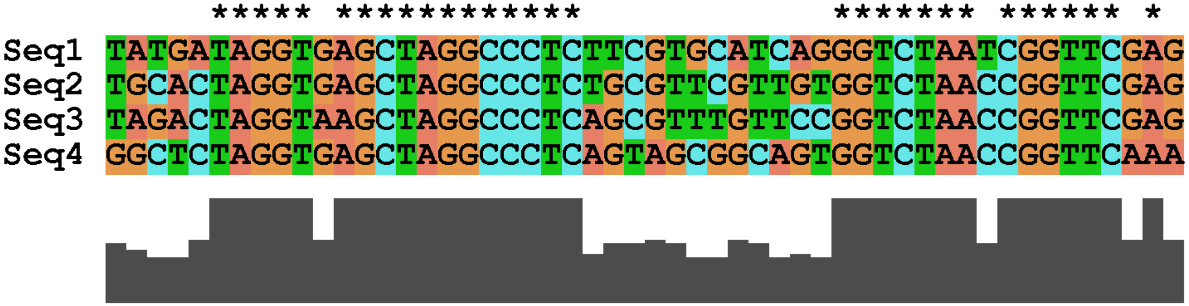


```
     *****  *************                *******  ******  *
Seq1 TAGGTGAGCTAGGCCCTCTATGATTCGTGCATCAGGGTCTAATCGGTTCGAG
Seq2 TAGGTGAGCTAGGCCCTCT--------------GTGGTCTAACCGGTTCGAG
Seq3 TAGGTAAGCTAGGCCCTCT--------------CCGGTCTAACCGGTTCGAG
Seq4 TAGGTGAGCTAGGCCCTCGGCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
```

```
     * **** * *** ** ** ** **     *** ** **  *** *** *
Seq1 TTAGGTGAGGCTAAGGGCTCCCTCTACAGGGTCCTTAATTCGGGTTCGAG
Seq2 T-AGGTGA-GCT-AG-GC-CC-TCT--GTGGT-CT-AAC-CGG-TTCGAG
Seq3 T-AGGTAA-GCT-AG-GC-CC-TCT--CCGGT-CT-AAC-CGG-TTCGAG
Seq4 TCAGGTGACGCTAAGGGCTCCCTCGTGATGGTGCTAAACTCGGCTTCAAA
```

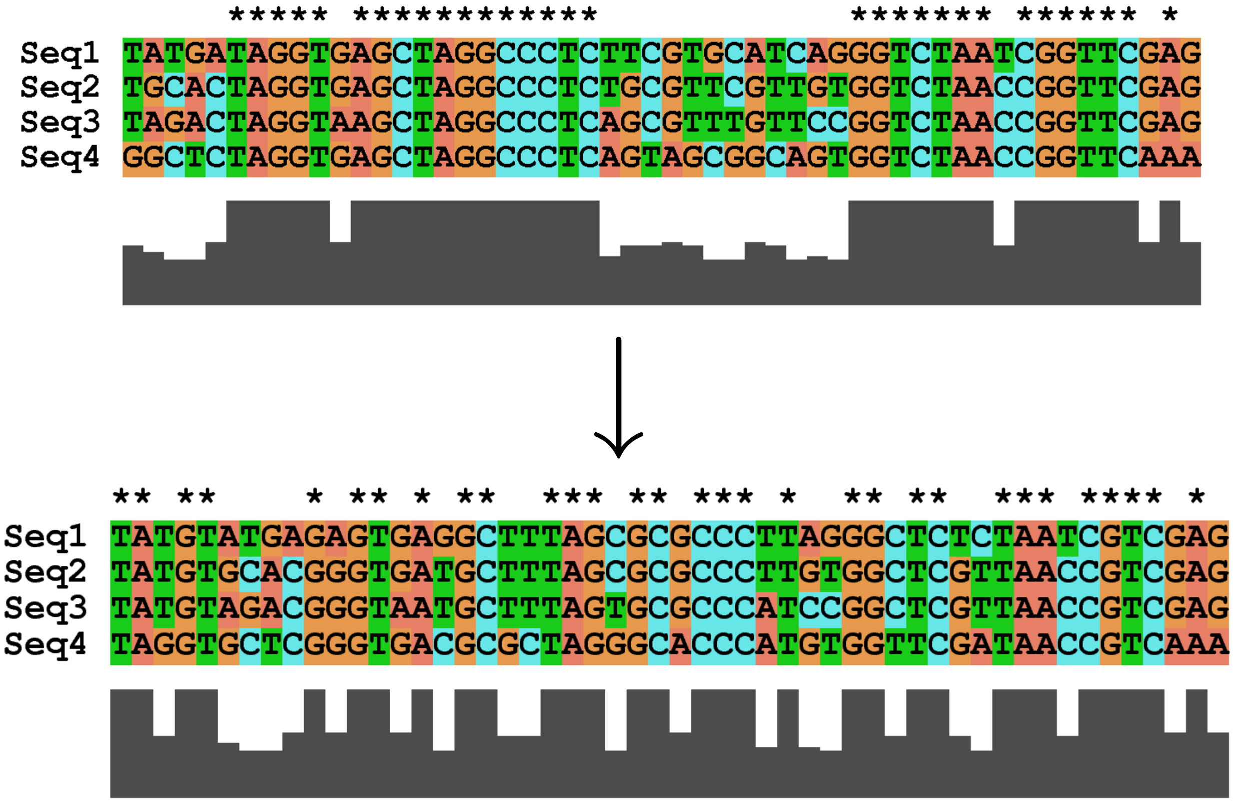
Gap structure is important

How not to shuffle a MSA II



```
          ***** ************            ******* ****** *
Seq1 TATGATAGGTGAGCTAGGCCCTCTTCGTGCATCAGGGTCTAATCGGTTCGAG
Seq2 TGCACTAGGTGAGCTAGGCCCTCTGCGTTCGTTGTGGTCTAACCGGTTCGAG
Seq3 TAGACTAGGTAAGCTAGGCCCTCAGCGTTTGTTCCGGTCTAACCGGTTCGAG
Seq4 GGCTCTAGGTGAGCTAGGCCCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
```

How not to shuffle a MSA II



```
          *****  ************            *******  ******  * 
Seq1 TATGATAGGTGAGCTAGGCCCTCTTCGTGCATCAGGGTCTAATCGGTTCGAG
Seq2 TGCACTAGGTGAGCTAGGCCCTCTGCGTTCGTTGTGGTCTAACCGGTTCGAG
Seq3 TAGACTAGGTAAGCTAGGCCCTCAGCGTTTGTTCCGGTCTAACCGGTTCGAG
Seq4 GGCTCTAGGTGAGCTAGGCCCTCAGTAGCGGCAGTGGTCTAACCGGTTCAAA
```

↓

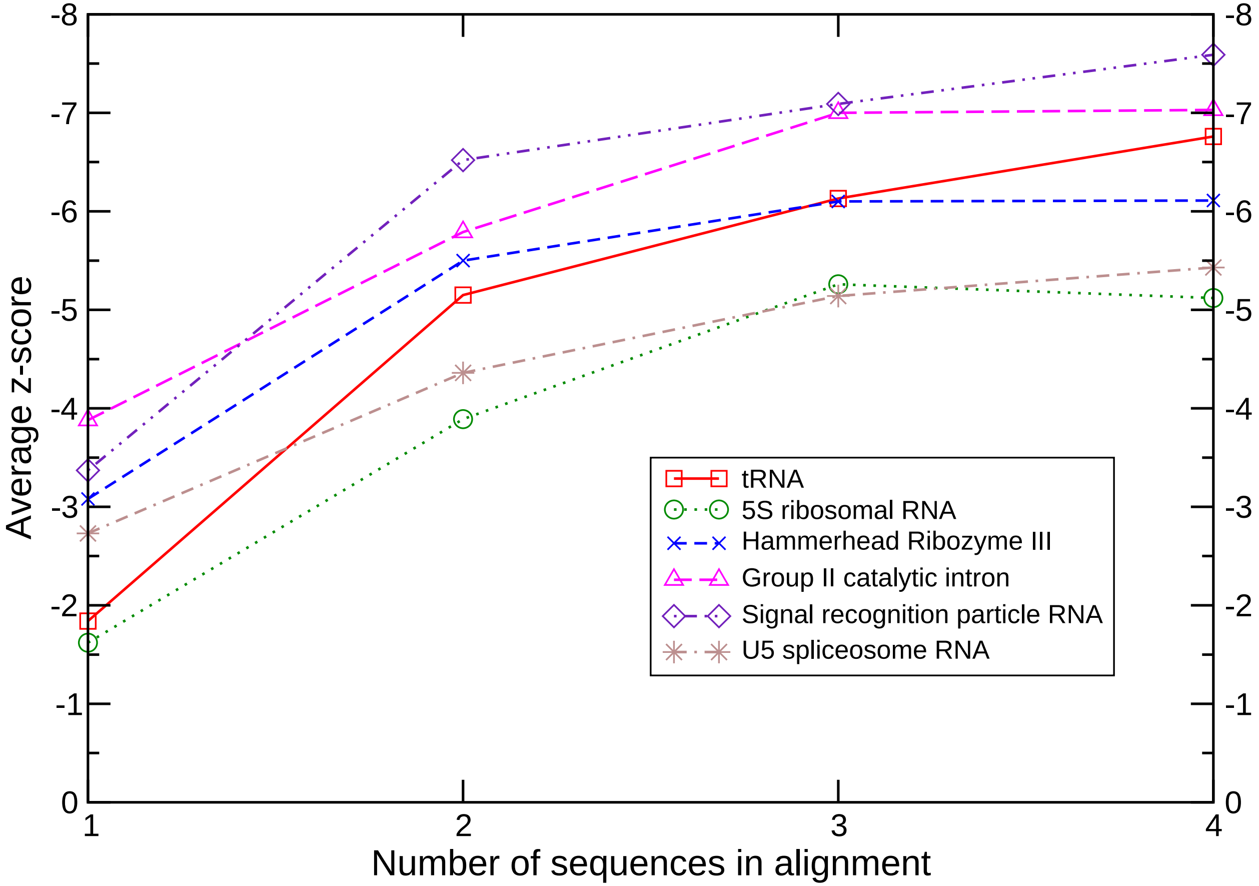
```
     ** **    * ** * **  *** ** *** *  ** **  *** **** * 
Seq1 TATGTATGAGAGTGAGGCTTTAGCGCGCCCTTAGGGCTCTCTAATCGTCGAG
Seq2 TATGTGCACGGGTGATGCTTTAGCGCGCCCTTGTGGCTCGTTAACCGTCGAG
Seq3 TATGTAGACGGGTAATGCTTTAGTGCGCCCATCCGGCTCGTTAACCGTCGAG
Seq4 TAGGTGCTCGGGTGACGCGCTAGGGCACCCATGTGGTTCGATAACCGTCAAA
```

Local conservation pattern is important

Conservative randomization of a MSA

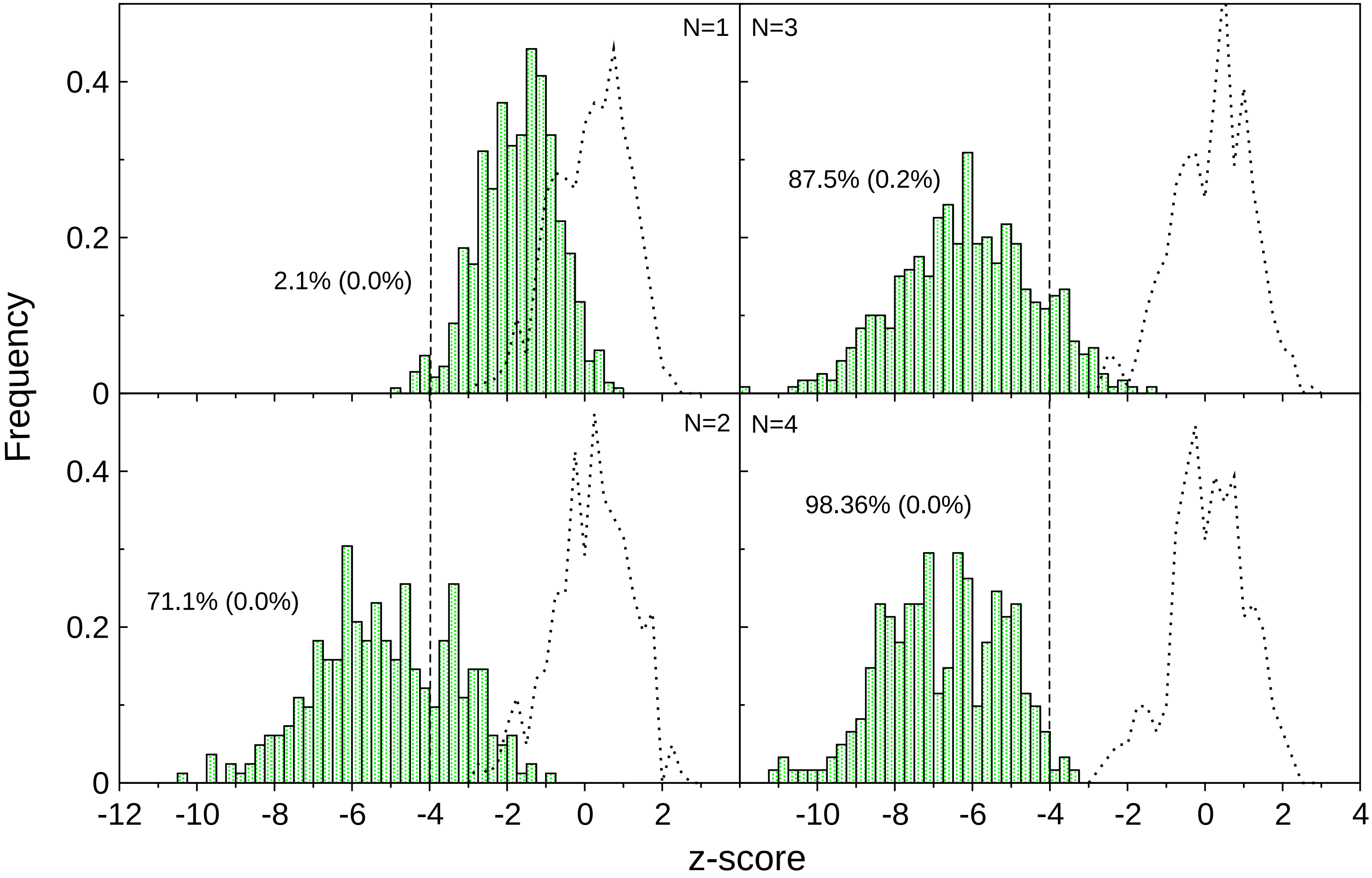
- A correct randomization procedure shuffles only columns of the same gap pattern and local conservation pattern.
- Considering this our algorithm produces alignments of the same
  - length
  - base composition
  - overall conservation
  - local conservation
  - gap structure
- This is the most conservative procedure possible. It is effective enough to remove correlations arising from secondary structures.

z-scores of RNAalifold MFEs



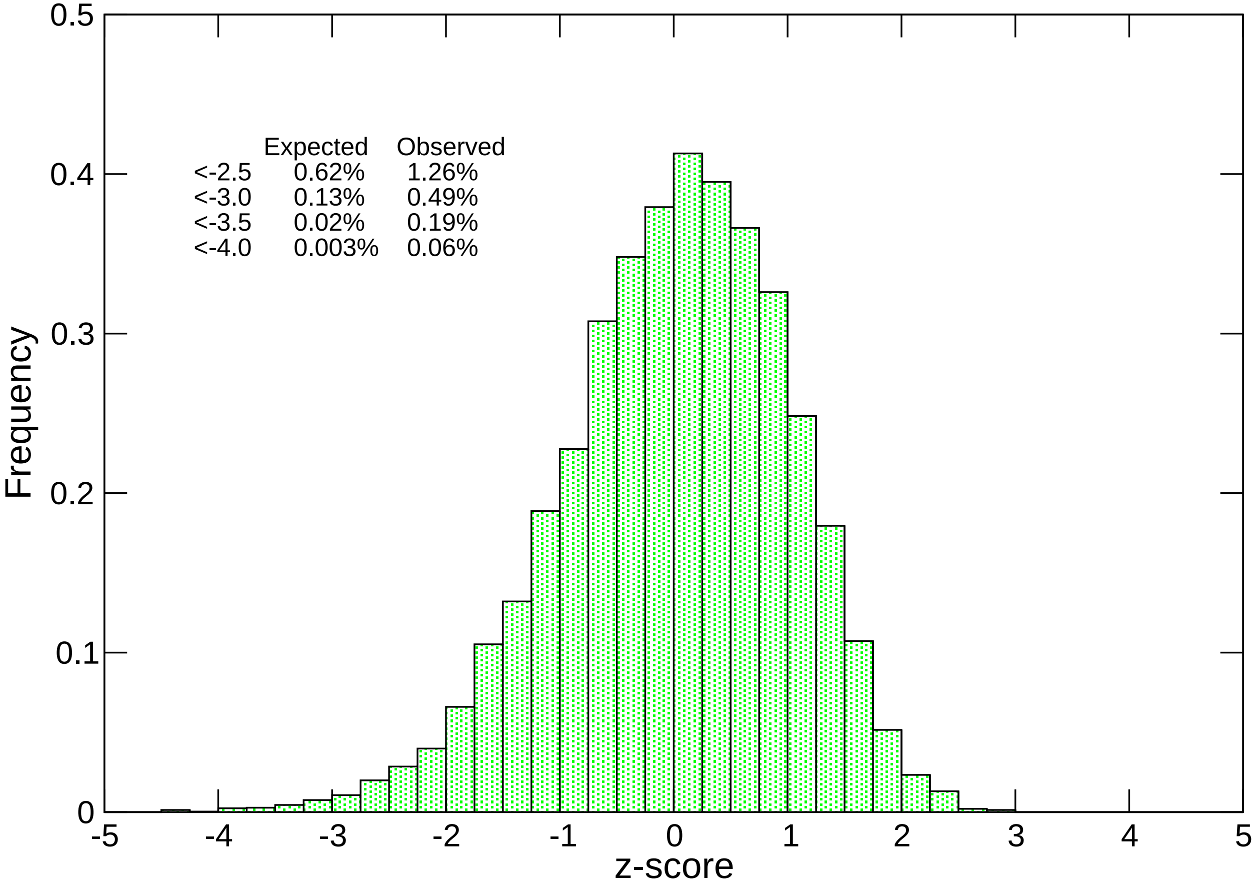
- We scored alignments with 2 to 4 sequences and mean pairwise identities between 65% and 85%.

z-score distribution for tRNA test sets



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

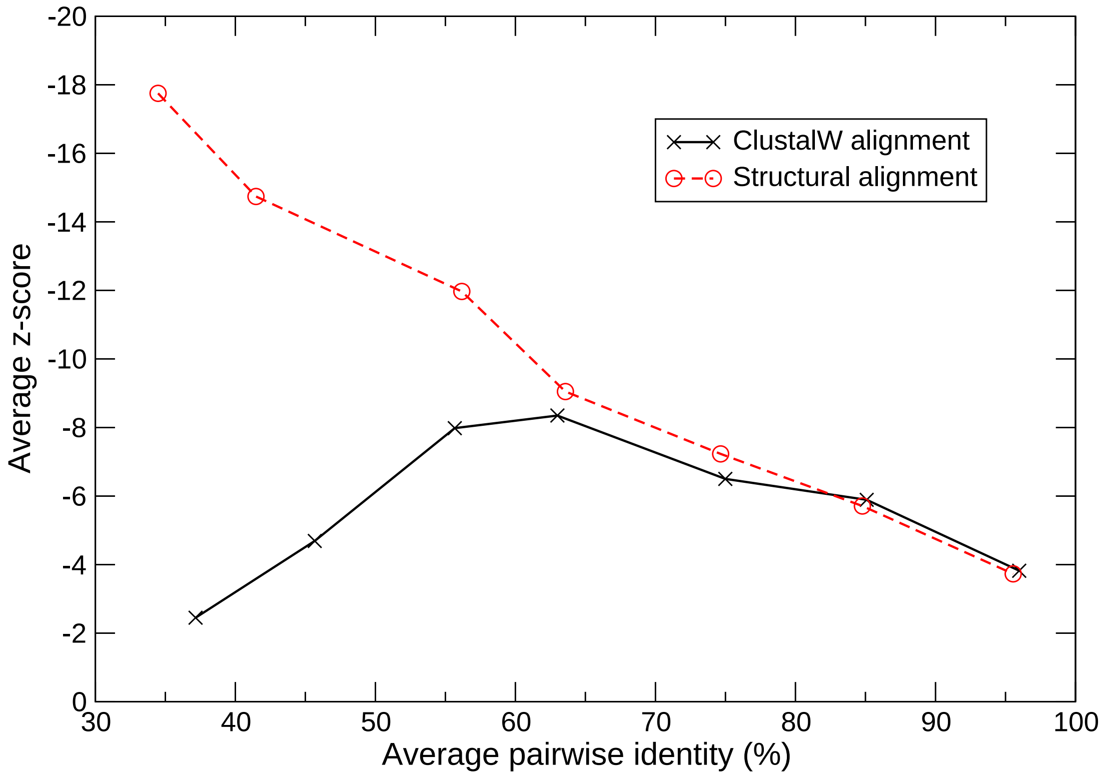
Distribution of 11633 random z-scores



	Expected	Observed
<-2.5	0.62%	1.26%
<-3.0	0.13%	0.49%
<-3.5	0.02%	0.19%
<-4.0	0.003%	0.06%

- z-scores of random alignments are well approximated by a standard normal distribution ($\mu =$0.01, $\sigma =$0.99) with a slight negative tail.

Structural vs. sequence based alignments



- 2083 pairwise alignments of SRP RNAs were scored.
- Above 60% there structural alignments and sequence based alignments are essentially the same.
- Our method scores best between 60% and 70%.

Genomic example: Saccharomyces sp.

				z-score	
ncRNA Type	Gene Name	No. of Seqs.	ID (%)	Single	Alignment
SRP RNA	SCR1	5	78.5	−2.2	−5.0
MRP RNA	NME1	7	81.5	−4.6	−8.9
RNAse P RNA	RPR1	7	72.3	−3.8	−6.7
U1 spliceosome RNA	snR19	5	82.9	−3.2	−6.7
U4 spliceosome RNA	snR14	7	88.0	−2.4	−4.2
U5 spliceosome RNA	snR7-L	5	88.0	−3.6	−4.5
	snR7-S	5	91.2	−3.3	−4.5
U6 spliceosome RNA	snR6	7	92.8	−1.9	−0.3
H/ACA snoRNA	snR9	5	88.5	−1.3	−3.2
	snR10	7	83.4	−2.1	−3.8
C/D snoRNA	snR4	5	77.3	−1.3	−1.6
	snR39	7	83.2	−0.4	−0.2

Genomic example: C.elegans/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

How to fold a complete genome?

- Straightforward approach: local predictions using a sliding window
- A sliding window has two major drawbacks:
  - Only for a step-size 1 all possible structures are considered. Realistic step sizes leave a “blind-spot”.
  - A fixed size window cannot predict all substructures of varying length optimally
- A local prediction algorithm is desirable
  - `QRNA` implements a local prediction algorithm.
  - Also standard algorithms for MFE predictions can be modified to smoothly scan a genome and predict all substructures smaller than a given maximum size: `RNAlfold`
  - In principle, this can be implemented also for `RNAalifold` without modification.

Is this feasible for complete genomes?

- Generally, `RNAalifold` is fast for moderate window sizes
- The Monte Carlo procedure to estimate statistical significance imposes a serious performance problem.
- A meaningful *ad hoc* score seems impossible. It would have to consider GC-content, degree of conservation, gap-pattern and length of the alignment.
- In theory, a genome has to be folded 200 times (sample size 100, forward and reverse strand)
- In practice, the number of calculations can be reduced drastically
  - Only conserved (=alignable) regions have to be analyzed
  - `RNAalifold` will not predict a consensus structure everywhere.
  - We are only interested 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. We can thus pre-estimate z-scores with lower sample size.

Summary

- The computational detection of non coding RNAs is a major goal of bioinformatics.
- Secondary structure predictions are of limited statistical significance.
- The same is true for other measures for single sequences (e.g. well-definedness)
- Comparative studies seem most promising but only few methods for comparative sequence analysis exist (`QRNA`).
- We have proposed a new procedure (z-scores of RNAalifold MFEs) to assess a multiple sequence alignment for the existence of a stable and/or conserved fold.
- Our method shows good sensitivity/selectivity in a variety of test cases, including real-life genomic examples.
- Our method is computationally demanding, but feasible if reduced to the essential.

What’s next?

1. Put all these ideas together into a (structural) RNA gene finder (“`RNAlalifoldz`”) as quickly as possible.
2. Convince people that this is the way to go and that `QRNA` sucks.
3. Start doing some biology.