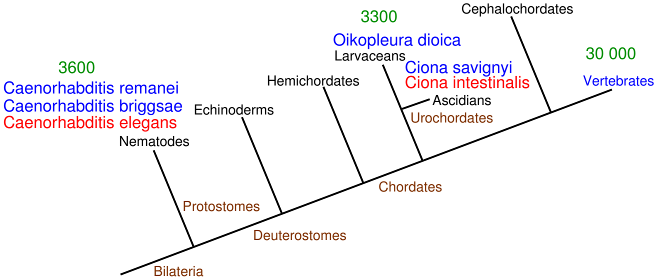
`RNAstrand` : Reading direction of structured RNAs in multiple sequence alignments

Kristin Reiche, Peter F. Stadler

Bioinformatics Group, Department of Computer Science
University of Leipzig

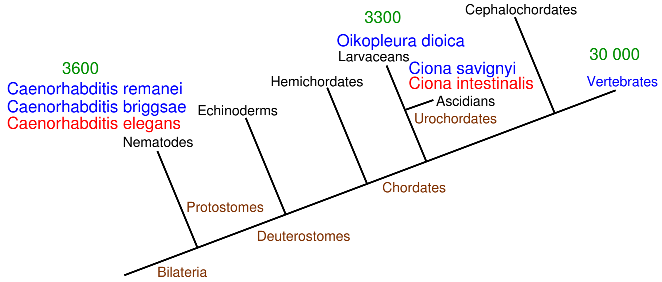
Bled 2007

Why strand predictor of structured RNAs?



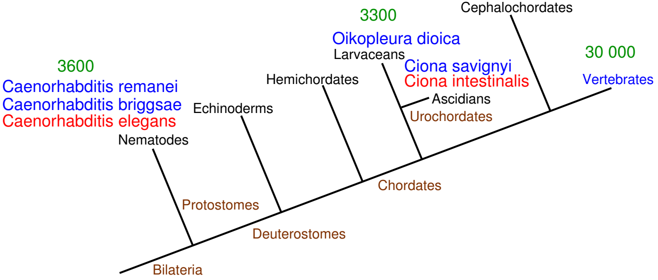
- Genome wide prediction of structured non-coding RNAs (`RNAz` , `EvoFold`)
- Annotation is next challenge:
  - RNAs with homologous secondary structures (`LocARNA`)
  - Detect specific RNA families (`RNAmicro` , `snoReport`)
  - intronic, intergenic or non-translated exon (`RNAstrand`)

Why strand predictor of structured RNAs?



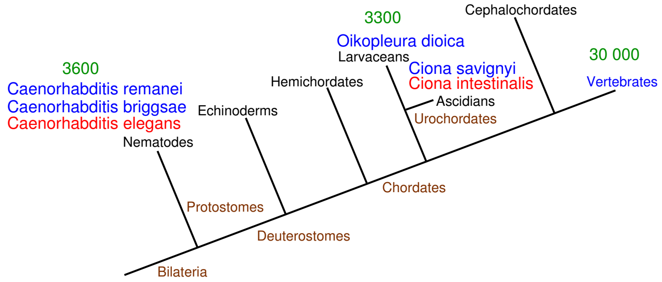
- Genome wide prediction of structured non-coding RNAs (`RNAz` , `EvoFold`)
- Annotation is next challenge:
  - RNAs with homologous secondary structures (`LocARNA`)
  - Detect specific RNA families (`RNAmicro` , `snoReport`)
  - intronic, intergenic or non-translated exon (`RNAstrand`)

Why strand predictor of structured RNAs?



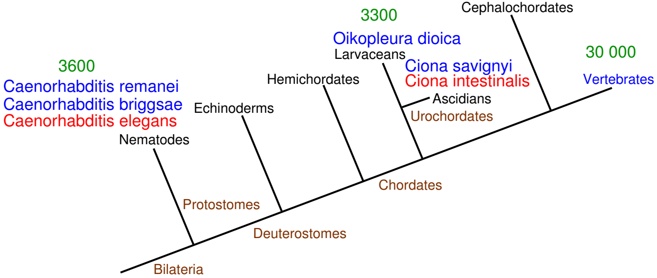
- Genome wide prediction of structured non-coding RNAs (`RNAz`, `EvoFold`)
- Annotation is next challenge:
  - RNAs with homologous secondary structures (`LocARNA`)
  - Detect specific RNA families (`RNAmicro`, `snoReport`)
  - intronic, intergenic or non-translated exon (`RNAstrand`)

Why strand predictor of structured RNAs?



- Genome wide prediction of structured non-coding RNAs (`RNAz` , `EvoFold`)
- Annotation is next challenge:
  - RNAs with homologous secondary structures (`LocARNA`)
  - Detect specific RNA families (`RNAmicro` , `snoReport`)
  - intronic, intergenic or non-translated exon (`RNAstrand`)

Why strand predictor of structured RNAs?



- Genome wide prediction of structured non-coding RNAs (`RNAz` , `EvoFold`)
- Annotation is next challenge:
  - RNAs with homologous secondary structures (`LocARNA`)
  - Detect specific RNA families (`RNAmicro` , `snoReport`)
  - intronic, intergenic or non-translated exon (`RNAstrand`)

Why strand predictor of structured RNAs?

- Naïve approach: Take strand which has higher `RNAz` probability or `EvoFold` score
- Problem: `RNAz` and `EvoFold` are not trained for strand prediction of RNA
- Accuracy: `RNAz` on miRNAs: 0.14, `EvoFold` on miRNAs: 0.84
- Can we do that better?

Why strand predictor of structured RNAs?

- Naïve approach: Take strand which has higher `RNAz` probability or `EvoFold` score
- Problem: `RNAz` and `EvoFold` are not trained for strand prediction of RNA
- Accuracy: `RNAz` on miRNAs: 0.14, `EvoFold` on miRNAs: 0.84
- Can we do that better?

Why strand predictor of structured RNAs?

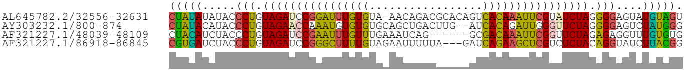
- Naïve approach: Take strand which has higher `RNAz` probability or `EvoFold` score
- Problem: `RNAz` and `EvoFold` are not trained for strand prediction of RNA
- Accuracy: `RNAz` on miRNAs: 0.14, `EvoFold` on miRNAs: 0.84
- Can we do that better?

Why strand predictor of structured RNAs?

- Naïve approach: Take strand which has higher `RNAz` probability or `EvoFold` score
- Problem: `RNAz` and `EvoFold` are not trained for strand prediction of RNA
- Accuracy: `RNAz` on miRNAs: 0.14, `EvoFold` on miRNAs: 0.84
- Can we do that better?

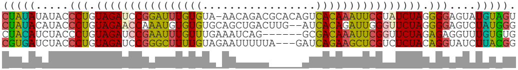
Task

Input alignment containing structured ncRNA



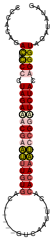
```
                        (((((.....(((.((((((((((((((((.................)))))))))))))))).))).....)))))).
AL645782.2/32556-32631  CUAUAUAUACCCUGUAGAUCCGGAUUUGUGUA-AACAGACGCACAGUCACAAAUUCGUAUCUAGGGGAGUAUGUAGU
AY303232.1/800-874      CUAUACAUACCCUGUAGAACCAAAUGUGUGUGCAGCUGACUUG--AUCACAGAUUGGGUUCUAGGGGAGUCUAUGGG
AF321227.1/48039-48109  CUACAUCUACCCUGUAGAUCCGAAUUUGUUUGAAAUCAG------GCGACAAAUUCGGUUCUAGAGAGGUUUGUGUG
AF321227.1/86918-86845  CGUGAUCUACCCUGUAGAUCCGGGCUUUUGUAGAAUUUUUA---GAUCAGAAGCUCGUCUCUACAGGUAUCUUACGG
```

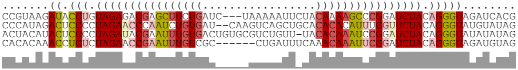
Structured ncRNA on reading direction of input alignment?



```
(((((.....(((.((((((((((((((((.................)))))))))))))))).))).....)))))).
CUAUAUAUACCCUGUAGAUCCGGAUUUGUGUA-AACAGACGCACAGUCACAAAUUCGUAUCUAGGGGAGUAUGUAGU
CUAUACAUACCCUGUAGAACCAAAUGUGUGUGCAGCUGACUUG--AUCACAGAUUGGGUUCUAGGGGAGUCUAUGGG
CUACAUCUACCCUGUAGAUCCGAAUUUGUUUGAAAUCAG------GCGACAAAUUCGGUUCUAGAGAGGUUUGUGUG
CGUGAUCUACCCUGUAGAUCCGGGCUUUUGUAGAAUUUUUA---GAUCAGAAGCUCGUCUCUACAGGUAUCUUACGG
```

Structured ncRNA on reverse complement of input alignment?



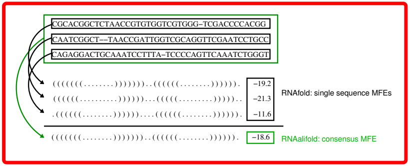


```
.......((.(((.((((((((((((((((.................)))))))))))))))).))))).........
CCGUAAGAUACCUGUAGAGACGAGCUUCUGAUC---UAAAAAUUCUACAAAAGCCCGGAUCUACAGGUAGAUCACG
CCCAUAGACUCCCCUAGAACCCAAUCUGUGAU--CAAGUCAGCUGCACACACAUUUGGUUCUACAGGGUAUGUAUAG
ACUACAUACUCCCCUAGAUACGAAUUUGUGACUGUGCGUCUGUU-UACACAAAUCCGGAUCUACAGGGUAUAUAUAG
CACACAAACCUCUCUAGAACCGAAUUUGUCGC------CUGAUUUCAAACAAAUUCGGAUCUACAGGGUAGAUGUAG
```

Approach

Approach

Reading direction of input alignment



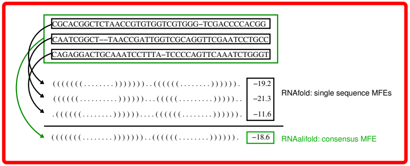
```
CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))).    –19.2
((((((..........))))))...(((((.........)))))..   –21.3
.((((((........))))))....((((((........)))))).   –11.6

(((((((........)))))))..((((((........)))))).    –18.6
```

RNAfold: single sequence MFEs

RNAalifold: consensus MFE

Approach

Reading direction of input alignment



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

RNAfold: single sequence MFEs

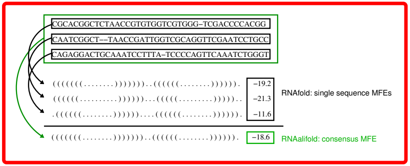
(((((((........)))))))..((((((........)))))). –18.6

RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

Approach

Reading direction of input alignment



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

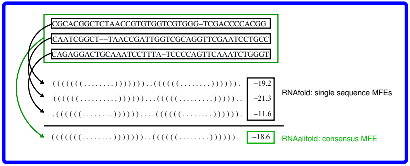
(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

(((((((........)))))))..((((((........)))))). –18.6

RNAfold: single sequence MFEs

RNAalifold: consensus MFE

Realigned reverse complement



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

(((((((........)))))))..((((((........)))))). –18.6

RNAfold: single sequence MFEs

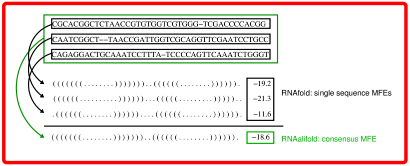
RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

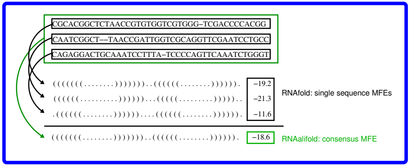
meanz

Approach

Reading direction of input alignment



Realigned reverse complement



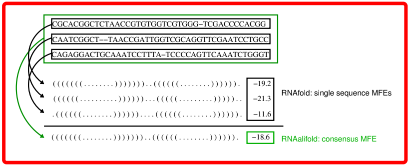
$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

meanz

Approach

Reading direction of input alignment



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

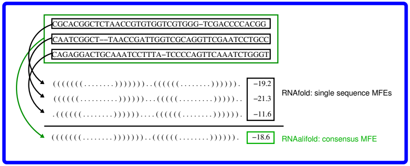
(((((((........)))))))..((((((........)))))). −19.2
((((((..........))))))...(((((.........))))).. −21.3
.((((((........))))))....((((((........)))))). −11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). −18.6

RNAalifold: consensus MFE

Realigned reverse complement



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))). −19.2
((((((..........))))))...(((((.........))))).. −21.3
.((((((........))))))....((((((........)))))). −11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). −18.6

RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

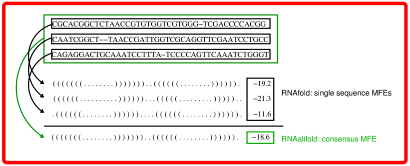
$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

meanz

meanmfe

Approach

Reading direction of input alignment



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

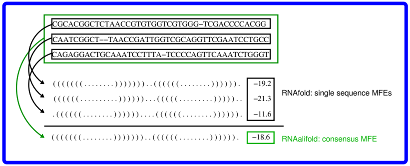
(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). –18.6

RNAalifold: consensus MFE

Realigned reverse complement



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). –18.6

RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

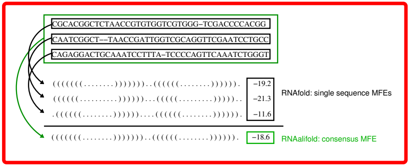
$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

meanz

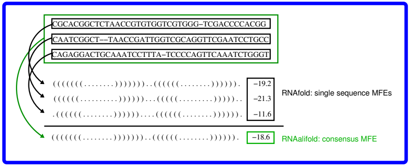
meanmfe

Approach

Reading direction of input alignment



Realigned reverse complement



$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

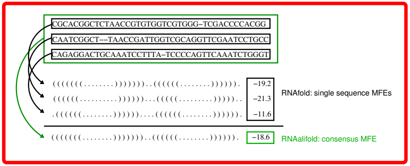
meanz

meanmfe

sci

Approach

Reading direction of input alignment



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

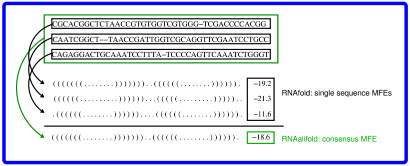
(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). –18.6

RNAalifold: consensus MFE

Realigned reverse complement



CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

(((((((........)))))))..((((((........)))))). –19.2
((((((..........))))))...(((((.........))))).. –21.3
.((((((........))))))....((((((........)))))). –11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). –18.6

RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

$$\text{consmfe} = \text{MFE of consensus sequence}$$

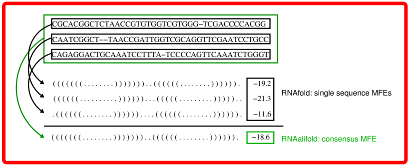
meanz

meanmfe

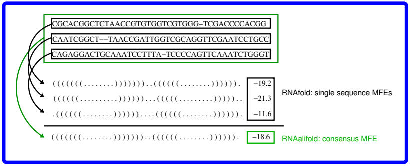
sci

Approach

Reading direction of input alignment



Realigned reverse complement



$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

$$\text{consmfe} = \text{MFE of consensus sequence}$$

meanz

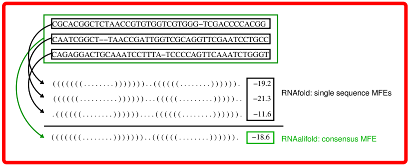
meanmfe

sci

consmfe

Approach

Reading direction of input alignment


CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

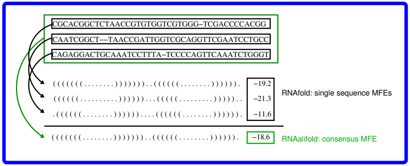
((((((((........)))))))..((((((........)))))). −19.2
((((((..........))))))...(((((.........))))).. −21.3
.((((((........))))))....((((((........)))))). −11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). −18.6

RNAalifold: consensus MFE

Realigned reverse complement


CGCACGGCTCTAACCGTGTGGTCGTGGG–TCGACCCCACGG
CAATCGGCT––TAACCGATTGGTCGCAGGTTCGAATCCTGCC
CAGAGGACTGCAAATCCTTTA–TCCCCAGTTCAAATCTGGGT

((((((((........)))))))..((((((........)))))). −19.2
((((((..........))))))...(((((.........))))).. −21.3
.((((((........))))))....((((((........)))))). −11.6

RNAfold: single sequence MFEs

(((((((........)))))))..((((((........)))))). −18.6

RNAalifold: consensus MFE

$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

$$\text{consmfe} = \text{MFE of consensus sequence}$$

$$\text{meanz} \longrightarrow \triangle \mathit{meanz}$$

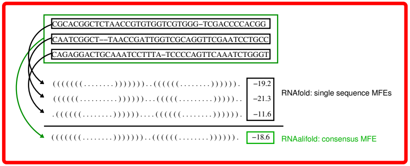
$$\text{meanmfe} \longrightarrow \triangle \mathit{meanmfe}$$

$$\text{sci} \longrightarrow \triangle \mathit{sci}$$

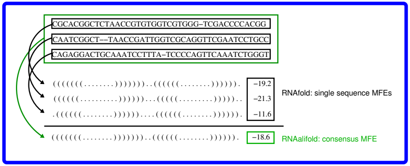
$$\text{consmfe} \longrightarrow \triangle \mathit{consmfe}$$

Approach

Reading direction of input alignment



Realigned reverse complement



$$\text{meanz} = \frac{\sum \text{single sequence z-score}}{\text{N}}$$

$$\text{meanmfe} = \frac{\sum \text{single sequence MFE}}{\text{N}}$$

$$\text{sci} = \frac{\text{consmfe}}{\text{meanmfe}}$$

$$\text{consmfe} = \text{MFE of consensus sequence}$$

$$\text{meanz} \rightarrow \triangle \mathit{meanz}$$

$$\text{meanmfe} \rightarrow \triangle \mathit{meanmfe}$$

$$\text{sci} \rightarrow \triangle \mathit{sci}$$

$$\text{consmfe} \rightarrow \triangle \mathit{consmfe}$$

Classification via support vector machine (SVM)

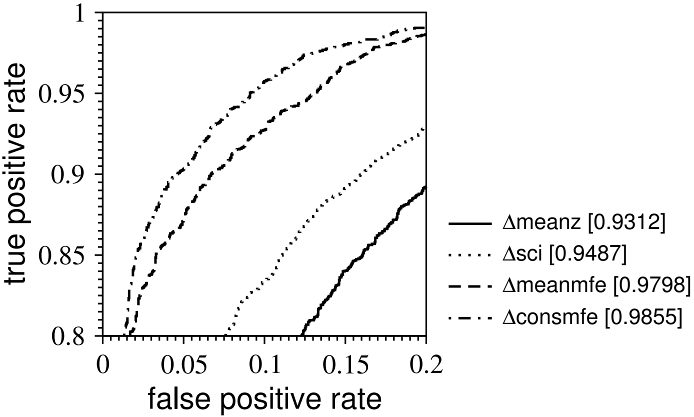
Which descriptors are the best?

Which descriptors are the best?

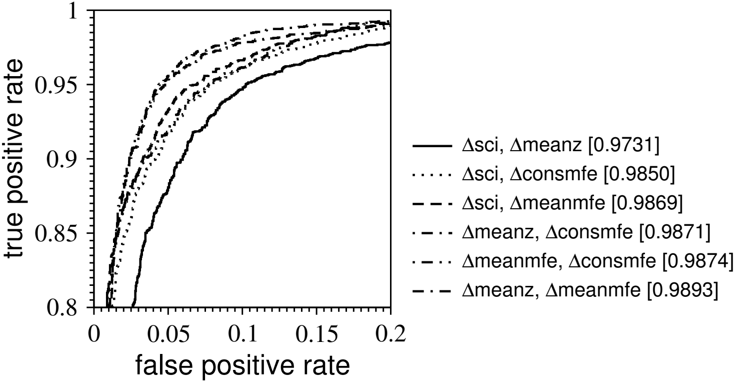
Receiver Operating Characteristics (5-fold cross validation):

Which descriptors are the best?

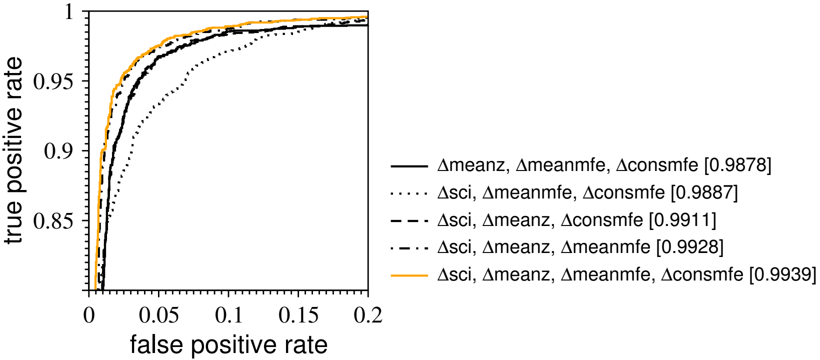
Receiver Operating Characteristics (5-fold cross validation):



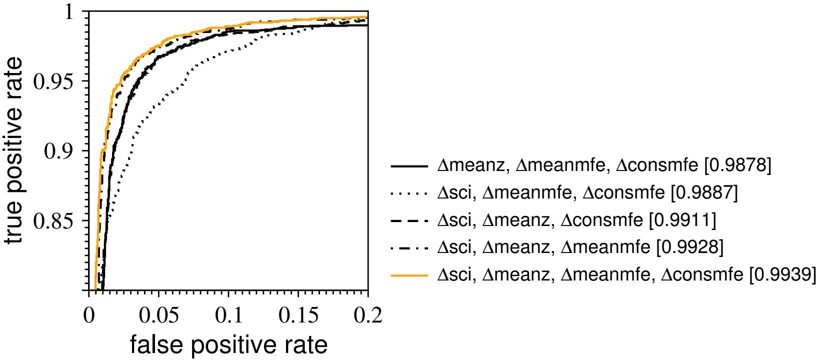
Combination of two



Combination of at least three



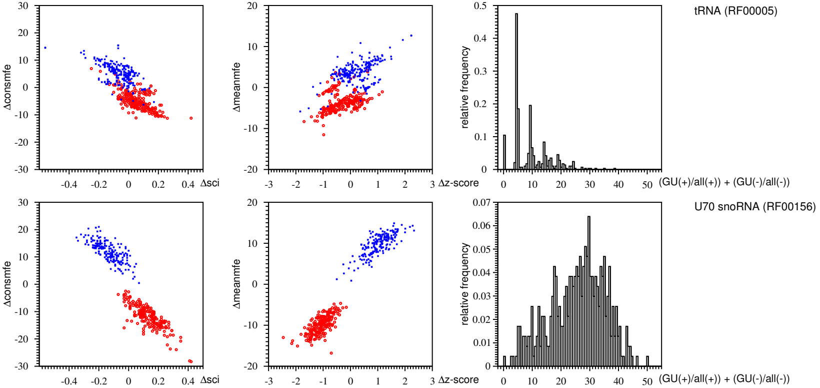
Combination of at least three



Maximal area under the curve (99.39%) if all four descriptors are taken

Additional descriptors

$\triangle$*meanmfe* , $\triangle$*consmfe* , $\triangle$*meanz* , $\triangle$*sci* depend on fraction of `GU` base pairs:



Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*
- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*
- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*
- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*

- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*

- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
    - $\triangle \mathit{meanz}$
    - $\triangle \mathit{meanmfe}$
- Differences in structure conservation
    - $\triangle \mathit{sci}$
    - $\triangle \mathit{consmfe}$

- $\frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}} + \frac{\text{GU pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*

Relevance of the strand differences are interpreted by:

- $\frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}} + \frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle$*meanz*
  - $\triangle$*meanmfe*
- Differences in structure conservation
  - $\triangle$*sci*
  - $\triangle$*consmfe*

Relevance of the strand differences are interpreted by:

- $\frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}} + \frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

Final set of descriptors

Strand differences are captured by:

- Differences in stability
  - $\triangle meanz$
  - $\triangle meanmfe$
- Differences in structure conservation
  - $\triangle sci$
  - $\triangle consmfe$

Relevance of the strand differences are interpreted by:

- $\frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}} + \frac{\texttt{GU}\text{ pairs in consensus}}{\text{all pairs in consensus}}$
- Average mean pairwise identity of alignment in both reading directions
- Number of sequences in alignment

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:
  $D > +c$: ncRNA in reading direction of input alignment
  $D < -c$: ncRNA is reverse complement of input alignment
  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

  $-c \leq D \leq +c$: No decision

SVM specifications

- SVM library `libsvm`
- Radial basis function kernel: $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$
- Attributes are scaled to -1 and 1
- Optimal parameters: penalty of error term $C = 128$, $\gamma = 0.5$
- Probability estimates P that alignment contains ncRNA in same reading direction
- `RNAstrand` score: $D = 2 * P - 1$, $D \in [-1, 1]$
- Different cutoffs c of score provide different prediction reliabilities:

  $D > +c$: ncRNA in reading direction of input alignment

  $D < -c$: ncRNA is reverse complement of input alignment

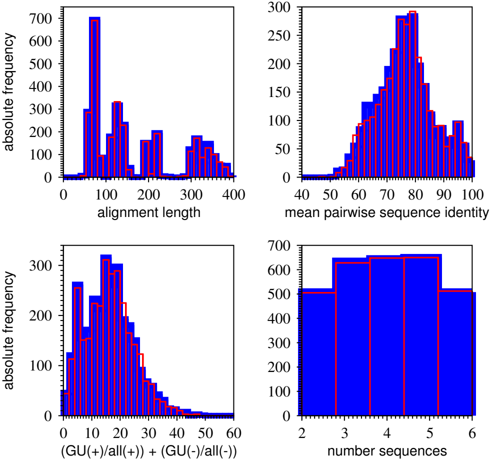
  $-c \leq D \leq +c$: No decision

Training of `RNAstrand`

5886 training alignments including representatives of rRNAs, snRNAs, snoRNAs, tRNAs, miRNAs, nuclear RNase P and SRP RNA

Training of `RNAstrand`

5886 training alignments including representatives of rRNAs, snRNAs, snoRNAs, tRNAs, miRNAs, nuclear RNase P and SRP RNA

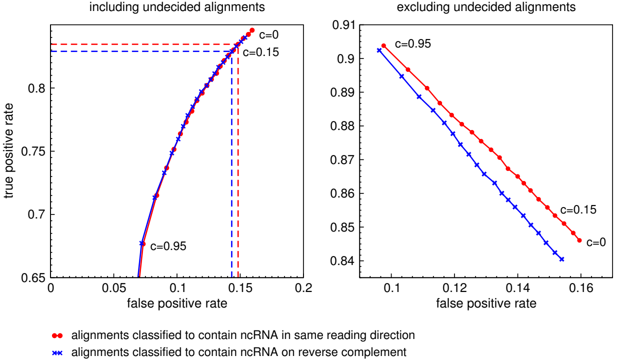


Validation of `RNAstrand`

Validate `RNAstrand` with 35766 automatically created `ClustalW` alignments of 313 non-coding RNA families found in `RFAM` 7.0

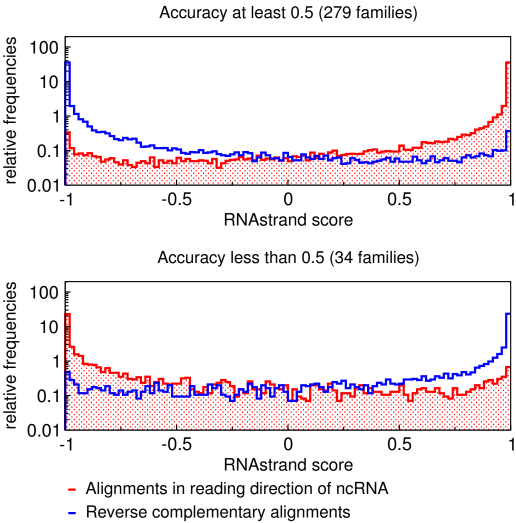
			$c = 0$			$c = 0.5$			$c = 0.9$	
ncRNA type	N	O	A_+	A_-	A	1-A-u	u	A	1-A-u	u
5S rRNA	860	12.6%	0.98	0.98	0.98	0.00	0.01	0.95	0.00	0.03
5.8S rRNA	146	-	0.93	0.93	0.89	0.05	0.05	0.73	0.02	0.24
tRNA	294	0.9%	0.94	0.94	0.88	0.01	0.09	0.62	0.00	0.36
miRNA	2496	4.5%	0.98	0.97	0.96	0.00	0.02	0.89	0.00	0.10
snoRNA (C/D)	204	9.9%	0.59	0.57	0.48	0.32	0.18	0.29	0.18	0.52
snoRNA (H/ACA)	1340	1.2%	0.98	0.98	0.97	0.01	0.01	0.94	0.00	0.05
spliceos. RNA	2878	35.0%	0.92	0.92	0.88	0.05	0.06	0.77	0.02	0.19
euk. SRP RNA	1000	8.1%	0.99	0.99	0.99	0.00	0.00	0.97	0.00	0.02
nucl. RNaseP	260	41.9%	0.93	0.93	0.92	0.04	0.03	0.85	0.01	0.12
RNase MRP	140	-	0.98	1.00	0.98	0.00	0.01	0.96	0.00	0.03
SECIS	76	-	0.65	0.64	0.51	0.25	0.22	0.32	0.19	0.48
7SK	184	-	0.04	0.03	0.02	0.91	0.05	0.01	0.80	0.18

Best cutoff *c*

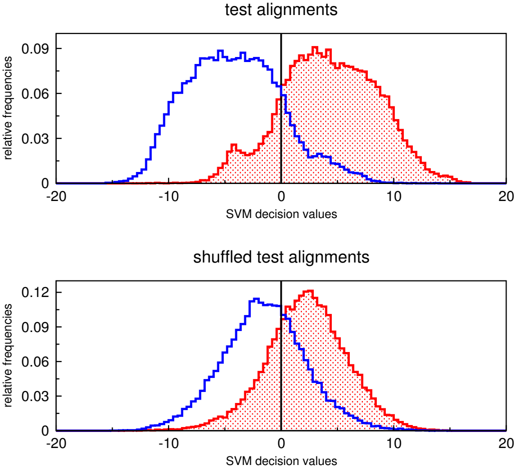


Maximal `Youden index` of 0.75 with cutoff $c = 0.15$

Distribution of `RNAstrand` scores



Distribution of SVM decision values



Are we better than naïve approach? (`RNAz`)

		Accuracy (`RNAstrand`: $c = 0$)	
ncRNA type	N	A(`RNAstrand`)	A(`RNAz`)
5S rRNA	860	0.98	0.97
5.8S rRNA	146	0.93	0.90
tRNA	294	0.94	0.53
miRNA	2496	0.97	0.14
snoRNA (C/D)	204	0.58	0.46
snoRNA (H/ACA)	1340	0.98	0.94
spliceos. RNA	2878	0.92	0.82
euk. SRP RNA	1000	0.99	0.84
nucl. RNaseP	260	0.93	0.82
RNase MRP	140	0.99	0.50
SECIS	76	0.65	0.48
7SK	184	0.04	0.03

Are we better than naïve approach? (RNAz)

RNAstrand	correct		incorrect	
correct	21536	21425	8711	8639
incorrect	1618	1729	3901	3973

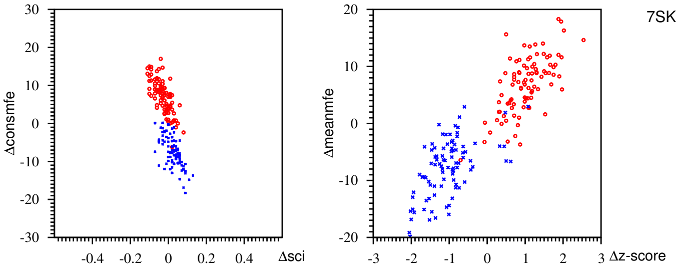
2-fold reduction of misclassification rate

Are we better than naïve approach? (EvoFold)

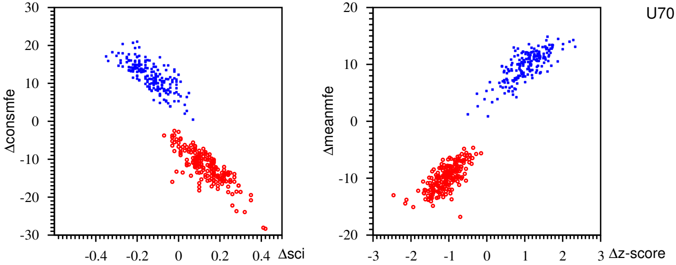
RNAstrand(fwd)	correct		incorrect	
correct	104	8	15	0
incorrect	16	0	7	2
RNAstrand(rev)	correct		incorrect	
correct	102	8	11	0
incorrect	18	0	11	2

Strand prediction of EvoFold comparable to RNAstrand .

Backup Slide - Classification of 7SK RNAs

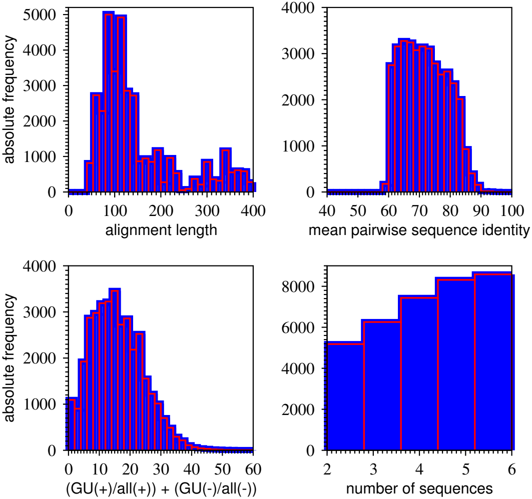


7SK (RF00100)



U70 snoRNA (RF00156)

Backup Slide - Test alignments



Backup Slide - Standard deviations

		$c = 0$		$c = 0.5$			$c = 0.9$		
ncRNA class	N	A_+	A_-	A	1-A-u	u	A	1-A-u	u
rRNA	2	0.02	0.02	0.04	0.02	0.01	0.11	0.01	0.10
miRNA	36	0.15	0.13	0.18	0.05	0.13	0.31	0.00	0.31
snoRNA (C/D)	23	0.38	0.40	0.38	0.36	0.15	0.37	0.28	0.35
snoRNA (H/ACA)	26	0.17	0.17	0.22	0.11	0.15	0.32	0.05	0.30
spliceos. RNA	6	0.24	0.25	0.29	0.20	0.08	0.29	0.12	0.17

Backup Slide - GU base pair fraction