Genomics of noncoding RNAs: status quo

- In human 97–98% of the transcriptional output is non-coding.
- High-throughput experimental approaches predict a large number of noncoding transcripts (30% of the genome is believed to be transcribed).
- The functions of these transcripts are unclear.
- Many evolutionary conserved non-coding elements await a functional interpretation.

“There is need for reliable experimental and computational methods for comprehensive identification of non-coding RNAs.”

Joint work with Stefan Washietl and Ivo Hofacker

Measuring thermodynamic stability of ncRNAs

- Naturally occurring structured RNAs have a lower folding energy compared to random sequences of the same size and base composition?

1. Calculate native MFE $m$.
2. Calculate mean $\mu$ and standard deviation $\sigma$ of MFEs of a large number of 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.
# z-scores of known ncRNAs

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

- Functional RNAs are clearly more stable than random sequences.
- However: The scores are too small to discriminate reliably in a genome-wide screens since the z-score distributions have heavy tails.
The structure conservation index

The SCI is an efficient and convenient measure for secondary structure conservation.
Efficient calculation of stability $z$-scores

- The mean $\mu$ and standard deviation $\sigma$ of random samples of a given sequence are functions of the length and the base composition:

$$\mu, \sigma(\text{length}, \frac{\text{GC}}{\text{AT}}, \frac{\text{G}}{\text{C}}, \frac{\text{A}}{\text{T}})$$

- It should therefore be possible to **calculate** $z$-scores by solving this 5 dimensional regression problem.
The regression problem is solved using a Support Vector Machine regression algorithm.

The SVM was trained on 10,000 synthetic sequences spaced evenly in the variable space.

The regression calculation is of the same accuracy as the sampling procedure.
Separation of native ncRNAs from random controls in two dimensions
Classification based on both scores
Classification based on both scores
Implementation and availability

- The approach is implemented in ANSI C in the program RNAz.
- The $z$-score regression is limited to 400 nucleotides.
- The classification model is currently limited to alignments of six sequences.
- At least an order of magnitude faster than other programs.
- RNAz is freely available:
  Download from www.tbi.univie.ac.at/~wash/RNAz
Comparison to other programs: Sensitivity/Specificity on RNAseP and SRP test sets

<table>
<thead>
<tr>
<th>Program</th>
<th>Number of sequences in alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>QRNA</td>
<td>27.8/98.5</td>
</tr>
<tr>
<td>ddbRNA</td>
<td>45.4/98.5</td>
</tr>
<tr>
<td>MSARi</td>
<td>—</td>
</tr>
<tr>
<td>RNAz</td>
<td>87.8/99.5</td>
</tr>
</tbody>
</table>
An integrated system for large scale analysis of complete genomes