RNA: applications of MFE
(Institute of Mathematical Sciences, National University of Singapore, 24 July 2007)

P. Clote
Biology and Computer Science
Boston College
Selenoprotein discovery

Kryukov, Kryukov, Gladyshev

Selenocysteine

- 21st amino acid
- Retranslation of stop codon UGA
- Selenocysteine insertion sequence (SECIS): stem-loop structure with specific nucleotide identities
- Selenoproteins in all kingdoms of life
- E. coli: 3 genes for selenoproteins
- M. jannaschii: 7 genes for selenoproteins
- Mammalian selenoproteins: 3 types of thioredoxin reductase, selenophosphate synthetase, etc.
Insufficient selenium levels associated with increased cancer incidence, decreased survival rate of HIV-infected patients, etc.
SECIS elements

- **>fdhA -16**
  - CGCCACCCUGCGAACCCGAAUUAAAUAAUAAAUUCAAGGGAGCAAGGUGGCG
  - length = 49
  - ((((((((((...(((((.................))).))).))))))).)))))))))
  - mfe = -20.53 kcal/mol

- **"FORMYL-MFR" +115**
  - AUGUUGAGGGGAACCCUGUAAGGGACCCUCCAACAU
  - length = 37
  - ((((((((((...((((........))).)))))))))))))))))))
  - mfe = -23.40 kcal/mol

- **>fruA +40**
  - CCUCGAGGGGAACCCGAAAGGGACCCGAGAGG
  - length = 32
  - ((((((.(((...((((((.....))).)))))))))))))))))))
  - mfe = -19.70 kcal/mol
SECIS element fruA of length 32 nt. Structure predicted by RNAfold of Vienna RNA package agrees with experimentally determined structure of A. Boeck (University of Munich).
PATSCAN filter

- Apply Overbeek's PATSCAN algorithm, and filter the dbEST using (essentially) the following PATSCAN search motif:

  - #Helix 2 allows for 1 mismatch, 1 insertion and 1 deletion
  - $r_1 = \{au, ua, gc, cg, gu, ug, ga, ag\}$
  - $p_1 = 4...4\ 2...6\ ATGA\ 1...1$ $p_2 = 9...12\ 0...3\ AAR\ 8...17$
  - $r_1 \sim p_2 [1, 1, 1]$
  - $BGA\ 1...1\ 4...9\ r_1 \sim p_1$
EMBL results (DNA) now contain the organism name.

1. **Choose the database to search against:** Protein: Swiss-Prot

2. **Select one (for DNA searches only):** Two DNA strand One DNA strand

3. **Enter the pattern to be searched for:**

4. **Allow overlapping hits?:** No Yes

5. **Maximum number of hits returned:** 2000

6. **Enter your email address:**

7. **Do you want us to store an HTML version of your output?** No Yes

   If you choose "Yes", an HTML document of your results will be saved on our server for four days and the URL for this document will be emailed to you. The HTML document uses the accession number (Swiss-Prot, EMBL and PDB) to link the results to the database entries.

Submit
Clear Form
Start over again

---

Screen shot of PATSCAN, pattern searching software written by Overbeek and Souza.
An example of SECISearch analysis of human ESTs

Primary sequence and secondary structure parameters in this search detected all selenoproteins with known SECIS elements in NR and dbEST. The free energy parameters in this search detected 12 out of 13 selenoproteins. SECIS elements in SelP did not satisfy the energy threshold for Helix I used in this search. The number of human ESTs that was selected by SECISearch with parameters sufficient to find 13 human selenoproteins was 6,055.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human ESTs in dbEST</td>
<td>1,253,123</td>
</tr>
<tr>
<td>Satisfied primary sequence consensus</td>
<td>179,614</td>
</tr>
<tr>
<td>ATGAN/11–12 nt/AA/18–27 nt/GA</td>
<td></td>
</tr>
<tr>
<td>Satisfied secondary structure consensus</td>
<td>32,652</td>
</tr>
<tr>
<td>Helix I, 7 nt, 1 mismatch and 1 bulge on each side are allowed;</td>
<td></td>
</tr>
<tr>
<td>internal loop, 3–7 nt, 5′ branch</td>
<td></td>
</tr>
<tr>
<td>ends with A; Quartet, TGAN...NGAN; Helix II, 11–12 nt, 3</td>
<td></td>
</tr>
<tr>
<td>mismatches and 1 bulge on each side are allowed; apical loop,</td>
<td></td>
</tr>
<tr>
<td>7–17 nt, starts with AA</td>
<td></td>
</tr>
<tr>
<td>Satisfied free energy threshold</td>
<td>974</td>
</tr>
<tr>
<td>ΔG for Helix I′ plus internal loop, &lt;7.4 kcal/mol;</td>
<td></td>
</tr>
<tr>
<td>ΔG for Helix II plus apical loop, &lt;11.0 kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Corresponded to known selenoprotein genes</td>
<td>678</td>
</tr>
<tr>
<td>Analyzed with MSGS</td>
<td>296</td>
</tr>
<tr>
<td>Corresponded to SelR gene</td>
<td>33</td>
</tr>
<tr>
<td>Corresponded to SelT gene</td>
<td>2</td>
</tr>
</tbody>
</table>

*Ten additional nucleotides upstream of the 5′ branch and 10 nucleotides downstream of the 3′ branch of predicted Helix I were considered in free energy estimation for Helix I plus the internal loop. This reflects the fact that, in most SECIS elements, Helix I is longer than the 7 nt assigned in the consensus secondary structure, and that Helix I could not be accurately predicted from that consensus.*

microRNA

- 22 nt. Mature miRNA processed from stem-loop precursor by DICER
- lin-4 and let-7 first identified in C. elegans
- Regulatory role (post-transcriptional regulation)
- Genes regulated concern developmental timing (lin-4, let-7), cell death, proliferation (bantam), cell death, fat storage (mir-14).
- Many-one and one-many relation of miRNA to mRNA target
RNAfold predicted stem-loop for precursor miRNA let-7 from C. elegans
Plant miRNA

- Precise or nearly precise complementary to mRNA target
- mRNA target usually in coding region
- miRNA directs cleavage and destruction of mRNA using RNA interference (RNAi)
Animal miRNA

- Imprecise complementary to mRNA target
- Critical base pairs and base pairing in positions 1-9 from 5’ end of miRNA
- mRNA target in 3’ UTR
- mRNA inhibits protein synthesis by unknown mechanism, which does not involve destruction of mRNA
Prediction of vertebrate miRNA genes

Lim et al.

Computational procedure

- Human/mouse BLAT alignments of noncoding regions
- Predict human stem-loop structures by RNAfold with 110 nt. window. Require at least 25 base pairs in stem and at most -25 kcal/mol folding energy. Obtain 800,000 hits.
- Retrieve homologous mouse regions using WU-BLAST, and fold using RNAfold
- Align using ClustalW and create consensus structure using Alidot
Score Alidot consensus structures using MiRscan, moving a 21 nt. window along conserved stems, using a log-likelihood score to measure similarity to 50 C. elegans experimentally validated miRNAs with their C. briggsae homologs

Obtain 15,651 human sequences in top 10% scores of MiRscan

For these human sequences, repeat procedure using Fugu genome
Results

- Obtain 188 human miRNA candidates with final MiRscan score above 10 (see figure)
- Previously 109 known human miRNA
- 107 new candidates of human miRNA
- Estimate of 255 miRNA in humans, hence 40 remain to be discovered
- miRNA genes represent 1% of all human genes
Fig. 1. Computational identification of vertebrate miRNA genes (6). The histogram represents the distribution of MiRscan scores for 15,133 human/Fugu consensus structures. Of the 109 reference-set loci, 91 were retained among these aligned segments (red), indicating that at least 80% of the human miRNAs are conserved in fish. The distribution peaks at the score of −4, with a count of 1198, but is truncated at a score of −4 and count of 200 to increase resolution at the high-scoring tail of the distribution. The 188 candidates with scores greater than 10.0 were examined further (expanded portion of the histogram): 81 were in the reference set of known loci (red), 14 were close paralogs of loci in the reference set (≤2 point substitutions within the miRNA) or represented cloned human miRNAs for which loci had not been previously reported (pink), and 38 were found in miRNA cDNA libraries made from zebrafish (purple) (6).

RNA switches

Giegerich, Haase, Rehmsmeier

RNA switches suspected or proved to be involved in processes:

- Regulation of gene expression in prokaryotes by attenuation
- Translation regulation of E. coli ribosomal protein S15
- Regulation of self-cleavage activity of Hepatitis Delta virus
- Translocation process in protein biosynthesis
- Splicing of pre-mRNA by spliceosomes
paRNAss

1. Sample structure space, by using mfold to obtain collection $S = \{s_1, \ldots, s_n\}$ of suboptimal structures (alternatively, use Sfold)
2. For all $s_i, s_j$ in $S$ compute pairwise distance using two distinct metrics $d_1, d_2$
3. Cluster to split $S$ into clusters $S_1, S_2$
4. Compute consensus structure for each cluster
Distance measures

The morphological distance $d_{MD}$ is a slightly modified version of a formula suggested by Zuker. Here structures are represented as sets of base pairs. $(i, j) \in s$ means that residues $i$ and $j$ form a base pair in $s$. For two sequences $s_1, s_2$ we define

$$d_{MD}(s_1, s_2) = \max\{d'_{MD}(s_1, s_2), d'_{MD}(s_2, s_1)\}, \text{ where}$$

$$d'_{MD}(s_1, s_2) = \sum_{(i_1, j_1) \in s_1} \min_{(i_2, j_2) \in s_2} \max\{|i_1 - i_2|, |j_1 - j_2|\}$$
The string edit distance $d_{SD}$ of two structures employs their string representation with dots and parantheses, e. g. (((...())), as used with the Vienna RNA package. We define

$$d_{SD}(s_1, s_2) = d_w(y_1, y_2)$$

where $y_i$ is the string representation of $s_i$, and $d_w$ is an edit distance (i.e. the score of an optimal alignment) on strings. Being defined via the edit distance model, $d_{SD}$ is a metric. This distance measure is provided with the Vienna RNA package.

Energy barrier distance

\[ d_{EB}(s_1, s_2) = \min \{ d'_{EB}(s_1, s_2), d'_{EB}(s_2, s_1) \}, \text{ where} \]
\[ d'_{EB}(s_1, s_2) = \min \{ e(p) | p \text{ is transition path from } s_1 \text{ to } s_2 \} \]
\[ e(p) = \max \{ e(s) - e(s_1) | s \text{ is intermediate structure in } p \} \]


Figure 1: Distance Plot showing a clear separation. Average energy barrier $d_{EB}$ between the two structure families is 20 kcal/mol; string distance $d_{SD}$ is about 10 edit operations.
Let $s_1 \cup s_2$ denote the union of two structures (i.e. base pair sets). Define

\[
pkDist(s_1, s_2) = \begin{cases} 
-1, & \text{if } s_1 \cup s_2 \text{ is planar,} \\
0, & \text{if } s_1 \cup s_2 \text{ contains a pseudoknot,} \\
k, & \text{if } k \text{ is the number of bases with conflicting pairings in } s_1 \cup s_2. 
\end{cases}
\]

(7) \quad (8) \quad (9)

paRNAss reports $pkDist(c_1, c_2)$ in addition to the above visualization.

Mono- and dinucleotide frequencies of classes of RNA

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Relative frequency</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.239922</td>
<td>0.07416</td>
<td>0.03895</td>
<td>0.07757</td>
<td>0.04822</td>
</tr>
<tr>
<td>C</td>
<td>0.253333</td>
<td>0.06195</td>
<td>0.07594</td>
<td>0.06135</td>
<td>0.05416</td>
</tr>
<tr>
<td>G</td>
<td>0.273618</td>
<td>0.08517</td>
<td>0.06274</td>
<td>0.08599</td>
<td>0.07173</td>
</tr>
<tr>
<td>U</td>
<td>0.231076</td>
<td>0.04865</td>
<td>0.07476</td>
<td>0.05670</td>
<td>0.05697</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Relative frequency</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.271796</td>
<td>0.06880</td>
<td>0.05457</td>
<td>0.06813</td>
<td>0.06231</td>
</tr>
<tr>
<td>C</td>
<td>0.247991</td>
<td>0.07495</td>
<td>0.06950</td>
<td>0.03441</td>
<td>0.06983</td>
</tr>
<tr>
<td>G</td>
<td>0.234521</td>
<td>0.06661</td>
<td>0.06317</td>
<td>0.05753</td>
<td>0.04719</td>
</tr>
<tr>
<td>U</td>
<td>0.245692</td>
<td>0.04415</td>
<td>0.06072</td>
<td>0.07446</td>
<td>0.06637</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Relative frequency</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.299366</td>
<td>0.09881</td>
<td>0.04586</td>
<td>0.05965</td>
<td>0.09401</td>
</tr>
<tr>
<td>C</td>
<td>0.185283</td>
<td>0.05654</td>
<td>0.04469</td>
<td>0.04333</td>
<td>0.04674</td>
</tr>
<tr>
<td>G</td>
<td>0.230897</td>
<td>0.06345</td>
<td>0.04698</td>
<td>0.06150</td>
<td>0.05897</td>
</tr>
<tr>
<td>U</td>
<td>0.284454</td>
<td>0.08558</td>
<td>0.04776</td>
<td>0.06643</td>
<td>0.08470</td>
</tr>
</tbody>
</table>

Figure 1: Upper Panel: Table of mononucleotide and dinucleotide frequencies, computed from 530 tRNAs from Sprinzl's collection. Middle Panel: Table of mononucleotide and dinucleotide frequencies, computed from 41 mRNAs previously investigated by [9, 14]. Lower Panel: Table of mononucleotide and dinucleotide frequencies, computed from 5 mRNAs considered in [14].

Generating random RNA using a walk on a Markov chain

Algorithm 1 (Markov) Input: An RNA sequence $a_1, \ldots, a_n$
Output: An RNA sequence $x_1, \ldots, x_n$ of the same expected dinucleotide frequency as $a_1, \ldots, a_n$.

1. compute the mono- and dinucleotide frequency of $a_1, \ldots, a_n$

2. generate $x_1$ by sampling from mononucleotide frequency

3. generate remaining nucleotides $x_2, \ldots, x_n$ by sampling from the conditional probabilities $Pr[X|Y]$, where $Pr[X|Y]$ equals the dinucleotide frequency that nucleotide $X$ follows $Y$ divided by mononucleotide frequency of nucleotide $Y$.

Dinucleotide shuffle algorithm

Algorithm 2 (Shuffle (Altschul-Erikson [1]))

**INPUT:** An RNA sequence $a_1, \ldots, a_n$

**OUTPUT:** An RNA sequence $x_1, \ldots, x_n$ of the same dinucleotide frequency as $a_1, \ldots, a_n$, where $x_1 = a_1$, $x_n = a_n$ (the Altschul-Erikson algorithm even produces the same number of dinucleotides of each type AA, AC, AG, AU, CA, CC, etc.).

1. For each nucleotide $x \in \{A, C, G, U\}$, create a list $L_x$ of edges $x \to y$ such that the dinucleotide $xy$ occurs in the input RNA.

2. For each nucleotide $x \in \{A, C, G, U\}$ distinct from the last nucleotide $x_n$, randomly choose an edge from the list $L_x$. Let $E$ be the set of chosen edges (note that $E$ contains at most three elements).

3. Let $G$ be the graph, whose edge set is $E$ and whose vertex set consists of those nucleotides $x, y$ such that $x \to y$ is an edge in $E$. If there is a vertex of $G$ which is not connected to the last nucleotide $a_n$, then return to (2).

4. For each nucleotide $x \in \{A, C, G, U\}$, permute the edges in $L_x - E$. Append to the end of each $L_x$ any edges from $E$ which had been removed.

5. for $i = 1$ to $n - 1$, generate $x_{i+1}$ by taking the next available nucleotide such that $x_i \to x_{i+1}$ belongs to the list $L_{x_i}$.

---

Z-scores of minimum free energy

- Z-score is \((x - \mu)/\sigma\) where \(\mu\) is mean minimum free energy of random RNA, and \(\sigma\) is standard deviation of minimum free energy of random RNA
Asymptotic limit \( \frac{\text{numBP}}{\text{seqLen}} \) for random RNA

<table>
<thead>
<tr>
<th>N</th>
<th>BP</th>
<th>StdDev</th>
<th>BP/N</th>
<th>Error</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.9600</td>
<td>0.8823</td>
<td>0.2960</td>
<td>0.0882</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>42.4200</td>
<td>2.2635</td>
<td>0.4242</td>
<td>0.0256</td>
<td>46</td>
<td>35</td>
</tr>
<tr>
<td>200</td>
<td>87.9100</td>
<td>2.3499</td>
<td>0.4398</td>
<td>0.0117</td>
<td>92</td>
<td>82</td>
</tr>
<tr>
<td>300</td>
<td>133.3600</td>
<td>2.6434</td>
<td>0.4468</td>
<td>0.0088</td>
<td>139</td>
<td>127</td>
</tr>
<tr>
<td>400</td>
<td>179.5300</td>
<td>3.1874</td>
<td>0.4488</td>
<td>0.0079</td>
<td>187</td>
<td>171</td>
</tr>
<tr>
<td>500</td>
<td>225.2100</td>
<td>2.9776</td>
<td>0.4504</td>
<td>0.0060</td>
<td>232</td>
<td>216</td>
</tr>
<tr>
<td>600</td>
<td>271.1300</td>
<td>3.1443</td>
<td>0.4519</td>
<td>0.0050</td>
<td>278</td>
<td>263</td>
</tr>
<tr>
<td>700</td>
<td>316.8800</td>
<td>3.3644</td>
<td>0.4527</td>
<td>0.0048</td>
<td>324</td>
<td>308</td>
</tr>
<tr>
<td>800</td>
<td>362.7400</td>
<td>4.0562</td>
<td>0.4534</td>
<td>0.0031</td>
<td>371</td>
<td>351</td>
</tr>
<tr>
<td>900</td>
<td>409.1800</td>
<td>3.8429</td>
<td>0.4546</td>
<td>0.0043</td>
<td>418</td>
<td>400</td>
</tr>
<tr>
<td>1000</td>
<td>455.0100</td>
<td>3.3439</td>
<td>0.4550</td>
<td>0.0036</td>
<td>463</td>
<td>443</td>
</tr>
</tbody>
</table>

Figure 11: Table of number \( BP \) of base pairs, ratio of base pairs to sequence length, etc. for random RNA sequences of length \( N \) generated by Algorithm 1 to have expected mononucleotide frequencies: \( q_A = q_C = q_G = q_U = 0.25 \). Our implementation of the Nussinov-Jacobson algorithm was used with Watson-Crick base pairs (no GU base pairs), threshold 0, and sequence length up to 1000. Average values were taken over 100 iterations, where error means Stddev/N; points are indicated along with error bars.

Clote, Kranakis, Krizanc, Stacho, to appear in *Discrete Applied Mathematics*
Asymptotic Z-scores

\[ X_{s,t} = \text{mfe}(x_s, \ldots, x_{t-1}), \text{ where mfe denotes minimum free energy as measured by Zuker’s algorithm. Then the limits} \]

\[
\lim_{n \to \infty} \frac{E[\text{mfe}(x_0, \ldots, x_n)]}{n} = \frac{E[X_{0,n}]}{n} = \mu(\bar{q}_{xy})
\]

and

\[
\lim_{n \to \infty} \sqrt{\frac{E[X_{0,n}^2]}{n^2} - (E[X_{0,n}])^2} = \sigma(\bar{q}_{xy})
\]

1. Compute minimum free energies for $m$ random RNAs, each of length $n$ nucleotides, as generated by Algorithm 3. In Figure 9, we used $m = 50$ and $n = 1000$.

2. Compute the mean and (sample) standard deviation for this collection, and divide both values by $n$ so as to normalize these values with respect to sequence length.

Since $m, n$ must be fixed for this computation, we denote the approximate mean by $\mu(\vec{q}_{xy}, m, n)$, and the approximate standard deviation by $\sigma(\vec{q}_{xy}, m, n)$. Thus, if $s_1, \ldots, s_m$ is a collection of $m$ random RNA sequences, each $s_i$ has length $n$ and is generated by Algorithm 3 from dinucleotide frequencies $\vec{q}_{xy}$, then

$$
\mu(\vec{q}_{xy}, m, n) = \frac{\sum_{k=1}^{m} \text{mfe}(s_i)/m}{n}
$$

$$
\sigma(\vec{q}_{xy}, m, n) = \sqrt{\frac{\sum_{k=1}^{m} \text{mfe}^2(s_i)}{m-1} - \left(\frac{\sum_{k=1}^{m} \text{mfe}(s_i)}{m}\right)^2 \cdot \frac{m}{m-1}}.
$$

Definition. Given RNA sequence $s$ of length $n_0$ (generally $n_0 \ll n$), compute the dinucleotide frequencies $q_{xy}$ of $s$. Define

$$Z^2_{m,n}(s) = \frac{mfe(s)/n_0 - \mu(q_{xy}, m, n)}{\sigma(q_{xy}, m, n)}$$
**Theorem 1 (Kingman)** Let $X_{s,t}$, for nonnegative integers $0 \leq s \leq t$, denote a family of doubly-indexed random variables which satisfy the following.

1. $X_{s,t} \leq X_{s,r} + X_{r,t}$ for all $s < r < t$.

2. The joint distribution of $X_{s,t}$ is the same as that of $X_{s+1,t+1}$ for all $0 \leq s \leq t$.

3. There exists $K < 0$ such that the expectation $E[X_{0,n}] = \mu_n$ exists and satisfies $\mu_n \geq K \cdot n$, for all natural numbers $n$.

Then there exists $\lambda$, for which $\lim_{n \to \infty} E[X_{0,n}]/n = \lambda$.

---

Theorem 2 (Clote et al.). Let $\mathcal{X}$ denote the infinite sequence of random variables $x_0, x_1, x_2, \ldots$ such that $x_0$ has the distribution $\vec{q}_x$, and for all $i$, $x_{i+1}$ has the distribution given by the conditional probabilities $\Pr[x_{i+1} = x] = \frac{q_{u,x}}{\Pr[x_i = u]}$. For all $0 \leq s \leq t$, define random variables $X_{s,t} = mfe(x_s, \ldots, x_{t-1})$, where mfe denotes minimum free energy as measured by Zuker’s algorithm. Then the limits

$$\lim_{n \to \infty} \frac{E[mfe(x_0, \ldots, x_n)]}{n} = E[X_{0,n}] = \frac{E[X_{0,n}]}{n} = \mu(\vec{q}_{xy})$$

and

$$\lim_{n \to \infty} \sqrt{\frac{E[X_{0,n}^2] - (E[X_{0,n}])^2}{n^2}} = \sigma(\vec{q}_{xy})$$

both exist and depend only on $\vec{q}_{xy}$. 

Proof Claim: The collection $X_{s,t}$ satisfies Kingman’s subadditive ergodicity Theorem 1.

From Zuker’s algorithm, RNA minimum free energy is subadditive, and hence condition (1) holds. For example:

$$mfe(ACGUACGUACGU) = -1.20$$
$$mfe(CAGUCCAUUUUGGG) = -0.90$$
$$mfe(ACGUACGUACGUCAGUCCAUUUUGGG) = -2.20$$

For condition (2), we show by induction that

\[ Pr[x_s = x] = Pr[x_0 = x] = q_x \]

holds for all \( s \) and for any given \( x \in \{ A, C, G, U \} \). Assume that \( Pr[x_s = x] = Pr[x_0 = x] = q_x \), and consider \( x_{s+1} \). Then

\[
Pr[x_{s+1} = x] = \sum_u Pr[x_s = u] \cdot Pr[x_{s+1} = x | x_s = u]
\]

\[
= \sum_u Pr[x_s = u] \cdot \frac{Pr[x_s = u, x_{s+1} = x]}{Pr[x_s = u]}
\]

\[
= \sum_u Pr[x_s = u, x_{s+1} = x]
\]

\[
= q_x
\]
We now establish condition (3) of Kingman’s theorem. Let $K_0$ be the minimum value, $-3.42$ kcal/mol, over all base stacking free energies from Turner’s current rules The nearest neighbor energy model with Turner’s experimentally measured energies is additive and there are at most $n/2$ base pairs in an RNA sequence of length $n + 1$ (going from $0$ to $n$), hence $K_0 \cdot n/2 \leq \mu_n$ for all $n$. It follows that condition (3) of Kingman’s theorem holds, so the existence of $\lim_{n \to \infty} \frac{E[mfe(x_0, \ldots, x_n)]}{n} = \mu(\vec{q}_{xy})$ follows.

Hence $Pr[x_s = u] = q_u$, for all $s$ and $u \in \{A, C, G, U\}$. Since the sequence $x_0, x_1, x_2, \ldots$ of random variables follows a first order Markov condition,

\[
Pr[x_{s+1} = y|x_s = x] = Pr[x_{s'} + 1 = y|x_s = x]
\]

so

\[
Pr[x_s = a_0, \ldots, x_{s+n} = a_n] = Pr[x_{s'} = a_0, \ldots, x_{s'+n} = a_n]
\]

Hence $X_{s,t}$ has the same joint distribution as that of $X_{s+1,t+1}$, for all $0 \leq s \leq t$. Thus condition (2) of Kingman’s theorem is satisfied.

Figure adapted from “A comprehensive comparison of comparative RNA structure prediction approaches”, Gardner, Giegerich. BMC Bioinformatics 5:140 (2004)
MSARi

  - Use multiple sequence alignment of orthologous genomic regions (7-12 species -- e.g. yeast genomes). Require diversity to see co varying base pairs.
  - Using covariation, determine base pairs from multiple sequence alignment of genomes. If a regional MSA has high score, indicating that RNA sequence displays similar secondary structure, then return TRUE.
PROBLEM: Automatic RNA MSA using (e.g.) ClustalW is error prone because alignment based on sequence and not secondary structure.
**SOLUTION:**

- Fix window of size 5 in MSA starting at positions i,j
- Determine all maximal subwindows (shift left/right) consistent with helix formation -- this approach corrects for small misalignments in MSA
- Determine probability for two windows to be reverse complementary.
- Give priority to windows having large measure of reverse complementarity such that |i-j| small (helical regions near loop region in hairpin loop).
- Greedy algorithm assembles secondary structure and scores MSA.
- For MSA of (say) bacterial genomes 300 nt. regions of high score indicate ncRNA gene.
• $X_i \parallel Y_j$ is event that positions $i,j$ co-vary

• $\Pr[ X_i \parallel Y_j ] = \sum \Pr[ X_i = x ] \cdot \Pr[ Y_j = y ]$
  where sum is over all Watson-Crick and GU base pairs $(x,y)$

• $\Pr[5\text{-tuple}] = \prod \Pr[X_i \parallel Y_{4-i}]$
  with product over $i=0,\ldots,4$

• $\Pr[4\text{-tuple}] = \sum \prod \Pr[X_{i+s} \parallel Y_{4-i+t}]$
  with sum over $s,t$ in $\{0,1\}$ and product over $i=0,\ldots,4$
Define $q(i,j)$ as product of tuple probabilities for maximum tuple length over each row.

Define score of MSA. Score of homologous ncRNA much higher than that of MSA obtained by random permutation of columns.
\[ q(i,j) = \text{Pr}[5\text{-tuple}] \times \text{Pr}[4\text{-tuple}]^2 \times \text{Pr}[3\text{-tuple}]^2 \]

RNAalifold

Hofacker, Fekete, Stadler, Washietl


Mutual information

\[ M_{i,j} = \sum_{X,Y} f_{i,j}(X,Y) \log \frac{f_{i,j}(X,Y)}{f_i(X)f_j(Y)} \]

Note: If column \( i \) consists only of \( G \) and column \( j \) consists only of \( G \) and \( U \), then \( M_{i,j} = 0 \). Indeed, \( f_i(G) = 1 \), \( f_{i,j}(G,C) = f_j(C) \) and \( f_{i,j}(G,U) = f_j(U) \), so expression in log is 1.

Alternative measure of covariance

In Nussinov-Jacobson model (maximum circular matching), for multiple alignment $A$ of RNAs, let

$$\Pi_{i,j}^\alpha = \begin{cases} 
1 & \text{if positions } i, j \text{ can base pair} \\
0 & \text{else}
\end{cases}$$

$$d_{i,j}^{\alpha,\beta} = 2 - \delta(a_i^\alpha, a_j^\beta) - \delta(a_j^\alpha, a_i^\beta)$$

$$C_{i,j}^\alpha = \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} d_{i,j}^{\alpha,\beta} \cdot \Pi_{i,j}^\alpha \cdot \Pi_{i,j}^\beta$$

$$C_{i,j} = \sum_{X,Y,X',Y'} f_{i,j}(X,Y)d_H(XY,X'Y')f_{i,j}(X',Y')$$

Let $q_{i,j}$ be the number of inconsistent base pairs in alignment $A$ at positions $i,j$ (gap on one side counted as inconsistent).

\[
B_{i,j} = C_{i,j} - \gamma_1 q_{i,j}
\]

\[
\Pi_{i,j}^A = \begin{dcases} 
1 & \text{if } B_{i,j} \geq B^* \\
0 & \text{if } B_{i,j} < B^* 
\end{dcases}
\]

\[
E_{i,j}^A = \frac{1}{N} \sum_{\alpha} E(a_i^\alpha, a_j^\beta) - \gamma_2 B_{i,j}
\]

Threshold $B^* = -1$, coefficients $\gamma_1 = 1 = \gamma_2$. Contribution of compensatory mutation is approximately same as adding one base pair to stem.

Minimum consensus free energy structure $S$ and associated minimum consensus free energy $E^A$ for alignment $A$ obtained by dynamic programming:

$$E^A = \min_S \sum_{(i,j) \in S} E_{i,j}$$

Time is $O(N \cdot n^2 + n^3)$ and space is $O(n^2)$. Can fill consensus structure for an individual sequence by using RNAdFold -C.


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

<table>
<thead>
<tr>
<th>ncRNA type</th>
<th>Single sequence</th>
<th>Number of sequences in alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Z_{mono}</td>
</tr>
<tr>
<td>tRNA</td>
<td>579</td>
<td>−1.84</td>
</tr>
<tr>
<td>SS rRNA</td>
<td>606</td>
<td>−1.62</td>
</tr>
<tr>
<td>Hammerh. III</td>
<td>251</td>
<td>−3.08</td>
</tr>
<tr>
<td>Gr. II 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</td>
<td>199</td>
<td>−2.73</td>
</tr>
</tbody>
</table>

n, number of sequences/alignments scored; ID, average mean pairwise identity; Z, average z-score; S, sensitivity (% below −4).
Table 2. The z-scores of ncRNAs in C. elegans aligned to homologs of C. briggsae

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

Detecting ncRNA genes

Washietl, Hofacker, Stadler

“Fast and reliable prediction of noncoding RNAs”, Washietl, Hofacker and Stadler, PNAS 102(7): 2454-2459 (2005)
Algorithm RNAZ

- **INPUT:** alignment A of RNAs, as produced by ClustalW or from Comparative Regulatory Genomics (CORG) database for functional RNAs
- **OUTPUT:** Score between 0 and 1, measuring likelihood that alignment contains structural RNAs

- **IDEA:** Combine average Z-scores of RNAs in alignment, computed using shuffling, together with a measure of commonality of secondary structure and covariation of base pairs
- Using RNAALIFOLD, compute energy $E_A$ of alignment $A$ using dynamic programming secondary structure prediction together with covariance term for compensatory and consistent mutations.
- Using RNAfold, compute average minimum free energy $<E>$ of RNAs in alignment $A$.
- $SCI = E_A / <E>$
- $Z =$ average Z-scores of RNAs in alignment $A$
Fig. 2. Classification based on z scores and SCI by using a SVM. Alignments of tRNAs and 5S rRNAs with two to four sequences per alignment and mean pairwise identities between 60% and 90% are shown. Green circles represent native alignments, and red crosses represent shuffled random controls. The background color ranging from red to green indicates the RNA class probability for different regions of the z-SCI plane.

“Fast and reliable prediction of noncoding RNAs”, Washietl, Hofacker and Stadler, PNAS 102(7): 2454-2459 (2005)
Using LIBSVM, train support vector machine (SVM) as binary classifier for structural RNA, given feature vectors:
- SCI
- Z
  - number of RNAs in alignment A
  - mean pairwise sequence identity of RNAs in A

In training set, positive (+1) training examples are alignments of structural RNA from same class in Rfam, negative (-1) training examples obtained by shuffling columns and applying ClustalW

SVM outputs score P. Compute sensitivity and specificity using threshold P
Novel procedure to compute Z-scores

- Generate 10,648 random sequences of lengths ranging from 50 to 400 nt. in increments of 50 nt. with base composition GC/AU, A/U, G/C ratios from 0.25 to 0.75 in increments of 0.05
- Using RNAfold, compute Z-scores of random sequences
- Using LIBSVM, train support vector machine (SVM) to determine 2 regression models: model for $\mu$ MFE, model for $\sigma$ MFE.
- SVM uses $\nu$ variant of regression and uses radial basis kernel.
“Fast and reliable prediction of noncoding RNAs”, Washietl, Hofacker and Stadler, PNAS 102(7): 2454-2459 (2005)
Results

- For cutoff value of 0.9 for P,
  - sensitivity 75.27%
  - Specificity 98.93%
- Sensitivity and specificity of 100% on MSARI test set with simple rule
  - SCI > 0.3 and Z < -1.5
- RNAZ used 2 aligned RNAs, rather than 14 aligned RNAs used in MSARI
Fast and reliable prediction of noncoding RNAs, Washietl, Hofacker and Stadler, PNAS 102(7): 2454-2459 (2005)

<table>
<thead>
<tr>
<th>ncRNA type</th>
<th>$N$</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S ribosomal RNA</td>
<td>297</td>
<td>81.48 (242)</td>
<td>96.63 (10)</td>
<td>68.69 (204)</td>
<td>99.33 (2)</td>
<td>33.00 (98)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td>tRNA</td>
<td>329</td>
<td>94.83 (312)</td>
<td>93.62 (21)</td>
<td>90.27 (297)</td>
<td>97.87 (7)</td>
<td>75.68 (249)</td>
<td>99.70 (1)</td>
</tr>
<tr>
<td>SRP RNA</td>
<td>464</td>
<td>100.00 (464)</td>
<td>96.55 (16)</td>
<td>96.55 (448)</td>
<td>98.92 (5)</td>
<td>66.16 (307)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td>RNAse P</td>
<td>291</td>
<td>98.97 (288)</td>
<td>96.22 (11)</td>
<td>84.19 (245)</td>
<td>99.31 (2)</td>
<td>56.70 (165)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td>U2 splicingosomal RNA</td>
<td>351</td>
<td>98.58 (346)</td>
<td>97.72 (8)</td>
<td>95.44 (335)</td>
<td>99.15 (3)</td>
<td>66.67 (234)</td>
<td>99.72 (1)</td>
</tr>
<tr>
<td>U5 splicingosomal RNA</td>
<td>285</td>
<td>91.58 (261)</td>
<td>98.25 (5)</td>
<td>81.75 (233)</td>
<td>100.00 (0)</td>
<td>70.53 (201)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td>U3 snoRNA</td>
<td>277</td>
<td>83.75 (232)</td>
<td>98.56 (4)</td>
<td>62.82 (174)</td>
<td>99.28 (2)</td>
<td>44.40 (123)</td>
<td>99.64 (1)</td>
</tr>
<tr>
<td>U70 snoRNA</td>
<td>363</td>
<td>61.16 (222)</td>
<td>96.69 (12)</td>
<td>35.54 (129)</td>
<td>98.90 (4)</td>
<td>17.91 (65)</td>
<td>99.72 (1)</td>
</tr>
<tr>
<td>Hammerhead III ribozyme</td>
<td>271</td>
<td>100.00 (271)</td>
<td>95.20 (13)</td>
<td>98.15 (266)</td>
<td>98.89 (3)</td>
<td>89.67 (243)</td>
<td>99.26 (2)</td>
</tr>
<tr>
<td>Group II catalytic intron</td>
<td>407</td>
<td>78.62 (320)</td>
<td>96.31 (15)</td>
<td>76.90 (313)</td>
<td>98.53 (6)</td>
<td>25.31 (103)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td>tmRNA</td>
<td>386</td>
<td>24.87 (96)</td>
<td>96.37 (14)</td>
<td>18.65 (72)</td>
<td>98.19 (7)</td>
<td>8.55 (33)</td>
<td>99.48 (2)</td>
</tr>
<tr>
<td>MicroRNA mir-10</td>
<td>380</td>
<td>100.00 (380)</td>
<td>95.26 (18)</td>
<td>97.63 (371)</td>
<td>99.21 (3)</td>
<td>62.37 (237)</td>
<td>100.00 (0)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>4,101</td>
<td>84.17 (3,452)</td>
<td>96.42 (147)</td>
<td>75.27 (3,087)</td>
<td>98.93 (44)</td>
<td>50.18 (2,058)</td>
<td>99.80 (8)</td>
</tr>
</tbody>
</table>

Results for alignments with two to four sequences and mean pairwise identities between 60% and 100% are shown. $N$ is the number of alignments in the test set. For each native alignment, one randomized alignment was produced, and randomized alignments classified as ncRNA were counted as false positives. Sensitivity and specificity are shown in percentage for three cutoffs of the RNA class probability predicted by the SVM. Absolute numbers of true positives and
Acknowledgements

For discussions and sharing of their papers, I’m indebted to the following and others not listed due to lack of space:

- Markus Bauer
- Ye Ding
- Sean Eddy
- Ivo Hofacker
- Gunnar Klaue
- David Mathews
- Elena Rivas
- Peter Stadler
- Martin Vingron
- Jerome Waldispuhl
- Eric Westhof
- Michael Zuker
Collaborators

- Behshad Behzadi, Google Zurich
- Eva Freyhult, Linnaeus Centre for Bioinformatics, Uppsala
- Vince Moulton, CS, Univ East Anglia
- Fabrizio Ferre, Harvard & Children’s Hospital
- Andy Lorenz, BC
- Yann Ponty, BC
- Jean-Marc Steyaert, Ecole Polytechnique
- Jerome Waldispuhl, MIT

- NSF DBI-0543506
bioinformatics.bc.edu/clotelab/
THANKS!