Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy  
Zasha Weinberg & W.L. Ruzzo  
Recomb ’04

Covariance Model
Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.

Rfam

IRE (partial seed alignment):

- Input (hand-tuned):
  - MSA
  - SS_cons
  - Score Thresh T
  - Window Len W
- Output:
  - CM
  - scan results

CM’s are good, but slow
Rfam Reality  
Our Work  
Rfam Goal

1 month, 1000 computers  
~2 months, 1000 computers  
10 years, 1000 computers
Oversimplified CM
(for pedagogical purposes only)

CM to HMM

25 emissions per state
5 emissions per state, 2x states

Key Issue: 25 scores \(\rightarrow\) 10

Viterbi/Forward Scoring

- Path \(\pi\) defines transitions/emissions
- Score(\(\pi\)) = product of “probabilities” on \(\pi\)
- NB: ok if “probs” aren’t, e.g. \(\sum \neq 1\)
- E.g. in CM, emissions are odds ratios vs 0th-order background
- For any nucleotide sequence \(x\):
  - Viterbi-score(\(x\)) = \(\max\{\text{score(\(\pi\))} \mid \pi \text{ emits } x\}\)
  - Forward-score(\(x\)) = \(\sum\{\text{score(\(\pi\))} \mid \pi \text{ emits } x\}\)

• Need: log Viterbi scores CM \(\leq\) HMM
**Key Issue: 25 scores → 10**

- Need: log Viterbi scores $\text{CM} \leq \text{HMM}$

  \[
  \begin{align*}
  P_{AA} &\leq L_A + R_A \\
  P_{AC} &\leq L_A + R_C \\
  P_{AG} &\leq L_A + R_G \\
  P_{AU} &\leq L_A + R_U \\
  P_{Au} &\leq L_A + R_\text{u} \\
  \end{align*}
  \]

  \[
  \begin{align*}
  P_{CA} &\leq L_C + R_A \\
  P_{CC} &\leq L_C + R_C \\
  P_{CG} &\leq L_C + R_G \\
  P_{CU} &\leq L_C + R_U \\
  P_{C\text{u}} &\leq L_C + R_\text{u} \\
  \end{align*}
  \]

  (NB: HMM not a prob. model)

**Rigorous Filtering**

- Any scores satisfying the linear inequalities give rigorous filtering

  \[
  \begin{align*}
  \text{CM} \text{ Viterbi path score} &\leq \text{“corresponding” HMM path score} \\
  &\leq \text{Viterbi HMM path score}
  \end{align*}
  \]

  (even if it does not correspond to any CM path)

**Some scores filter better**

- $P_{UA} = 1 \leq L_U + R_A$
- $P_{UG} = 4 \leq L_U + R_G$

  **Option 1:**
  \[
  \begin{align*}
  L_U &= R_A = R_G = 2 \\
  \text{Opt 1:} \quad L_U + (R_A + R_G)/2 &= 4
  \end{align*}
  \]

  **Option 2:**
  \[
  \begin{align*}
  L_U &= 0, R_A = 1, R_G = 4 \\
  \text{Opt 2:} \quad L_U + (R_A + R_G)/2 &= 2.5
  \end{align*}
  \]

  Assuming ACGU = 25%

**Optimizing filtering**

- For any nucleotide sequence $x$:

  \[
  \begin{align*}
  \text{Viterbi-score}(x) &= \max\{ \text{score}(\pi) | \pi \text{ emits } x \} \\
  \text{Forward-score}(x) &= \sum \{ \text{score}(\pi) | \pi \text{ emits } x \}
  \end{align*}
  \]

  \[
  \begin{align*}
  E(L_i, R_i) &= \sum_x \text{Forward-score}(x) \cdot \text{Pr}(x) \\
  \text{– NB: } E \text{ is a function of } L_i, R_i \text{ only}
  \end{align*}
  \]

  **Expected Forward Score**

  **Under 0th-order background model**

  \[
  \begin{align*}
  \text{Optimization:} \\
  \text{Minimize } E(L_i, R_i) \text{ subject to score L.I.s} \\
  \text{– This is heuristic (”forward down ⇒ Viterbi down ⇒ filter down”)}
  \end{align*}
  \]

  \[
  \begin{align*}
  \text{– But still rigorous because ”subject to score L.I.s”}
  \end{align*}
  \]
Calculating $E(L_i, R_i)$

$E(L_i, R_i) = \sum_x \text{Forward-score}(x) \cdot \text{Pr}(x)$

- Forward-like: for every state, calculate expected score for all paths ending there, easily calculated from expected scores of predecessors & transition/emission probabilities/scores.

Minimizing $E(L_i, R_i)$

- Calculate $E(L_i, R_i)$ symbolically, in terms of emission scores, so we can do partial derivatives for numerical convex optimization algorithm:

$$\frac{\partial E(L_1, L_2, \ldots)}{\partial L_i}$$

Estimated Filtering Efficiency

(139 Rfam 4.0 families)

<table>
<thead>
<tr>
<th>Filtering fraction</th>
<th># families (compact)</th>
<th># families (expanded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 10^{-4}</td>
<td>105</td>
<td>110</td>
</tr>
<tr>
<td>10^{-4} - 10^{-2}</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>.01 - .10</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>.10 - .25</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>.25 - .99</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>.99 - 1.0</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

Results: buried treasures

<table>
<thead>
<tr>
<th>Name</th>
<th># found BLAST + CM</th>
<th># found rigorous filter + CM</th>
<th># new</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrococcus snoRNA</td>
<td>57</td>
<td>180</td>
<td>123</td>
</tr>
<tr>
<td>Iron response element</td>
<td>201</td>
<td>322</td>
<td>121</td>
</tr>
<tr>
<td>Histone 3' element</td>
<td>1004</td>
<td>1106</td>
<td>102</td>
</tr>
<tr>
<td>Purine riboswitch</td>
<td>69</td>
<td>123</td>
<td>54</td>
</tr>
<tr>
<td>Retron mar</td>
<td>11</td>
<td>69</td>
<td>48</td>
</tr>
<tr>
<td>Hammerhead I</td>
<td>167</td>
<td>193</td>
<td>26</td>
</tr>
<tr>
<td>Hammerhead III</td>
<td>251</td>
<td>264</td>
<td>13</td>
</tr>
<tr>
<td>U4 snRNA</td>
<td>283</td>
<td>290</td>
<td>7</td>
</tr>
<tr>
<td>S-box</td>
<td>128</td>
<td>131</td>
<td>3</td>
</tr>
<tr>
<td>U6 snRNA</td>
<td>1462</td>
<td>1484</td>
<td>2</td>
</tr>
<tr>
<td>U5 snRNA</td>
<td>199</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>U7 snRNA</td>
<td>312</td>
<td>313</td>
<td>1</td>
</tr>
</tbody>
</table>
## Results: With additional work

<table>
<thead>
<tr>
<th></th>
<th># with BLAST+CM</th>
<th># with rigorous filter series + CM</th>
<th># new</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfam tRNA</td>
<td>58609</td>
<td>63767</td>
<td>5158</td>
</tr>
<tr>
<td>Group II intron</td>
<td>5708</td>
<td>6039</td>
<td>331</td>
</tr>
<tr>
<td>tRNAscan-SE (human)</td>
<td>608</td>
<td>729</td>
<td>121</td>
</tr>
<tr>
<td>tmRNA</td>
<td>226</td>
<td>247</td>
<td>21</td>
</tr>
<tr>
<td>Lysine riboswitch</td>
<td>60</td>
<td>71</td>
<td>11</td>
</tr>
</tbody>
</table>

And more…