BMI/CS 776
Lecture #23 -
Applications of Lightweight Stochastic Context Free Grammars for RNA Analysis

Colin Dewey
April 15, 2008
Recall Nussinov

- Finds structure with maximum number of base pairs
- Recursion has four cases:

SCFG:

\[ S \rightarrow aS \mid cS \mid gS \mid uS \]

\[ S \rightarrow aS \mid cS \mid gS \mid uS \]

\[ S \rightarrow Sa \mid Sc \mid Sg \mid Su \]

\[ S \rightarrow SS \]
Nussinov Algorithm

- let \( \delta(i, j) = \begin{cases} 
1 & \text{if } x_i \text{ and } x_j \text{ are complementary} \\
0 & \text{otherwise}
\end{cases} \)

- initialization: \( \gamma(i, i-1) = 0 \) for \( i = 2 \) to \( L \)
  \( \gamma(i, i) = 0 \) for \( i = 1 \) to \( L \)

- recursion
  \( \gamma(i, j) = \max \begin{cases} 
\gamma(i+1, j) \\
\gamma(i, j-1) \\
\gamma(i+1, j-1) + \delta(i, j) \\
\max_{i<k<j}[\gamma(i, k) + \gamma(k+1, j)]
\end{cases} \)

max # of paired bases in Subsequence \([i, j]\)
Nussinov algorithm $\rightarrow$ CYK

- initialization:
  $$\gamma(i, i - 1) = -\infty \quad \text{for } i = 2 \text{ to } L$$
  $$\gamma(i, i) = \max(\log p(x_i S), \log p(S x_i)) \quad \text{for } i = 1 \text{ to } L$$

- recursion
  $$\gamma(i, j) = \max \left\{ \begin{array}{l}
  \gamma(i + 1, j) + \log p(x_i S) \\
  \gamma(i, j - 1) + \log p(S x_j) \\
  \gamma(i + 1, j - 1) + \log p(x_i S x_j) \\
  \max_{i < k < j} \gamma(i, k) + \gamma(k + 1, j) + \log p(SS) \end{array} \right\}$$

  log probability of most likely structure of subsequence $[i,j]$
Searching Sequence for a Secondary Structure

Given

- a single RNA sequence with its secondary structure
- another RNA query sequence

ACGGCUUCGGCCUUGGCGAGACCC

Determine if the query sequence has “same” secondary structure
Searching Sequence for a Secondary Structure

- this is analogous to pairwise alignment with primary sequences
- we take into account substitutions, insertions/deletions, and base-pair substitutions

ACG GCUU C CGG CCU U GG C G A G A C C

Diagram:

\[
\begin{align*}
\text{A} & \text{G} & \text{A} & \text{G} & \text{U} & \text{C} & \text{A} & \text{G} & \text{G} & \text{G} & \text{G} & \text{C} & \text{C} & \text{C} & \text{C} & \text{A} & \text{C} \\
\text{U} & \text{C} & \text{G} & \text{G} & \text{G} & \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{G} & \text{G} & \text{C} & \text{A} & \text{G} \\
\text{U} & \text{G} & \text{C} & \text{G} & \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{G} & \text{G} & \text{C} & \text{A} & \text{G} & \text{C} \\
\text{G} & \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{G} & \text{G} & \text{C} & \text{A} & \text{G} & \text{C} & \text{G} & \text{A} & \text{G} \\
\end{align*}
\]
The RIBOSUM Matrices [Klein & Eddy]

observed frequency of \( i \) aligned to \( j \) in homologous RNAs

\[
s_{ij} = \log_2 \frac{f_{ij}}{g_i g_j}
\]

background frequency of \( i \)

\[
s'_{ijkl} = \log_2 \frac{f'_{ijkl}}{g_i g_j g_k g_l}
\]

observed frequency of two base pairs \( i-j \) and \( k-l \) aligned to each other in homologous RNAs
Using a Lightweight SCFG to Search for Secondary Structure

given a structure can construct a simple grammar characterizing it can add productions to allow for variation

\[ s \rightarrow C_{s_1}G \]
\[ s_1 \rightarrow A_{s_2}U \]
\[ s_2 \rightarrow b_1l_1 \]
\[ l_1 \rightarrow b_1b_2 \]
\[ b_1 \rightarrow U \]
\[ b_2 \rightarrow U \]
\[ b_3 \rightarrow C \]

\[ s \rightarrow Us_1A \]
\[ s \rightarrow As_1U \]
\[ s \rightarrow Gs_1C \]

- base pair substitutions
- insertions
- single base substitutions
Setting the Parameters in the Grammar

- Infer them from the parameters from the RIBOSUM matrices (taking into account the latter are log-odds scores)

\[
\begin{align*}
  s & \rightarrow Cs_1G \\
  s_1 & \rightarrow As_2U \\
  s_2 & \rightarrow b_1l_1 \\
  l_1 & \rightarrow b_1b_2 \\
  b_1 & \rightarrow U \\
  b_2 & \rightarrow U \\
  b_3 & \rightarrow C
\end{align*}
\]
RSEARCH: Searching Sequence for a Secondary Structure

• the RSEARCH algorithm [Klein & Eddy, *BMC Bioinformatics* 2003] implements this idea

• but uses a somewhat different SCFG formulation – covariance models (see section 10.3 in Durbin et al.)

• an RSEARCH case study: finding 6S genes in bacterial genomes
  – used E. coli 6S as the query structure
  – searched 14 other genomes with known 6S genes
    ~ 5,000 intergenic sequences on average
  – the top-scoring RSEARCH hit in all 14 genomes was the known 6S gene
6S RNA Secondary Structure

E. coli

H. influenzae

B. subtilis
Given: a pairwise alignment of homologous sequences

Identify novel RNA genes in the sequences
key idea: the pattern of substitutions in the two sequences provides evidence about the role of the sequence

substitutions tend to be in the 3\textsuperscript{rd} codon (wobble) position

substitutions tend to preserve complementary base pairings

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001
RNA Gene Detection

- Illustrative examples of emission scores for three models
  (numbers before parens are log-odds with respect to a model of no alignment)

<table>
<thead>
<tr>
<th></th>
<th>$p_{OTH}(\begin{array}{c}G \ G \end{array})$</th>
<th>$p_{OTH}(\begin{array}{c}C \ C \end{array})$</th>
<th>$p_{OTH}(\begin{array}{c}U \ C \end{array})$</th>
<th>$p_{OTH}(\begin{array}{c}A \ U \end{array})$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OTH</strong></td>
<td>+0.76(-3.20)</td>
<td>+0.72(-3.52)</td>
<td>-0.19(-4.41)</td>
<td>-0.53(-4.45)</td>
</tr>
<tr>
<td><strong>COD</strong></td>
<td>$p_{COD}(\begin{array}{c}A A C \ A A C \end{array})$</td>
<td>$p_{COD}(\begin{array}{c}A A C \ A A U \end{array})$</td>
<td>$p_{COD}(\begin{array}{c}A A C \ A U C \end{array})$</td>
<td>$p_{COD}(\begin{array}{c}A U C \ A G C \end{array})$</td>
</tr>
<tr>
<td></td>
<td>+3.31(-8.19)</td>
<td>+3.31(-8.19)</td>
<td>-0.52(-12.31)</td>
<td>+1.29(-10.95)</td>
</tr>
<tr>
<td><strong>RNA</strong></td>
<td>$p_{RNA}(\begin{array}{c}G \cdots C \ G \cdots C \end{array})$</td>
<td>$p_{RNA}(\begin{array}{c}G \cdots U \ G \cdots C \end{array})$</td>
<td>$p_{RNA}(\begin{array}{c}G \cdots A \ G \cdots A \end{array})$</td>
<td>$p_{RNA}(\begin{array}{c}C \cdots G \ G \cdots G \end{array})$</td>
</tr>
<tr>
<td></td>
<td>+3.81(-4.37)</td>
<td>+1.36(-6.82)</td>
<td>-8.82(-16.42)</td>
<td>+2.43(-5.76)</td>
</tr>
</tbody>
</table>

Figure from Rivas & Eddy, BMC Bioinformatics, 2001
RNA Gene Detection via Comparative Sequence Analysis

- given sequences $x$ and $y$, want a model that can distinguish
  - homologous RNA subsequences
  - homologous coding subsequences
  - “other” homologous subsequences
  - non-homologous subsequences
- allow these to be interleaved, have gaps
RNA Gene Detection: The IID Model

• models non-homologous sequences, $x$ and $y$

$\begin{align*}
1-\eta&\quad&1-\eta&\quad&1-\eta
\end{align*}$

emits a base in sequence $x$

denotes an IID submodel

emits a base in sequence $y$

• $S$, $K$ and $T$ are silent states

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001
RNA Gene Detection: The “Other” Homologous Sequence Model

$F_L$, $F_J$ and $F_R$ are IID submodels

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001
RNA Gene Detection: The Coding Sequence Model

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

- **$O_B$**, **$O_J$** and **$O_E$** are “other” submodels
- **$C_B$** emits a codon in $x$ only
- **$C_E$** emits codons in $x$ and $y$
- **$O_J$** emits a codon in $y$ only
RNA Gene Detection: The RNA Model

$O_B$, $O_J$ and $O_E$ are “other” submodels

- here, the RNA box is a “lightweight” pairwise SCFG

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001
Summary of RNA Analysis Tasks

• given a sequence, predict its secondary structure
• given a set of related RNA sequences, construct a model of the set
  • parameter learning (Inside-Outside)
  • structure refinement
• given a model of an RNA class, find sequences that belong to the class (Inside or CYK)
• given a sequence/structure, find other sequences with similar structure
• given a pair of related genomic sequences, find subsequences that seem have similar secondary structure (RNA gene finding)