Cocke-Kasami-Younger Parser

Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$
(by mechanically transforming grammar)

Given $x = x_1...x_n$, want $M_{i,j} = \{ A | A \Rightarrow^* x_{i+1}...x_j \}$

For $j=2$ to $n$

$M[j-1,j] = \{A | A \rightarrow x_j$ is a rule$\}$
for $i = j-1$ down to $1$

$M[i,j] = \bigcup_{i < k < j} M[i,k] \otimes M[k,j]$

Where $X \otimes Y = \{A | A \rightarrow BC, B \in X, \text{ and } C \in Y \}$

Time: $O(n^3)$
And now for something completely different

CFGs beyond compilers
An RNA Structure
An RNA Sensor & On/Off Switch

L19 absent: Gene On       L19 present: Gene Off
mRNA leader

mRNA leader switch?
An RNA Grammar

\[ S \rightarrow LS \mid L \]
\[ L \rightarrow s \mid "dFd" \]
\[ F \rightarrow LS \mid "dFd" \]

“s” means unpaired; “dFd” means paired
(Watson–Crick:
\[
aFu \mid uFa \mid gFc \mid cFg
\]
paren-like nesting)

\[ \begin{align*}
S & \rightarrow LS \rightarrow LLLLdSSSs \\
& \rightarrow ssLssssss \rightarrow ssdFdssssss \\
& \rightarrow ssddddFdddssssss \\
& \rightarrow ssddddLSddddssssss \\
& \rightarrow ssddddLLLLddddssssss \\
& \rightarrow ssddddssssddddssssss
\end{align*} \]

\[ F \rightarrow dFd \rightarrow dddFddd \rightarrow dddLSddd \\
& \rightarrow dddLLddd \rightarrow dddLsddd \rightarrow ddddFdsdd \]
Actually, a **Stochastic** CFG

Associate *probabilities* with rules, e.g.:

\[
S \rightarrow LS \quad (p = 0.87)
\]

\[
S \rightarrow L \quad (p = 0.13)
\]

\[
\ldots
\]

Now we can ask, not only “Does S generate w?” But also “How likely is it?”
Cocke-Kasami-Younger Parser

Suppose all rules of form \( A \rightarrow BC \) or \( A \rightarrow a \) (by mechanically transforming grammar)

Given \( x = x_1 \ldots x_n \), want \( M_{i,j} = \{ A \mid A \Rightarrow^* x_{i+1} \ldots x_j \} \)

For \( j = 2 \) to \( n \)

\[
M[j-1,j] = \{ A \mid A \rightarrow x_j \text{ is a rule} \}
\]

for \( i = j - 1 \) down to 1

\[
M[i,j] = \bigcup_{i < k < j} M[i,k] \otimes M[k,j]
\]

Where \( X \otimes Y = \{ A \mid A \rightarrow BC , B \in X, \text{ and } C \in Y \} \)

Time: \( O(n^3) \)
“Inside” Algorithm for SCFG

Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$
(by mechanically transforming grammar)

Given $x = x_1 \ldots x_n$, want $M^A_{i,j} = p(A \Rightarrow^* x_{i+1} \ldots x_j)$

For $j=2$ to $n$

$M^A[j-1,j] = p( \text{rule } A \rightarrow x_j )$

for $i = j-1$ down to $1$

$M^A[i,j] = \sum_{A \rightarrow BC, \ i < k < j} M^B[i,k] \times M^C[k,j]$

I.e., probability of $A$ in $M[i,j]$, instead of its possibility

Time: $O(n^3)$
ncRNA Discovery in Bacteria


Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker. Identification of 22 candidate structured RNAs in bacteria using the confirmed riboswitch (+2 more).

ncRNA Discovery in Humans

Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin

Genome Research, Jan ‘08
Experimental Validation

<table>
<thead>
<tr>
<th>B</th>
<th>Brain</th>
<th>Adrenal gland</th>
<th>Bone marrow</th>
<th>Spinal cord</th>
<th>Kidney</th>
<th>Lung</th>
<th>Salivary gland</th>
<th>Skeletal muscle</th>
<th>Spleen</th>
<th>Testis + Prostate</th>
<th>Uterus + Placenta</th>
<th>Thymus + Thrd gld</th>
<th>Trachea</th>
<th>Colon</th>
<th>Small intestine</th>
<th>Fetal liver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cdt. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-actin RT+ control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-actin RT- control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C

<table>
<thead>
<tr>
<th>C</th>
<th>Whole Brain</th>
<th>Fetal Brain</th>
<th>Cerebellum</th>
<th>Hippocampus</th>
<th>Spinal cord</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cdt. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>cdt. 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-actin RT+ control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-actin RT- control</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bottom Line

CFG technology is a key tool for RNA description, discovery and search
A very active research area. (Some call RNA the “dark matter” of the genome.)

Huge compute hog: results above represent hundreds of CPU-years, and smart algorithms can have a big impact
More?

Check out CSE 427