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 $M^A_{i,j} = \{1 \text{ if } (A \Rightarrow * x_i, \ldots, x_j) \text{ else } 0\}$

For $j=2$ to $n$

$M^A[j-1,j] = \{1 \text{ if } (A \rightarrow x_j \text{ is a rule}) \text{ else } 0\}$
for $i = j-1$ down to 1

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

Time: $O(n^3)$

And now for something completely different …

CFGs beyond compilers

An RNA Structure

An RNA Computer!
Sensor & On/Off Switch

L19 absent: Gene On       L19 present: Gene Off
Actually, a *Stochastic* CFG

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

\[
\begin{align*}
  S &\rightarrow LS \quad (p = 0.87) \\
  S &\rightarrow L \quad (p = 0.13) \\
  \ldots
\end{align*}
\]

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 \( M^A_{i,j} = \{1 \text{ if } (A \Rightarrow^* x_{i+1} \ldots x_j) \text{ else } 0\} \)

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

\[
M^A[i,j] = \begin{cases} 
1 & \text{if } (A \rightarrow x_j \text{ is a rule}) \text{ else } 0 \\
\end{cases}
\]

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

\[
M^A[i,j] = \bigvee_{A \rightarrow BC, i < k < j} M^B[i,k] \land M^C[k,j]
\]

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

Suppose all rules of form A → BC or A → 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

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

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

Time: \( O(n^3) \)

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; smart algorithms have a big impact

(Recall the \( O(n^3) \)...)

More?

Check out CSE 427/428: “Comp Bio”