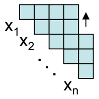
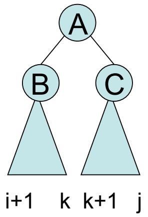
Cocke-Kasami-Younger Parser

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

Given $x = x_1 \ldots x_n$, want $M^A_{i,j}$ = {1 if (A $\Rightarrow^* x_{i+1} \ldots x_j$) else 0}

```
For j=2 to n
    $M^A[j-1,j]$ = {1 if (A $\rightarrow x_j$ is a rule) else 0}
    for i = j-1 down to 1
        $M^A[i,j] = \bigvee_{A \rightarrow BC,\ i<k<j} M^B[i,k] \wedge M^C[k,j]$
```



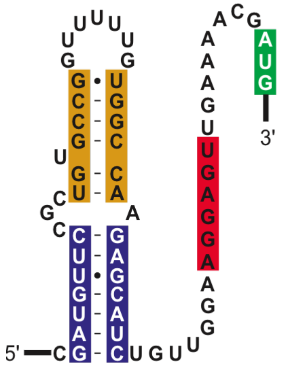


Time: $O(n^3)$

And now for something completely different …

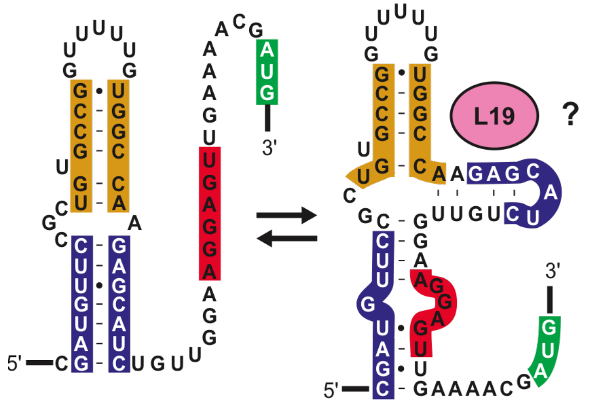
CFGs beyond compilers

An RNA Structure



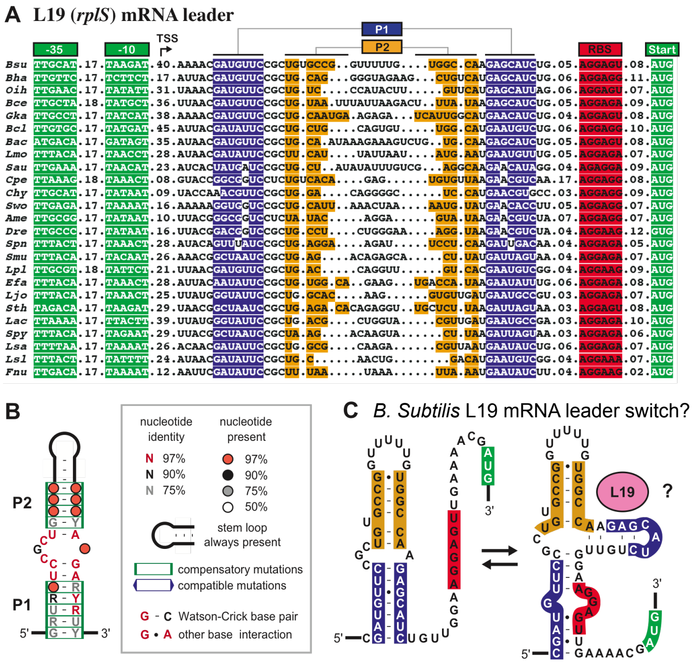
An RNA Computer!

Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off



A CFG for RNA

$S \rightarrow LS \quad | \; L$

$L \rightarrow$ "dFd" $| \; s$

$F \rightarrow$ "dFd" $| \; LS$

"s" means unpaired;

"dFd" means paired (Watson–Crick:

aFu | uFa | gFc | cFg

paren-like nesting)

$S \Rightarrow LS \Rightarrow^* LLLLLLLS$

$\Rightarrow LLLLLLLL$

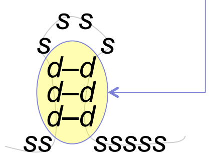
$\Rightarrow^* ssLsssss$

$\Rightarrow ssdFdsssss$

$\Rightarrow ssddFddsssss$

$\Rightarrow ssdddFdddsssss$

$\Rightarrow \ldots$



Actually, a *Stochastic* CFG

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

$S \rightarrow LS$ (p = 0.87)

$S \rightarrow L$ (p = 0.13)

. . .

Now we can ask, not only

"Does S generate w?"

But also

"How likely is it?"

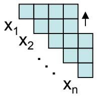
Cocke-Kasami-Younger Parser

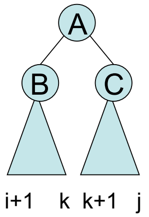
Suppose all rules of form A → BC or A → a

(by mechanically transforming grammar)

Given $x = x_1 \ldots x_n$, want $M^A_{i,j}$ = {1 if ($A \Rightarrow^* x_{i+1} \ldots x_j$) else 0}

```
For j=2 to n
    $M^A[j-1,j]$ = {1 if (A → $x_j$ is a rule) else 0}
    for i = j-1 down to 1
        $M^A[i,j] = \bigvee_{A \to BC,\ i<k<j} M^B[i,k] \wedge M^C[k,j]$
```





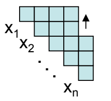
Time: $O(n^3)$

“Inside” Algorithm for SCFG

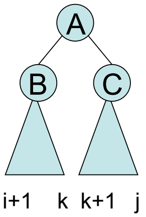
Suppose all rules of form A → BC or A → a
(by mechanically transforming grammar)

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

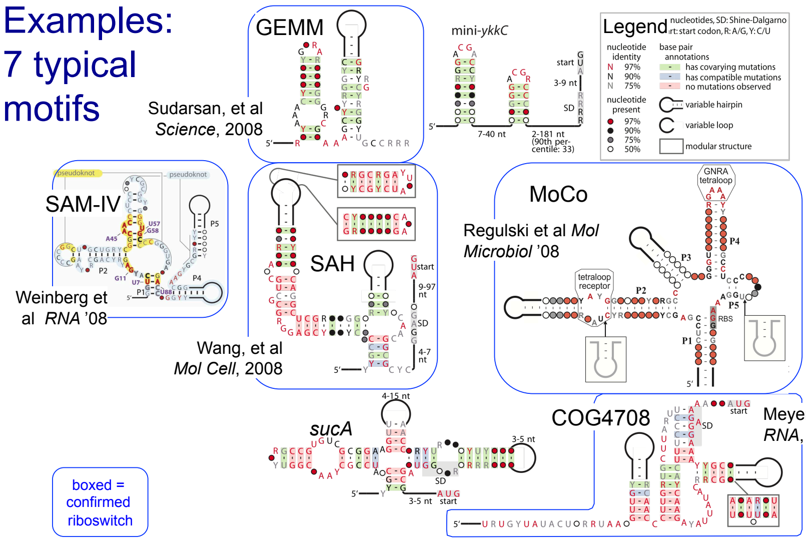


I.e., *probability* of A in M[i,j], instead of its *possibility*



Time: $O(n^3)$

Examples: 7 typical motifs



Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo, Breaker. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucl. Acids Res.*, July 2007 35: 4809-4819.

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”