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 $M_{i,j} = \{ A \mid A \Rightarrow^* x_{i+1} \ldots x_j \}$

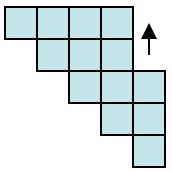
For j=2 to n

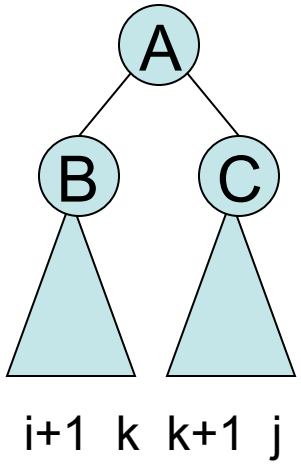
M[j-1,j] = {A | A $\rightarrow x_j$ is a rule}

for i = j-1 down to 1

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

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



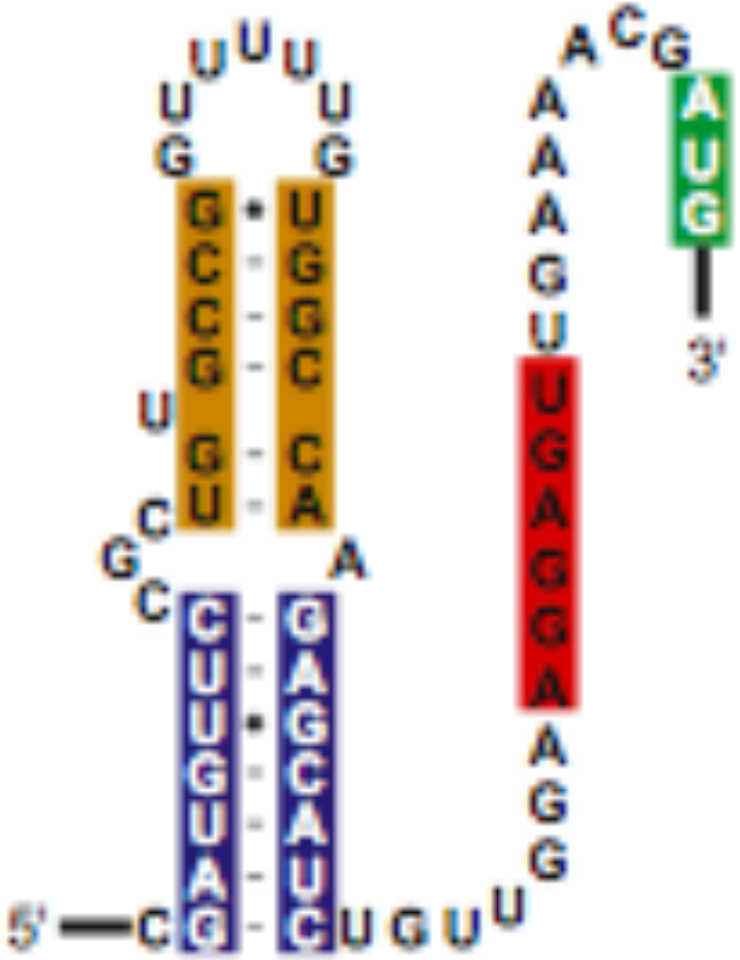


Time: $O(n^3)$

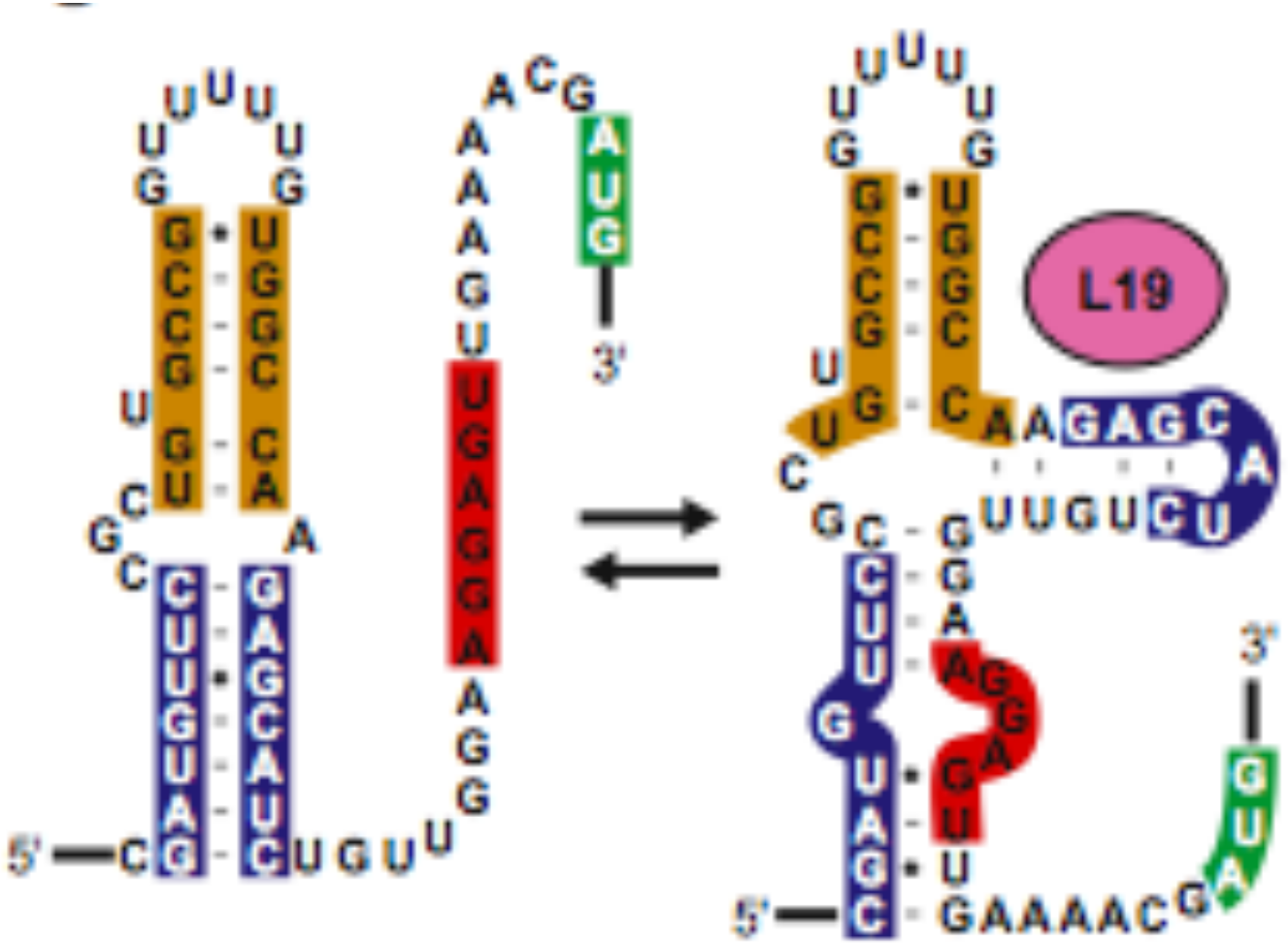
And now for something completely different

CFGs beyond compilers

An RNA Structure

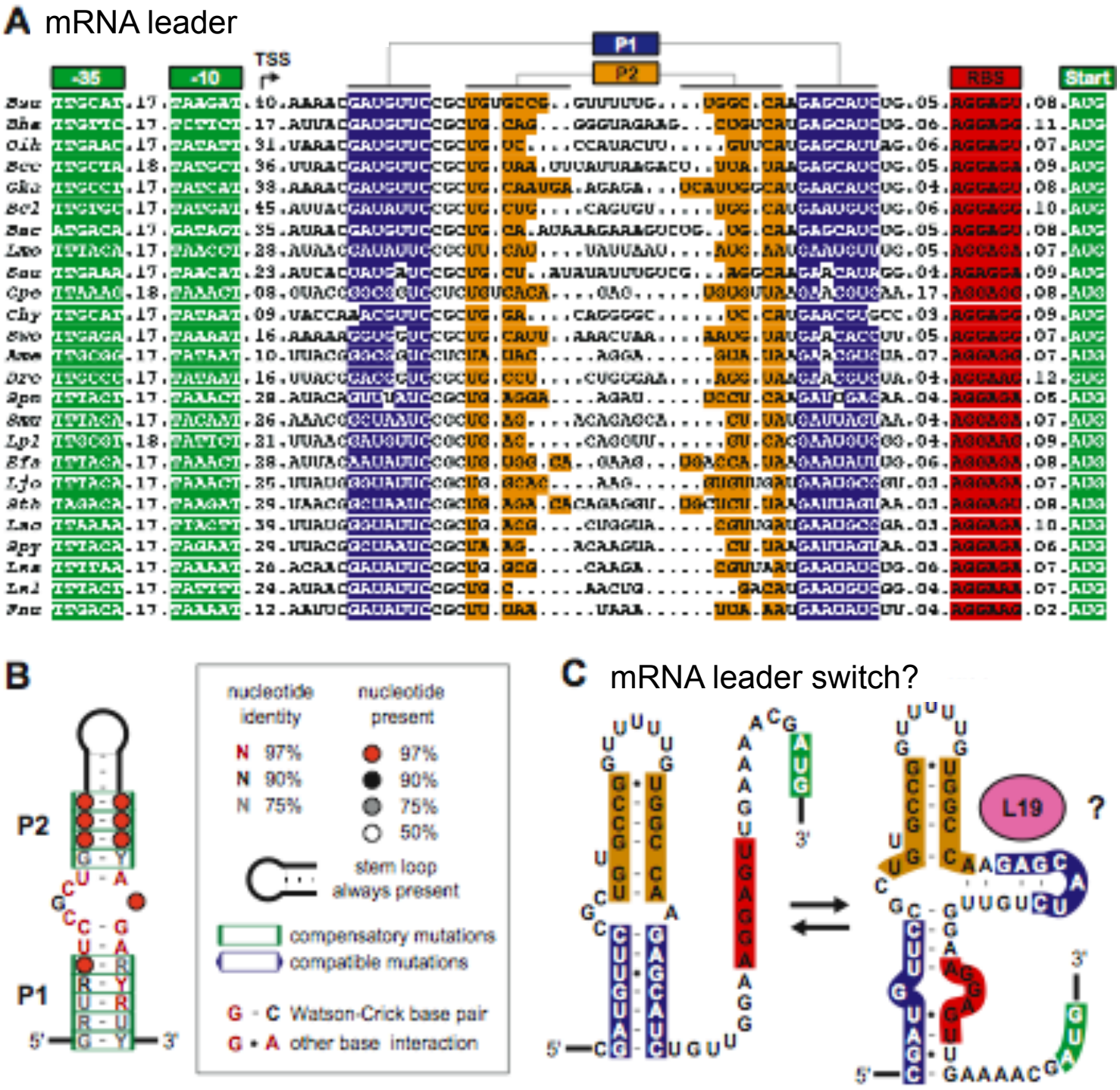


An RNA Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off


A mRNA leader
P1
P2
TSS
-35
-10
RBS
Start
B
P2
P1
5'
3'
nucleotide identity
nucleotide present
N 97%
N 90%
N 75%
97%
90%
75%
50%
stem loop always present
compensatory mutations
compatible mutations
G - C Watson-Crick base pair
G • A other base interaction
C mRNA leader switch?
L19
?
5'
3'
GAAAAC

An RNA Grammar

$S \rightarrow LS \mid L$

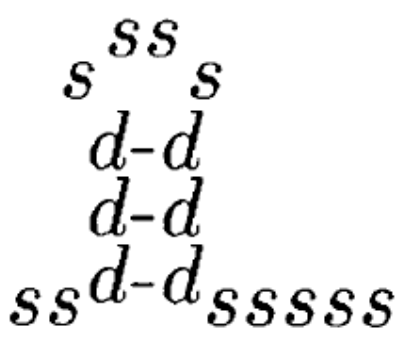
$L \rightarrow s \mid$ "dFd"

$F \rightarrow LS \mid$ "dFd"

"s" means unpaired;

"dFd" means paired (Watson–Crick: *aFu* | *uFa* | *gFc* | *cFg* paren-like nesting)

$$\begin{aligned} S &\rightarrow LS \rightarrow LLLLLLLS \rightarrow LLLLLLLL \\ &\rightarrow ssLssssss \rightarrow ssdFdssssss \\ &\rightarrow ssdddFddddssssss \\ &\rightarrow ssdddLSddddssssss \\ &\rightarrow ssdddLLLLddddssssss \\ &\rightarrow ssddddssssddddssssss \end{aligned}$$



$$\begin{aligned} F &\rightarrow dFd \rightarrow ddFdd \rightarrow ddLSdd \\ &\rightarrow ddLLdd \rightarrow ddLsdd \rightarrow dddFdsdd \end{aligned}$$

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 $A \rightarrow BC$ or $A \rightarrow a$
(by mechanically transforming grammar)

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

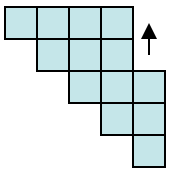
For j=2 to n

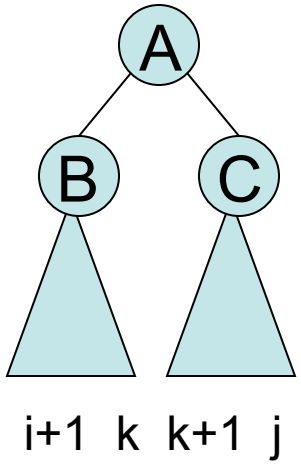
M[j-1,j] = {A | A $\rightarrow x_j$ is a rule}

for i = j-1 down to 1

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

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)

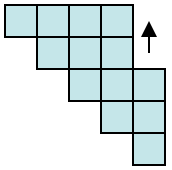
Given $x = x_1 \ldots x_n$, want $M^A_{i,j} = p(A \Rightarrow^* x_{i+1} \ldots x_j)$

For j=2 to n

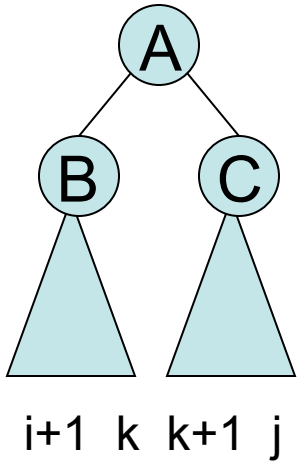
$M^A[j-1,j] = p(\text{ rule } A \rightarrow x_j)$

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*



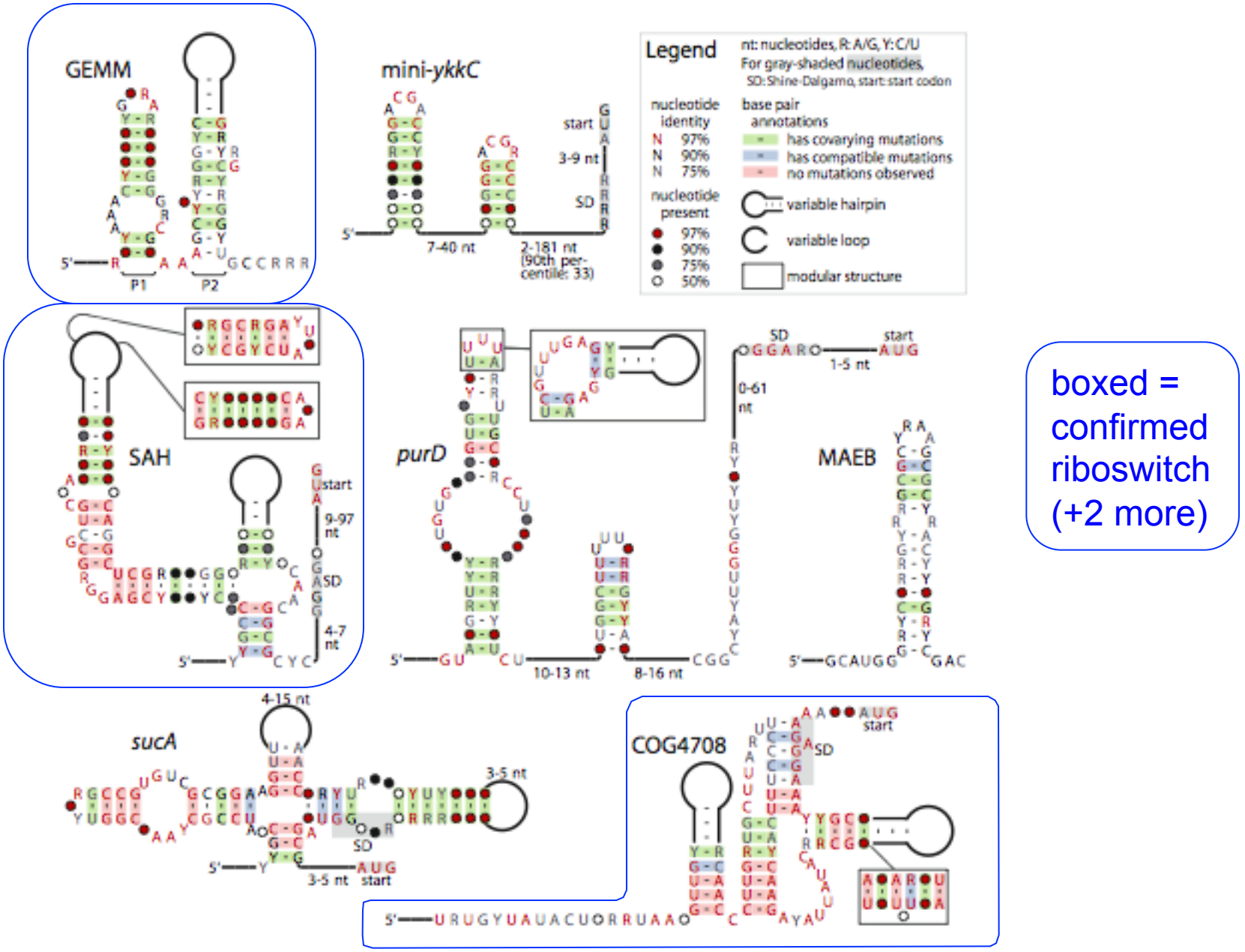
Time: $O(n^3)$

ncRNA Discovery in Bacteria

Cmfinder--A Covariance Model Based RNA Motif Finding Algorithm, Yao, Weinberg, Ruzzo,
Bioinformatics, 2006, 22(4): 445-452,

A Computational Pipeline for High Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo**.**
PLoS Comput Biol. 3(7): e126, July 6, 2007.

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



Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker. *Nucl. Acids Res.,* July 2007

Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa,

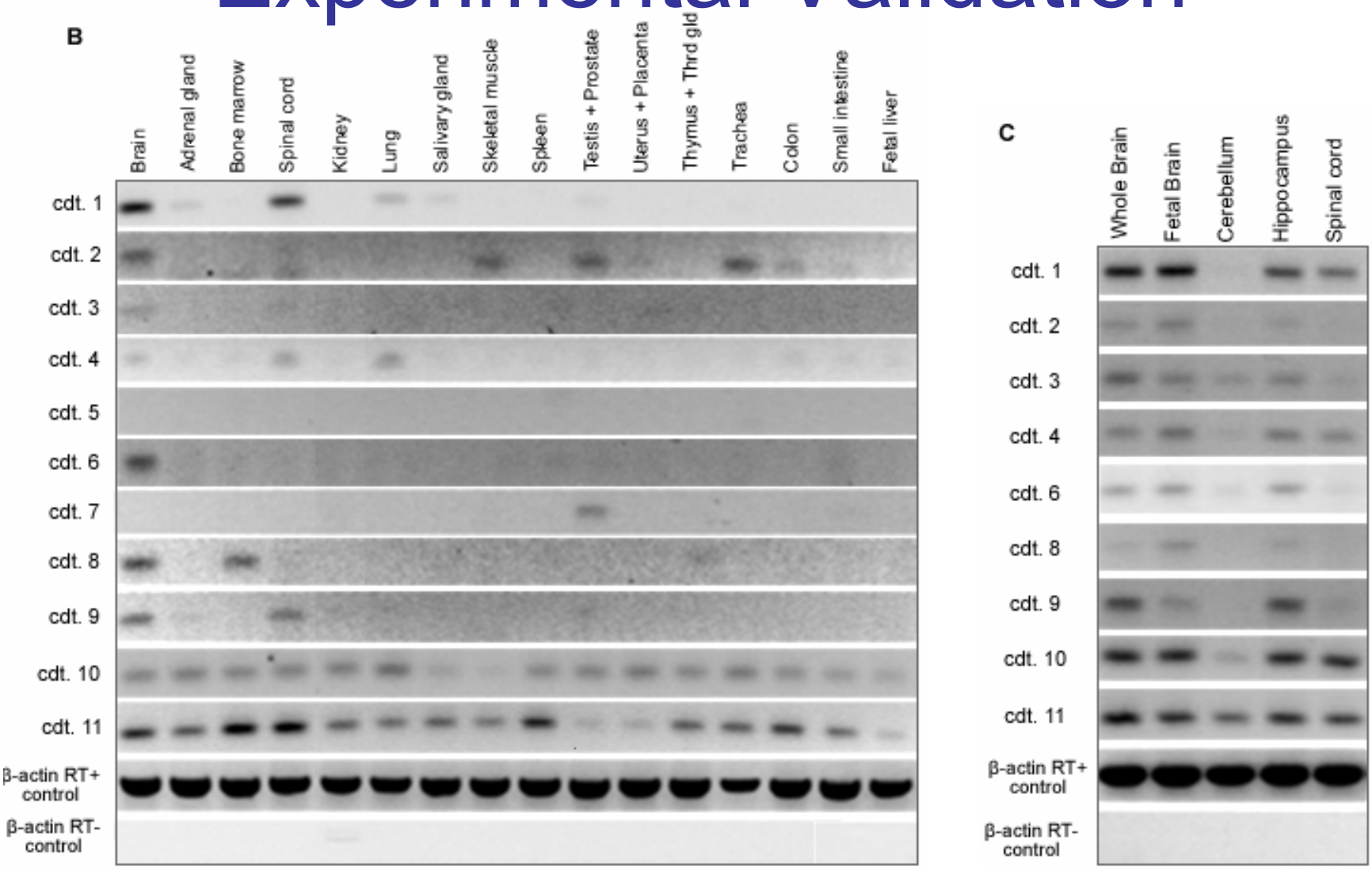
ncRNA Discovery in Humans

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



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