Alt proof: $\{a^n b^n c^n \mid n \geq 0\}$ not CFL:

v & y can't have both a & c

Then case 1 not c

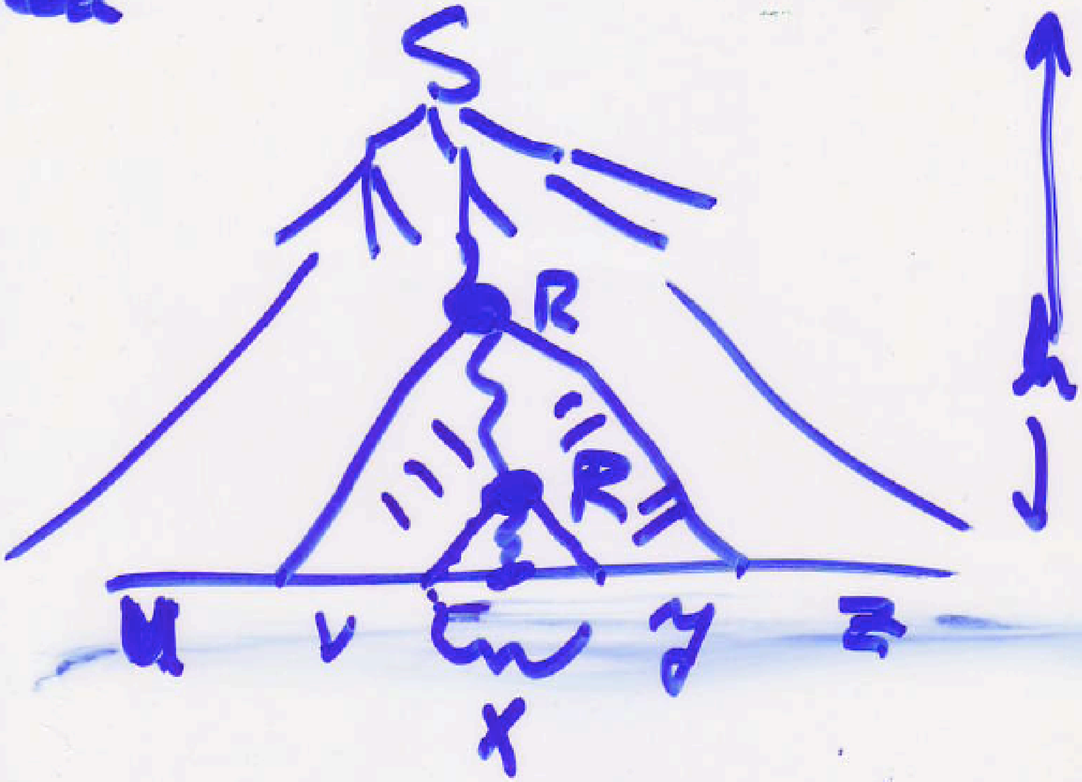
$u v^2 x y^2 z$ too few c's

Similarly

case 2 not a

... too few a's

Idea



Repeating shaded part i times gives $u v^i x y^i z \in A$

$S \Rightarrow^* uRz$

$R \Rightarrow^* vRy$

$R \Rightarrow^* x$

$R \to RE$

$E \to \varepsilon$

$p = b^{|V|+1}$

Subtlety 1 : tree w/ fewest nodes $vy \neq \varepsilon$

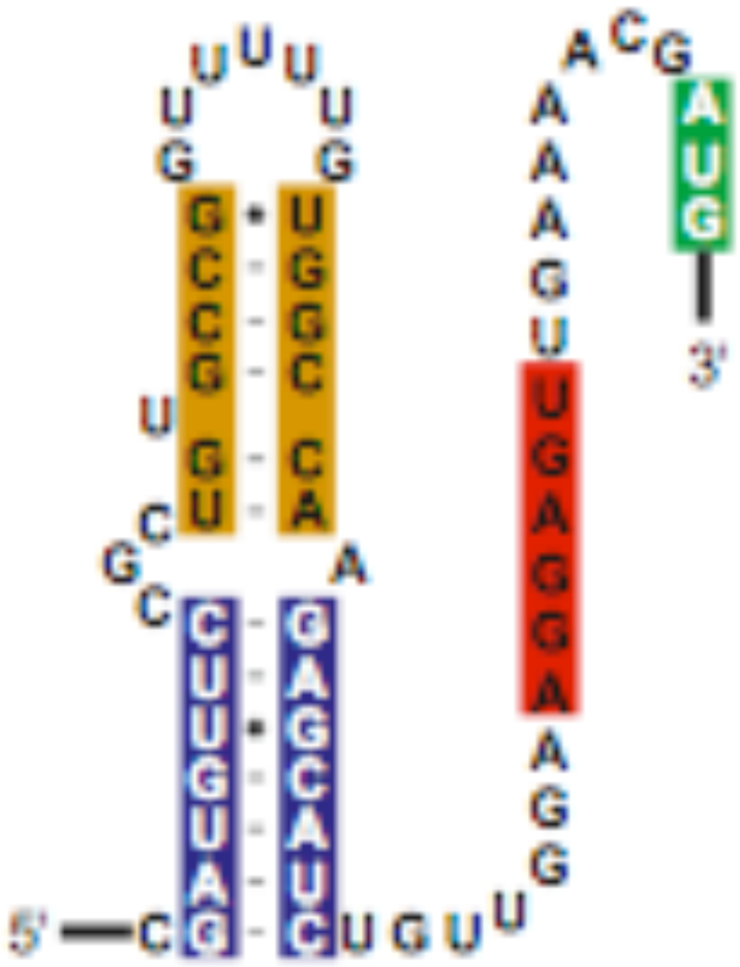
Subtlety 2 : pick rep. nearest leaves.

$|vxy| \le p$

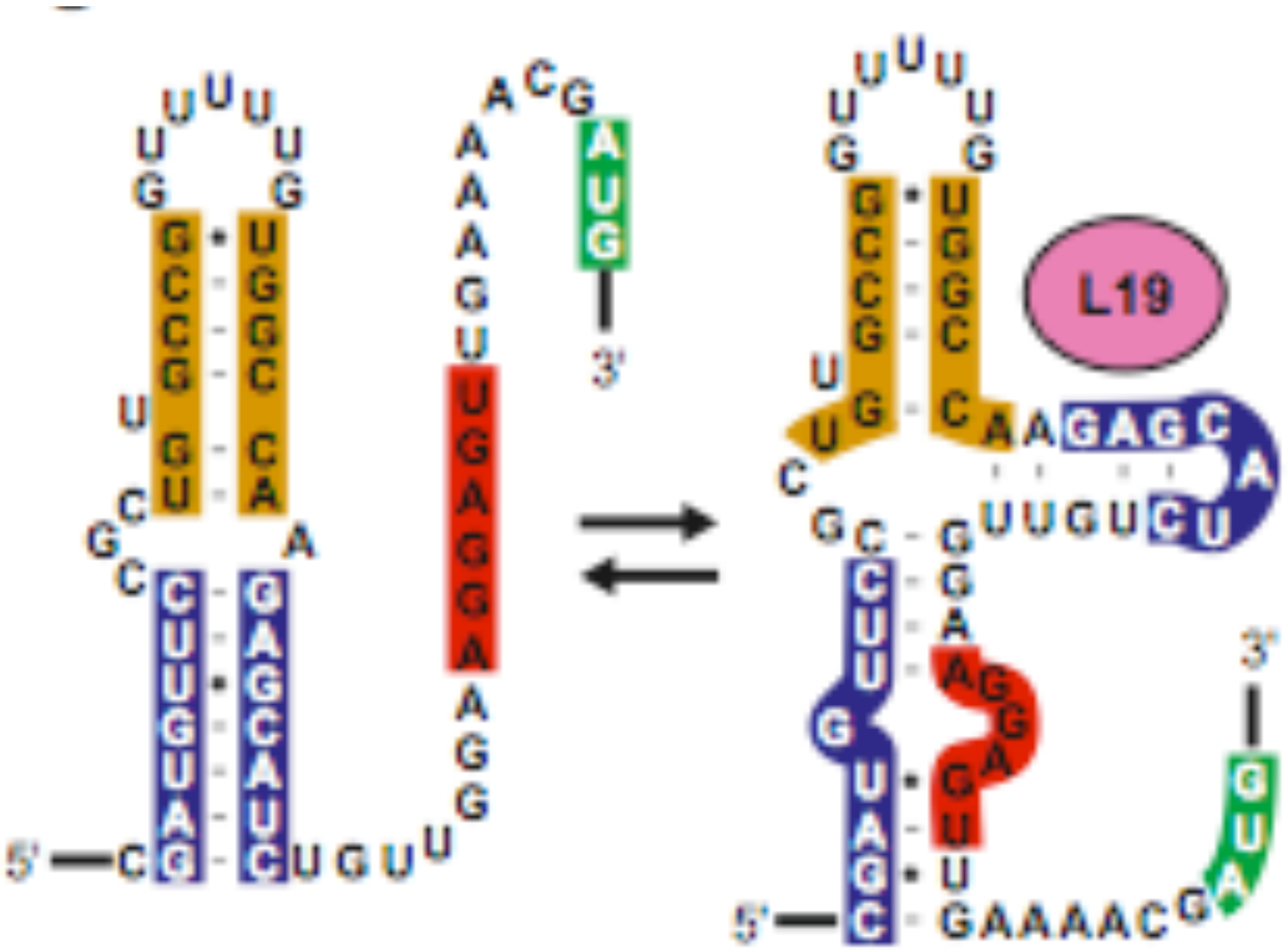
And now for something completely different

CFG utility beyond compilers

An RNA Structure

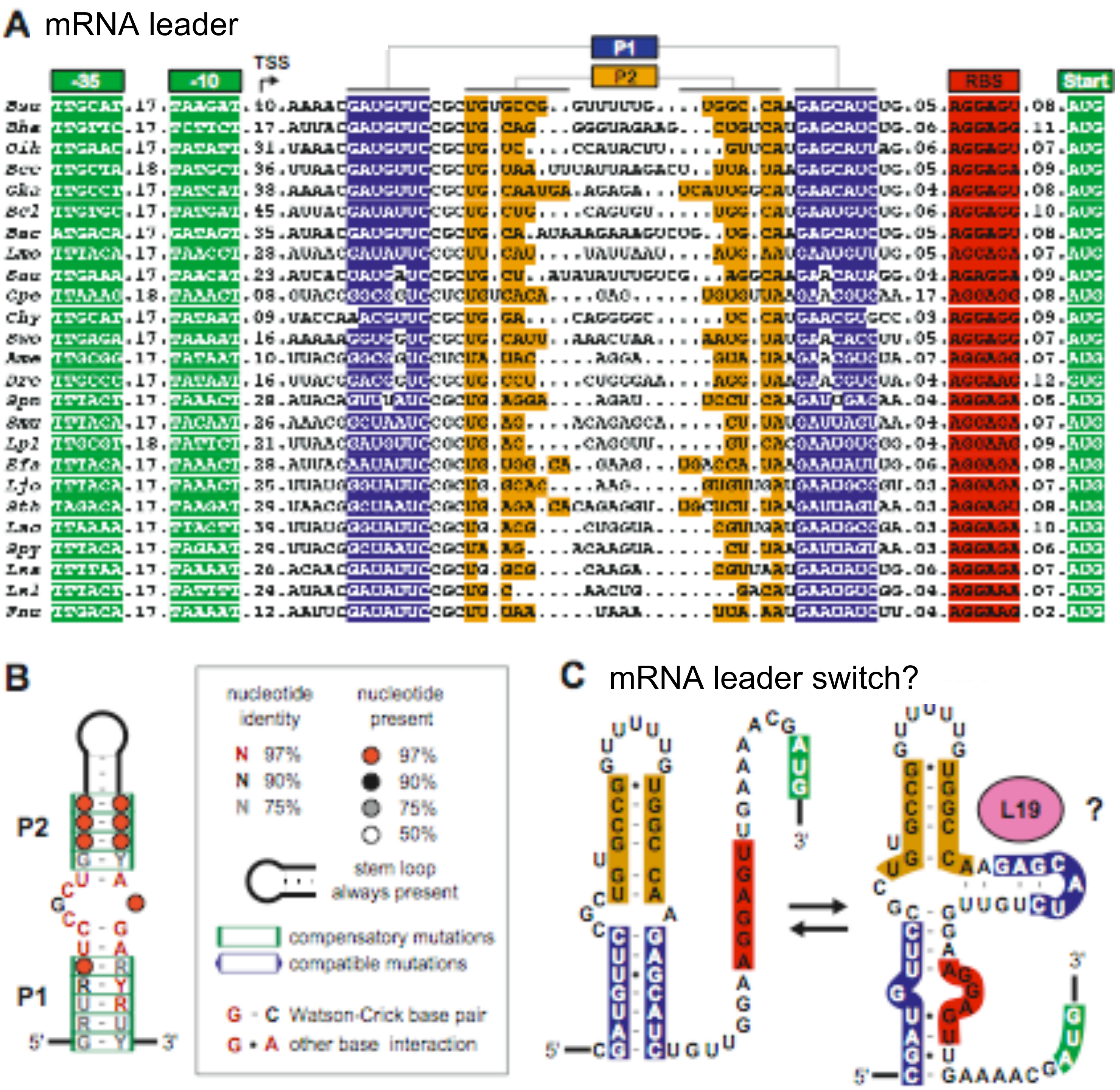

5'
3'
C G C U U G U A G C C G G C U G G U U U U U G
U G G C C A A G A G C A U C U G U U
G G A A G G A G U U G A A A A C G A U G

An RNA Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off


A mRNA leader
-35
-10
TSS
P1
P2
RBS
Start
Bsu
Bha
Oih
Bce
Gka
Bcl
Bac
Lmo
Sau
Cpe
Chy
Swo
Ame
Dre
Spn
Smu
Lpl
Efa
Ljo
Sth
Lac
Spy
Lsa
Lsl
Fnu
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'

An RNA Grammar

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

$L \rightarrow s \quad | \; \text{“}dFd\text{”}$

$F \rightarrow LS \quad | \; \text{“}dFd\text{”}$

“dFd” means Watson-Crick base pair:

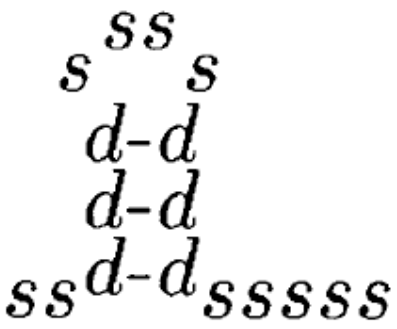
aFu | *uFa* | *gFc* | *cFg*

paren-like nesting

a)

$$\begin{aligned} S &\rightarrow LS \rightarrow LLLLLLLS \rightarrow LLLLLLLL \\ &\rightarrow ssLsssss \rightarrow ssdFdsssss \\ &\rightarrow ssdddFdddsssss \\ &\rightarrow ssdddLSdddsssss \\ &\rightarrow ssdddLLLLdddsssss \\ &\rightarrow ssdddssssdddsssss \end{aligned}$$

b)



c)

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

Actually, a *Stochastic* CFG

Associate probabilities with rules:

$$\begin{array}{llll} S \rightarrow LS & \texttt{(0.87)} & \mid L & \texttt{(0.13)} \\ L \rightarrow S & \texttt{(0.89*p(s))} & \mid dFd & \texttt{(0.11*p(dd))} \\ F \rightarrow LS & \texttt{(0.21)} & \mid dFd & \texttt{(0.79*p(dd))} \end{array}$$

Where p(s) & p(dd) are the probabilities of the specific single/paired nucleotides, perhaps from empirical data or a model of sequence evolution

What SCFG Gives

“Prior” probabilities for

- frequencies of nucleotides/pairs
- fraction paired vs unpaired
- average lengths of each, etc.

Result: a probability distribution on sequences/structures

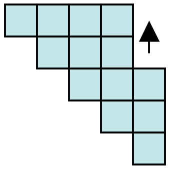
- E.g., is my sequence more likely to arise under this RNA model or a simple “background” model, say where A/C/G/T = 1/4?

Cocke-Kasami-Younger Parser

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

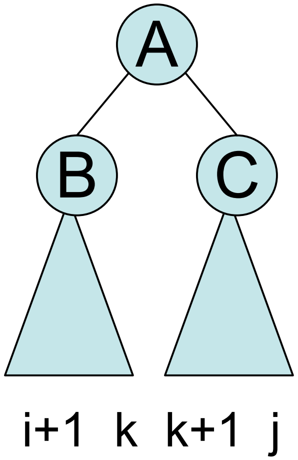
Given $x = x_1 \ldots x_n$, want $M_{i,j} = \{ A \mid A \rightarrow x_{i+1} \ldots x_j \}$

```
For j=2 to n
    M[j-1,j] = {A | A → $x_j$ 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

Just like CKY, but instead of just recording *possibility* of A in M[i,j], record its *probability*:

For each A, do sum instead of union, over all possible k and all possible $A \rightarrow BC$ rules, of products of their respective probabilities.

Result: for each i, j, A, have $\Pr(A \Rightarrow^* x_{i+1} \ldots x_j)$

The SCFG “Viterbi” algorithm

Like inside, but use max instead of sum;

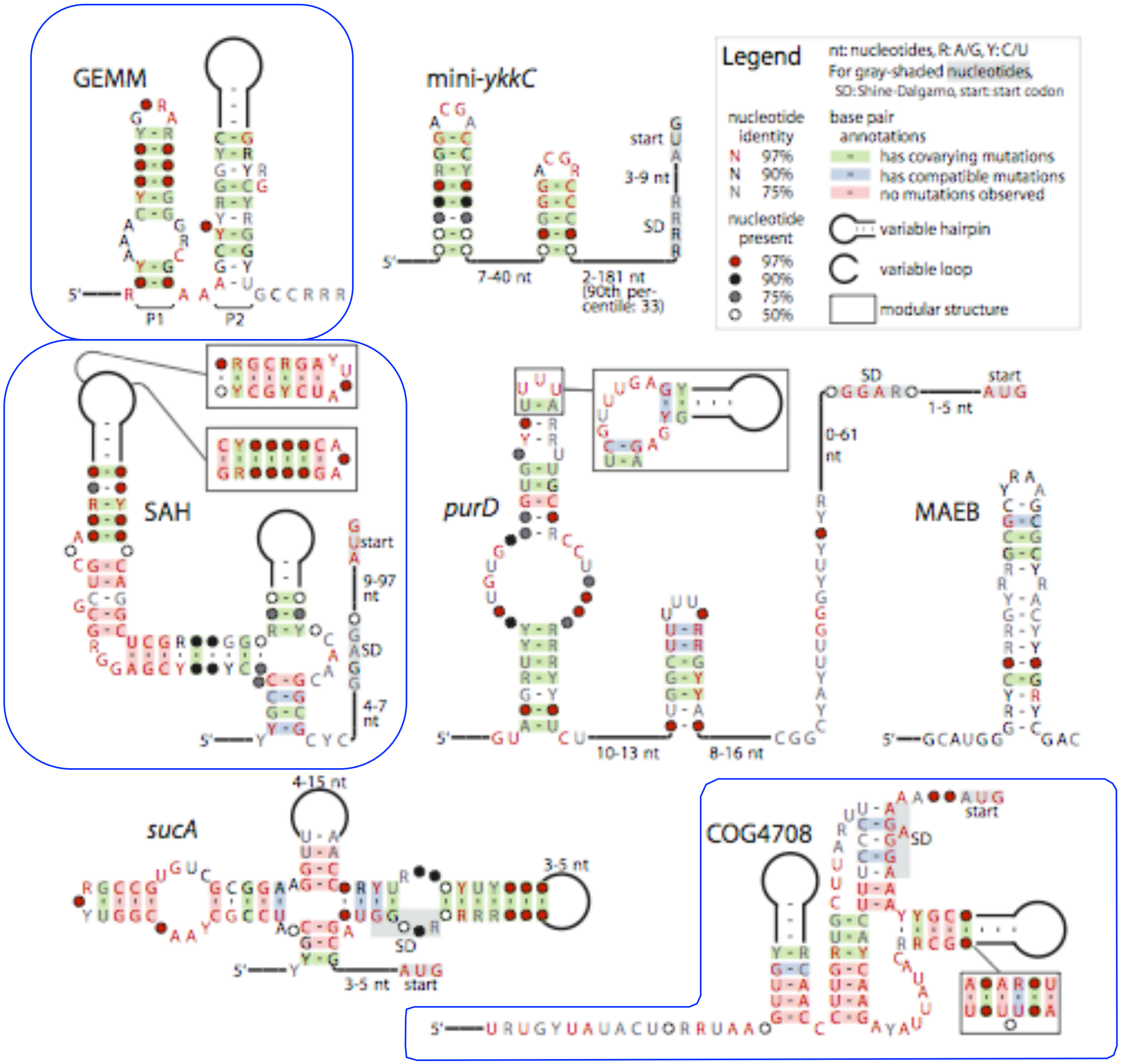
Gives probability of the *single* parse tree having max probability; (inside sums probability over *all* legal trees)

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.



boxed = confirmed riboswitch (+2 more)

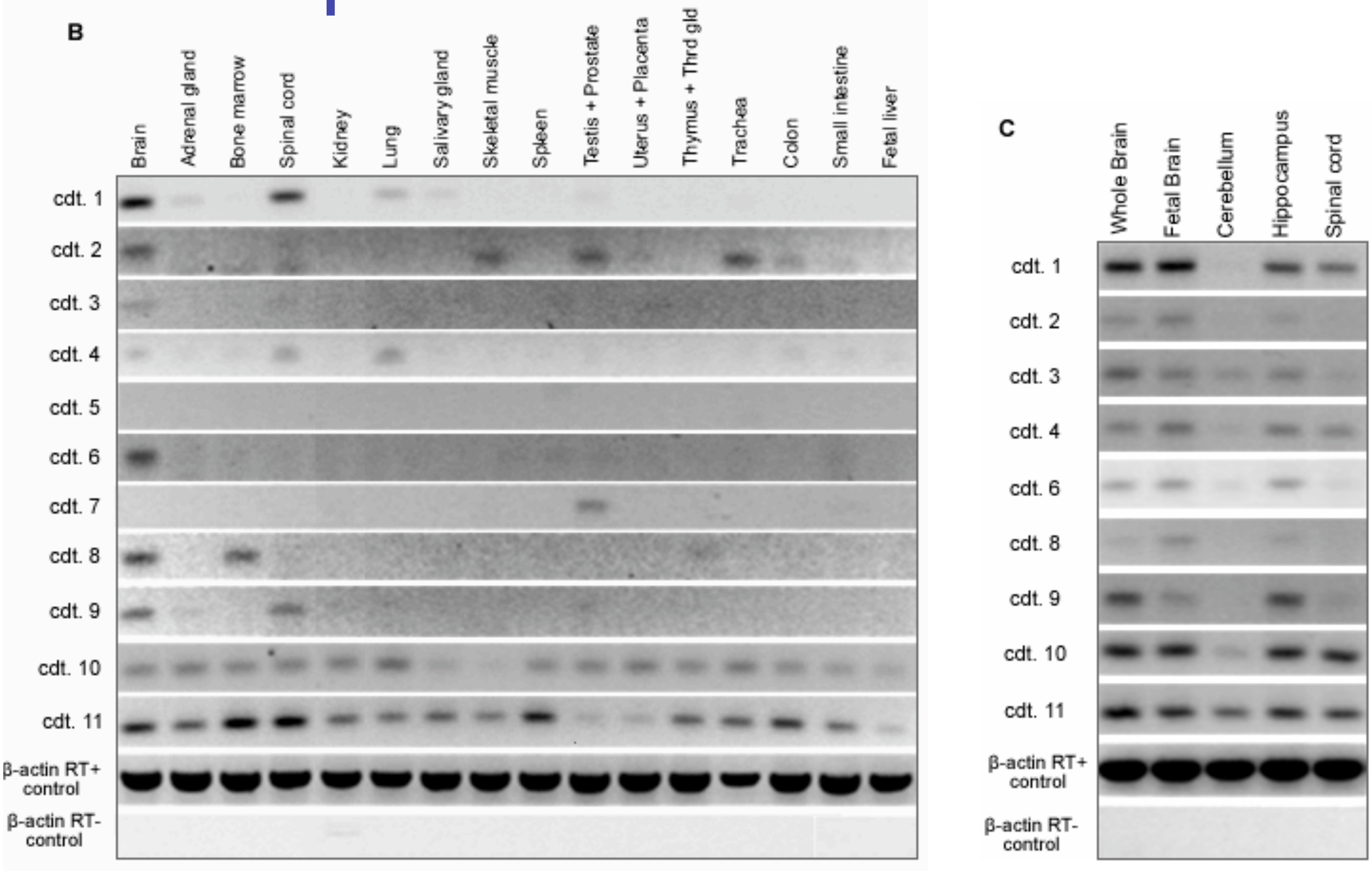
ncRNA Discovery in Vertebrates

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

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

Genome Research, to appear

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