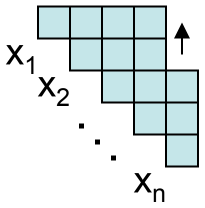
Cocke-Kasami-Younger Parser

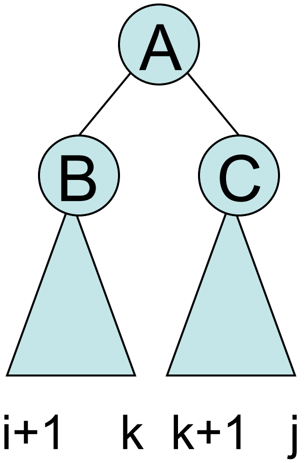
Suppose all rules of form $A \rightarrow BC$ or $A \rightarrow a$

(by mechanically transforming grammar)

Given $x = x_1 \ldots x_n$, want $M^A_{i,j} = \{1$ 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
        $M^A[i,j] = \bigvee_{A \rightarrow BC,\ i < k < j} M^B[i,k] \wedge M^C[k,j]$
```



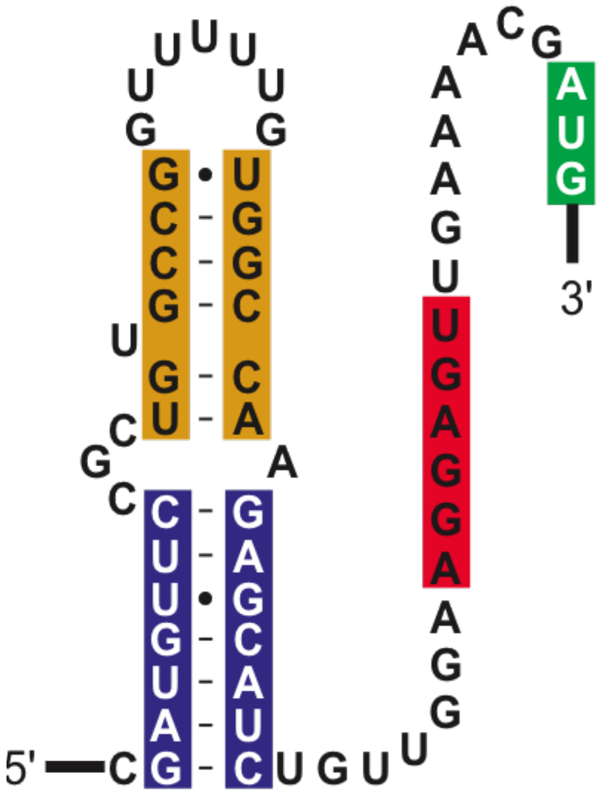


Time: $O(n^3)$

And now for something completely different …

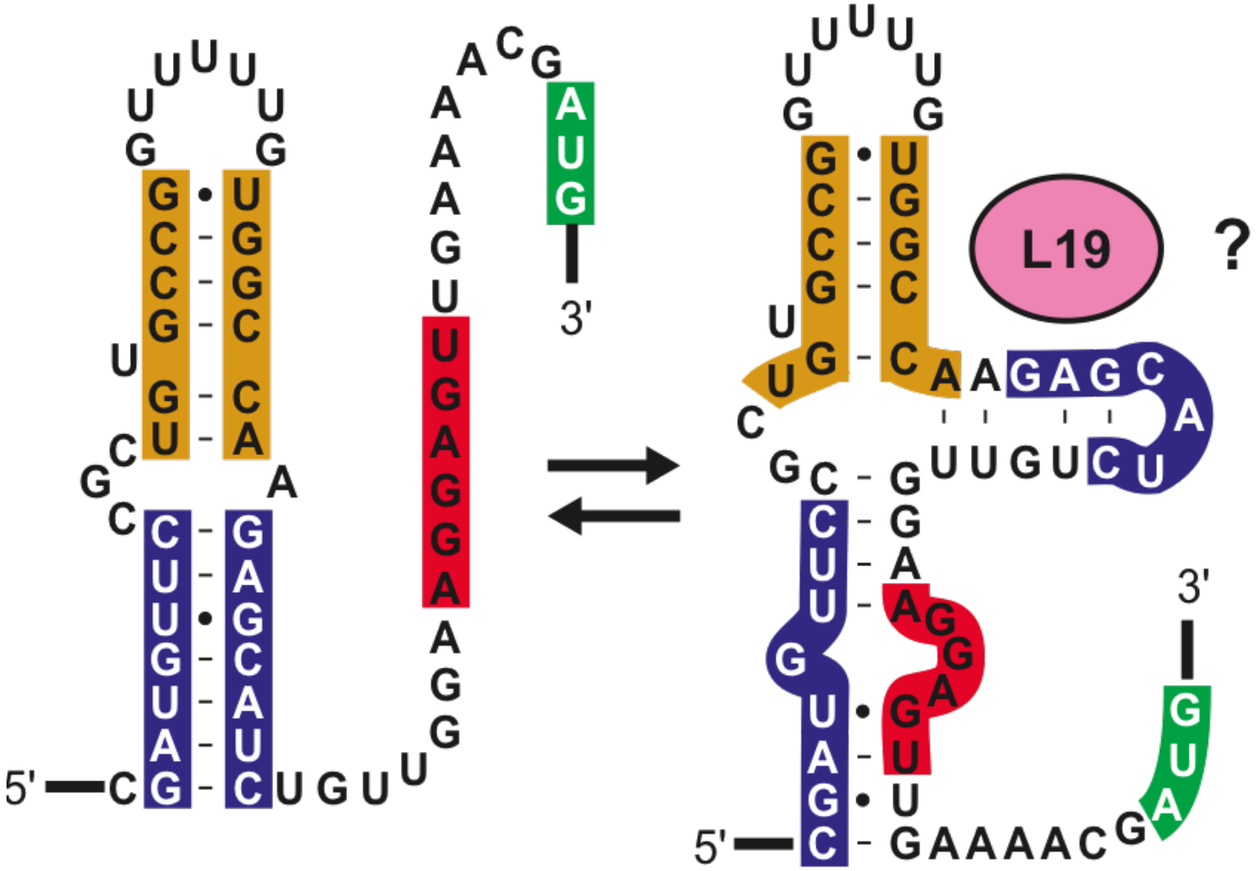
CFGs beyond compilers

An RNA Structure



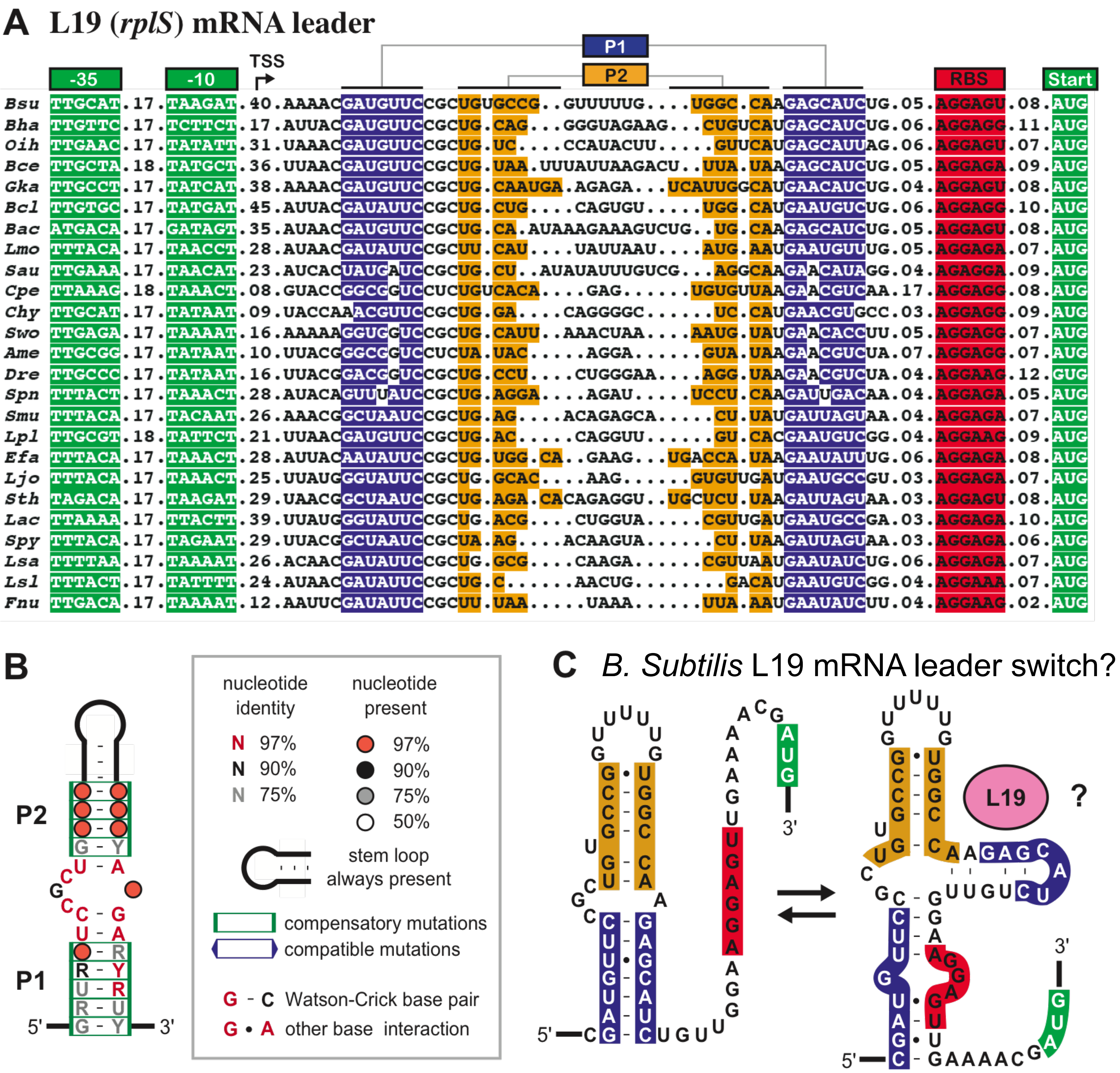
An RNA Computer!

Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off



A L19 (*rplS*) mRNA leader

```
Bsu TTGCAT.17.TAAGAT.40.AAAACGAUGUUCCGCUGUGCCG..GUUUUUG....UGGC.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Bha TTGTTC.17.TCTTCT.17.AUUACGAUGUUCCGCUG.CAG...GGGUAGAAG...CUGUCAUGAGCAUCUG.06.AGGAGG.11.AUG
Oih TTGAAC.17.TATATT.31.UAAACGAUGUUCCGCUG.UC....CCAUACUU.....GUUCAUGAGCAUUAG.06.AGGAGU.07.AUG
Bce TTGCTA.18.TATGCT.36.UUAACGAUGUUCCGCUG.UAA.UUUAUUAAGACU..UUA.UAAGAGCAUCUG.05.AGGAGA.09.AUG
Gka TTGCCT.17.TATCAT.38.AAAACGAUGUUCCGCUG.CAAUGA.AGAGA...UCAUUGGCAUGAACAUCUG.04.AGGAGU.08.AUG
Bcl TTGTGC.17.TATGAT.45.AUUACGAUAUUCCGCUG.CUG....CAGUGU.....UGG.CAUGAAUGUCUG.06.AGGAGG.10.AUG
Bac ATGACA.17.GATAGT.35.AUAACGAUGUUCCGCUG.CA.AUAAAGAAAGUCUG..UG.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Lmo TTTACA.17.TAACCT.28.AUAACGAUAUUCCGCUU.CAU....UAUUAAU....AUG.AAUGAAUGUUUG.05.AGGAGA.07.AUG
Sau TTGAAA.17.TAACAT.23.AUCACUAUGAUCCGCUG.CU..AUAUAUUUGUCG...AGGCAAGAACAUAGG.04.AGAGGA.09.AUG
Cpe TTAAAG.18.TAAACT.08.GUACCGGCGGUCCUCUGUCACA....GAG......UGUGUUAAGAACGUCAA.17.AGGAGG.08.AUG
Chy TTGCAT.17.TATAAT.09.UACCAAACGUUCCGCUG.GA....CAGGGGC......UC.CAUGAACGUGCC.03.AGGAGG.09.AUG
Swo TTGAGA.17.TAAAAT.16.AAAAAGGUGGUCCGCUG.CAUU..AAACUAA....AAUG.UAUGAACACCUU.05.AGGAGG.07.AUG
Ame TTGCGG.17.TATAAT.10.UUACGGGCGGUCCUCUA.UAC.....AGGA......GUA.UAAGAACGUCUA.07.AGGAGG.07.AUG
Dre TTGCCC.17.TATAAT.16.UUACGGACGGUCCGCUG.CCU....CUGGGAA....AGG.UAAGAACGUCUA.04.AGGAAG.12.GUG
Spn TTTACT.17.TAAACT.28.AUACAGUUUAUCCGCUG.AGGA....AGAU.....UCCU.CAAGAUUGACAA.04.AGGAGA.05.AUG
Smu TTTACA.17.TACAAT.26.AAACGGCUAAUCCGCUG.AG....ACAGAGCA.....CU.UAUGAUUAGUAA.04.AGGAGA.07.AUG
Lpl TTGCGT.18.TATTCT.21.UUAACGAUGUUCCGCUG.AC.....CAGGUU......GU.CACGAAUGUCGG.04.AGGAAG.09.AUG
Efa TTTACA.17.TAAACT.28.AUUACAAUAUUCCGCUG.UGG.CA..GAAG...UGACCA.UAAGAAUAUUUG.06.AGGAGA.08.AUG
Ljo TTTACA.17.TAAACT.25.UUAUGGGUAUUCCGCUG.GCAC....AAG......GUGUUGAUGAAUGCCGU.03.AGGAGA.07.AUG
Sth TAGACA.17.TAAGAT.29.UAACGGCUAAUCCGCUG.AGA.CACAGAGGU..UGCUCU.UAAGAUUAGUAA.03.AGGAGU.08.AUG
Lac TTAAAA.17.TTACTT.39.UUAUGGGUAUUCCGCUG.ACG....CUGGUA.....CGUUGAUGAAUGCCGA.03.AGGAGA.10.AUG
Spy TTTACA.17.TAGAAT.29.UUACGGCUAAUCCGCUA.AG....ACAAGUA......CU.UAAGAUUAGUAA.03.AGGAGA.06.AUG
Lsa TTTTAA.17.TAAAAT.26.ACAACGAUAUUCCGCUG.GCG....CAAGA......CGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG
Lsl TTTACT.17.TATTTT.24.AUAACGAUAUUCCGCUG.C......AACUG........GACAUGAAUGUCGG.04.AGGAAA.07.AUG
Fnu TTGACA.17.TAAAAT.12.AAUUCGAUAUUCCGCUU.UAA.....UAAA......UUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG
```

B

nucleotide identity

- N 97%
- N 90%
- N 75%

nucleotide present

- ● 97%
- ● 90%
- ● 75%
- ○ 50%

stem loop always present

compensatory mutations

compatible mutations

G - C Watson-Crick base pair

G • A other base interaction

C *B. Subtilis* L19 mRNA leader switch?

A CFG for RNA

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

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

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

“s” means unpaired;

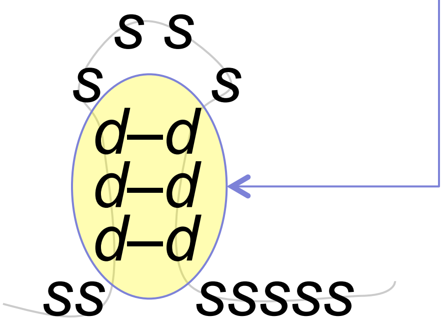
“dFd” means paired

(Watson–Crick:

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

paren-like nesting)

$$
\begin{aligned}
S &\Rightarrow \; LS \Rightarrow^{*} LLLLLLLLS \\
&\Rightarrow \; LLLLLLLL \\
&\Rightarrow^{*} ssLsssss \\
&\Rightarrow \; ssdFdsssss \\
&\Rightarrow \; ssddFddsssss \\
&\Rightarrow \; ssdddFdddsssss \\
&\Rightarrow \; \ldots
\end{aligned}
$$



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

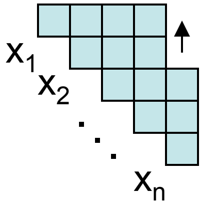
"How likely is it?"

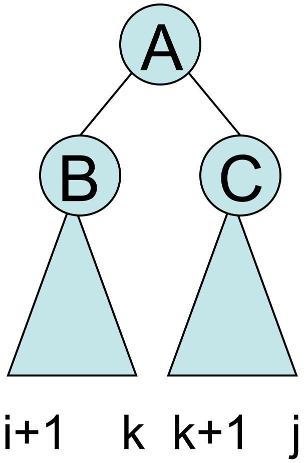
Cocke-Kasami-Younger Parser

Suppose all rules of form A → 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
        $M^A[i,j] = \bigvee_{A \rightarrow BC,\ i < k < j} M^B[i,k] \wedge M^C[k,j]$
```





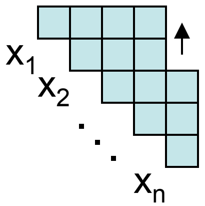
Time: $O(n^3)$

“Inside” Algorithm for SCFG

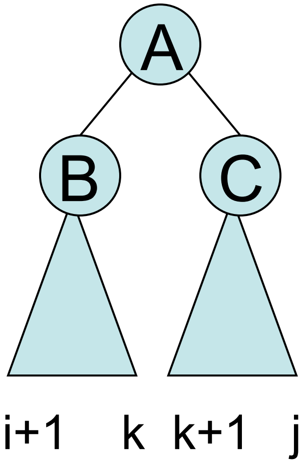
Suppose all rules of form A → BC or A → a
(by mechanically transforming grammar)

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*

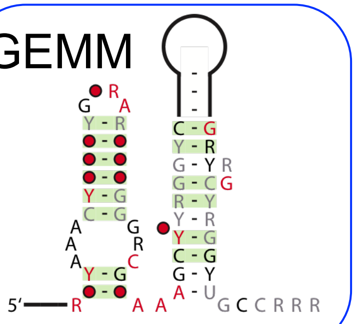


Time: $O(n^3)$

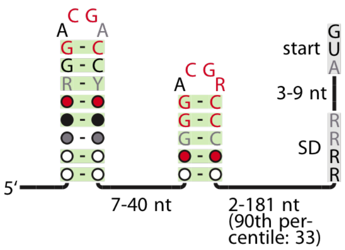
Examples: 7 typical motifs

GEMM

Sudarsan, et al *Science*, 2008



mini-*ykkC*



Legend nucleotides, SD: Shine-Dalgarno
rt: start codon, R: A/G, Y: C/U

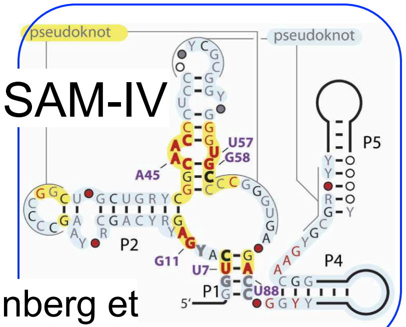
nucleotide identity
N 97%
N 90%
N 75%

nucleotide present
● 97%
● 90%
● 75%
○ 50%

base pair annotations
🟩 has covarying mutations
🟦 has compatible mutations
🟥 no mutations observed
variable hairpin
variable loop
□ modular structure

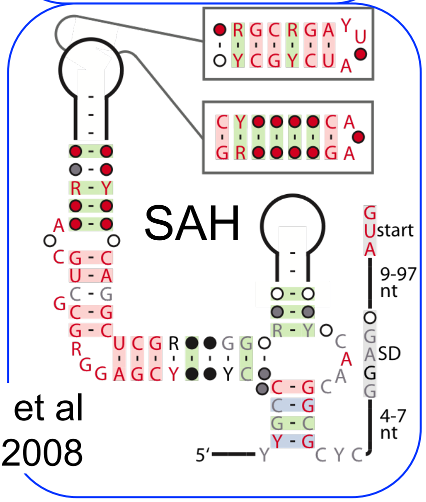
SAM-IV

Weinberg et al *RNA* ’08



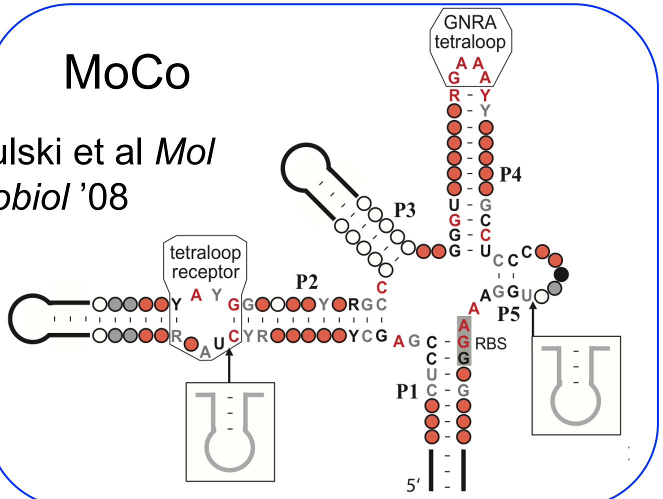
SAH

Wang, et al *Mol Cell*, 2008

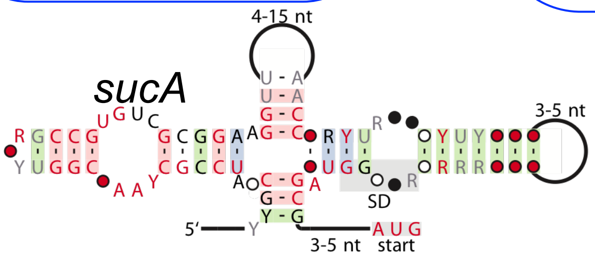


MoCo

Regulski et al *Mol Microbiol* ’08

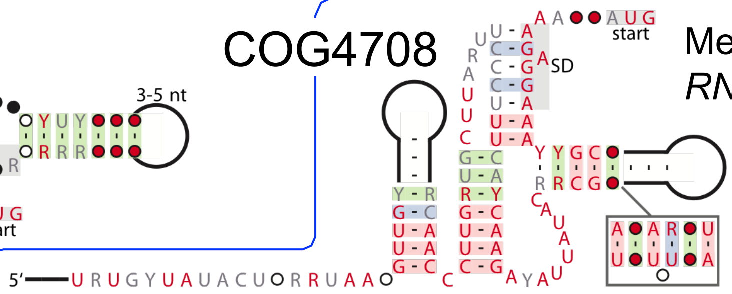


sucA



COG4708

Meyer, et al *RNA*, 2008



boxed = confirmed riboswitch

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”