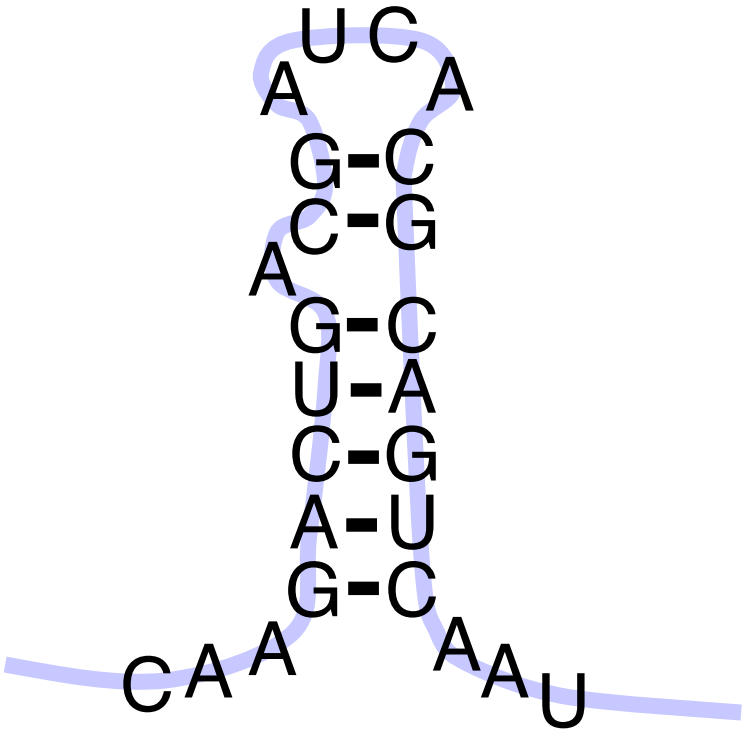
# CSE 427
## Winter 2008

## RNA
### Secondary Structure Prediction

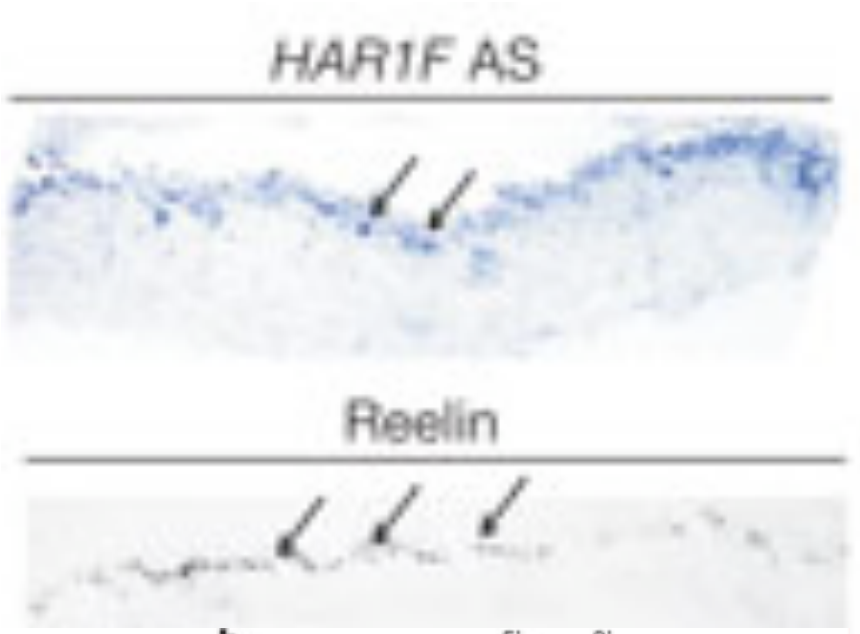
# RNA Secondary Structure:
## RNA makes helices too

Base pairs

A-U
C-G



# Fastest Human Gene?




**a**

| Position | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| Human | AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT | | | |
| Chimpanzee | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Gorilla | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Orang-utan | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Macaque | AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT | | | |
| Mouse | AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT | | | |
| Dog | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Cow | AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT | | | |
| Platypus | ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT | | | |
| Opossum | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Chicken | AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT | | | |
| Fold | ((((((((.....)))))).....))[[[[[.((((.(((....))))).)))) | | | |
| Pair symbol | lmnopqr rqpon ml rstuvwx xwvutsr | | | |

**b**

# Origin of Life?

Life needs

information carrier: DNA

molecular machines, like enzymes: Protein

making proteins needs DNA + RNA + proteins

making (duplicating) DNA needs proteins

Horrible circularities! How could it have arisen in an abiotic environment?

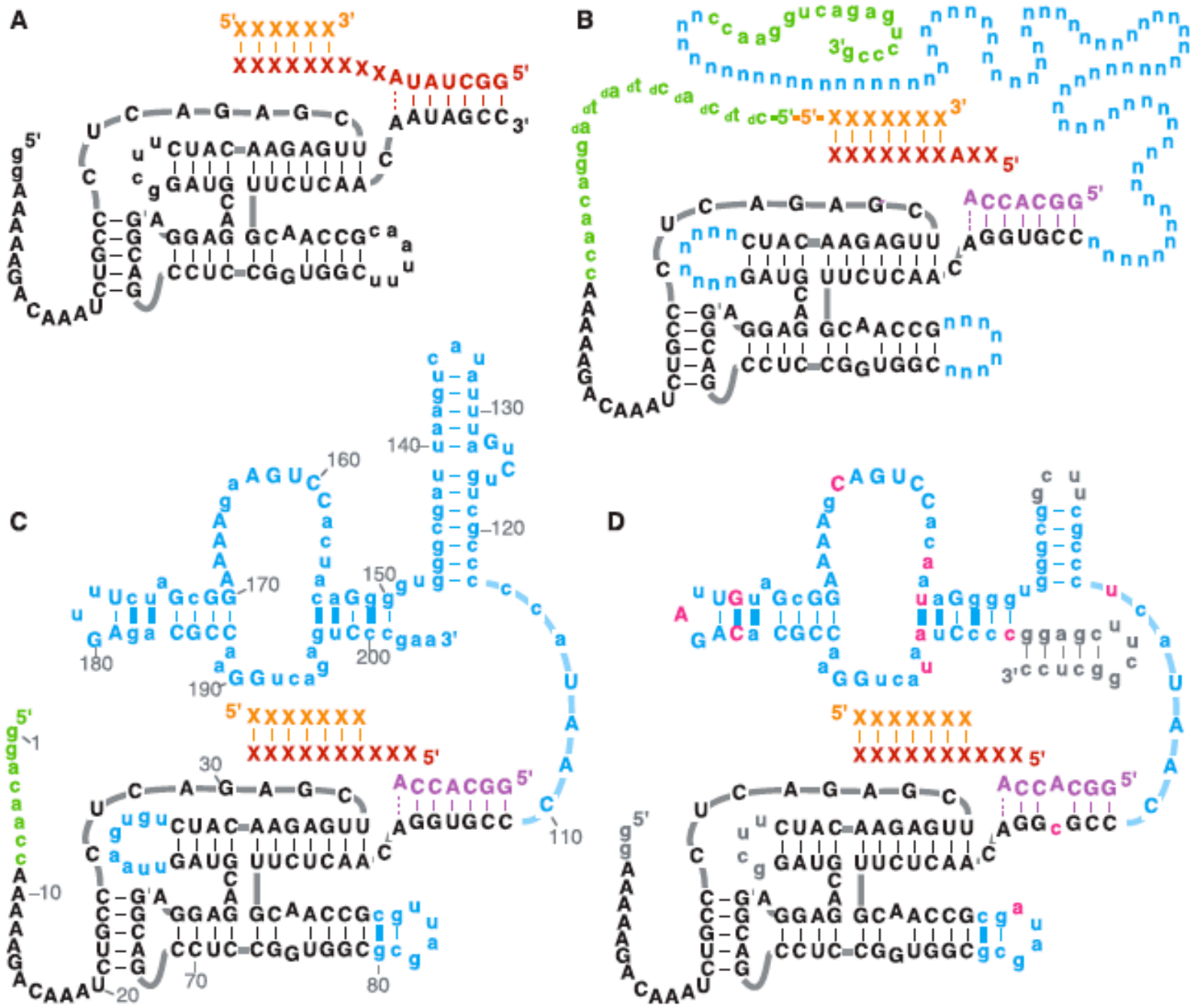
# Origin of Life?

RNA can carry information too
(RNA double helix)

RNA can form complex structures

RNA enzymes exist (ribozymes)

The "RNA world" hypothesis:
1st life was RNA-based


A
B
C
D

| Round | Mutagenesis | Template RNA | NTPs | Time (hour) | Selection criteria |
|---|---|---|---|---|---|
| 1 | Synthesis | 3′-GGUCAGAUU | $^{4S}$UTP (2 mM) | 36 | $^{4S}$U |
| 2 | None | 3′-GGUCAGAACC | $^{4S}$UTP (2 mM) | 20 | $^{4S}$U |
| 3 | None | 3′-GGUCAGAA | $^{4S}$UTP (2 mM) | 20 | $^{4S}$U |
| 4 | None | 3′-CUUAGUUCAUU | $^{4S}$UTP (2 mM) | 19 | $^{4S}$U |
| 5 | None | 3′-CUUAGUUCAUU | $^{4S}$UTP (2 mM) | 1 | $^{4S}$U |
| 6 | None | 3′-GGUCAGAUU | $^{4S}$UTP, $^{B}$ATP (1 mM each) | 14 | $^{B}$A, $^{4S}$U |
| 7 | None | 3′-CUUAGUUCAUU | $^{4S}$UTP, $^{B}$ATP (1 mM each) | 17 | $^{B}$A, $^{4S}$U |
| 8 | None | 3′-GGUCAGAUU | $^{4S}$UTP, $^{B}$ATP (1 mM each) | 17 | $^{B}$A, $^{4S}$U |
| 9 | None | 3′-GGUCAGAUU | $^{4S}$UTP, $^{B}$ATP (1 mM each) | 4 | $^{B}$A, $^{4S}$U |
| 10 | None | 3′-CUUAGUUCAUU | $^{4S}$UTP (1 mM) | 20 | $^{4S}$U |
| 11 | Synthesis | 3′-UCGACGGAACC | $^{4S}$UTP (1 mM) | 4 | 2 $^{4S}$U |
| 12 | None | 3′-ACCUGAGAAGG | $^{4S}$UTP (1 mM) | 4 | 2 $^{4S}$U |
| 13 | None | 3′-CAAGUCCAACC | $^{4S}$UTP (1 mM) | 0.2 | 2 $^{4S}$U |
| 14 | None | 3′-UCGACGGAACC | $^{4S}$UTP (1 mM) | 0.2 | 2 $^{4S}$U |
| 15 | PCR | 3′-UCGACGG$^{2N}$p$^{2N}$PCCUGCGUC | $^{4S}$UTP (0.1 mM), Comp. NTPs | 20 | 2 $^{4S}$U |
| 16 | PCR | 3′-CAAGUCC$^{2N}$p$^{2N}$PUGAUCGUA | $^{4S}$UTP (0.1 mM), Comp. NTPs | 4 | 2 $^{4S}$U |
| 17 | PCR | 3′-ACCUGAG$^{2N}$p$^{2N}$PGUGUAUGU | $^{4S}$UTP (0.1 mM), Comp. NTPs | 2 | 2 $^{4S}$U |
| 18 | None | 3′-UCGACGG$^{2N}$p$^{2N}$PCCUGCGUC | $^{4S}$UTP (0.1 mM), Comp. NTPs | 0.1 | 2 $^{4S}$U |

# Outline

Biological roles for RNA

What is “secondary structure?

How is it represented?

Why is it important?

Examples

Approaches

# RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

# RNA Pairing

Watson-Crick Pairing

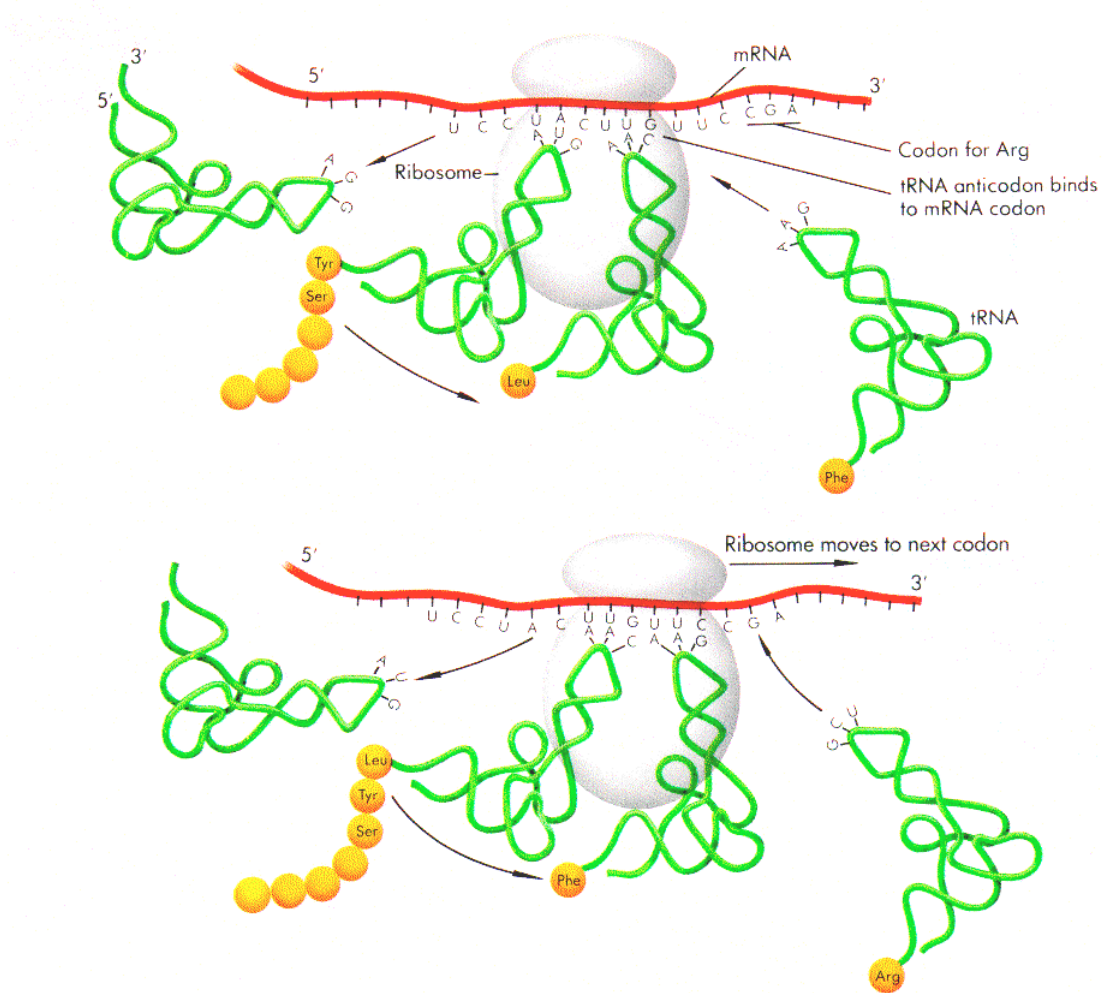
C - G ~ 3 kcal/mole

A - U ~ 2 kcal/mole

"Wobble Pair" G - U ~1 kcal/mole

Non-canonical Pairs (esp. if modified)

# Ribosomes




Watson, Gilman, Witkowski, & Zoller, 1992

# Ribosomes

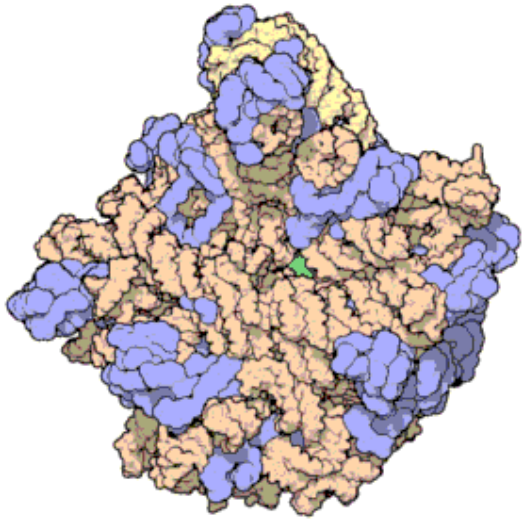
1974 Nobel prize to Romanian biologist George Palade for discovery in mid 50's

50-80 proteins

3-4 RNAs (half the mass)

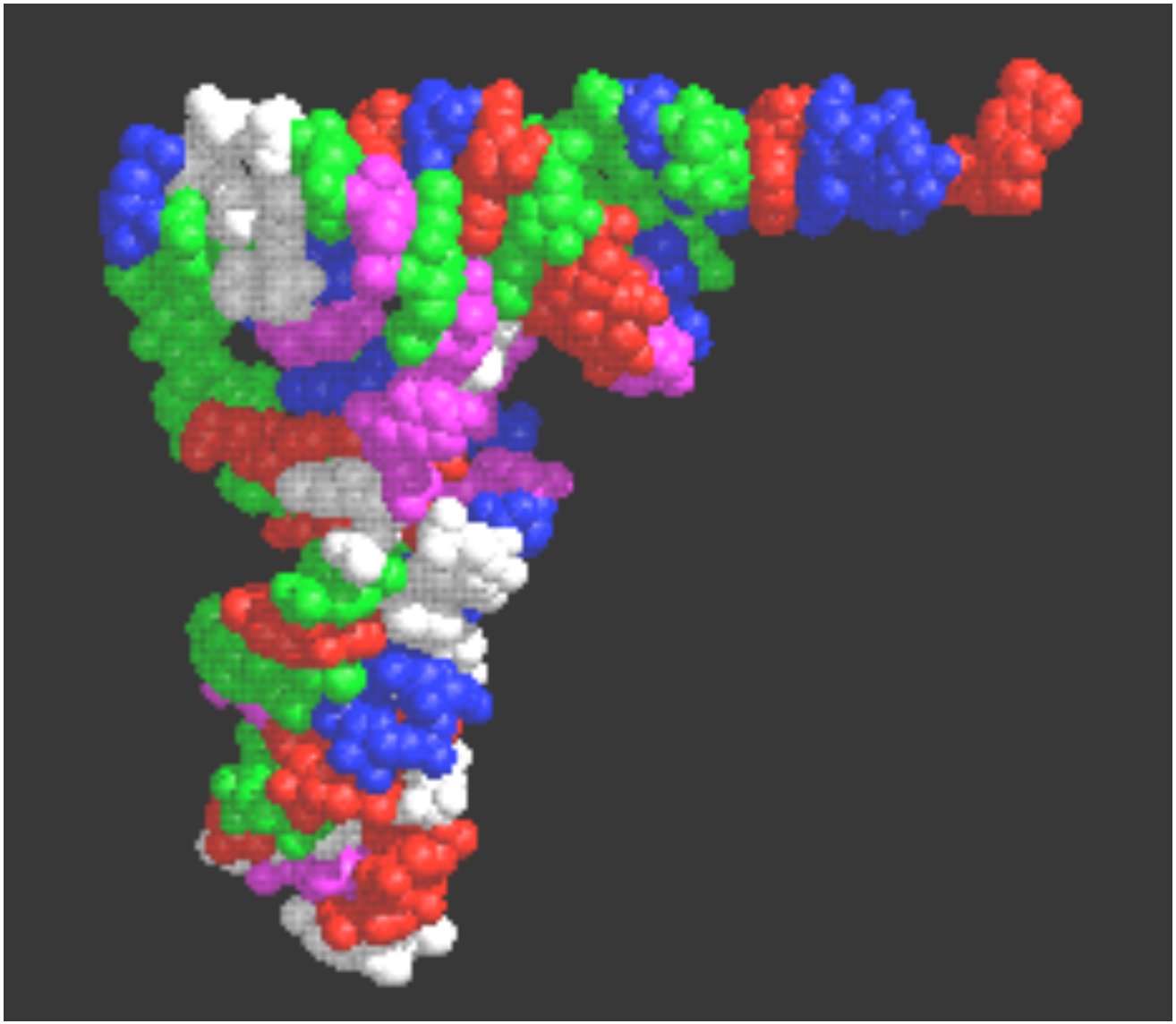
Catalytic core is RNA

Of course, mRNAs and tRNAs (messenger & transfer RNAs) are critical too

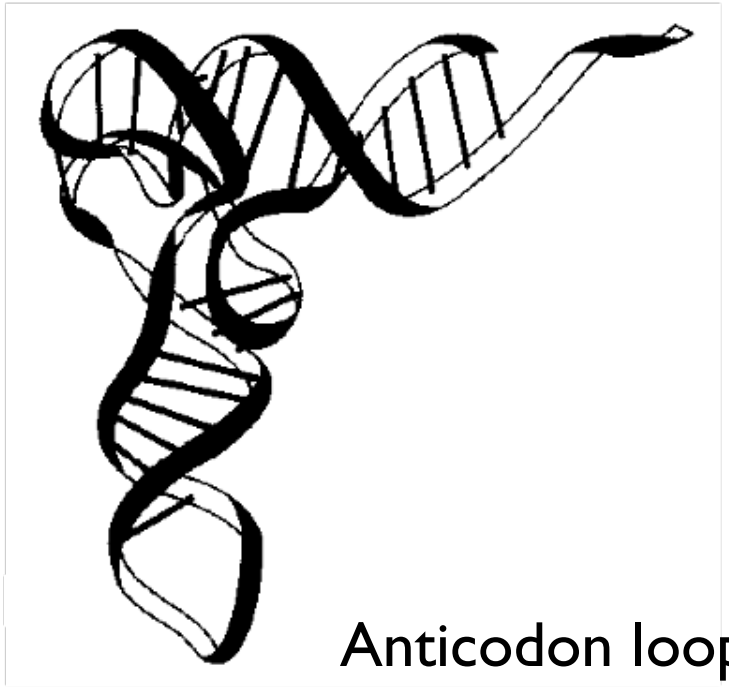


Atomic structure of the 50S Subunit from *Haloarcula marismortui*. Proteins are shown in blue and the two RNA strands in orange and yellow. The small patch of green in the center of the subunit is the active site.
- Wikipedia

# tRNA 3d Structure



# tRNA - Alt. Representations



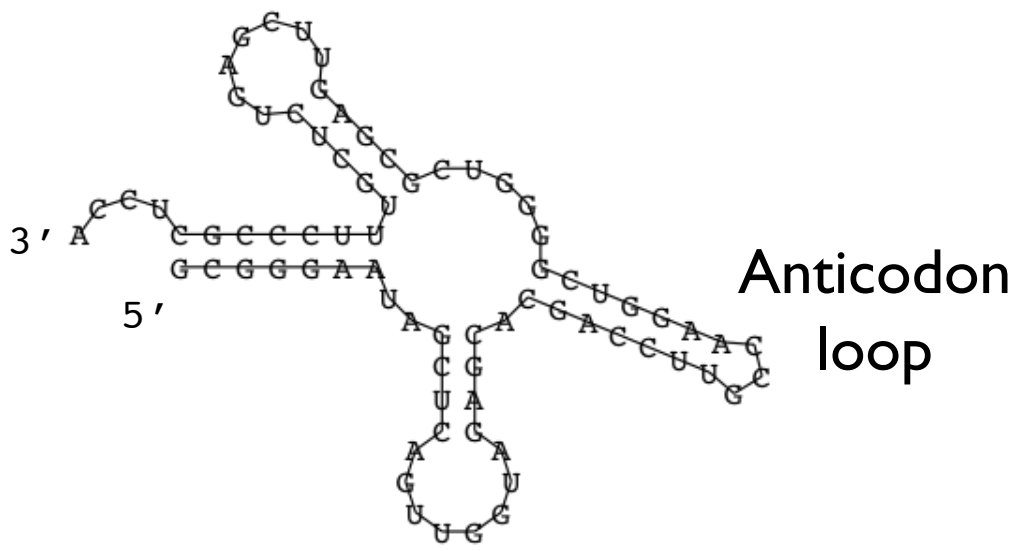
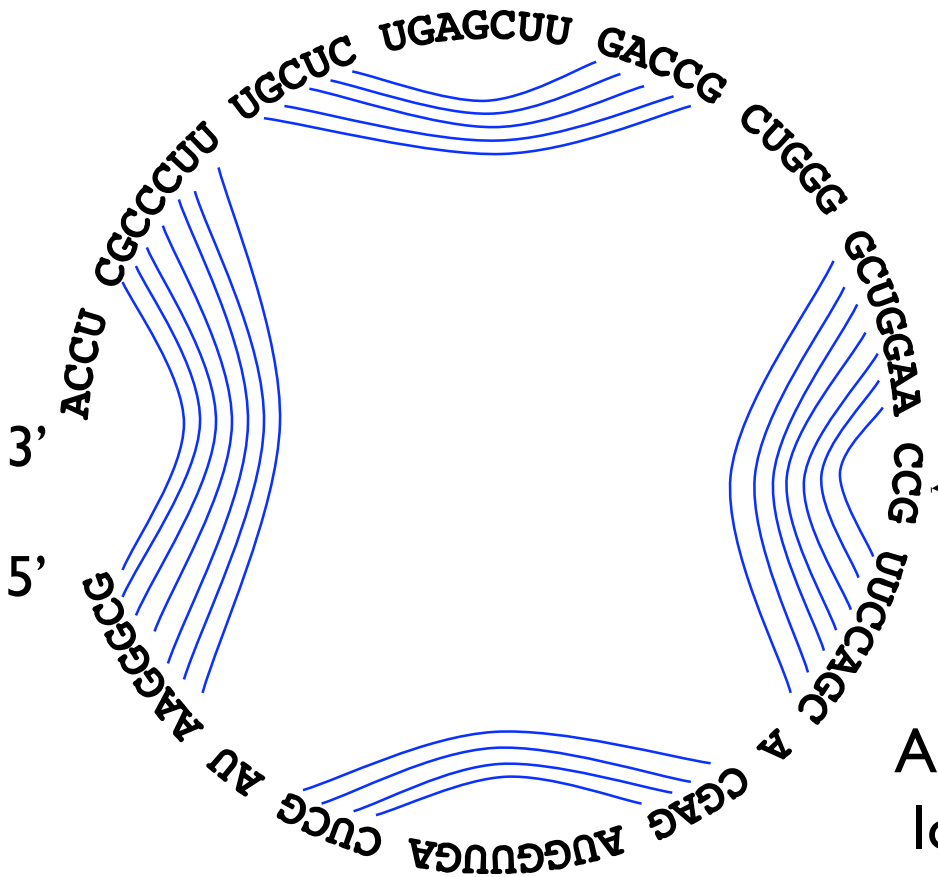



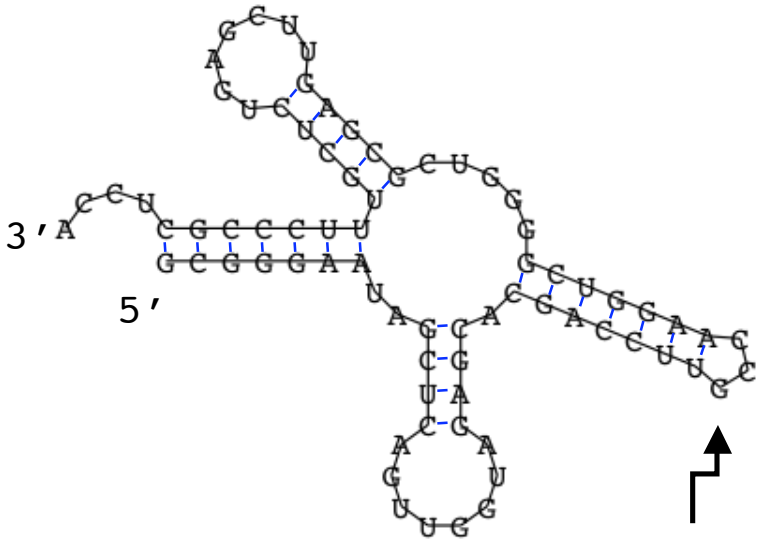

**Figure 1: a)** The spatial structure of the phenylalanine tRNA form yeast

**b)** The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.

# tRNA - Alt. Representations




Anticodon loop




Anticodon loop

# "Classical" RNAs

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

RNase MRP - rRNA processing; mito. rep. (~225 nt)

SRP - signal recognition particle; membrane targeting (~100-300 nt)

SECIS - selenocysteine insertion element (~65nt)

6S - ? (~175 nt)

# Semi-classical RNAs
## (discovery in mid 90's)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)

# Recent discoveries

microRNAs (Nobel prize 2006, Fire & Mello)

riboswitches

many ribozymes

regulatory elements

…

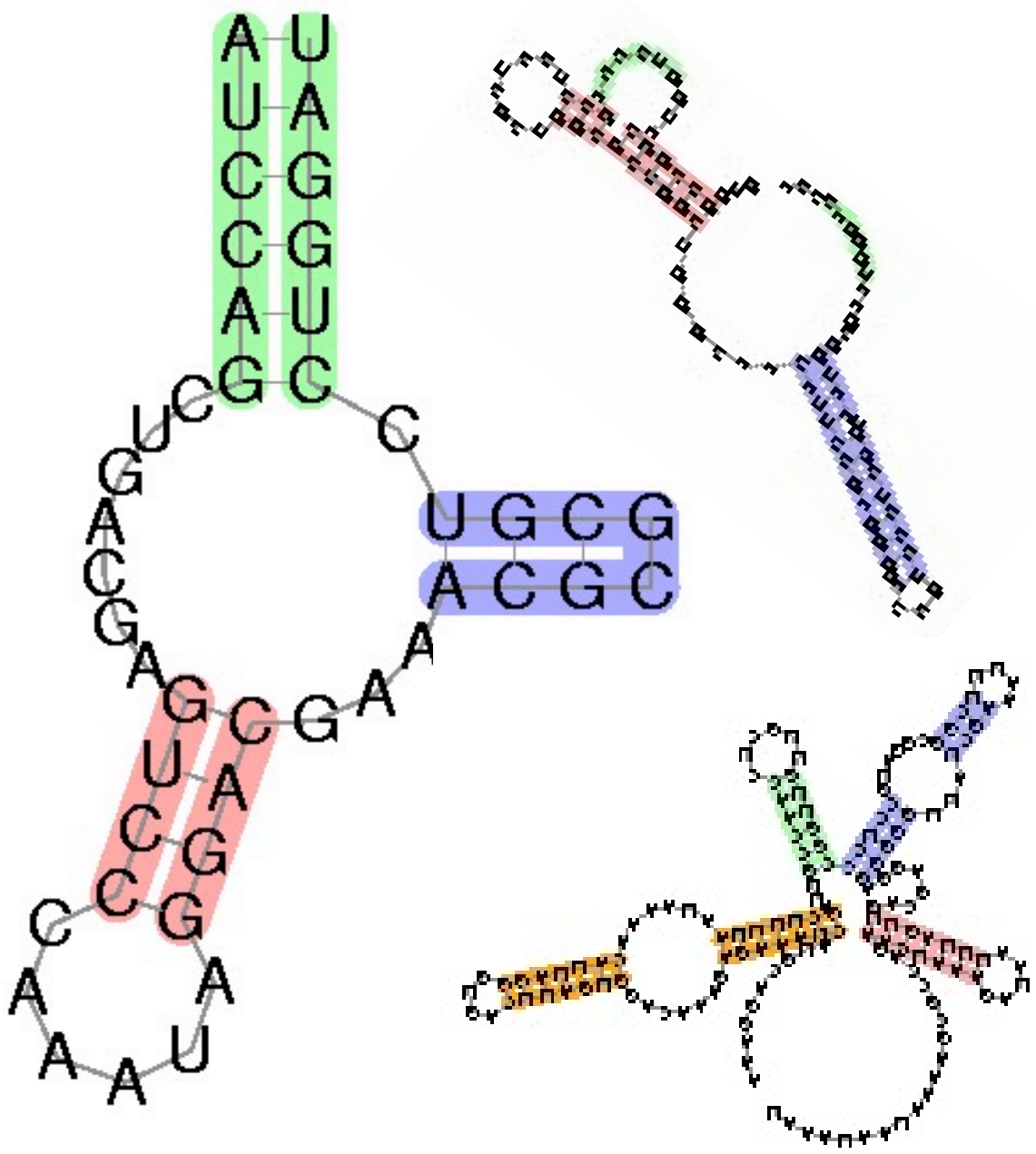
*Hundreds* of families

Rfam release 1, 1/2003: 25 families, 55k instances
Rfam release 7, 3/2005: 503 families, 300k instances

# Why?

RNA's fold, and function

Nature uses what works




# Noncoding RNAs

Dramatic discoveries in last 5 years

*100s* of new families

Many roles: Regulation, transport, stability, catalysis, …

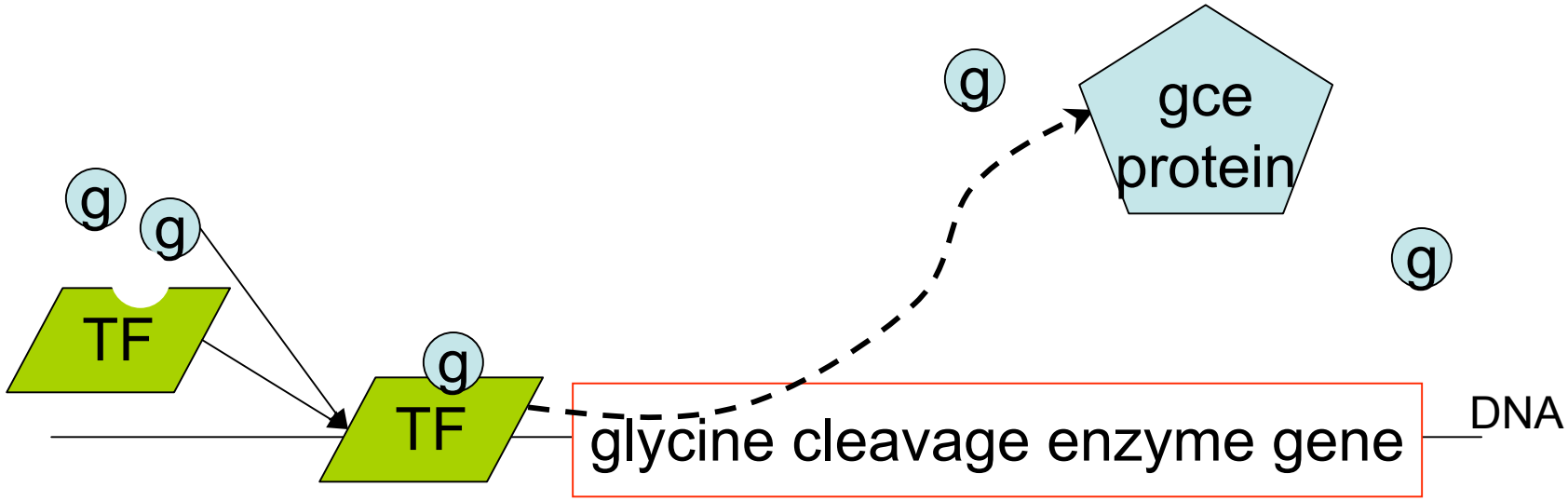
*1% of DNA codes for protein, but 90% of it is copied into RNA, i.e. ncRNA >> mRNA*

*Significance unclear, controversial*

# Example: Glycine Regulation

How is glycine level regulated?

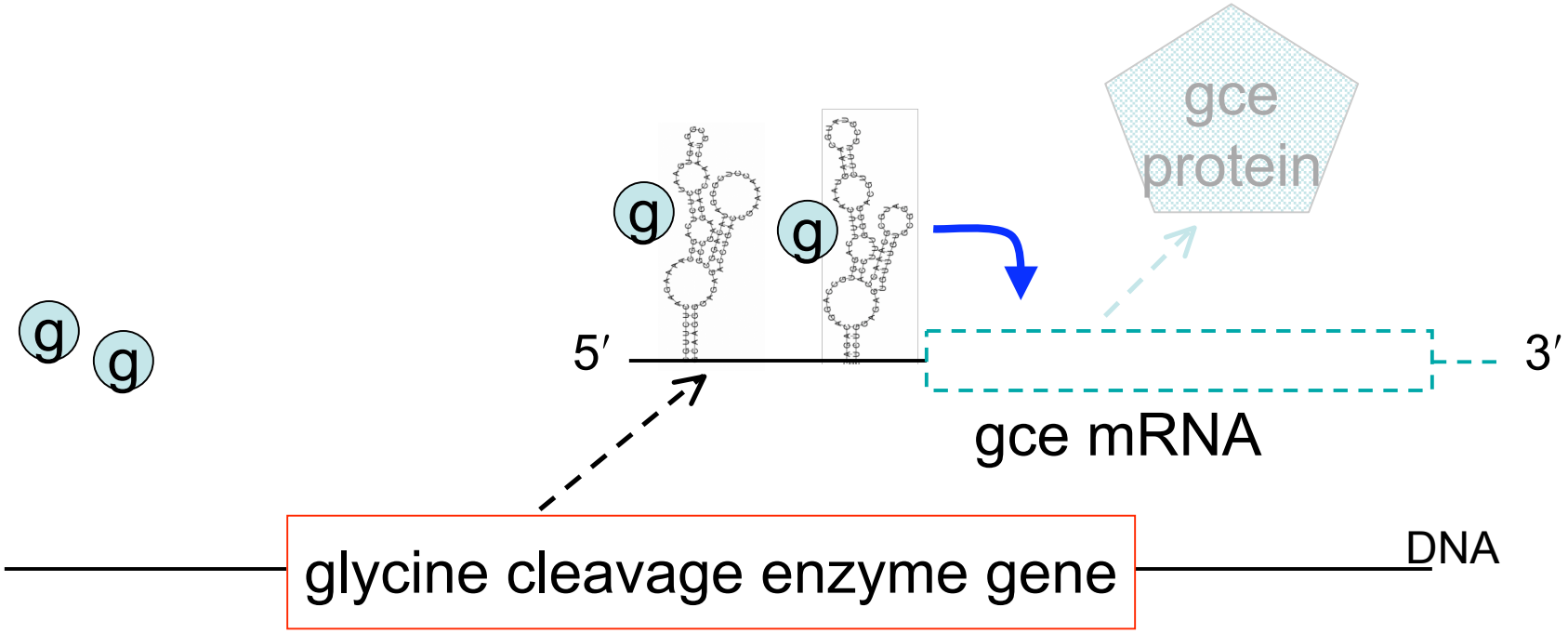
Plausible answer:



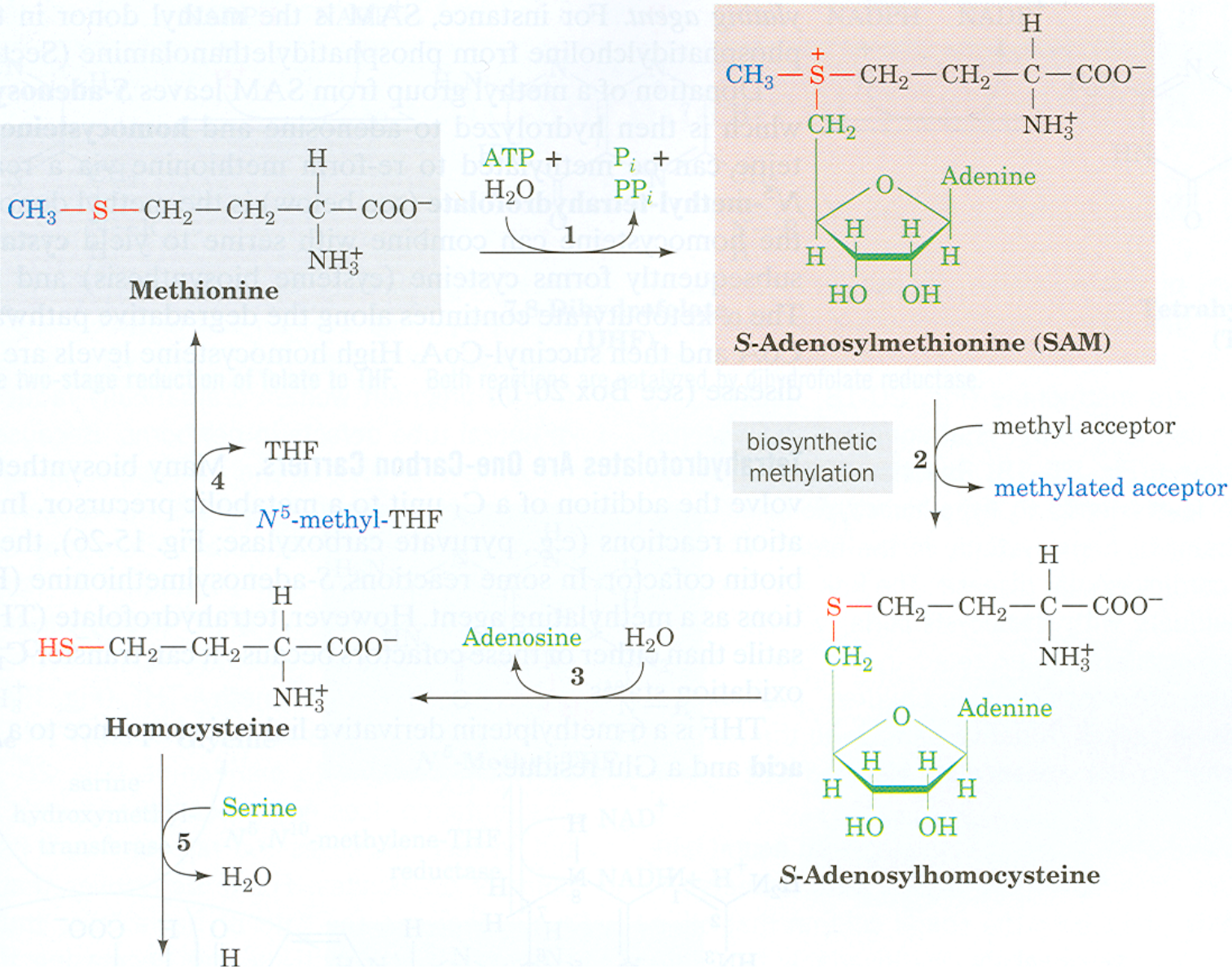

transcription factors (proteins) bind to DNA to turn nearby genes on or off

# The Glycine Riboswitch

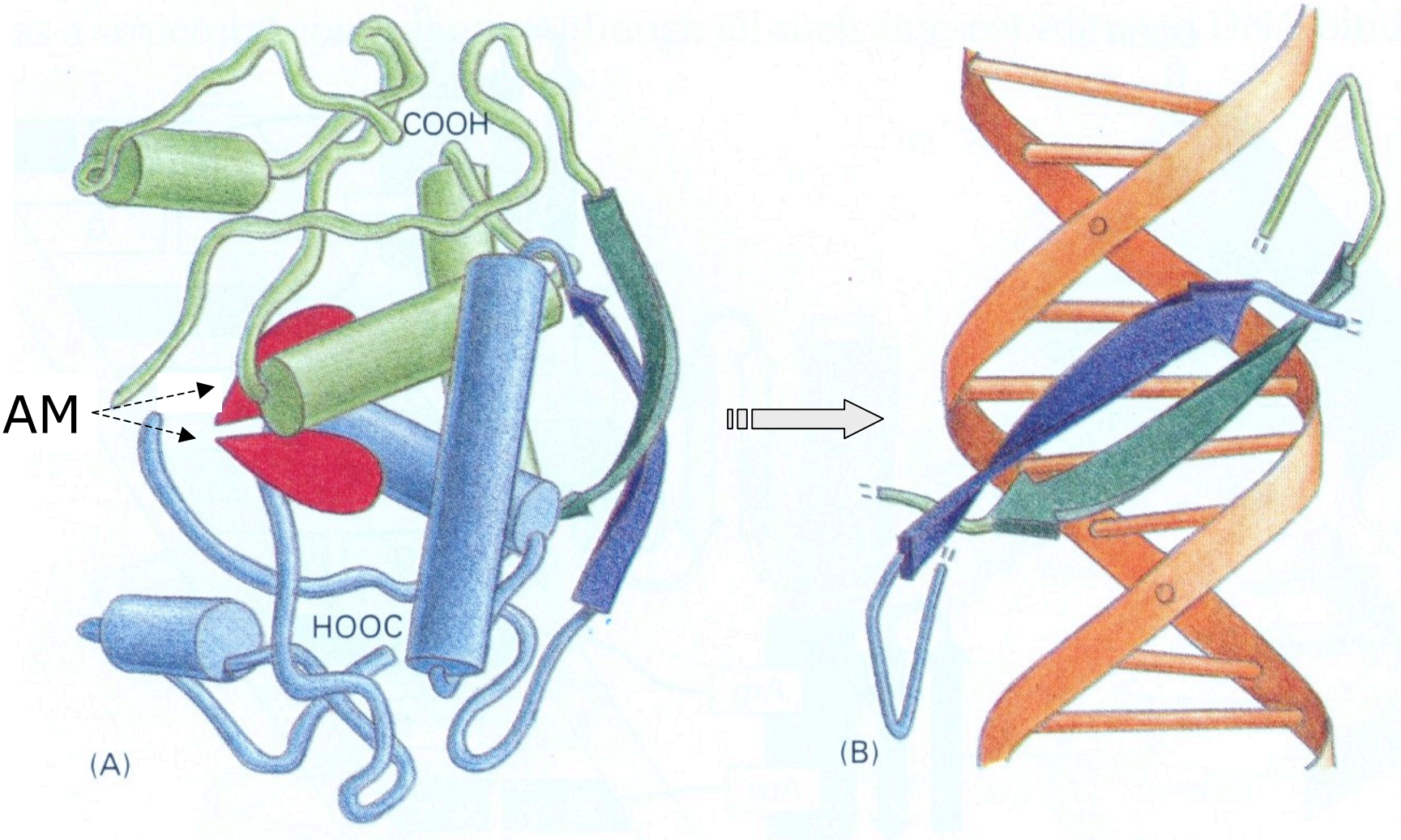
Actual answer (in many bacteria):




Mandal et al. Science 2004


Methionine
ATP + H2O
Pi + PPi
1
S-Adenosylmethionine (SAM)
Adenine
biosynthetic methylation
2
methyl acceptor
methylated acceptor
THF
4
N5-methyl-THF
Homocysteine
Adenosine
H2O
3
S-Adenosylhomocysteine
Serine
5
H2O

# Gene Regulation: The MET Repressor

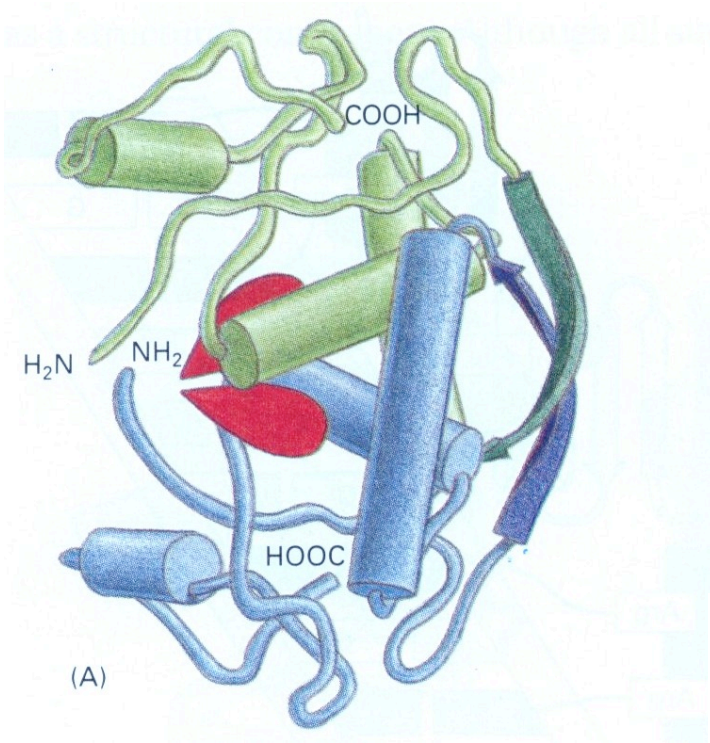


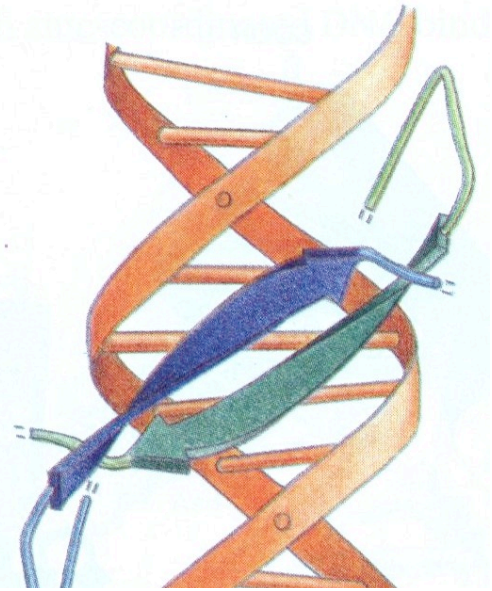

Protein

Alberts, et al, 3e.

DNA

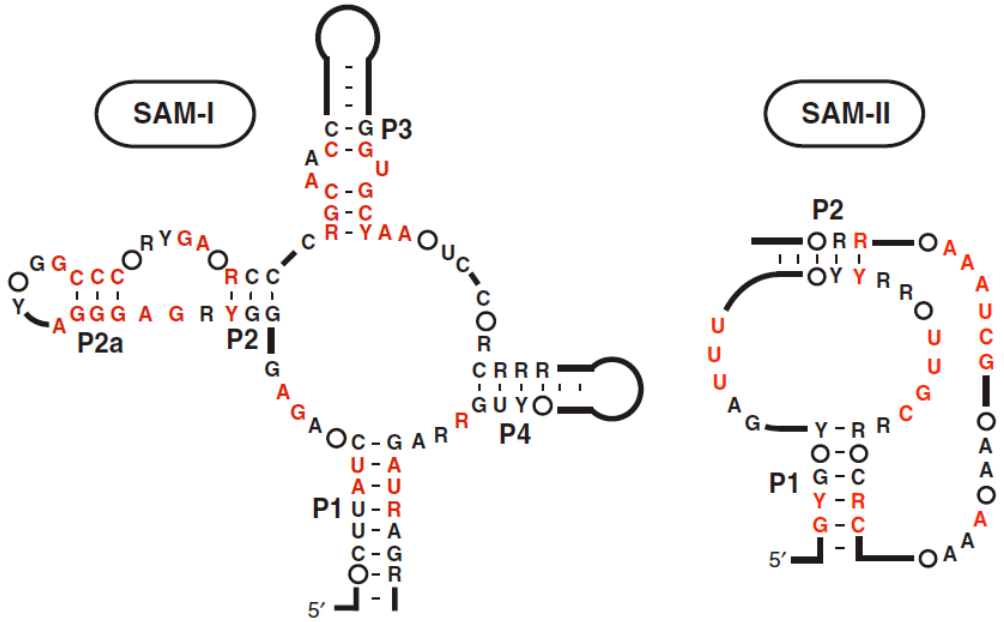
Alberts, et al, 3e.



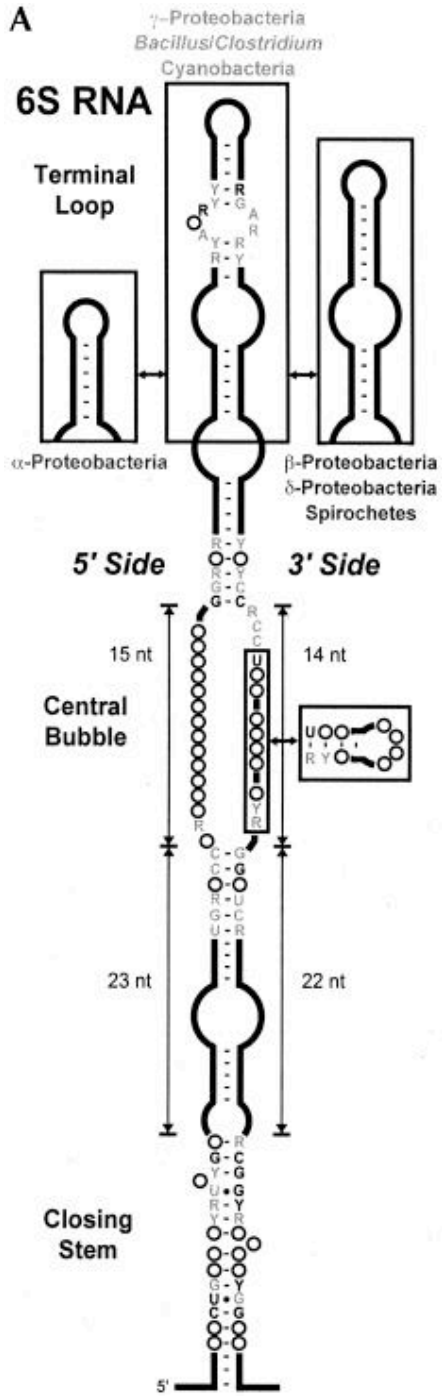
The protein way

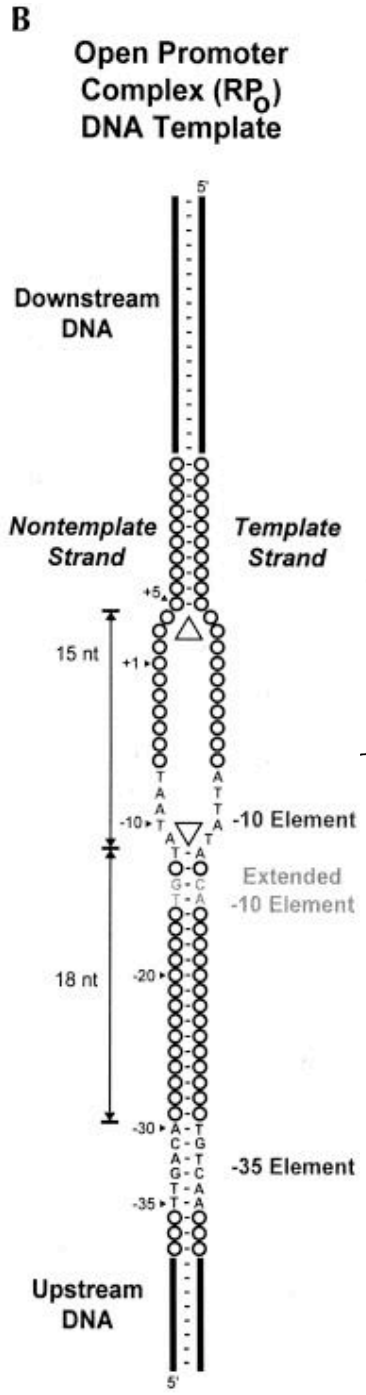
Riboswitch alternatives

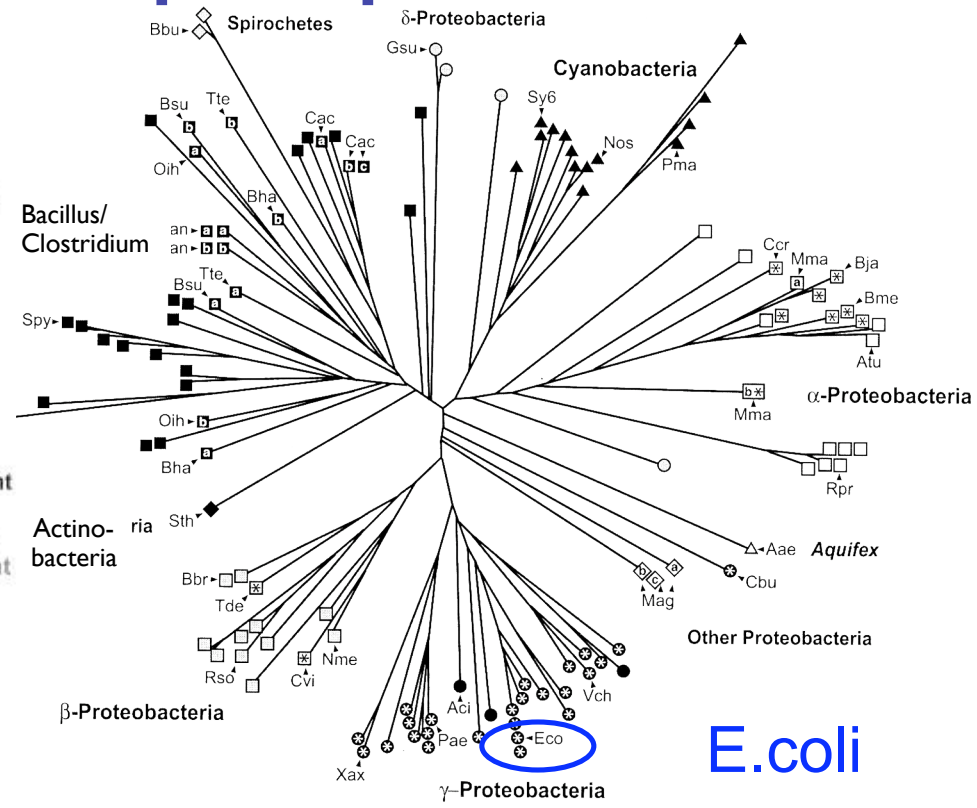



Corbino et al., Genome Biol. 2005

# 6S mimics an open promoter



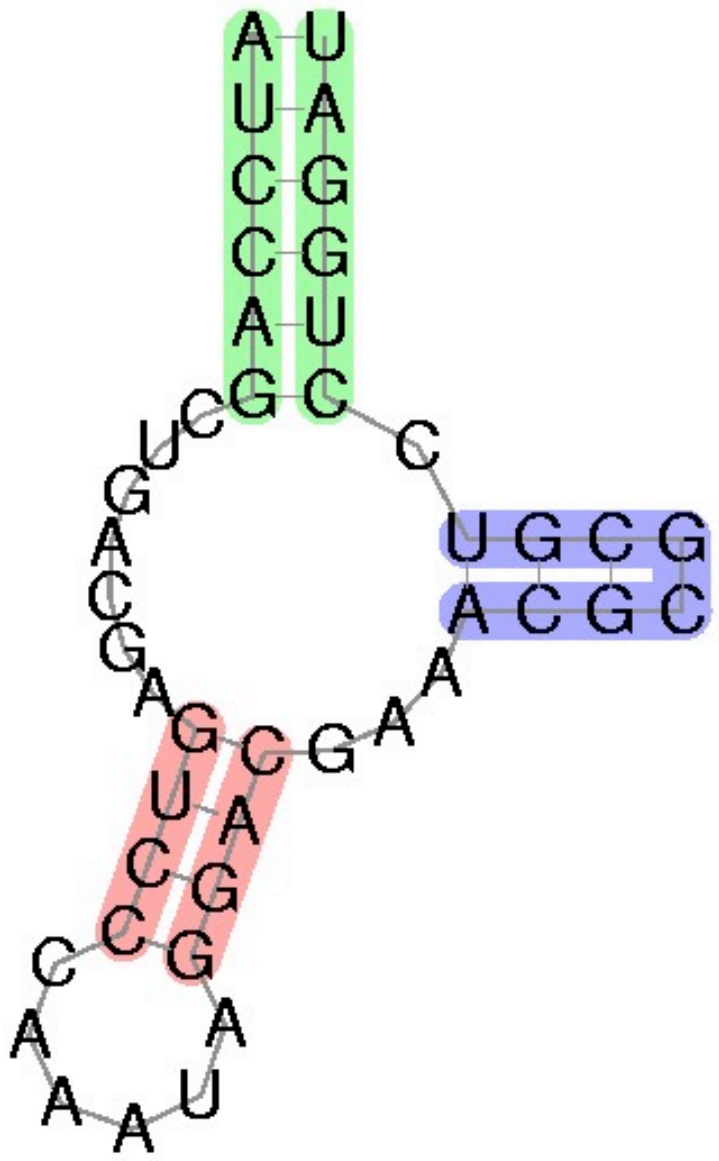




Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005

# The Hammerhead Ribozyme

Involved in "rolling circle replication" of viruses.



# Wanted

Good structure prediction tools

Good motif descriptions/models

Good, fast search tools

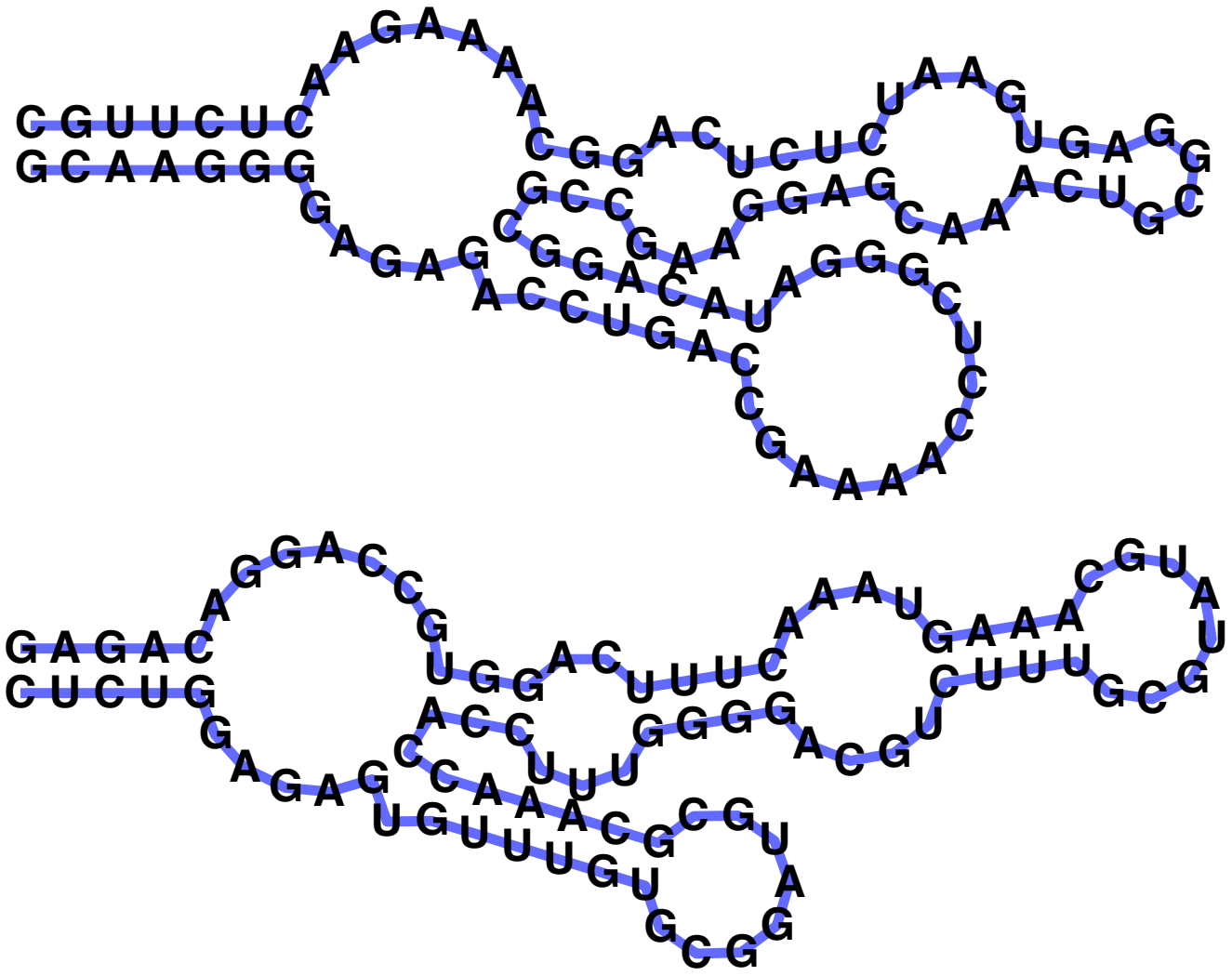
("RNA BLAST", etc.)

Good, fast motif discovery tools

("RNA MEME", etc.)

Importance of structure makes last 3 hard

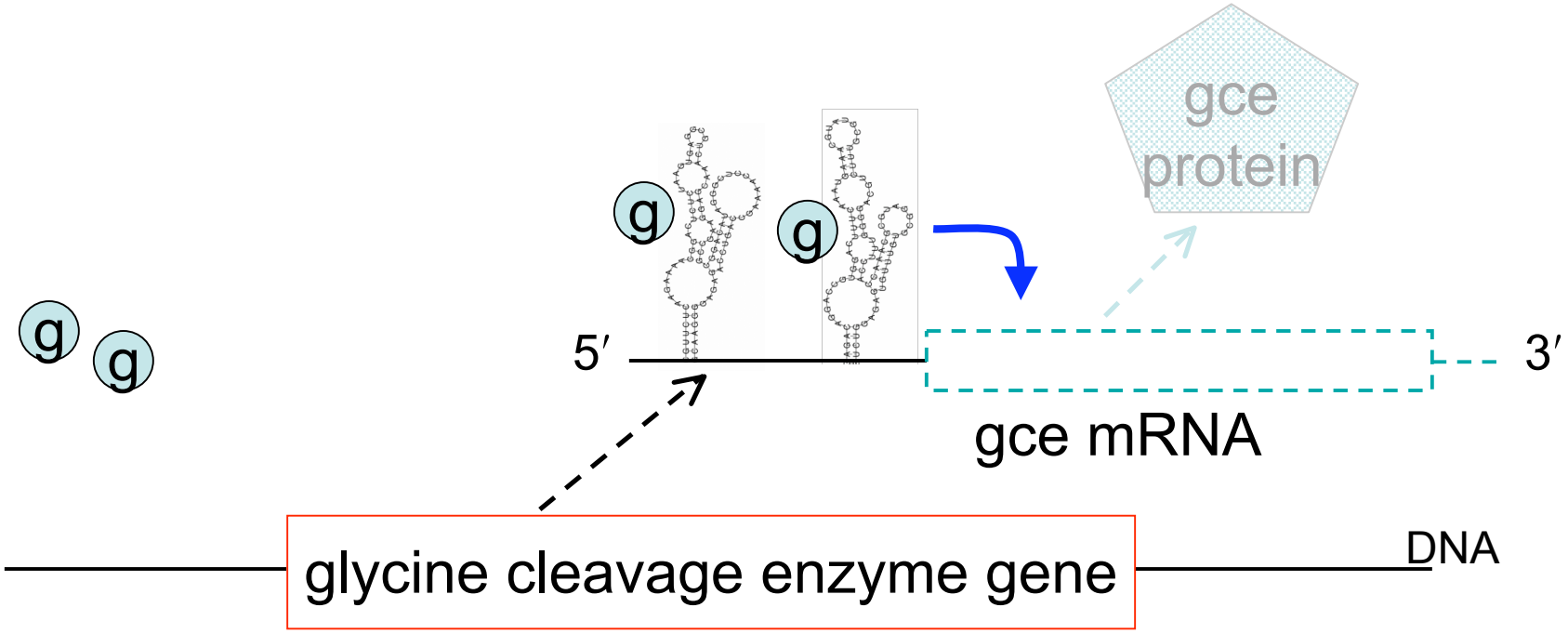
# Why is RNA hard to deal with?



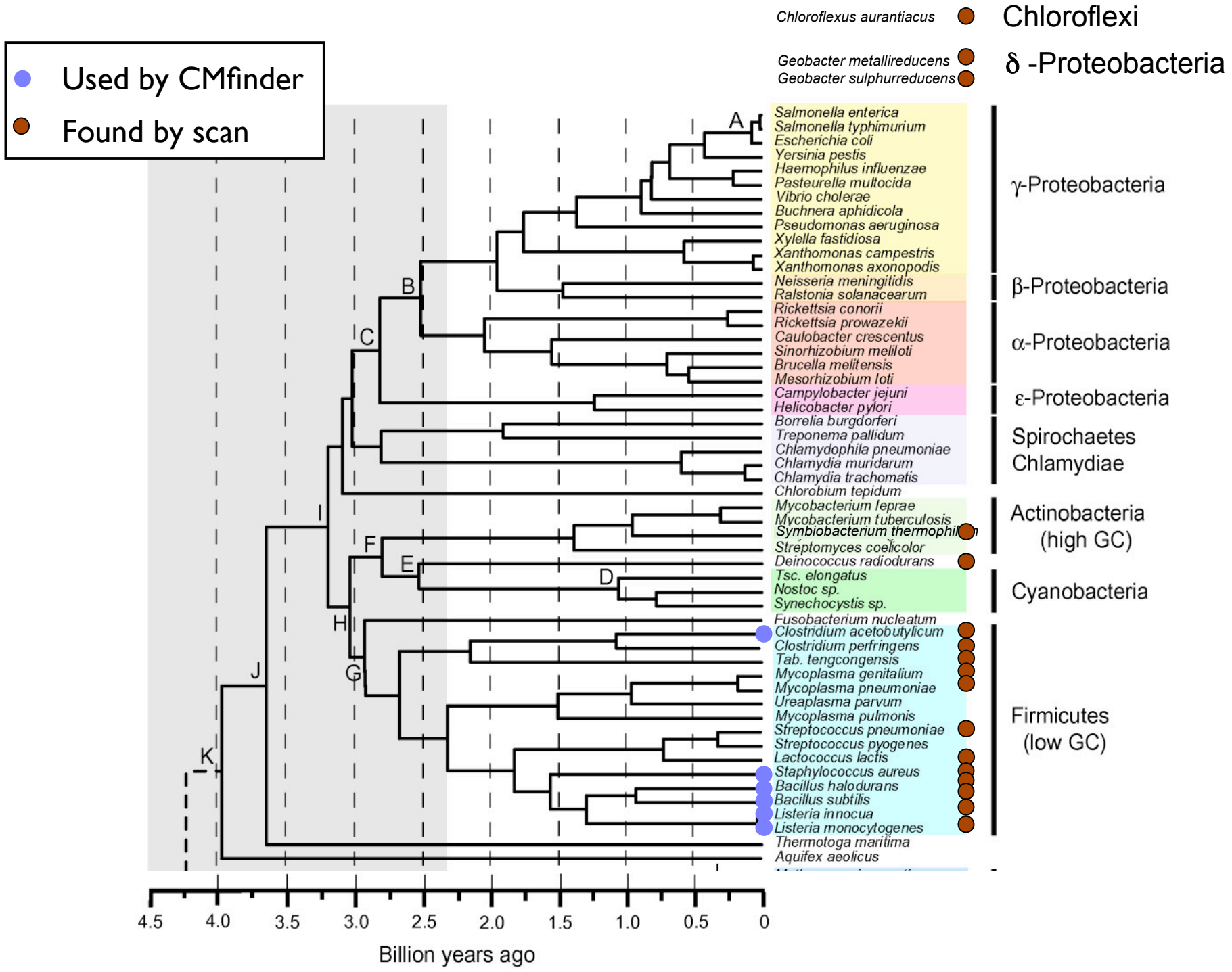
A: *Structure* often more important than *sequence*

# The Glycine Riboswitch

Actual answer (in many bacteria):




Mandal et al. Science 2004


Used by CMfinder
Found by scan
Chloroflexus aurantiacus
Chloroflexi
Geobacter metallireducens
Geobacter sulphurreducens
δ -Proteobacteria
Salmonella enterica
Salmonella typhimurium
Escherichia coli
Yersinia pestis
Haemophilus influenzae
Pasteurella multocida
Vibrio cholerae
Buchnera aphidicola
Pseudomonas aeruginosa
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
γ-Proteobacteria
Neisseria meningitidis
Ralstonia solanacearum
β-Proteobacteria
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Sinorhizobium meliloti
Brucella melitensis
Mesorhizobium loti
α-Proteobacteria
Campylobacter jejuni
Helicobacter pylori
ε-Proteobacteria
Borrelia burgdorferi
Treponema pallidum
Chlamydophila pneumoniae
Chlamydia muridarum
Chlamydia trachomatis
Spirochaetes
Chlamydiae
Chlorobium tepidum
Mycobacterium leprae
Mycobacterium tuberculosis
Symbiobacterium thermophil
Streptomyces coelicolor
Actinobacteria
(high GC)
Deinococcus radiodurans
Tsc. elongatus
Nostoc sp.
Synechocystis sp.
Cyanobacteria
Fusobacterium nucleatum
Clostridium acetobutylicum
Clostridium perfringens
Tab. tengcongensis
Mycoplasma genitalium
Mycoplasma pneumoniae
Ureaplasma parvum
Mycoplasma pulmonis
Streptococcus pneumoniae
Streptococcus pyogenes
Lactococcus lactis
Staphylococcus aureus
Bacillus halodurans
Bacillus subtilis
Listeria innocua
Listeria monocytogenes
Firmicutes
(low GC)
Thermotoga maritima
Aquifex aeolicus
A
B
C
D
E
F
G
H
I
J
K
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
Billion years ago

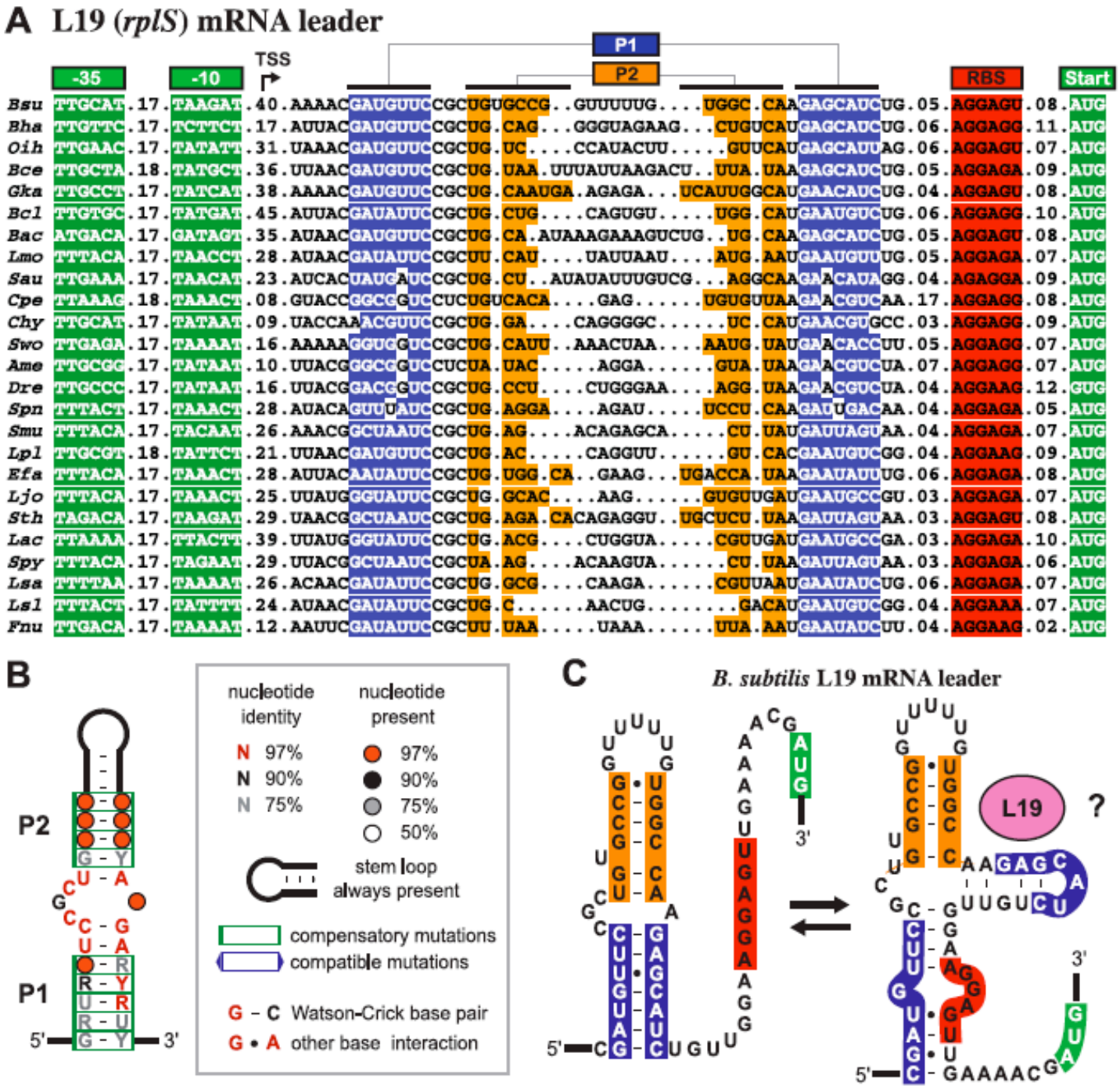
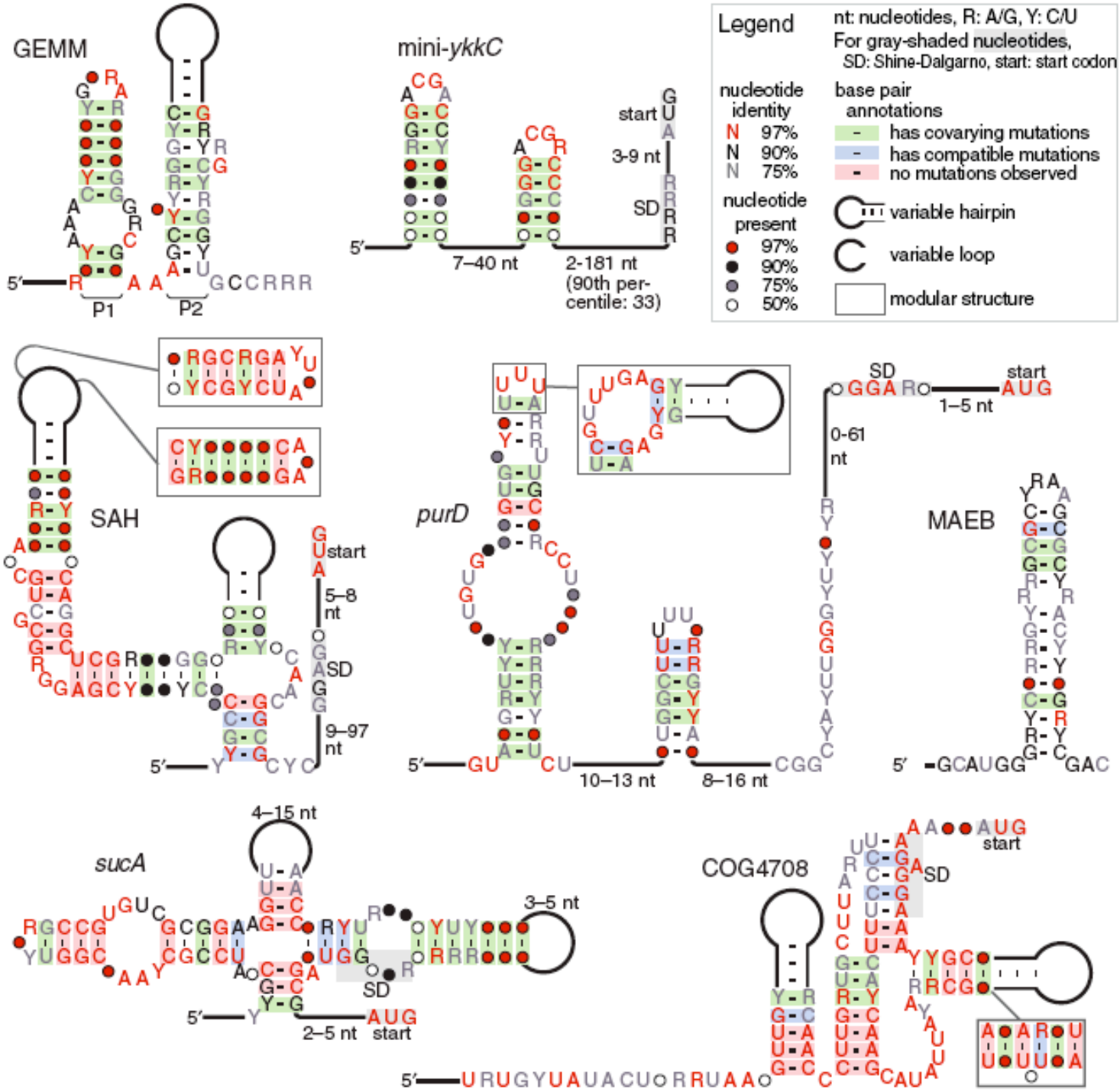



**Figure 3.** Putative Autoregulatory Structure in L19 mRNA Leaders


GEMM
P1
P2
mini-ykkC
start
3-9 nt
SD
7–40 nt
2-181 nt
(90th per-
centile: 33)
Legend
nt: nucleotides, R: A/G, Y: C/U
For gray-shaded nucleotides,
SD: Shine-Dalgarno, start: start codon
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
SAH
5–8 nt
9–97 nt
purD
10–13 nt
8–16 nt
SD
1–5 nt
0-61 nt
MAEB
sucA
4–15 nt
3–5 nt
2–5 nt
COG4708

# Task 1:
# Structure Prediction

# RNA Pairing

Watson-Crick Pairing

C - G ~ 3 kcal/mole

A - U ~ 2 kcal/mole

“Wobble Pair” G - U ~ 1 kcal/mole

Non-canonical Pairs (esp. if modified)

# Definitions

Sequence $^{5'}\ r_1\ r_2\ r_3 \ldots r_n\ ^{3'}$ in {A, C, G, T}

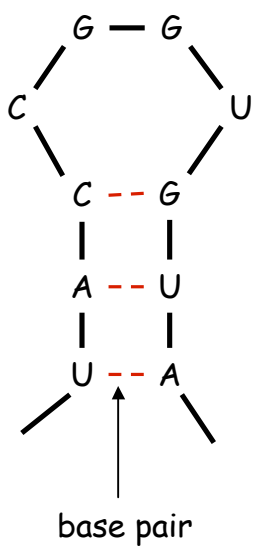
A Secondary Structure is a set of pairs i•j s.t.

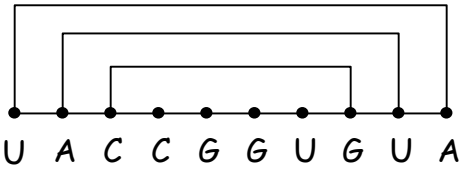
- i < j-4, and } no sharp turns
- if i•j & i'•j' are two different pairs with i ≤ i', then
  - j < i', or
  - i < i' < j' < j

  } 2nd pair follows 1st, or is nested within it; no "pseudoknots."

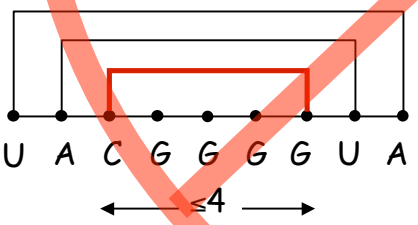
# RNA Secondary Structure: Examples

## Examples.

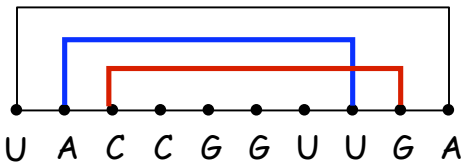


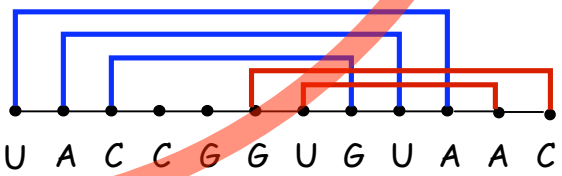




ok

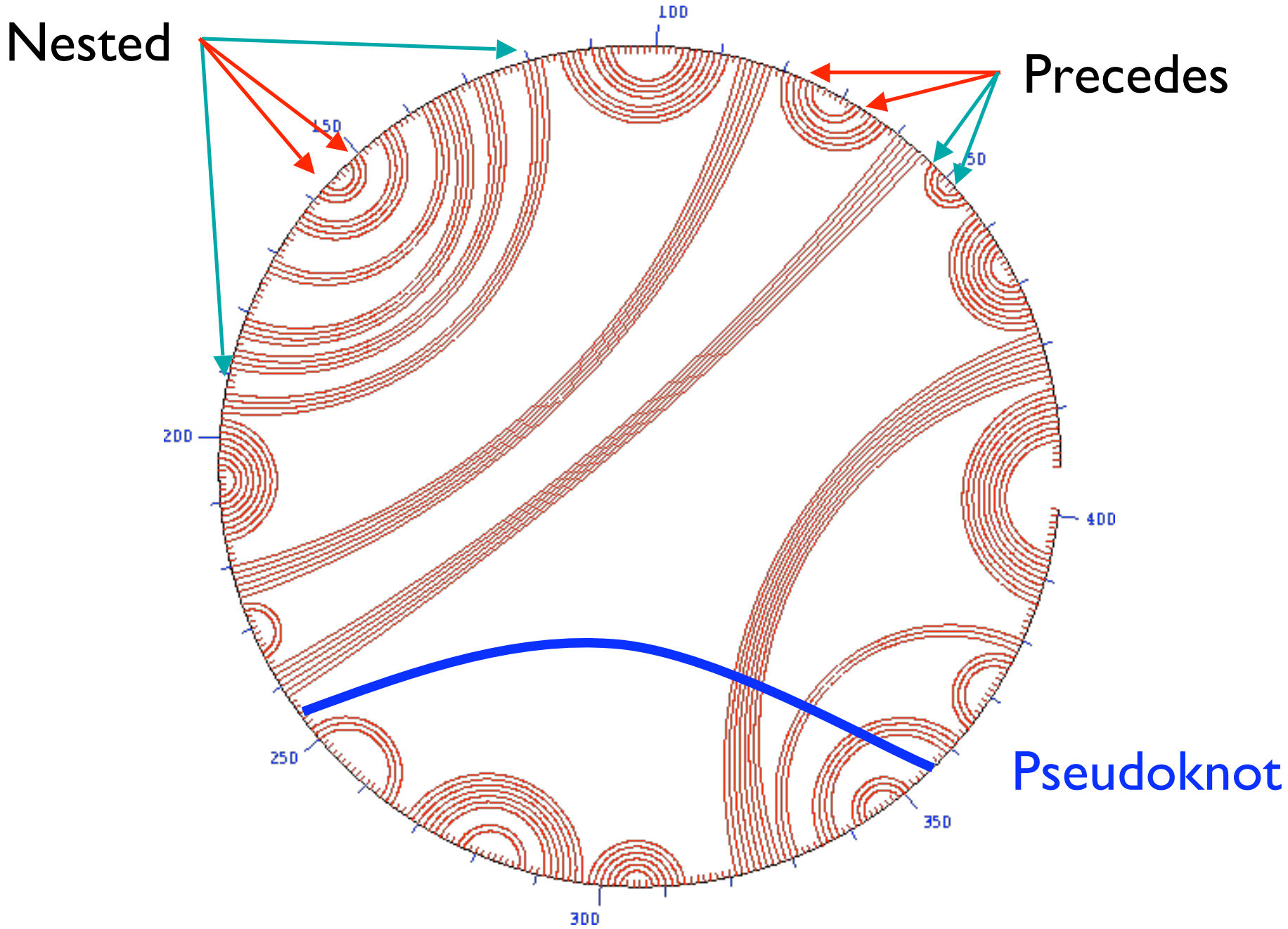



sharp turn

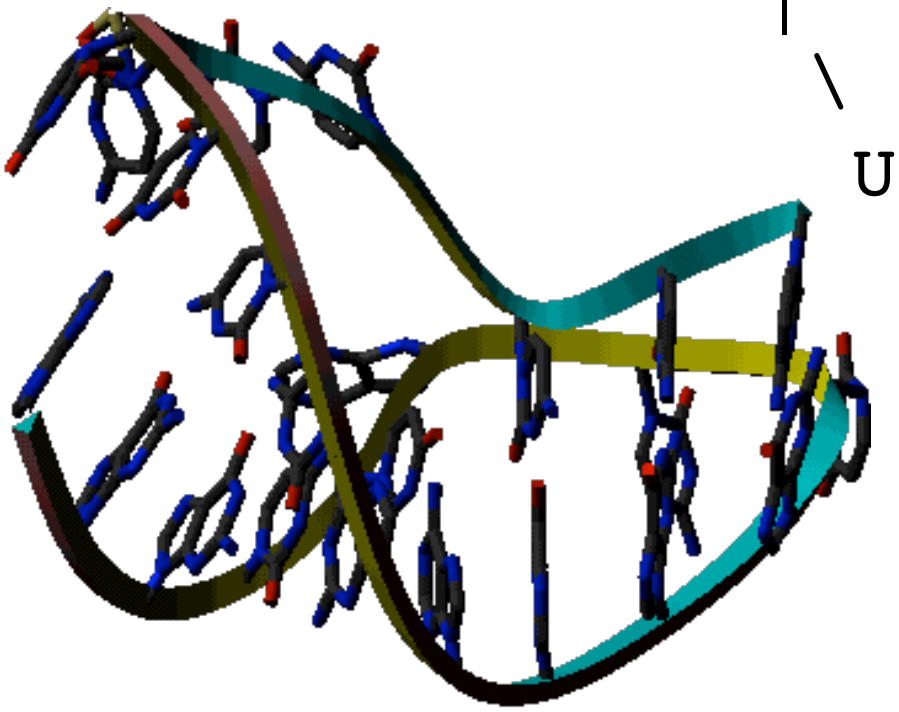



crossing


Nested
Precedes
Pseudoknot
100
150
50
200
400
250
350
300

# A Pseudoknot



```
                    A-C
                   /   \
3' - A-G-G-C-U      U
     U-C-C-G-A-G-G-G
     |           C-C-C - 5'
      \         /
       U-C-U-C
```

# Approaches to Structure Prediction

Maximum Pairing

+ works on single sequences
+ simple
- too inaccurate

Minimum Energy

+ works on single sequences
- ignores pseudoknots
- only finds “optimal” fold

Partition Function

+ finds all folds
- ignores pseudoknots

# RNA Secondary Structure (somewhat oversimplified)

Secondary structure. A set of pairs $S = \{ (b_i, b_j) \}$ that satisfy:

- [Watson-Crick.]
  - S is a *matching* , i.e. each base pairs with at most one other, and
  - each pair in S is a Watson-Crick pair: A-U, U-A, C-G, or G-C.
- [No sharp turns.] The ends of each pair are separated by at least 4 intervening bases. If $(b_i, b_j) \in S$, then $i < j - 4$.
- [Non-crossing.] If $(b_i, b_j)$ and $(b_k, b_l)$ are two pairs in S, then we cannot have $i < k < j < l$. (Violation of this is called a *pseudoknot.*)

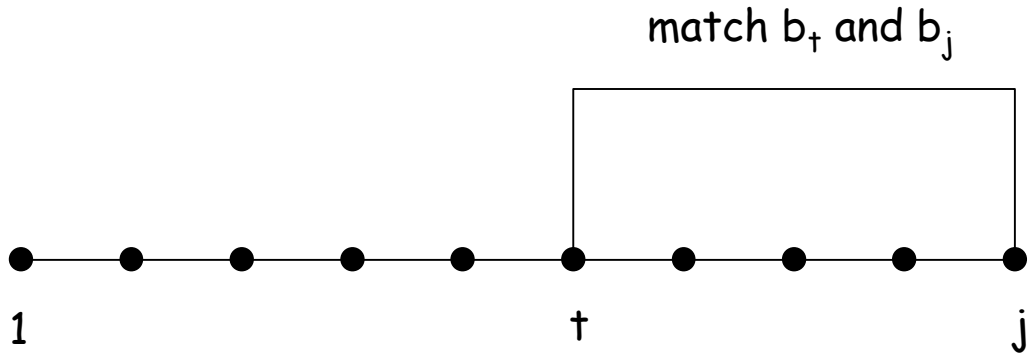
Free energy. Usual hypothesis is that an RNA molecule will form the secondary structure with the optimum total free energy.

approximate by number of base pairs

Goal. Given an RNA molecule $B = b_1 b_2 \ldots b_n$, find a secondary structure S that maximizes the number of base pairs.

# RNA Secondary Structure: Subproblems

First attempt. OPT[j] = maximum number of base pairs in a secondary structure of the substring $b_1 b_2 \ldots b_j$.




Difficulty. Results in two sub-problems.

- Finding secondary structure in: $b_1 b_2 \ldots b_{t-1}$. ← OPT(t-1)
- Finding secondary structure in: $b_{t+1} b_{t+2} \ldots b_{j-1}$. ← not OPT of anything; need more sub-problems

# Dynamic Programming Over Intervals: (R. Nussinov's algorithm)

Notation. OPT[i, j] = maximum number of base pairs in a secondary structure of the substring $b_i b_{i+1} \ldots b_j$.

- Case 1. If $i \geq j - 4$.
  - OPT[i, j] = 0 by no-sharp turns condition.
- Case 2. Base $b_j$ is not involved in a pair.
  - OPT[i, j] = OPT[i, j-1]
- Case 3. Base $b_j$ pairs with $b_t$ for some $i \leq t < j - 4$.
  - non-crossing constraint decouples resulting sub-problems
  - $OPT[i, j] = 1 + \max_t \{ OPT[i, t-1] + OPT[t+1, j-1] \}$

take max over t such that $i \leq t < j-4$ and $b_t$ and $b_j$ are Watson-Crick complements

Key point: Either last base is unpaired (case 1,2) or paired (case 3)

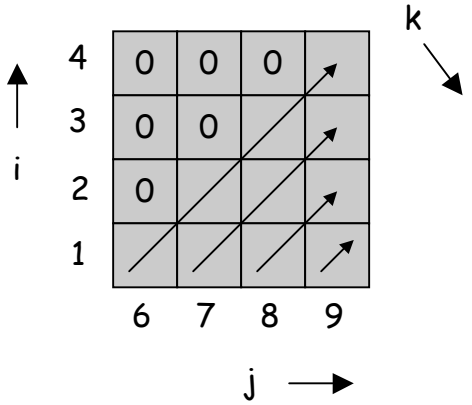
Remark. Same core idea in CKY algorithm to parse context-free grammars.

# Bottom Up Dynamic Programming Over Intervals

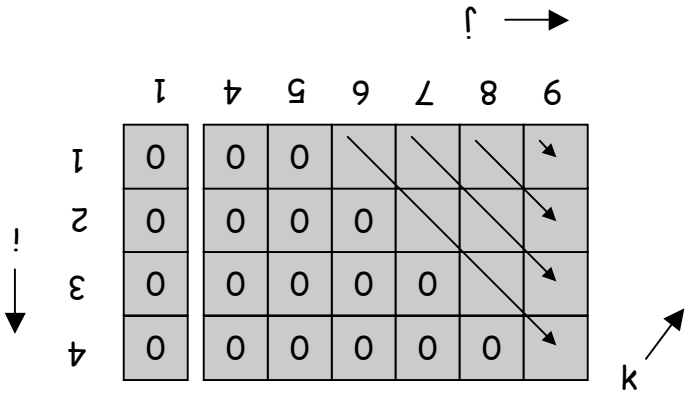
Q. What order to solve the sub-problems?

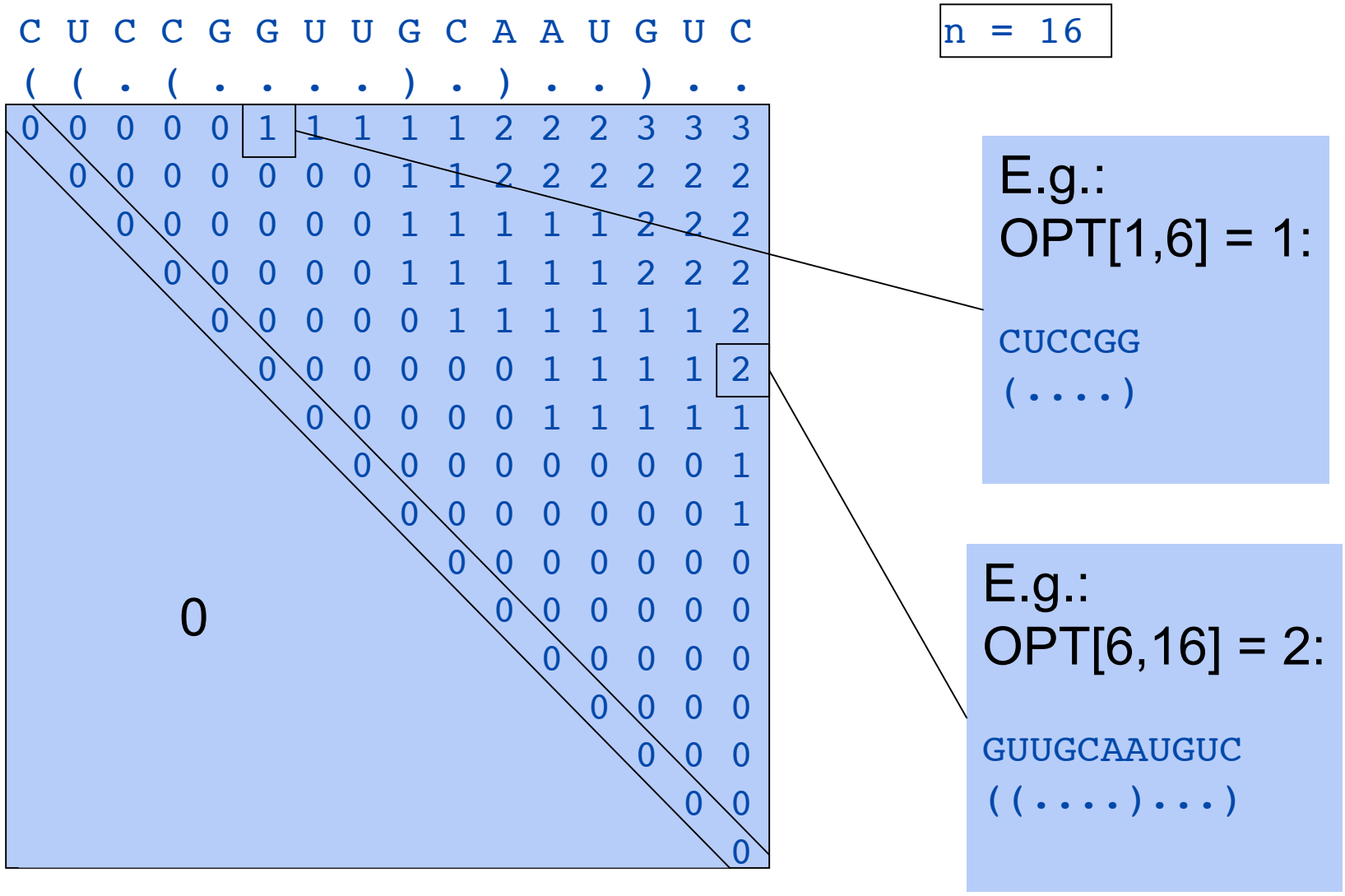
A. Do shortest intervals first.

```
RNA(b1,…,bn) {
    for k = 5, 6, …, n-1
        for i = 1, 2, …, n-k
            j = i + k
            Compute OPT[i, j]
                                  ↖ using recurrence
    return OPT[1, n]
}
```




Running time. $O(n^3)$.






n = 16

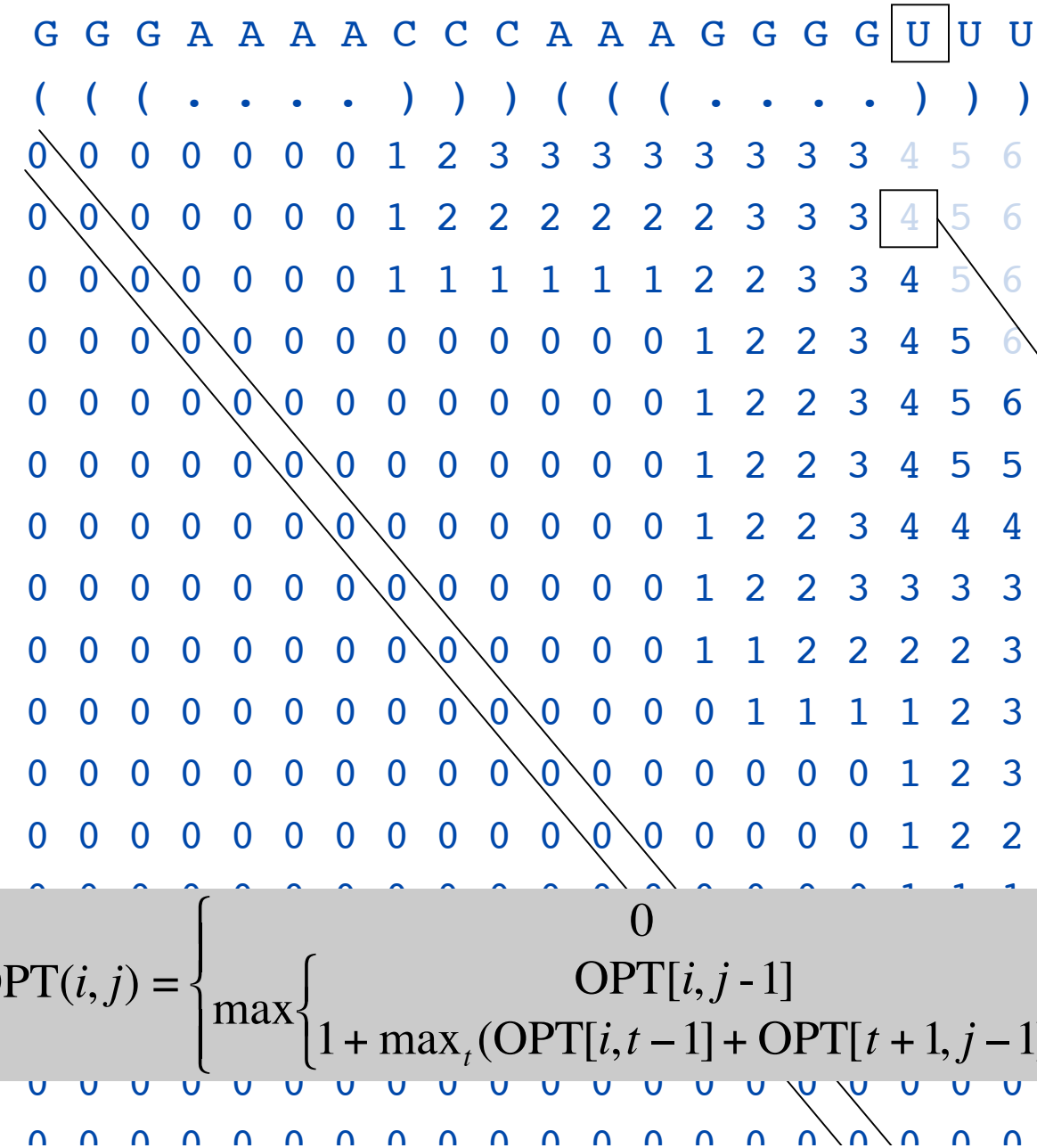
E.g.:
OPT[1,6] = 1:

CUCCGG
(....)

E.g.:
OPT[6,16] = 2:

GUUGCAAUGUC
((....)...)

# Computing one cell: OPT[2,18] = ?



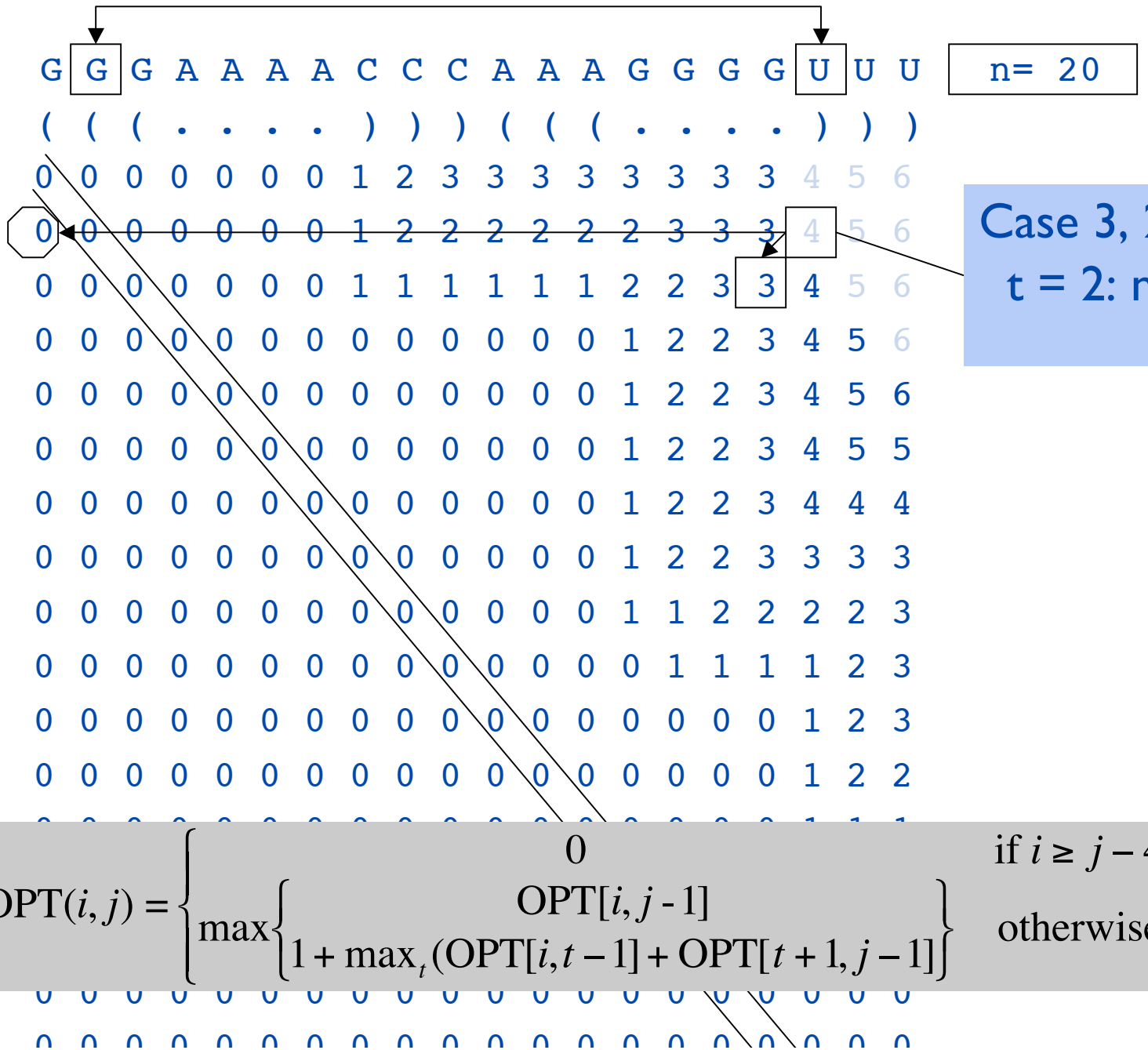

Case 1:
2 ≥ 18-4? no.

Case 2:
$B_{18}$ unpaired?
Always a possibility;
then OPT[2,18] ≥ 3

GGAAAACCCAAAGGGGU
( ( . . . . ) ) ( . . . . ) . . .

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

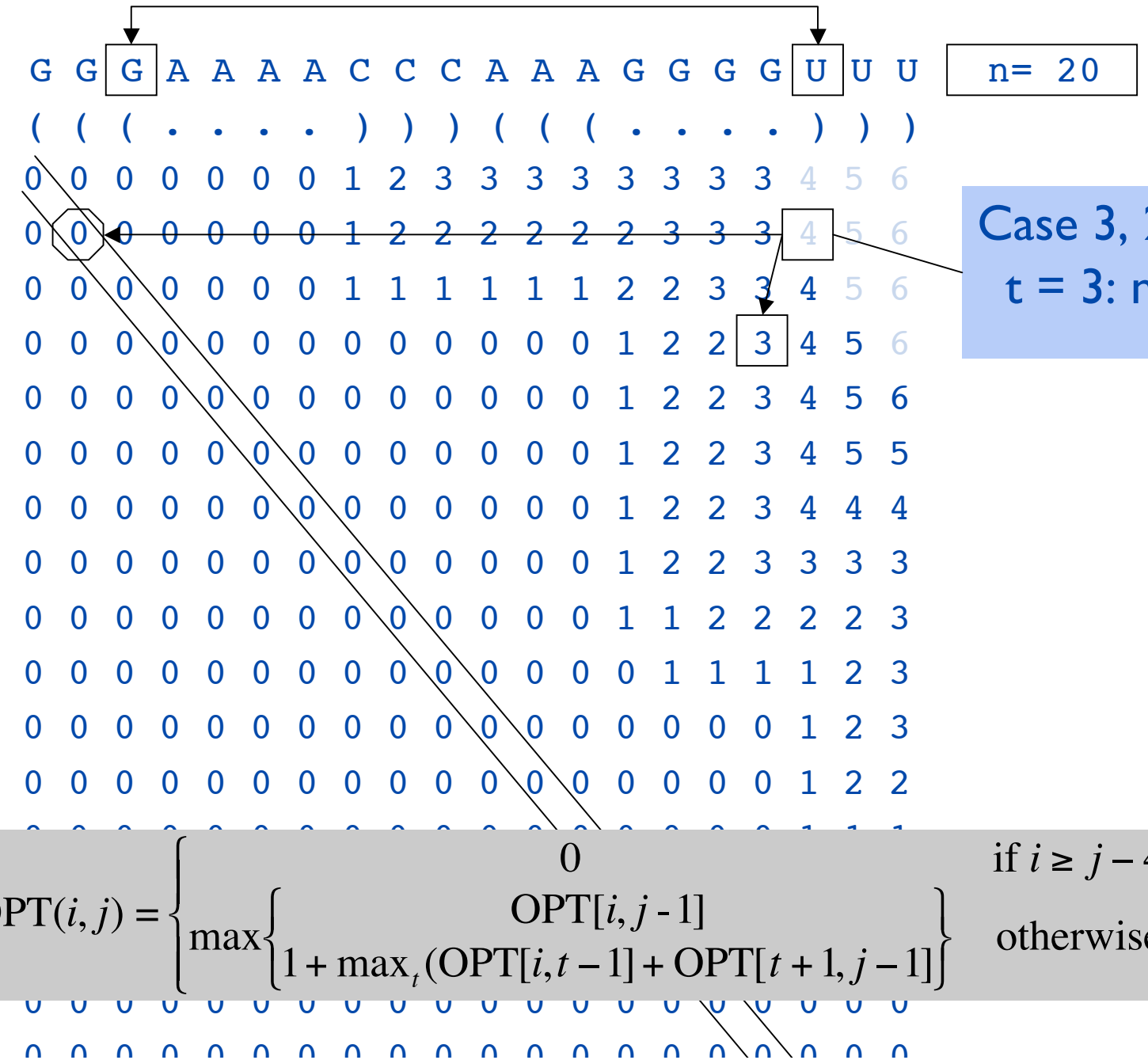
# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:
$t = 2$: no pair

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

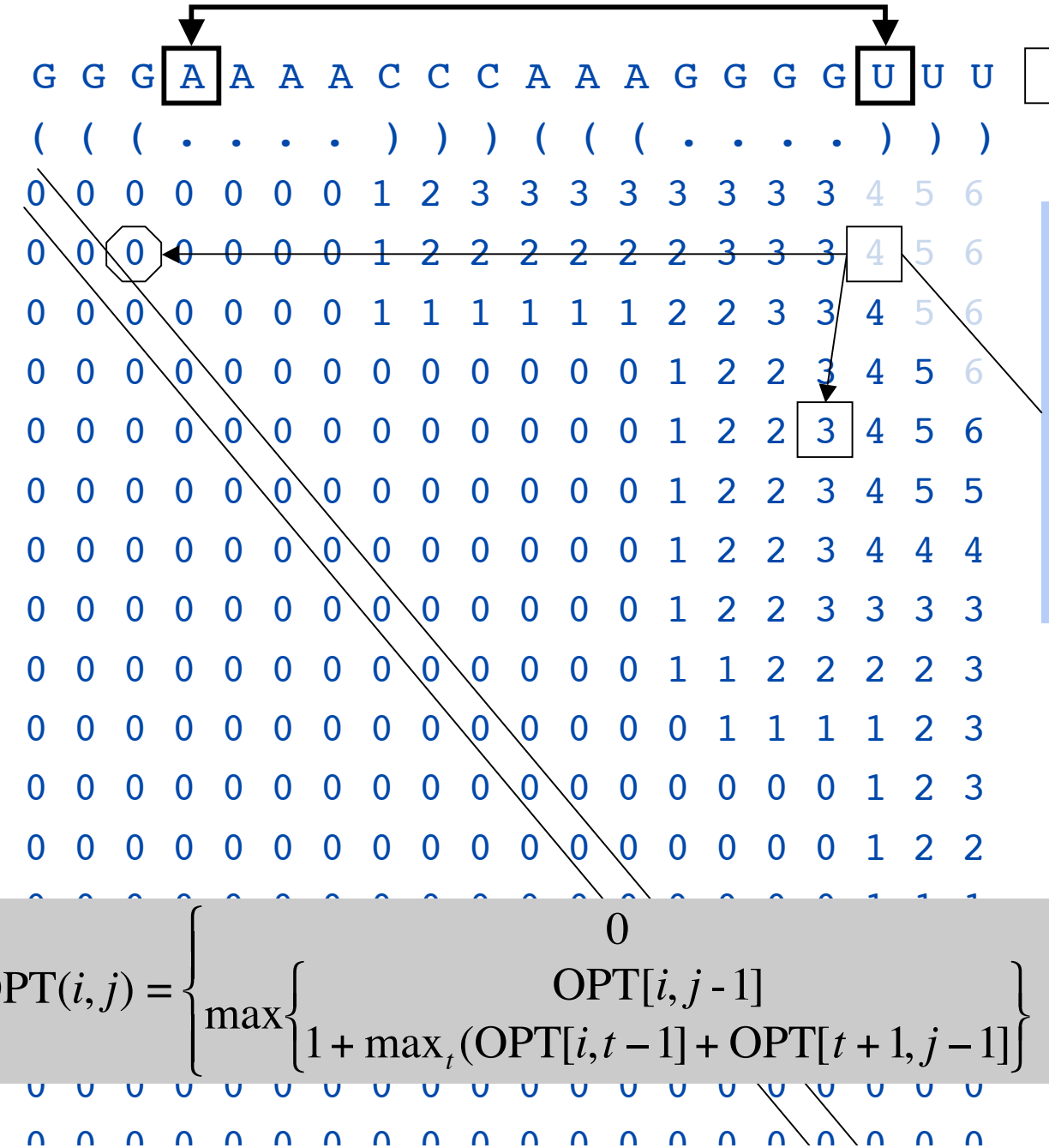
# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:
$t = 3$: no pair

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:

t = 4: yes pair

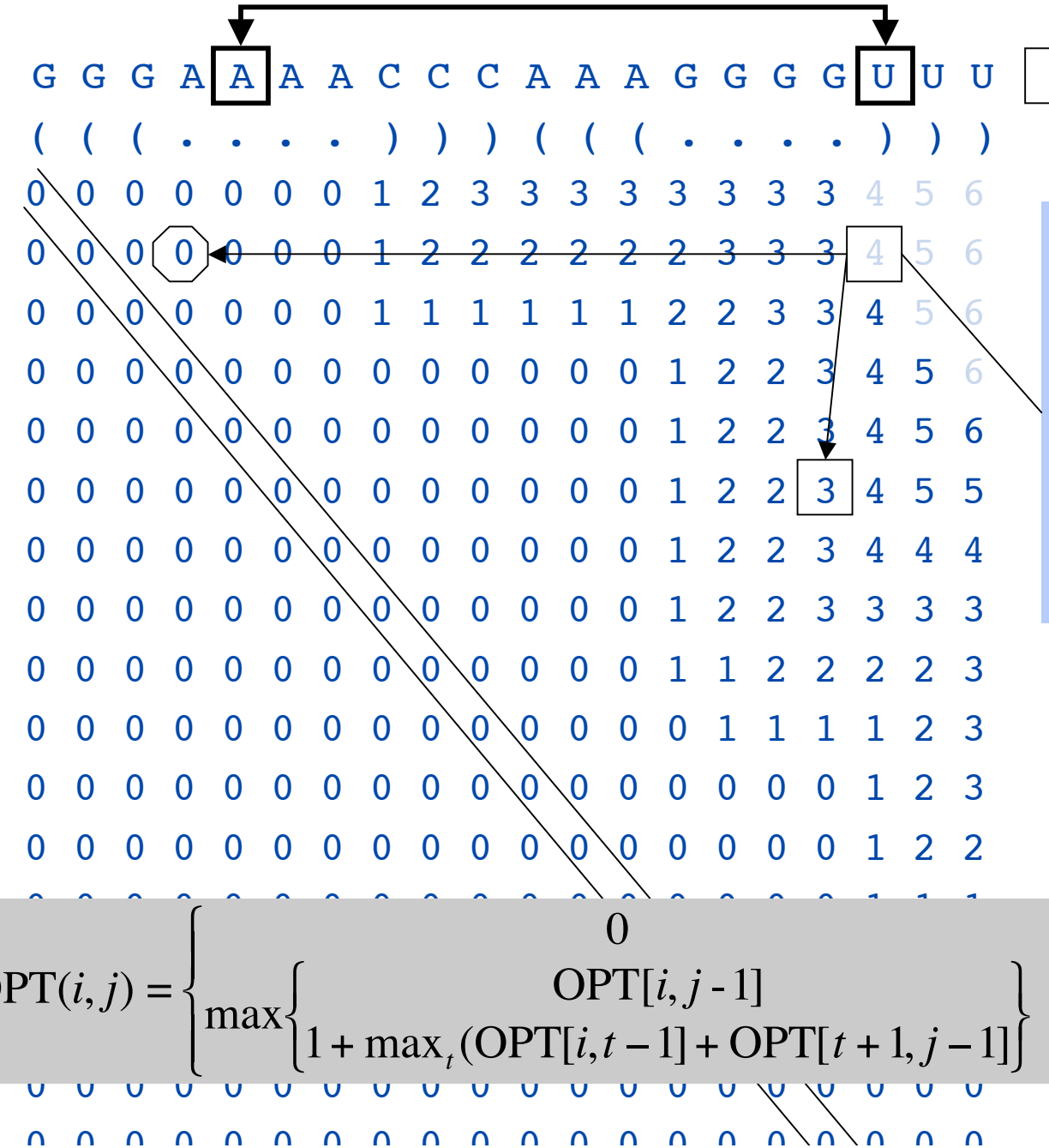
OPT[2,18]≥1+0+3

GG**A**AAACCCAAAGGGG**U**

..(...(((....)))).

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:

t = 5: yes pair

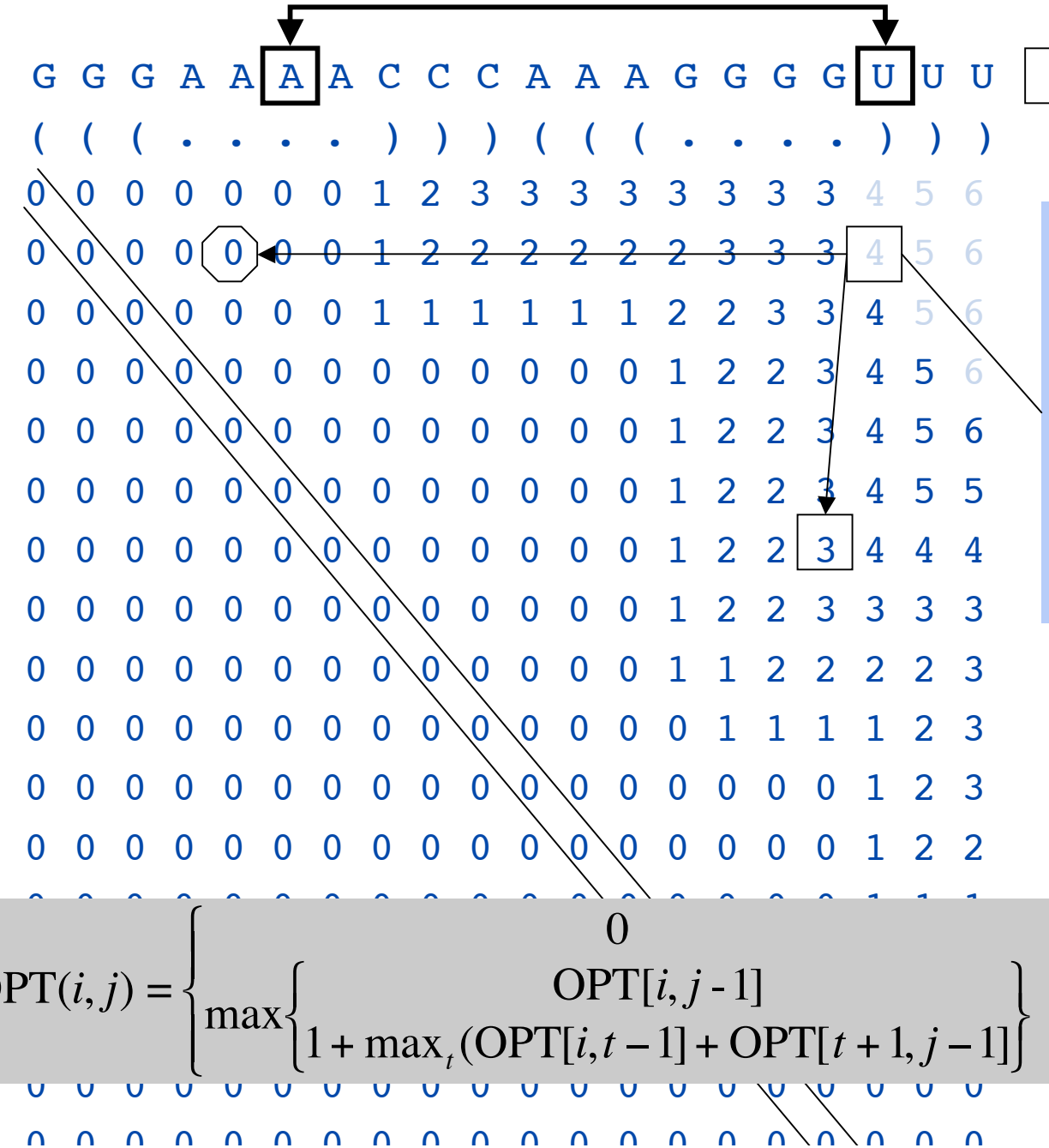
OPT[2,18]≥1+0+3

GGA**A**AACCCAAAGGGG**U**

...(..(((....)))

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:

t = 6: yes pair

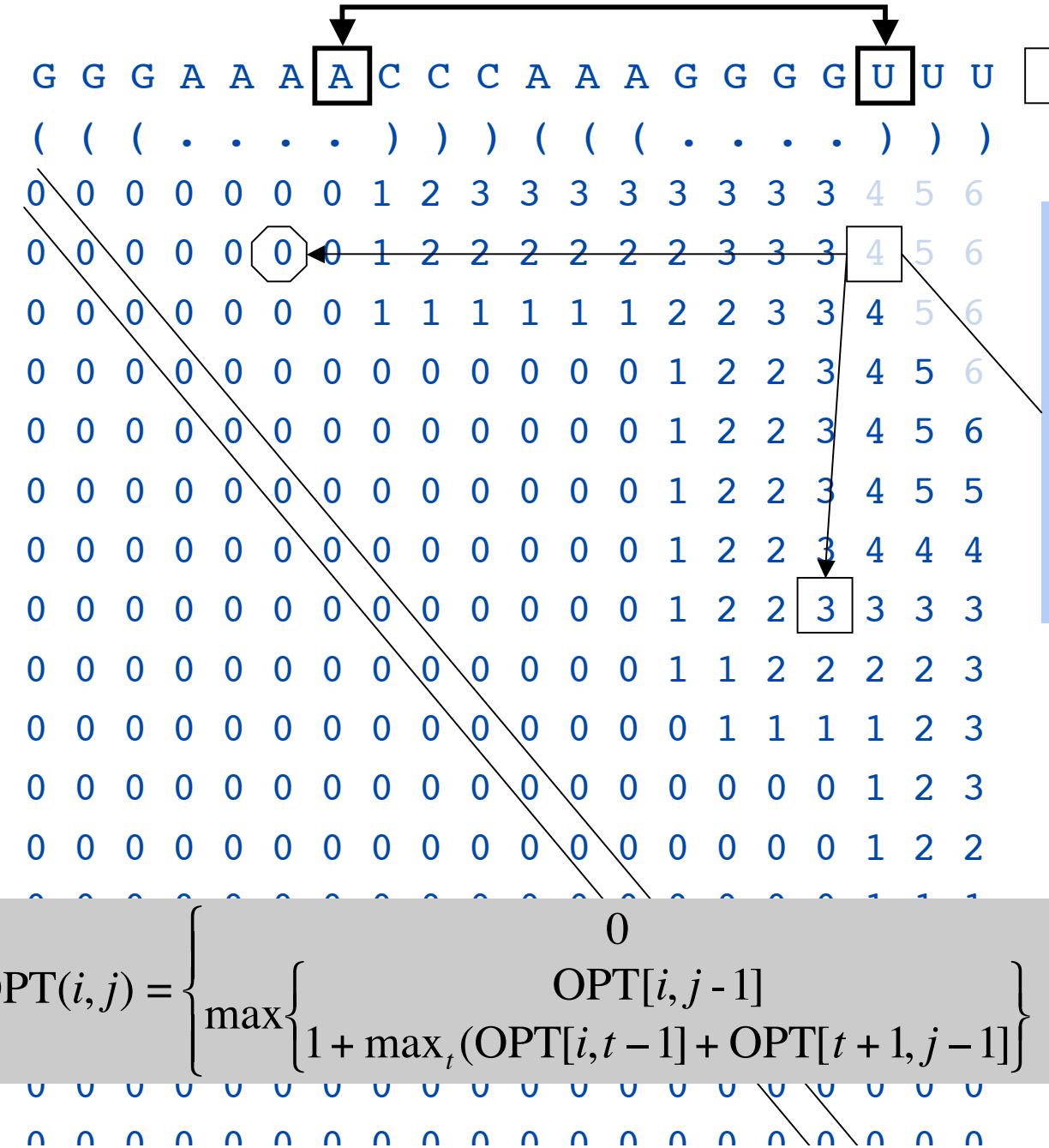
OPT[2,18]≥1+0+3

GGAA**A**ACCCAAAGGGG**U**

....(.(((....)))).

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:

t = 7: yes pair

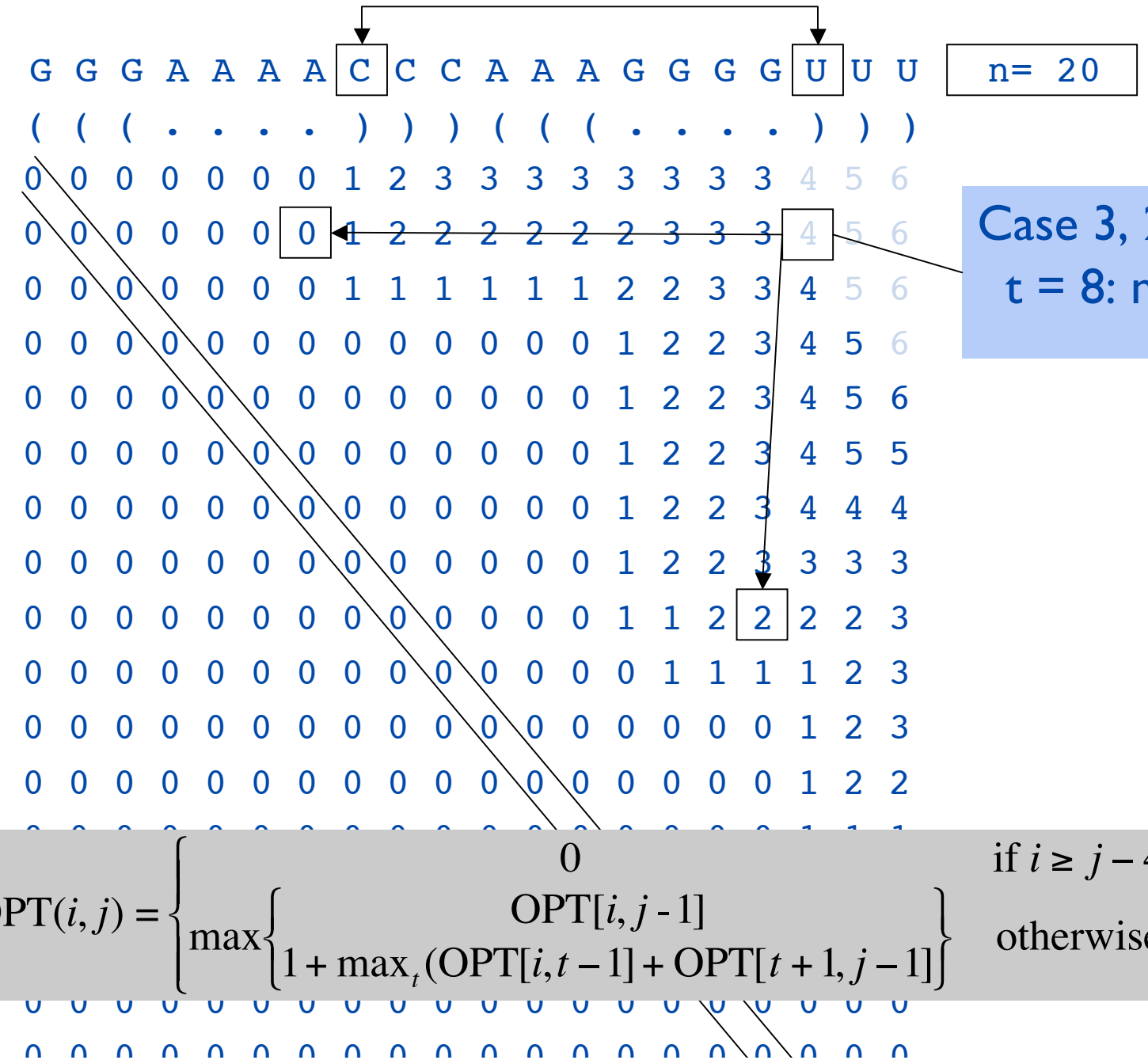
OPT[2,18]≥1+0+3

GGAAA**A**CCCAAAGGGG**U**

.....((((....))))

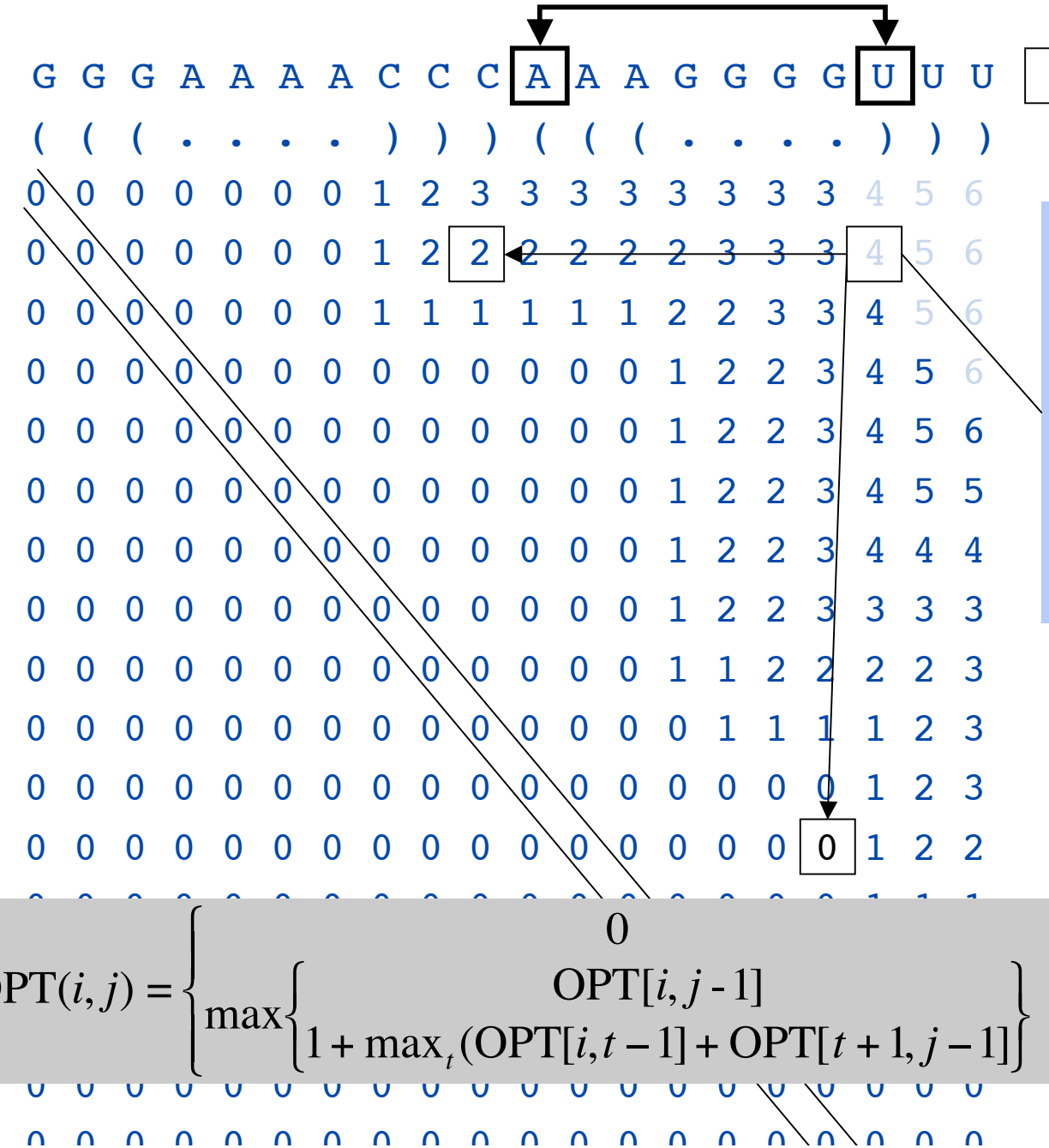
$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Computing one cell: OPT[2,18] = ?




Case 3, $2 \le t < 18-4$:

t = 11: yes pair

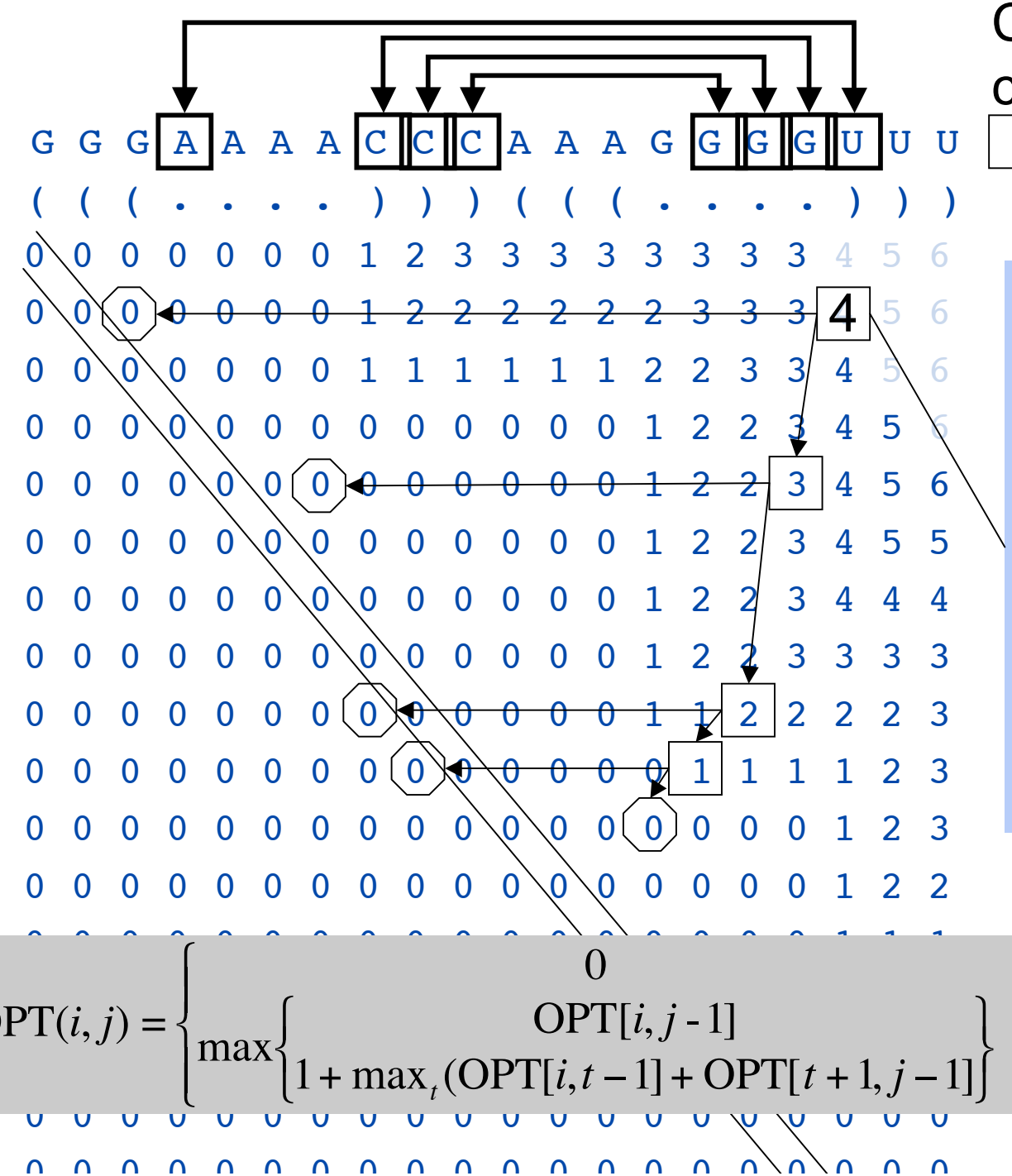
OPT[2,18]≥1+2+0

GGAAAACCC**A**AAGGGG**U**

((.....))(......)

(not shown: t=9,10, 12,13)

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \ge j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$




# Computing one cell: OPT[2,18] = 4

n= 20

## Overall, Max = 4 several ways, e.g.:

GGAAAACCCAAAGGGGU

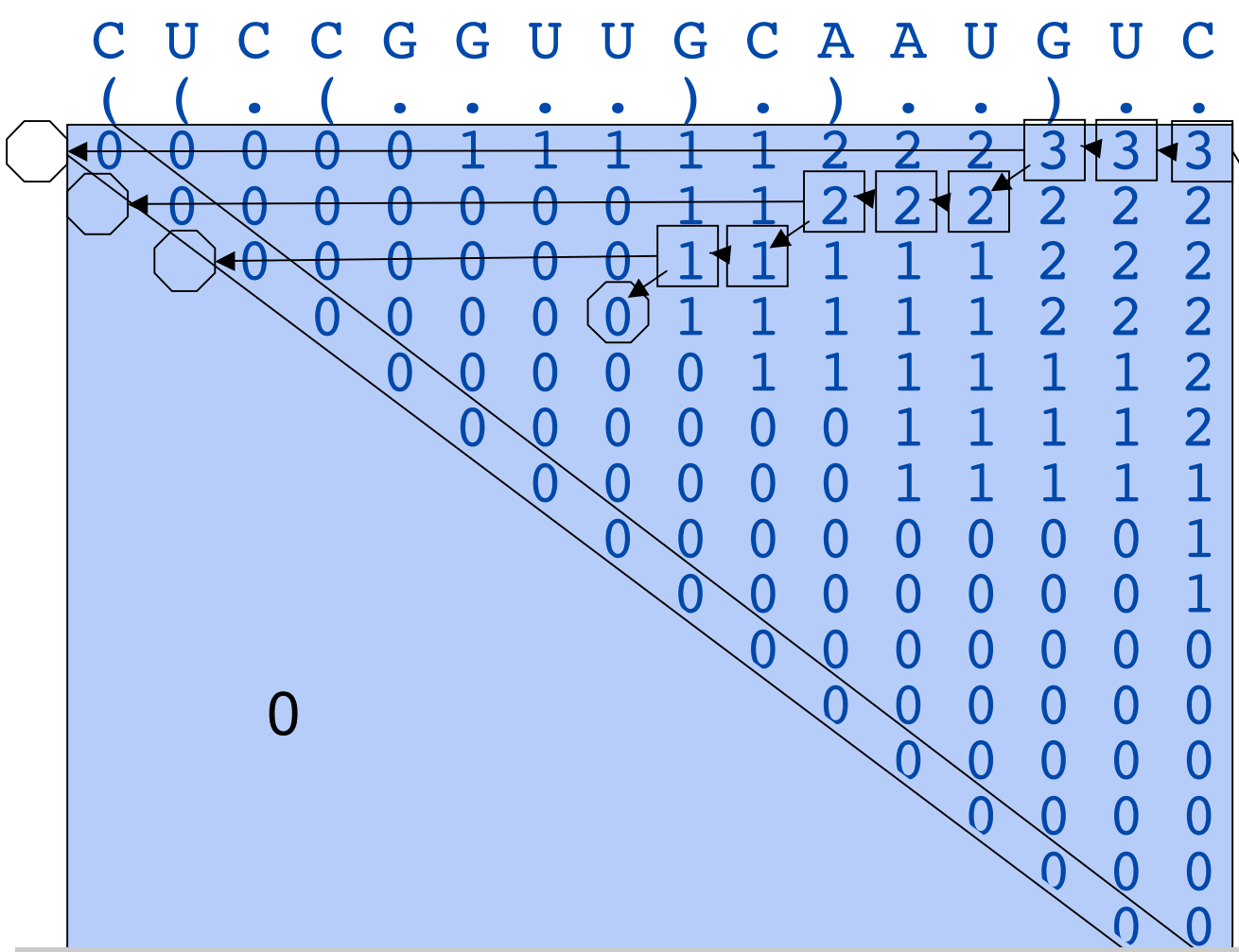
..(...(((....)))) 

tree shows trace back:

- square = case 3
- octagon = case 1

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

# Another Trace Back Example




E.g.:
OPT[1,16] = 3:

CUCCGGUUGCAAUGUC
((.(....).)..)..

$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

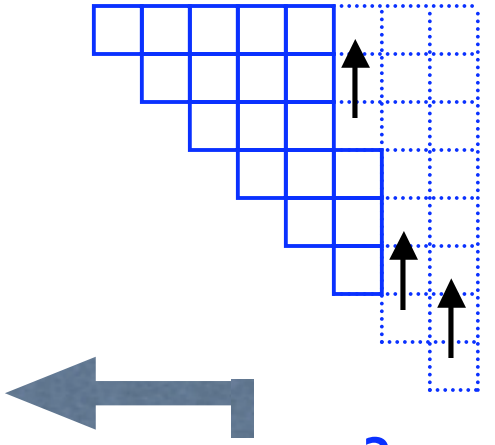
# Nussinov: Max Pairing

B(i,j) = # pairs in optimal pairing of $r_i$ ... $r_j$

B(i,j) = 0 for all i, j with i ≥ j-4; otherwise

B(i,j) = max of:

$$\begin{cases} B(i,j-1) \\ \max \{ B(i,k-1)+1+B(k+1,j-1) \mid \\ \quad i \le k < j-4 \text{ and } r_k\text{-}r_j \text{ may pair} \} \end{cases}$$



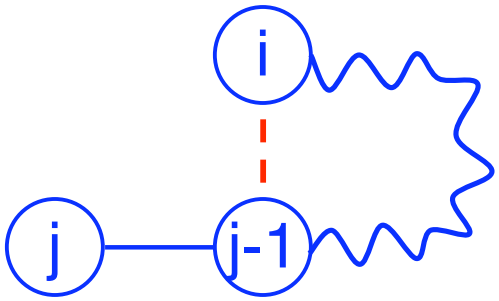
Time: $O(n^3)$

# "Optimal pairing of $r_i$ ... $r_j$"

## Two possibilities

J Unpaired:

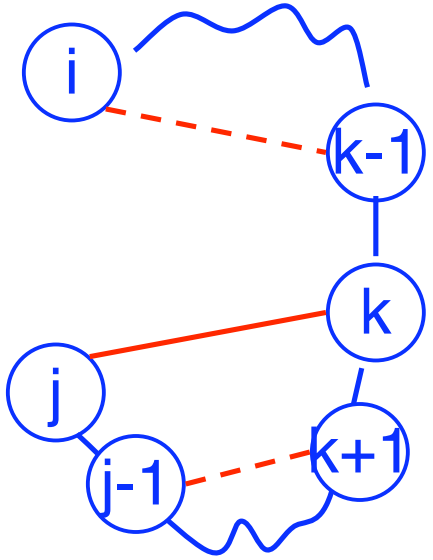
Find best pairing of $r_i$ ... $r_{j-1}$




J Paired (with some k):

Find best $r_i$ ... $r_{k-1}$ +

best $r_{k+1}$ ... $r_{j-1}$ plus 1




Why is it slow?

Why do pseudoknots matter?

# Pair-based Energy Minimization

$E(i,j)$ = energy of pairs in optimal pairing of $r_i$ ... $r_j$

$E(i,j) = \infty$ for all i, j with $i \geq j-4$; otherwise

$E(i,j)$ = min of:

$$\begin{cases} E(i,j-1) \\ \min \{ E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) \mid i \leq k < j-4 \} \end{cases}$$

energy of j-k pair

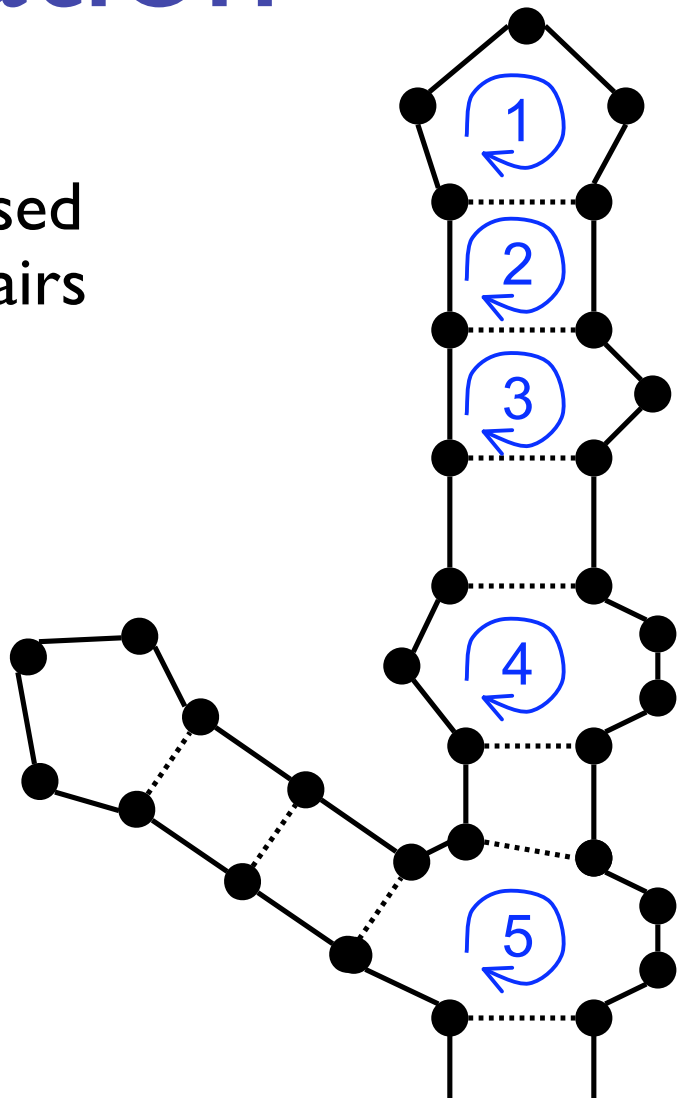
Time: $O(n^3)$

# Loop-based Energy Minimization

Detailed experiments show it's more accurate to model based on loops, rather than just pairs

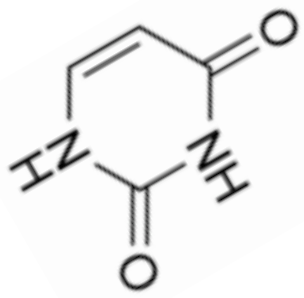
Loop types

1. Hairpin loop
2. Stack
3. Bulge
4. Interior loop
5. Multiloop

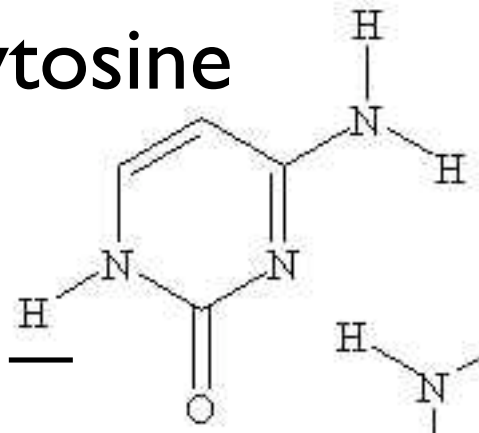


# Base Pairs and Stacking

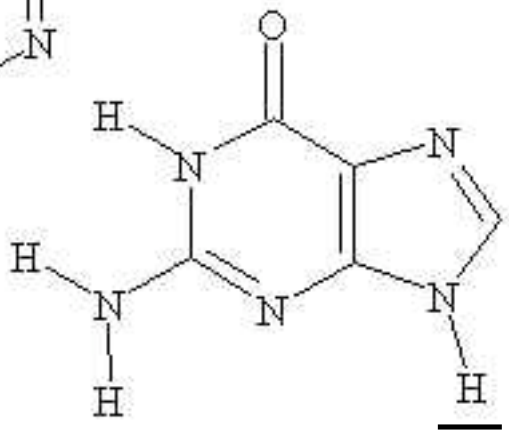
uracil



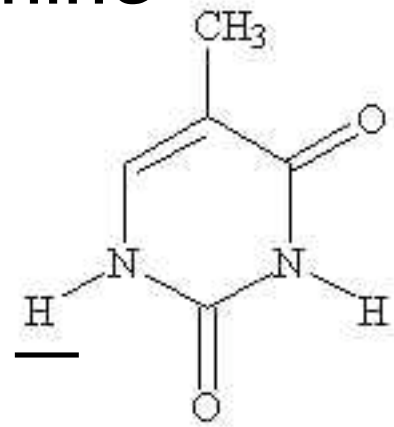
cytosine



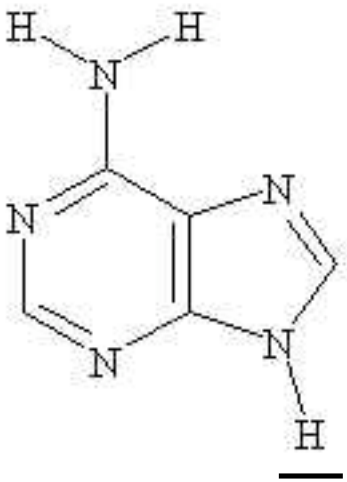
guanine



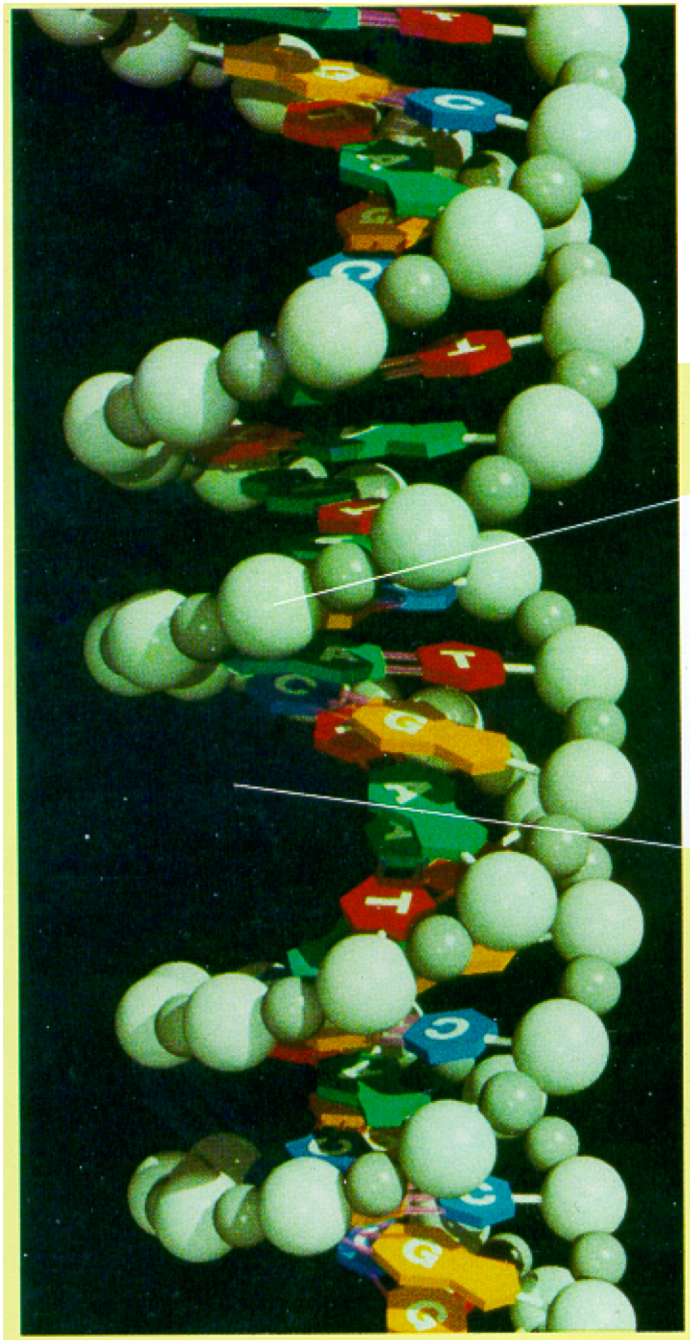
thymine

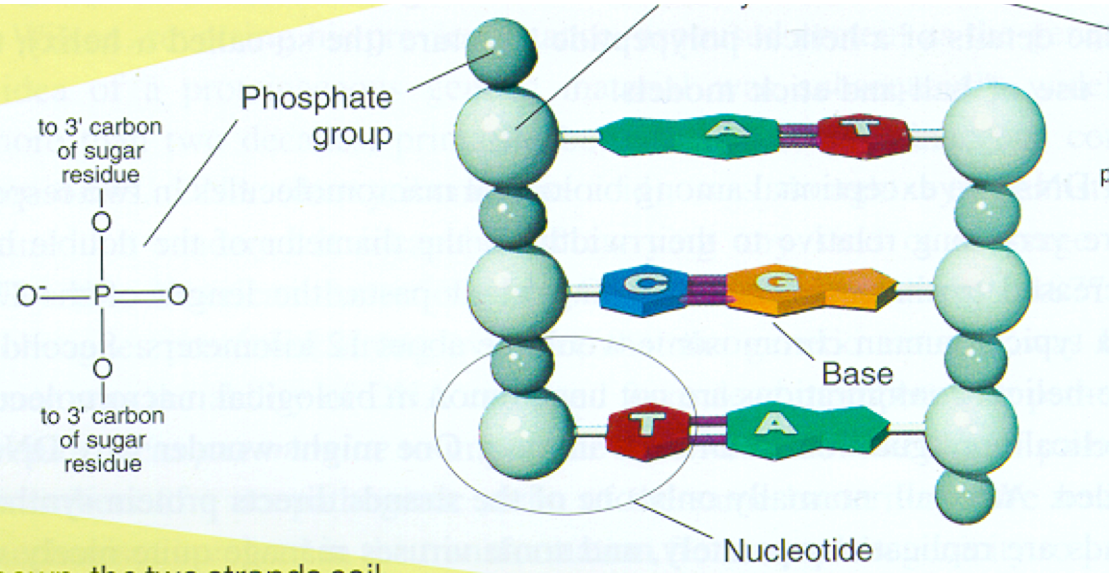


adenine



# The Double Helix

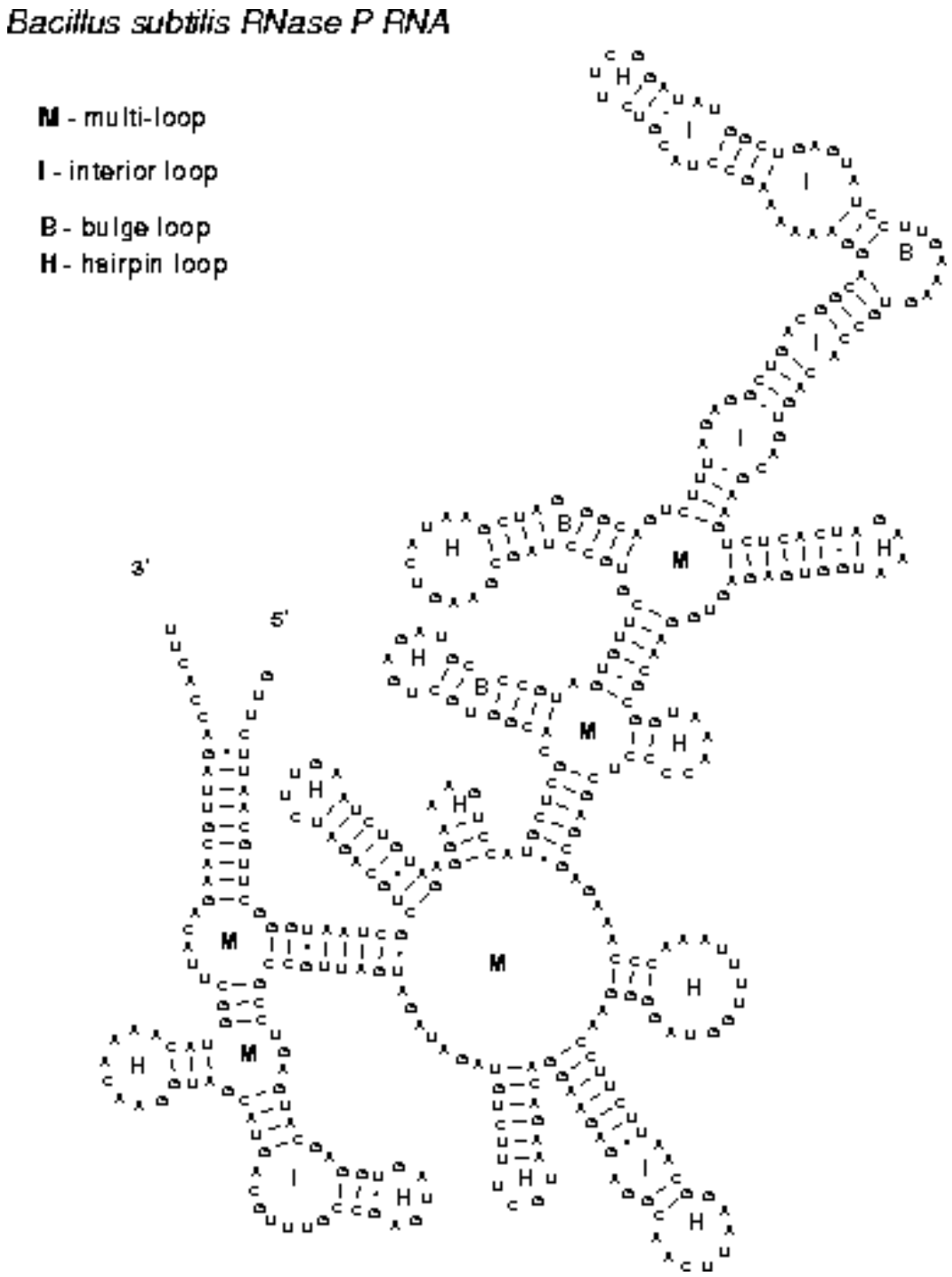





As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

# Loop Examples



# Zuker: Loop-based Energy, I

W(i,j) = energy of optimal pairing of $r_i$ ... $r_j$

V(i,j) = as above, but forcing pair i•j

W(i,j) = V(i,j) = ∞ for all i, j with i ≥ j-4

W(i,j) = min(W(i,j-1),
min { W(i,k-1)+V(k,j) | i ≤ k < j-4 }
)

# Zuker: Loop-based Energy, II

hairpin stack bulge/ interior multi- loop

$$V(i,j) = \min(eh(i,j),\ es(i,j)+V(i+1,j-1),\ VBI(i,j),\ VM(i,j))$$

$$VM(i,j) = \min \{\ W(i,k)+W(k+1,j) \mid i < k < j\ \}$$

$$VBI(i,j) = \min \{\ ebi(i,j,i',j') + V(i', j') \mid i < i' < j' < j \ \&\ i'-i+j-j' > 2\ \}$$

bulge/ interior

Time: $O(n^4)$

$O(n^3)$ possible if ebi(.) is "nice"

# Energy Parameters

Q. Where do they come from?

A1. Experiments with carefully selected synthetic RNAs

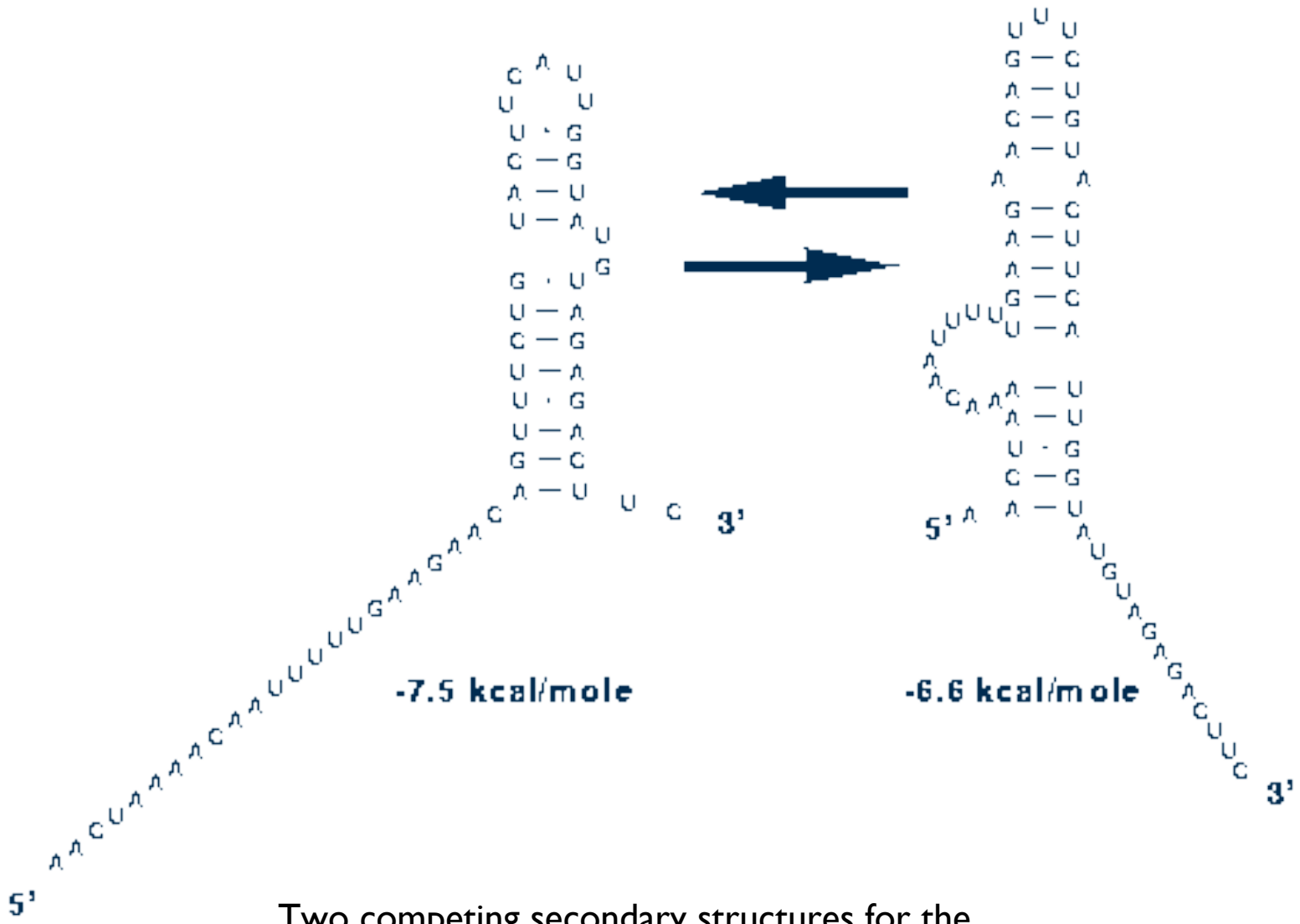
A2. Learned algorithmically from trusted alignments/structures

# Suboptimal Energy

There are always alternate folds with near-optimal energies. Thermodynamics: populations of identical molecules will exist in different folds; individual molecules even flicker among different folds

Mod to Zuker's algorithm finds subopt folds

McCaskill: more elaborate dyn. prog. algorithm calculates the "partition function," which defines the probability distribution over all these states.
(Key addition: recurrence must count each possibility exactly once.)




Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.

# Accuracy

Latest estimates suggest ~50-75% of base pairs predicted correctly in sequences of up to ~300nt

Definitely useful, but obviously imperfect

# Approaches to Structure Prediction

Maximum Pairing

- \+ works on single sequences
- \+ simple
- \- too inaccurate

Minimum Energy

- \+ works on single sequences
- \- ignores pseudoknots
- \- only finds “optimal” fold

Partition Function

- \+ finds all folds
- \- ignores pseudoknots

# Approaches, II

Comparative sequence analysis

+ handles all pairings (incl. pseudoknots)
- requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars

Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

# Summary

RNA has important roles beyond mRNA

- Many unexpected recent discoveries

Structure is critical to function

- True of proteins, too, but they're easier to find, due, e.g., to codon structure, which RNAs lack

RNA secondary structure can be predicted (to useful accuracy) by dynamic programming

Next time: RNA "motifs" (seq + 2-ary struct) well-captured by "covariance models"