# CSE 421: Intro Algorithms

W. L. Ruzzo

## Dynamic Programming: String alignment and RNA Folding

# Outline

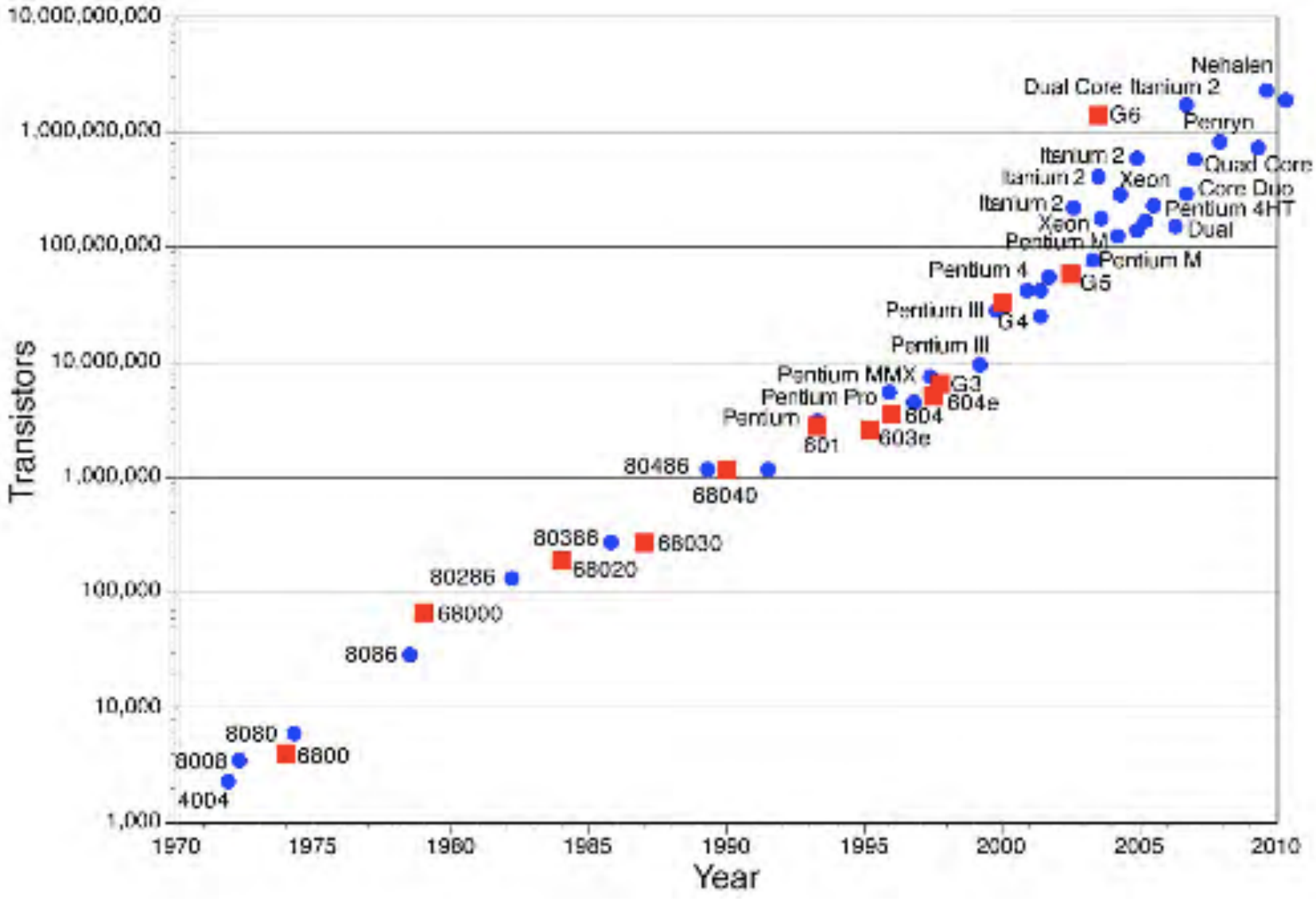
A few slides on *applications* of dynamic programming in biology

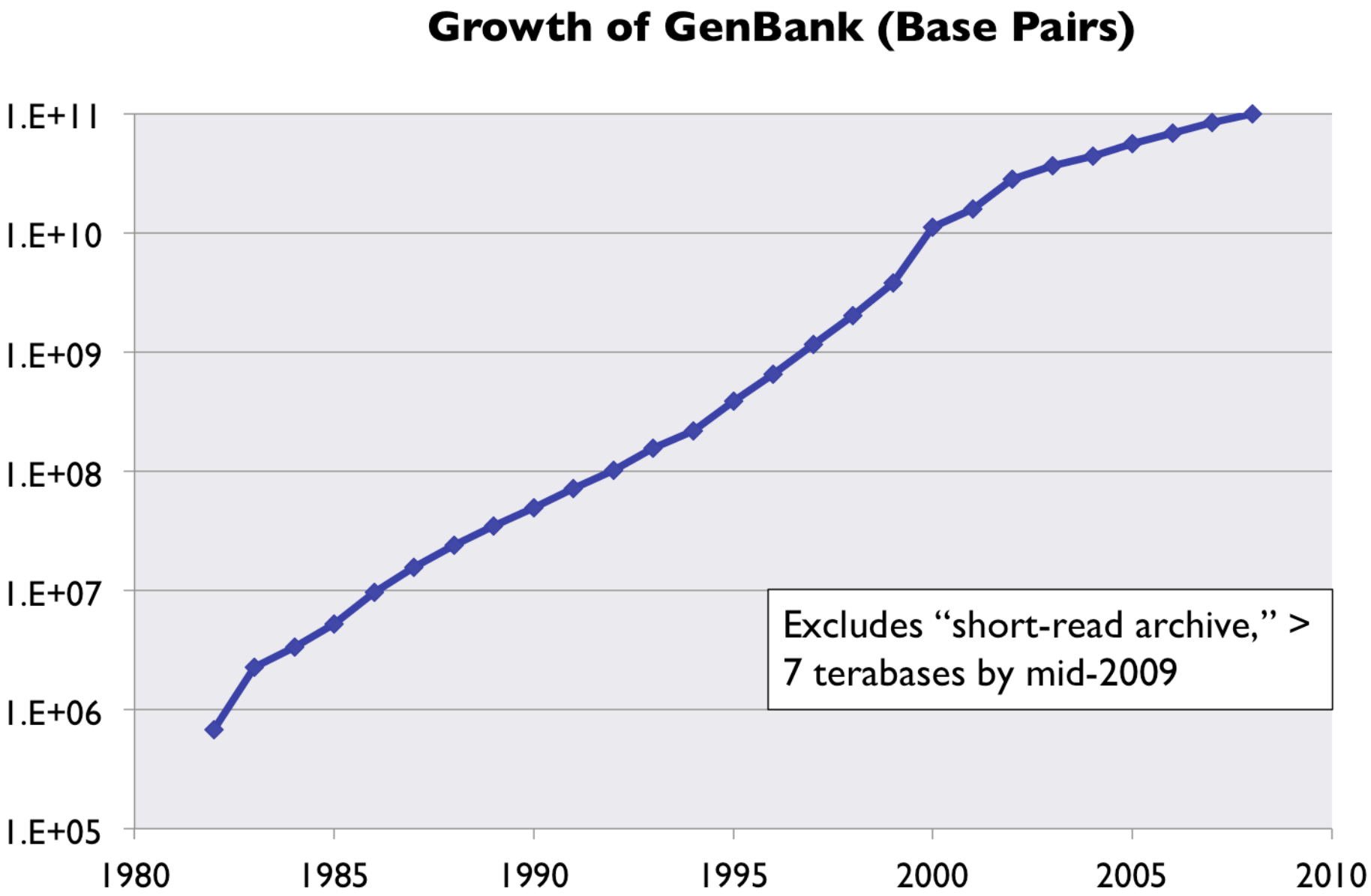
- Sequence alignment
- RNA structure

Algorithms for RNA structure

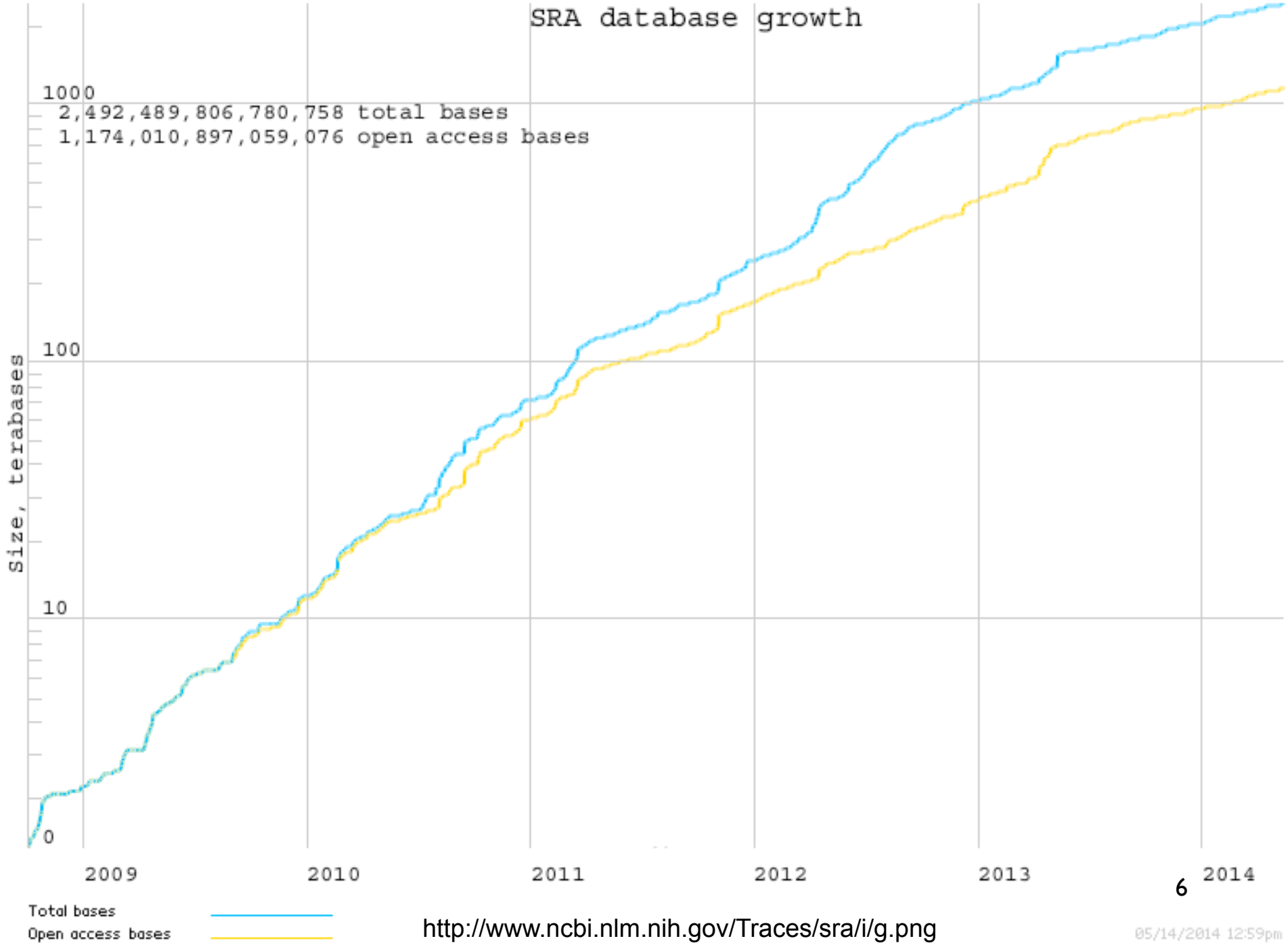
# Application: Sequence Search

# Moore's Law


10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Transistors
1970
1975
1980
1985
1990
1995
2000
2005
2010
Year
Nehalem
Dual Core Itanium 2
G6
Penryn
Itanium 2
Quad Core
Itanium 2
Xeon
Core Duo
Itanium 2
Pentium 4HT
Xeon
Dual
Pentium M
Pentium M
Pentium 4
G5
Pentium III
G4
Pentium III
Pentium MMX
G3
Pentium Pro
604e
Pentium
604
603e
601
80486
68040
80386
68030
68020
80286
68000
8086
8080
8008
6800
4004



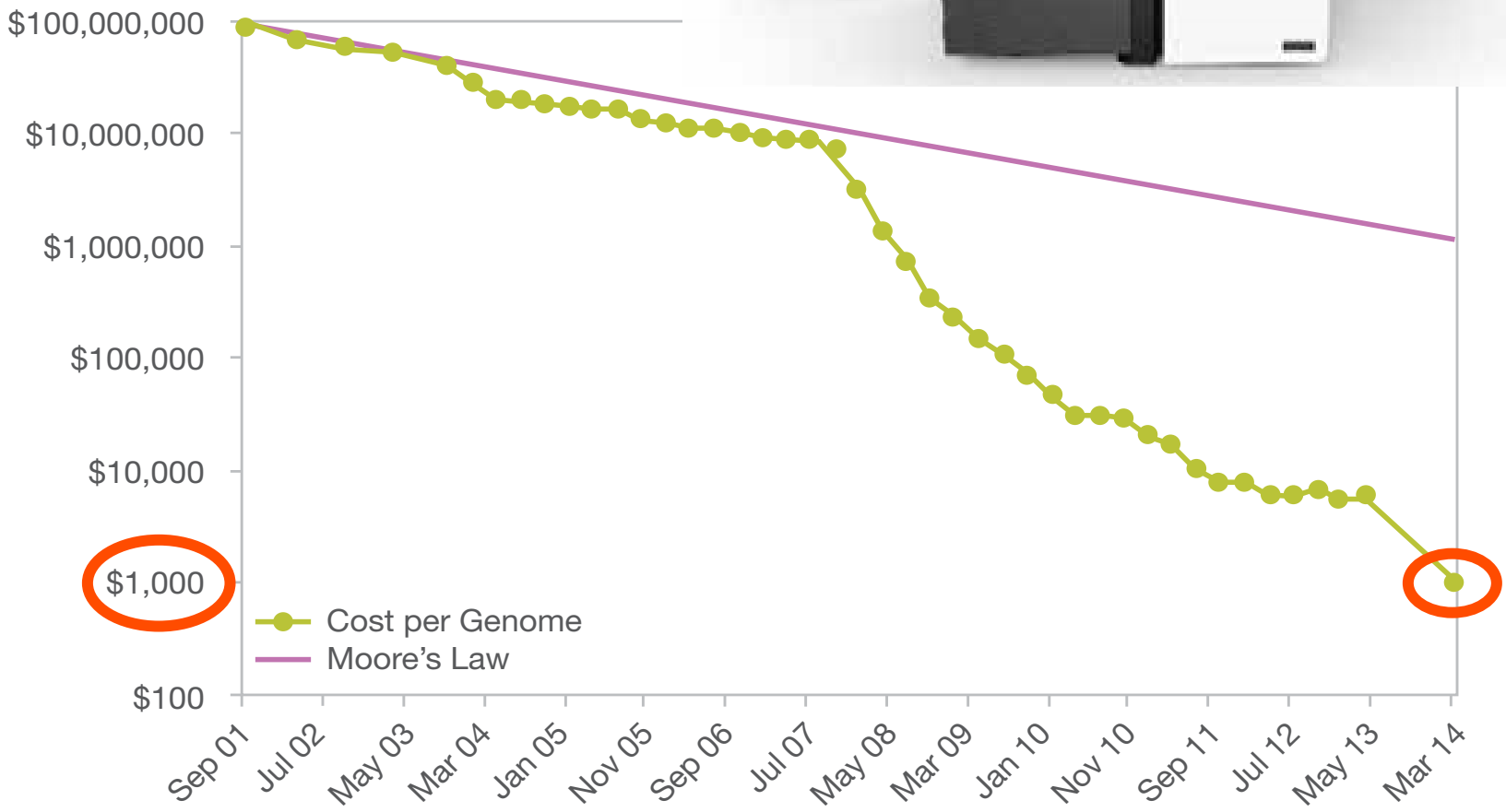

Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html






http://www.ncbi.nlm.nih.gov/Traces/sra/i/g.png

# Sequencing Costs Outpace Moore's Law



# A Database Search

go to, e.g., http://www.uniprot.org/, "blast" tab, and paste in this:

```
>sp|P15172|MYOD1_HUMAN Myoblast determination protein 1 OS=Homo
   sapiens GN=MYOD1 PE=1 SV=3
MELLSPPLRDVDLTAPDGSLCSFATTDDFYDDPCFDSPDLRFFEDLDPRLMHVGALLKPE
EHSHFPAAVHPAPGAREDEHVRAPSGHHQAGRCLLWACKACKRKTTNADRRKAATMRERR
RLSKVNEAFETLKRCTSSNPNQRLPKVEILRNAIRYIEGLQALLRDQDAAPPGAAAAFYA
PGPLPPGRGGEHYSGDSDASSPRSNCSDGMMDYSGPPSGARRRNCYEGAYYNEAPSEPRP
GKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGESSGDPTQS
PDAAPQCPAGANPNPIYQVL
```

# A Few seconds Later…

## Graphical overview

**Color code for identity 0-100% =**

| Accession | Entry name | 0Query hit320 | 0Match hit (sqrt scale)17392 | Name (Organism) |
|---|---|---|---|---|
| P15172 | MYOD1_HUMAN | human | Myoblast determination protein 1 (Homo sapiens) | |
| B2RC72 | B2RC72_HUMAN | human | cDNA, FLJ95884, highly similar to Hom... (Homo sapiens) | |
| E2RT59 | E2RT59_CANFA | dog | Uncharacterized protein (Canis familiaris) | |
| P49811 | MYOD1_PIG | pig | Myoblast determination protein 1 (Sus scrofa) | |
| D2KPI9 | D2KPI9_PIG | pig | Myogenic differentiation 1 (Sus scrofa) | |
| F1S9A9 | F1S9A9_PIG | pig | Uncharacterized protein (Sus scrofa) | |
| D2I0V4 | D2I0V4_AILME | panda | Putative uncharacterized protein (Ailuropoda melanoleuca) | |
| P29331 | MYOD1_SHEEP | sheep | Myoblast determination protein 1 (Ovis aries) | |
| D2SP11 | D2SP11_BUBBU | water buffalo | Myogenic factor MYOD1 (Bubalus bubalis) | |
| Q0VBX9 | Q0VBX9_BOVIN | cow | Myogenic differentiation 1 (Bos taurus) | |
| Q7YS82 | MYOD1_BOVIN | cow | Myoblast determination protein 1 (Bos taurus) | |
| Q8C6B1 | Q8C6B1_MOUSE | mouse | Myogenic differentiation 1 (Mus musculus) | |
| A0JPK9 | A0JPK9_RAT | rat | Myogenic differentiation 1 (Rattus norvegicus) | |
| Q02346 | MYOD1_RAT | rat | Myoblast determination protein 1 (Rattus norvegicus) | |
| P10085 | MYOD1_MOUSE | mouse | Myoblast determination protein 1 (Mus musculus) | |
| Q6DTY5 | Q6DTY5_PIG | pig | Eukaryotic myogenic factor MYF-3 (Sus scrofa) | |
| P21572 | MYOD1_COTJA | quail | Myoblast determination protein 1 homolog (Coturnix coturnix japonica) | |
| Q6DV59 | Q6DV59_MELGA | turkey | MyoD (Meleagris gallopavo) | |
| P16075 | MYOD1_CHICK | chicken | Myoblast determination protein 1 homolog (Gallus gallus) | |
| C5J072 | C5J072_CHICK | chicken | Myogenic differentiation 1 (Gallus gallus) | |
| C3U0I1 | C3U0I1_ANAPL | duck | Myogenic differentiation 1 (Anas platyrhynchos) | |
| F1NHM3 | F1NHM3_CHICK | chicken | Uncharacterized protein (Gallus gallus) | |
| F1NXM5 | F1NXM5_CHICK | chicken | Uncharacterized protein (Gallus gallus) | |
| P13904 | MYODA_XENLA | frog | Myoblast determination protein 1 homolog A (Xenopus laevis) | |
| Q8AVZ0 | Q8AVZ0_XENLA | frog | Myod1-a protein (Xenopus laevis) | |
| Q7T109 | Q7T109_XENTR | | MyoD protein (Xenopus tropicalis) | |



…And 100's more…

| Accession | Entry name | Status | Protein names | Organism | Length |
|---|---|---|---|---|---|
| Q7T109 | Q7T109_XENTR | ★ | **MyoD protein** | Xenopus tropicalis (Western clawed frog) (Silurana tropicalis) | 288 |

Some Details from #25

## Alignment 1 against Q7T109

| | | | |
|---|---|---|---|
| **Score** | 964 | **E-value** | 1.0 ×10 $^{-102}$ |
| **Identity** | 64.0% | **Positives** | 74.0% |
| **Query length** | 320 | **Match length** | 288 |
| **Position** | **Q7T109** matches from 1 to 288 (288AA), in the query sequence from 1 to 320 (320AA) | | |
| **Graphical** | | | |

```
1     MELLSPPLRDVDLTAPDGSLCSFATTDDFYDDPCFDSPDLRFFEDLDPRLMHVGALLKPE   60 P15172
      MELL PPLRD+++T  +GSLCSF T DDFYDDPCF++ D+ FFEDLDPRL+HV ALLKPE
1     MELLPPPLRDMEVT--EGSLCSFPTPDDFYDDPCFNTSDMSFFEDLDPRLVHV-ALLKPE   57 Q7T109

61    EHSHFPAAVHPAPGAREDEHVRAPSGHHQAGRCLLWACKACKRKTTNADRRKAATMRERR  120 P15172
      +  H             EDEHVRAPSGHHQAGRCLLWACKACKRKTTNADRRKAATMRERR
58    DPHH-----------NEDEHVRAPSGHHQAGRCLLWACKACKRKTTNADRRKAATMRERR  106 Q7T109

121   RLSKVNEAFETLKRCTSSNPNQRLPKVEILRNAIRYIEGLQALLRDQDAAPPGAAAAFYA  180 P15172
      RLSKVNEAFETLKRCTS+NPNQRLPKVEILRNAIRYIE LQ+LLR Q+         +FY
107   RLSKVNEAFETLKRCTSTNPNQRLPKVEILRNAIRYIESLQSLLRGQE-------ESFY-  158 Q7T109

181   PGPLPPGRGGEHYSGDSDASSPRSNCSDGMMDYSGPPSGARRRNCYEGAYYNEAPSEPRP  240 P15172
        P+       EHYSGDSDASSPRSNCSDGM DYS PP G+RRRN Y+ ++Y+++P+  R
159   --PVL-----EHYSGDSDASSPRSNCSDGMTDYS-PPCGSRRRNSYDSSFYSDSPNGLRL  210 Q7T109

241   GKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGES---SGDP  297 P15172
      GKS+ +SSLDCLSSIVERISTESP  P +  AD  SE  P       +P +GE+   SG
211   GKSSVISSLDCLSSIVERISTESPVCPVIPAADSGSEGSP-----CSPLQGETLSESGII  265 Q7T109
```

Filter · Overview · Results · Job information Customize order « Previous | Page 10 of 10

| Alignments | Entry | Entry name | Status | Protein names | Organism | Length | Identity | Score | E-value | Gene names |
|---|---|---|---|---|---|---|---|---|---|---|
| | B3LY60 | B3LY60_DROAN | ☆ | **GF18746** | Drosophila ananassae (Fruit fly) | 334 | 42.0% | 344 | $4.0\times10^{-30}$ | **GF18746** Dana\GF18746 Dana_GF18746 |
| | Q4RGJ6 | Q4RGJ6_TETNG | ☆ | **Chromosome undetermined SCAF15099, whole geno...** | Tetraodon nigroviridis (Spotted green pufferfish) (Chelonodon nigroviridis) | 126 | 57.0% | 343 | $5.0\times10^{-30}$ | GSTENG00034775001 |
| | F1NFS6 | F1NFS6_CHICK | ☆ | **Uncharacterized protein** | Gallus gallus (Chicken) | 237 | 46.0% | 343 | $5.0\times10^{-30}$ | **Gga.378** |
| | Q91151 | Q91151_NOTVI | ☆ | **Myogenic regulatory factor; transcription fac...** | Notophthalmus viridescens (Eastern newt) (Triturus viridescens) | 219 | 44.0% | 342 | $6.0\times10^{-30}$ | **MRF-4** |
| | Q29BN7 | Q29BN7_DROPS | ☆ | **GA10192** | Drosophila pseudoobscura pseudoobscura (Fruit fly) | 330 | 42.0% | 342 | $6.0\times10^{-30}$ | **GA10192** Dpse\GA10192 Dpse_GA10192 |
| | B4GP81 | B4GP81_DROPE | ☆ | **GL13832** | Drosophila persimilis (Fruit fly) | 330 | 42.0% | 342 | $6.0\times10^{-30}$ | **GL13832** Dper\GL13832 Dper_GL13832 |
| | Q92020 | MYF6_XENLA | ★ | **Myogenic factor 6** | Xenopus laevis (African clawed frog) | 240 | 44.0% | 340 | $1.0\times10^{-29}$ | **myf6** mrf4 |
| | B7ZQB0 | B7ZQB0_XENLA | ☆ | **MRF4a** | Xenopus laevis (African clawed frog) | 240 | 44.0% | 340 | $1.0\times10^{-29}$ | **MRF4** |
| | A7UCI1 | A7UCI1_XENLA | ☆ | **MRF4a** | Xenopus laevis (African clawed frog) | 240 | 44.0% | 340 | $1.0\times10^{-29}$ | **MRF4A** MRF4 |
| | F7FIX8 | F7FIX8_MONDO | ☆ | **Uncharacterized protein** | Monodelphis domestica (Gray short-tailed opossum) | 243 | 47.0% | 339 | $1.0\times10^{-29}$ | **MYF6** |
| | D9IV56 | D9IV56_EPICO | ☆ | **Myogenin** | Epinephelus coioides (Orange-spotted grouper) (Epinephelus nebulosus) | 250 | 47.0% | 338 | $2.0\times10^{-29}$ | |
| | Q6SYV5 | MYF6_TAKRU | ★ | **Myogenic factor 6** | Takifugu rubripes (Japanese pufferfish) (Fugu rubripes) | 239 | 46.0% | 337 | $2.0\times10^{-29}$ | **myf6** mrf4 |
| | G3WJP0 | G3WJP0_SARHA | ☆ | **Uncharacterized protein** | Sarcophilus harrisii (Tasmanian devil) (Sarcophilus laniarius) | 243 | 47.0% | 337 | $2.0\times10^{-29}$ | **MYF6** |
| | G7Y452 | G7Y452_CLOSI | ☆ | **Transcription factor SUM-1** | Clonorchis sinensis (Chinese liver fluke) | 946 | 64.0% | 337 | $2.0\times10^{-29}$ | CLF_100742 |
| | G4LXJ1 | G4LXJ1_SCHMA | ☆ | **Myogenic factor, putative** | Schistosoma mansoni (Blood fluke) | 864 | 68.0% | 337 | $2.0\times10^{-29}$ | Smp_167400 |
| | G1EN33 | G1EN33_DUGJA | ☆ | **Myogenic determinant factor** | Dugesia japonica (Planarian) | 498 | 42.0% | 336 | $3.0\times10^{-29}$ | **MDF** |
| | Q8MWP6 | Q8MWP6_SCHMD | ☆ | **MyoD-like protein** | Schmidtea mediterranea (Freshwater planarian flatworm) | 466 | 69.0% | 335 | $4.0\times10^{-29}$ | |
| | Q6PUV5 | MYF6_TETNG | ★ | **Myogenic factor 6** | Tetraodon nigroviridis (Spotted green pufferfish) (Chelonodon nigroviridis) | 239 | 45.0% | 333 | $7.0\times10^{-29}$ | **myf6** mrf4 GSTENG00021536001 |
| | Q2PZ12 | Q2PZ12_SALSA | ☆ | **Myogenin** | Salmo salar (Atlantic salmon) | 254 | 55.0% | 333 | $7.0\times10^{-29}$ | |
| | Q91207 | Q91207_ONCMY | ☆ | **TMyogenin protein** | Oncorhynchus mykiss (Rainbow trout) (Salmo gairdneri) | 254 | 45.0% | 332 | $9.0\times10^{-29}$ | **TMyogenin** |

hits at rank ~ 250 still extremely good matches, even though very distantly related organisms

# The foregoing search capability is a *huge* deal

millions of searches daily

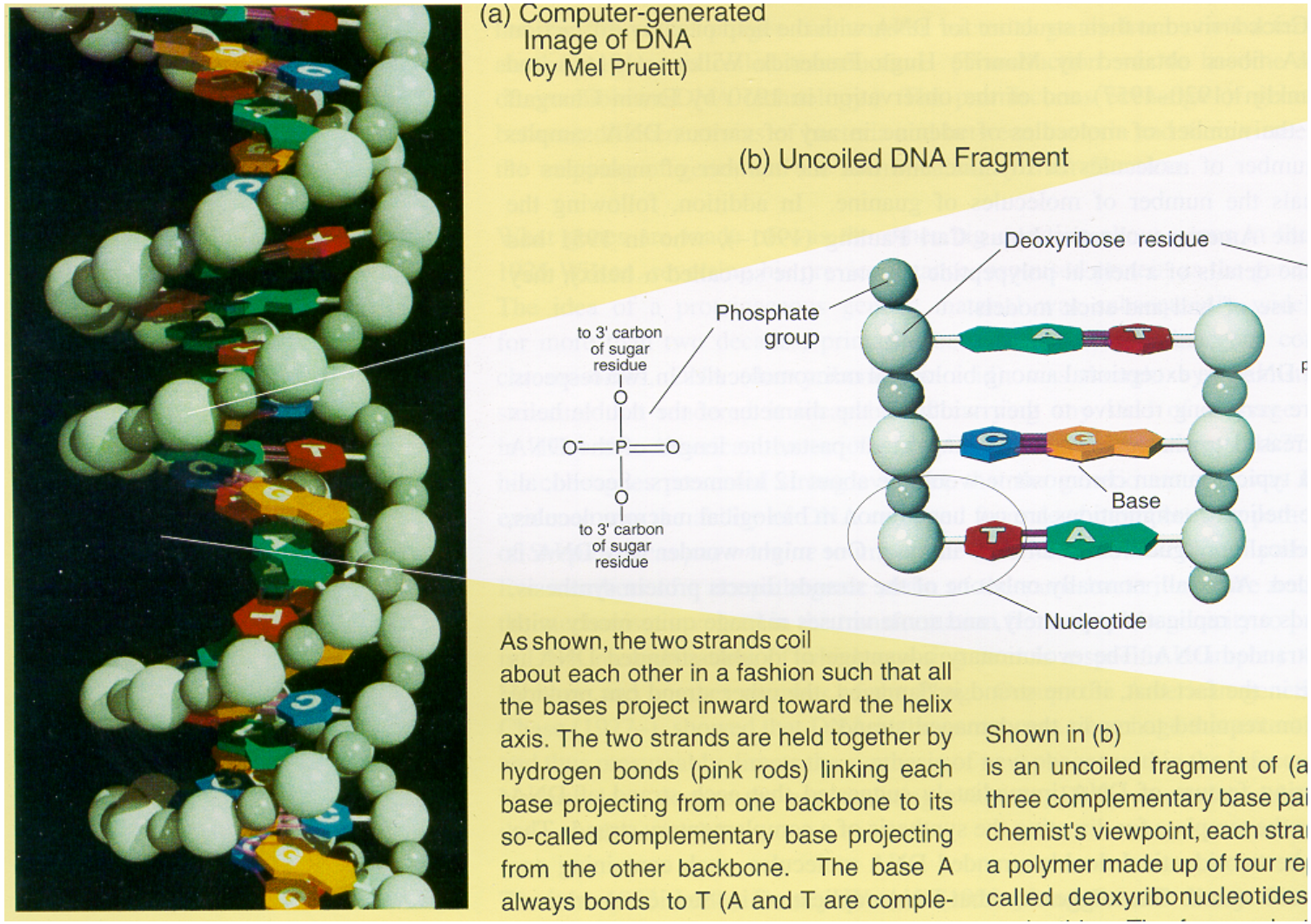
biologists (not just "computational" biologists) use this routinely

it connects information about *all* living things

More on algorithm later …

# Application: RNA structure

# The Double Helix




Los Alamos Science

http://www.rcsb.org/pdb/explore.do?structureId=1GAT

NATURE VOL. 227 AUGUST 8 1970

# Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

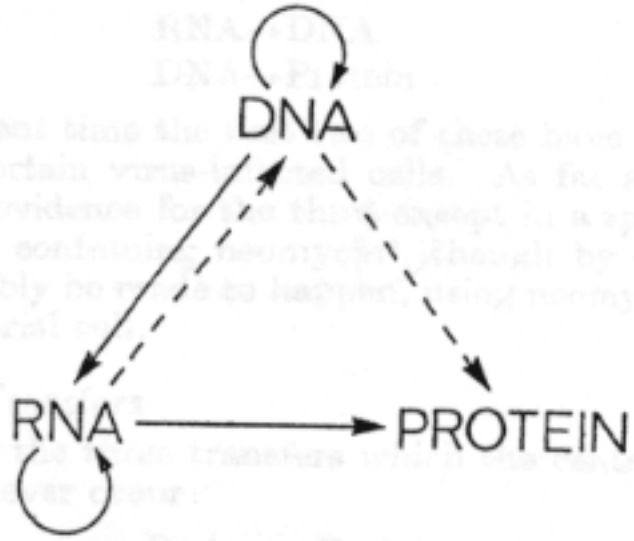



Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

# Non-coding RNA

Messenger RNA - codes for proteins

Non-coding RNA - all the rest

- Before, say, mid 1990's, 1-2 dozen known
  (critically important, but narrow roles: e.g., tRNA)

Since mid 90's dramatic discoveries

- Regulation, transport, stability/degradation
- E.g. "miRNA": >1000 in humans; regulate >50% of genes
- E.g. "riboswitches": 10000's in bacteria

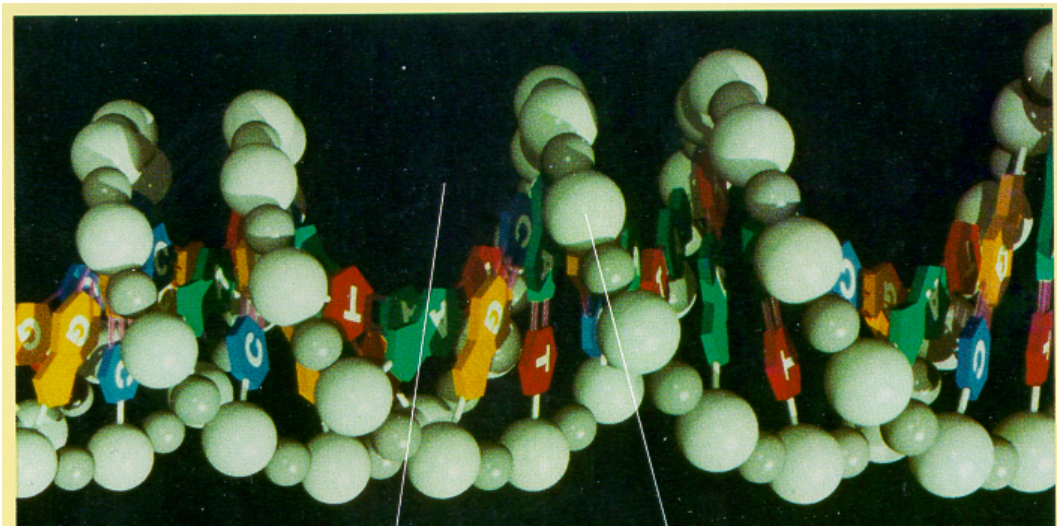
*By some estimates, ncRNA >> mRNA*

# DNA structure: dull

5'…ACCGCTAGATG…3'

| | | | | | | | | | |

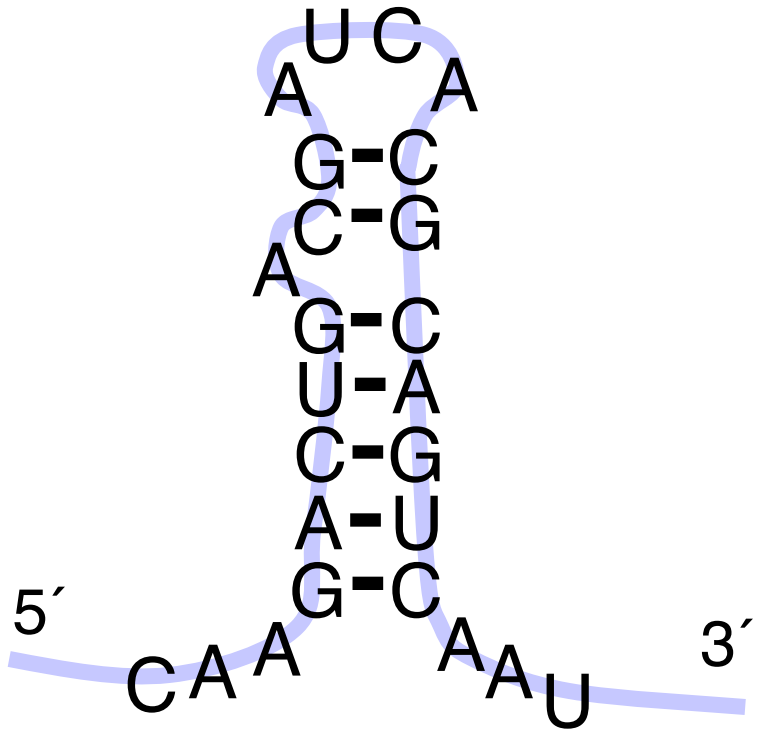
3'…TGGCGATCTAC…5'



# RNA Secondary Structure:
## RNA makes helices too

Base pairs

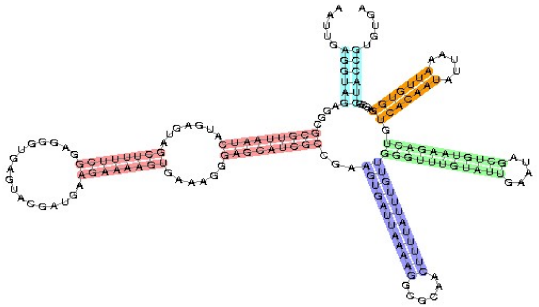
A-U
C-G

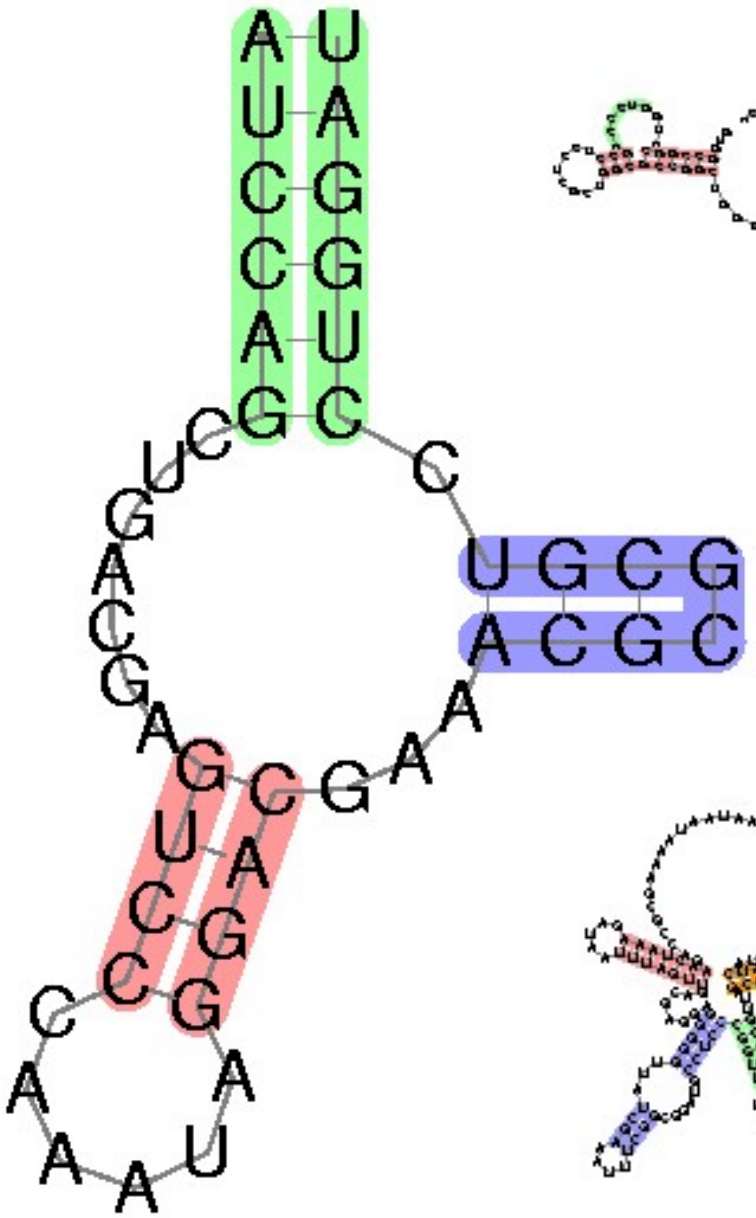


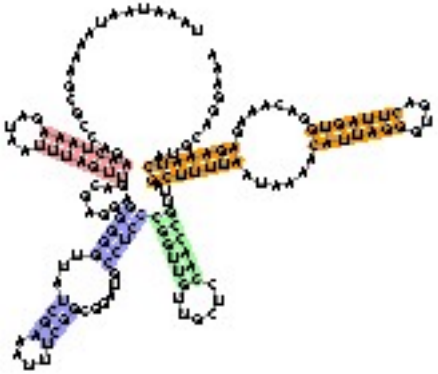

Usually *single* stranded

# RNA Secondary Structure:

Not everything, but important, easier than 3d



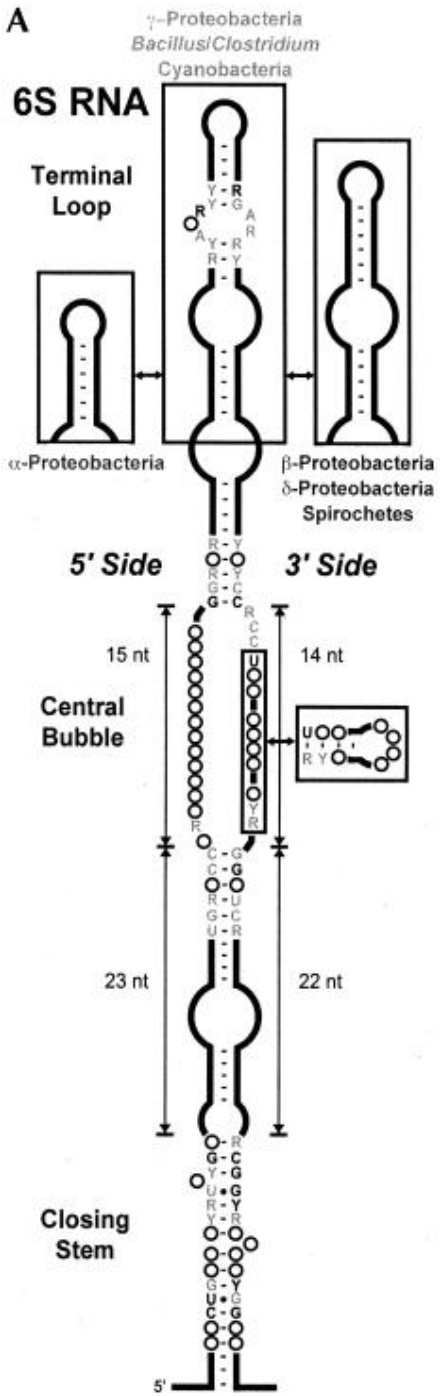


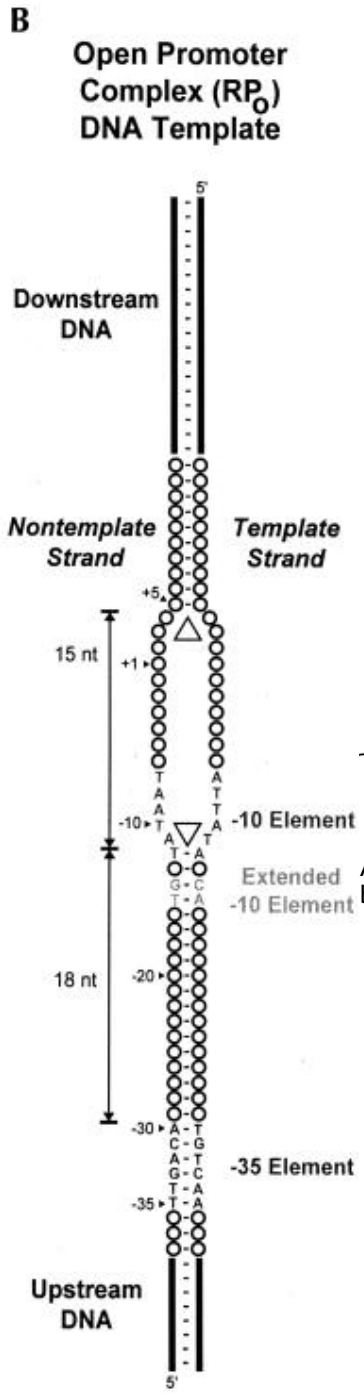


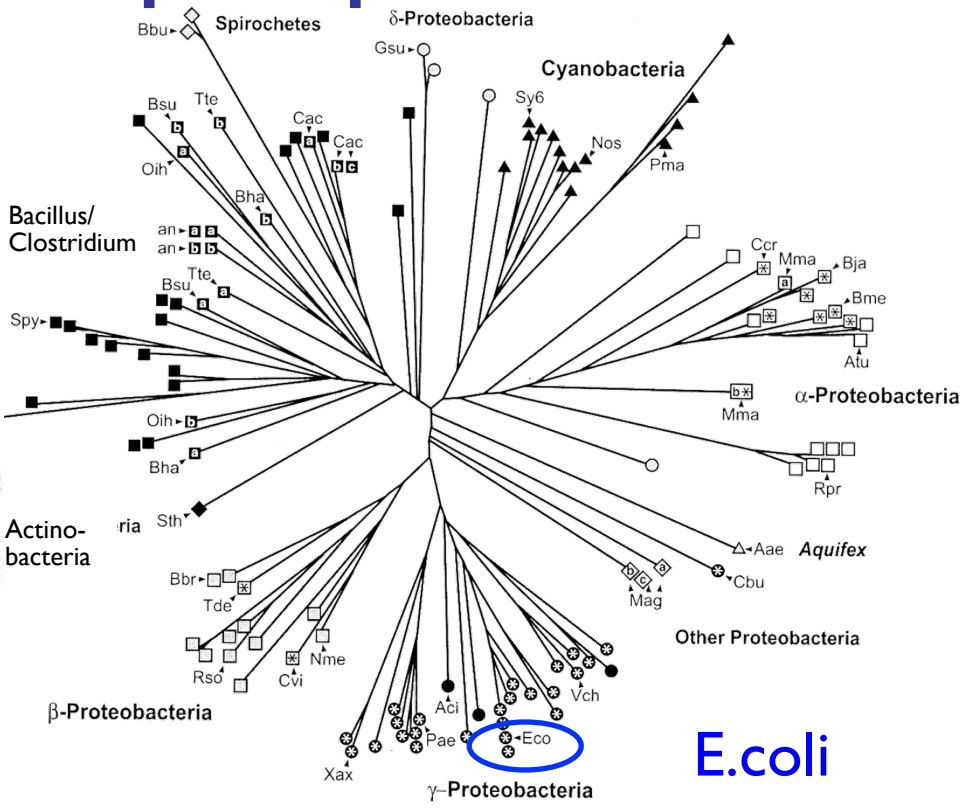
# Why is structure important?

- For protein-coding, similarity in sequence is a powerful tool for finding related sequences
  - e.g. "hemoglobin," "MyoD" and many others are easily recognized in all animals
- For many non-coding RNAs, *different sequences* can have the *same structure*, and structure is most important for function.
  - So, using structure plus sequence, can find related sequences at much greater evolutionary distances

# 6S mimics an open promoter

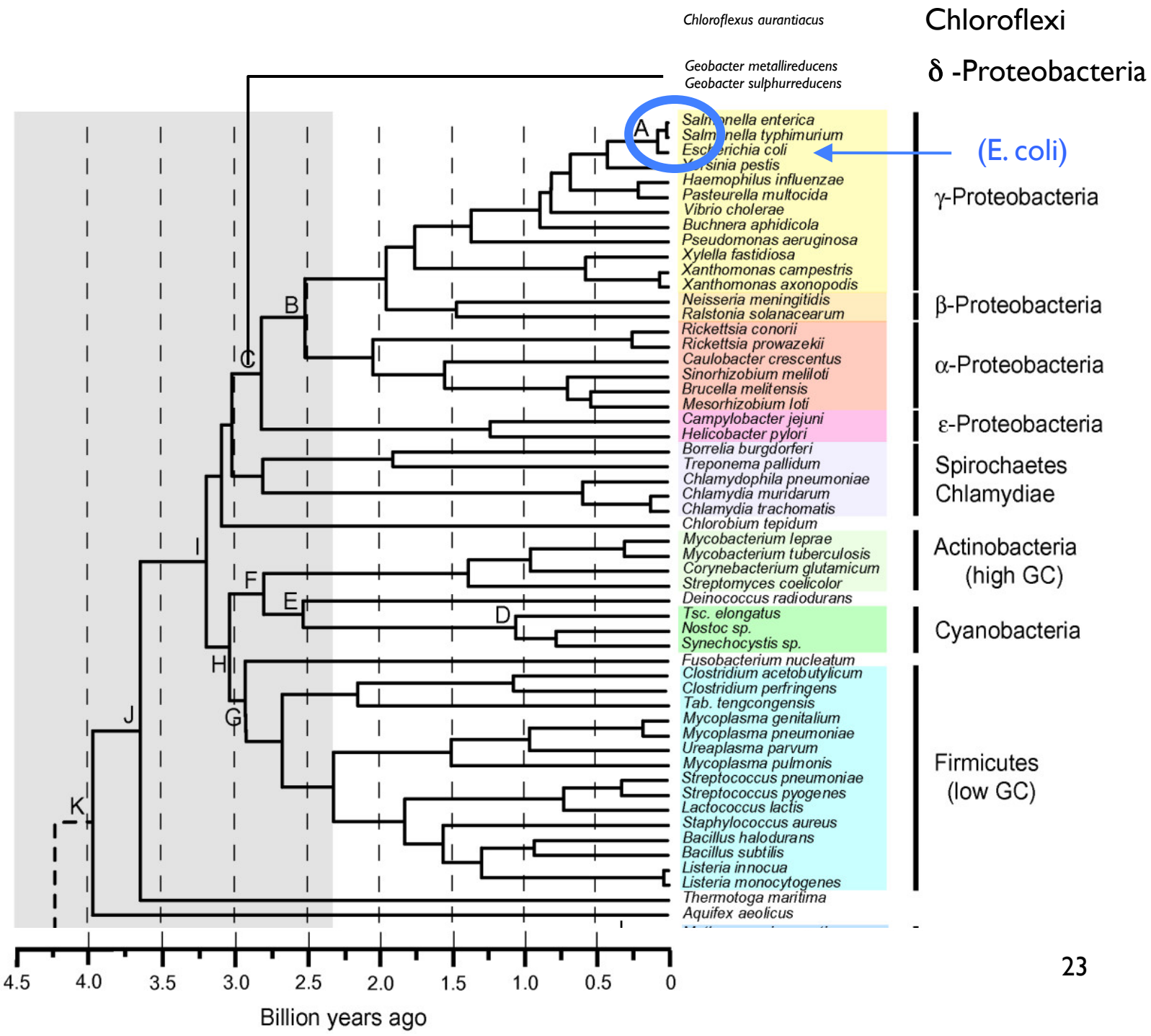


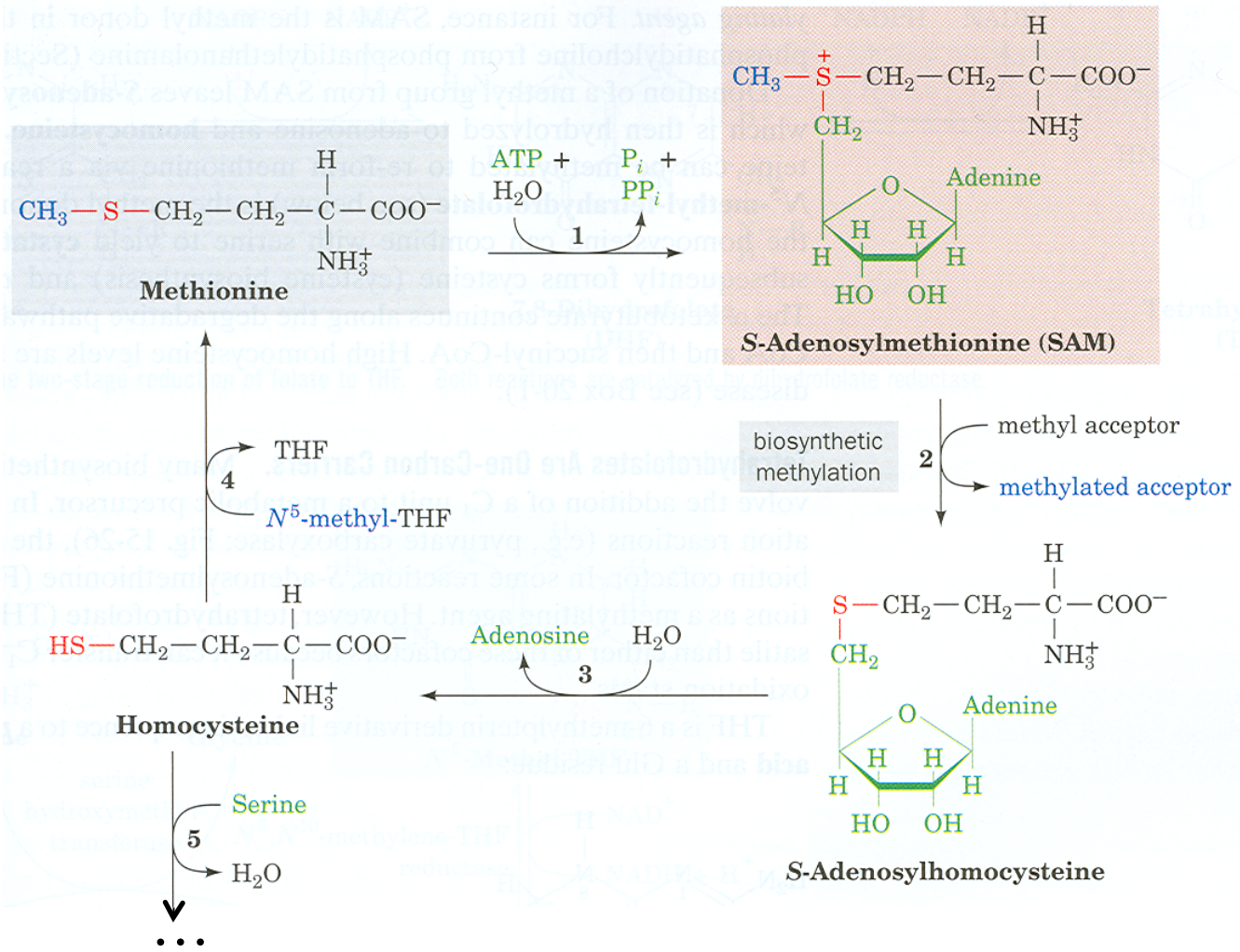




E.coli

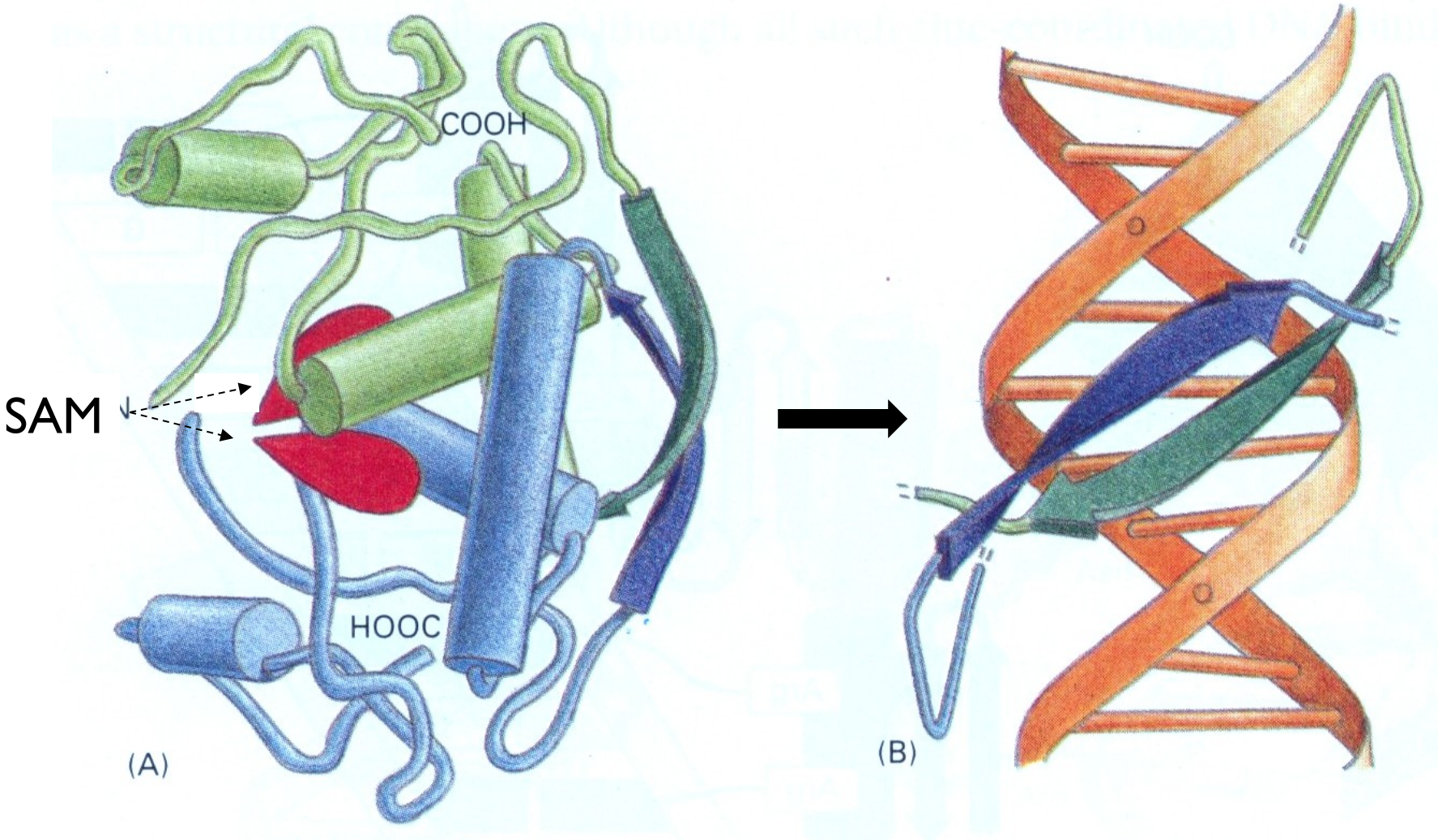
Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005


Chloroflexus aurantiacus
Geobacter metallireducens
Geobacter sulphurreducens
Salmonella enterica
Salmonella typhimurium
Escherichia coli
Yersinia pestis
Haemophilus influenzae
Pasteurella multocida
Vibrio cholerae
Buchnera aphidicola
Pseudomonas aeruginosa
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
Neisseria meningitidis
Ralstonia solanacearum
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Sinorhizobium meliloti
Brucella melitensis
Mesorhizobium loti
Campylobacter jejuni
Helicobacter pylori
Borrelia burgdorferi
Treponema pallidum
Chlamydophila pneumoniae
Chlamydia muridarum
Chlamydia trachomatis
Chlorobium tepidum
Mycobacterium leprae
Mycobacterium tuberculosis
Corynebacterium glutamicum
Streptomyces coelicolor
Deinococcus radiodurans
Tsc. elongatus
Nostoc sp.
Synechocystis sp.
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
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
Chloroflexi
δ -Proteobacteria
(E. coli)
γ-Proteobacteria
β-Proteobacteria
α-Proteobacteria
ε-Proteobacteria
Spirochaetes
Chlamydiae
Actinobacteria
(high GC)
Cyanobacteria
Firmicutes
(low GC)

# In Bacteria: A typical biosynthetic cycle around a critical metabolite ("SAM")



# Gene Regulation: The MET Repressor



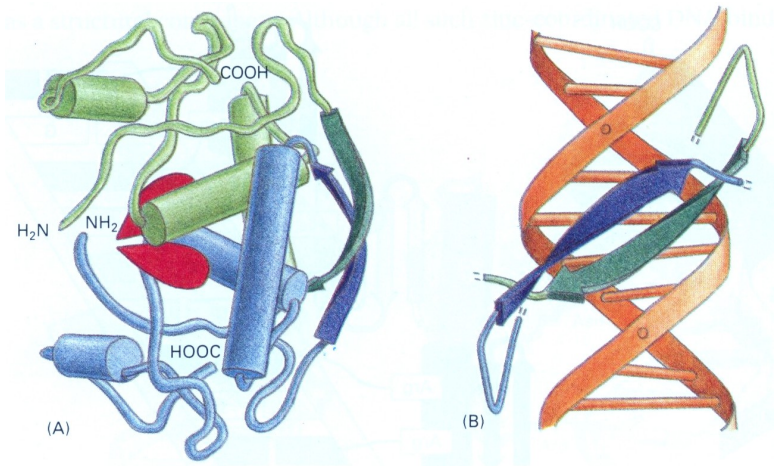

Protein

Alberts, et al, 3e.

DNA

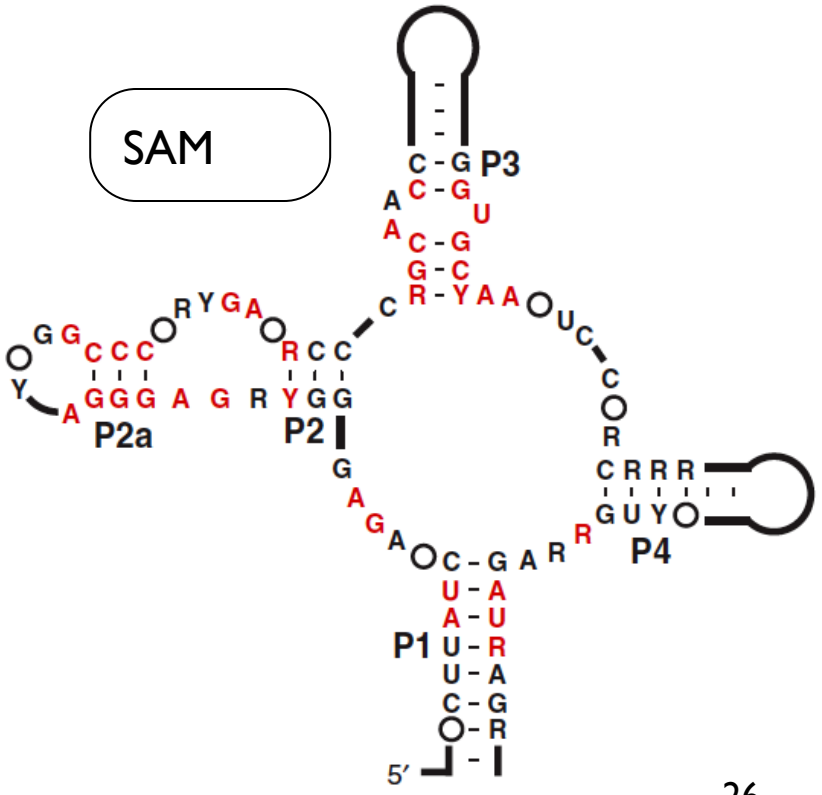
# Not the only way!

Alberts, et al, 3e.




Protein way

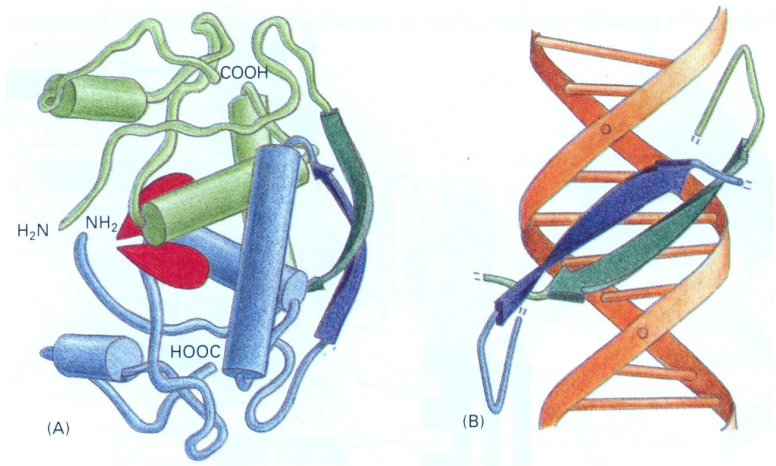
Riboswitch alternative




Grundy & Henkin, Mol. Microbiol 1998
Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003

# Not the only way!

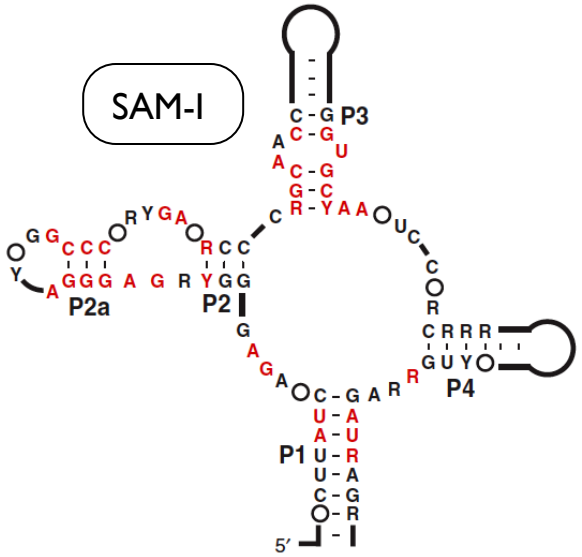
Alberts, et al, 3e.




Protein way

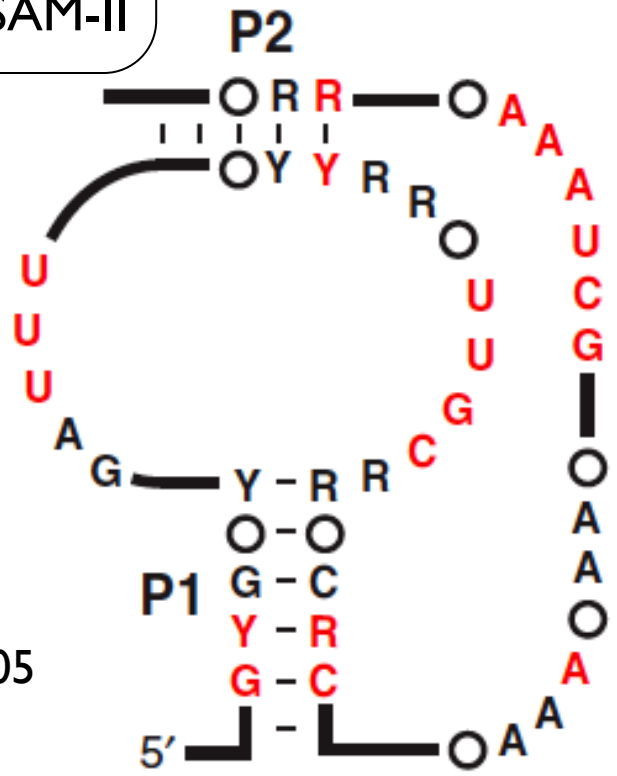
Riboswitch alternatives

SAM-I




Grundy, Epshtein, Winkler et al., 1998, 2003

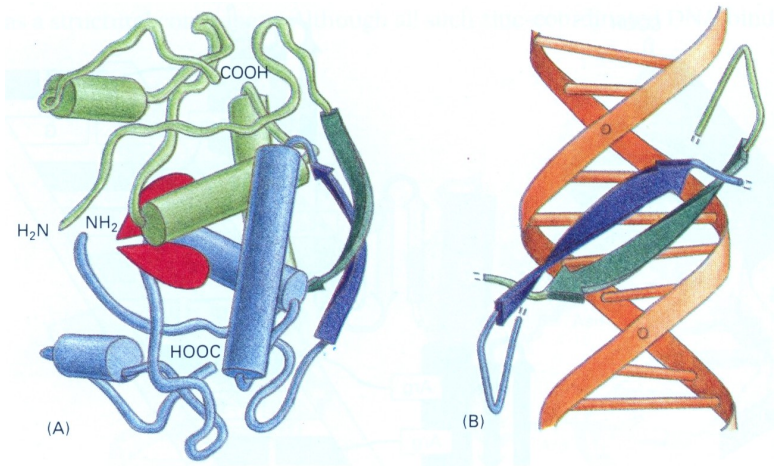
SAM-II




Corbino et al., Genome Biol. 2005

# Not the only way!

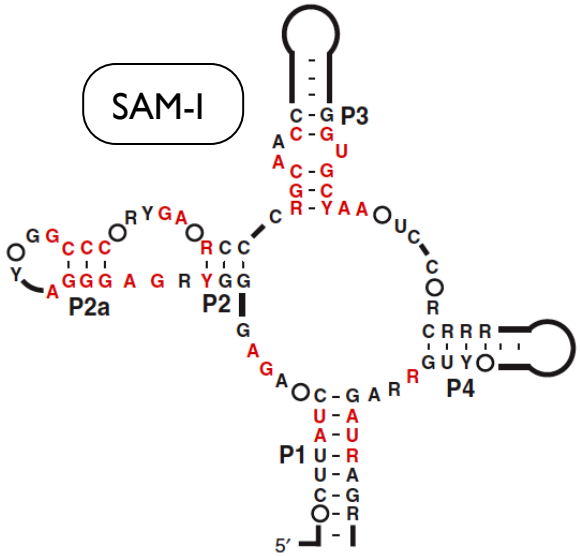
Alberts, et al, 3e.




Protein way

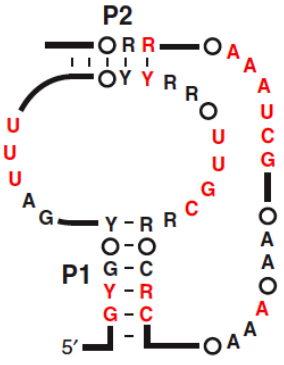
Riboswitch alternatives

SAM-I



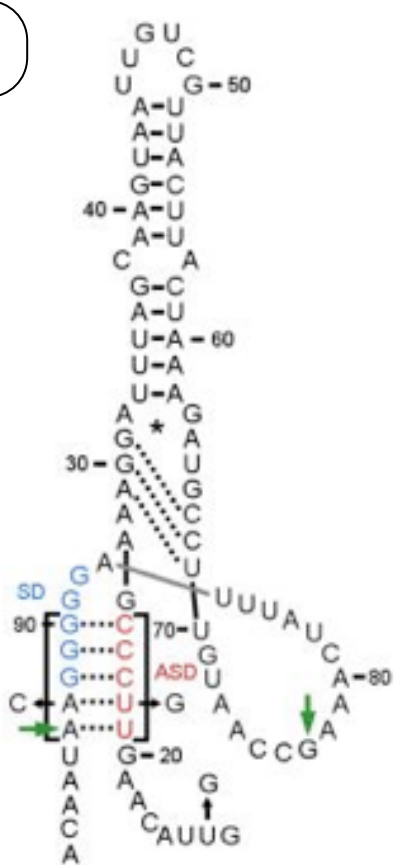

Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



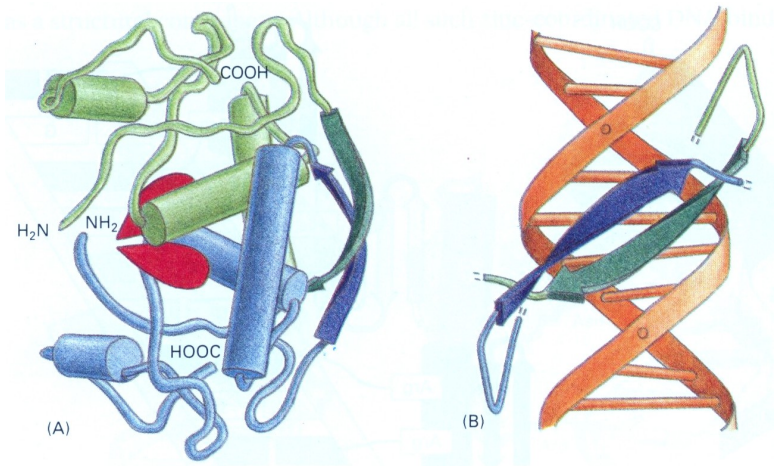

Corbino et al., Genome Biol. 2005

SAM-III




Fuchs et al., NSMB 2006

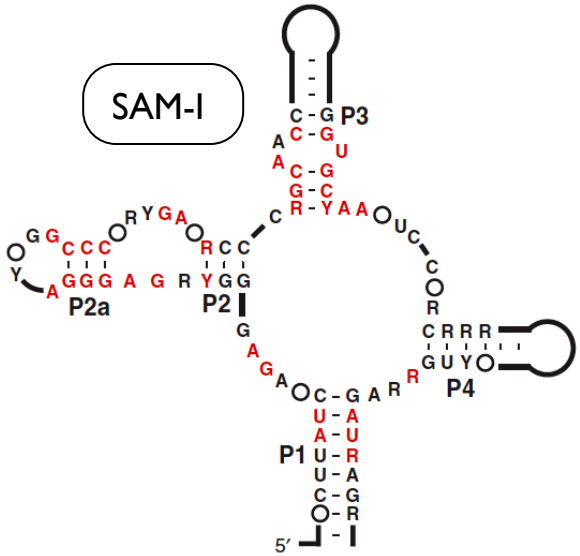
# Not the only way!



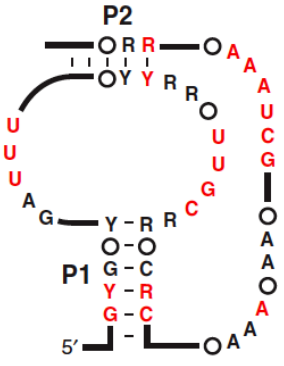

Alberts, et al, 3e.

← Protein way

Riboswitch alternatives ↓

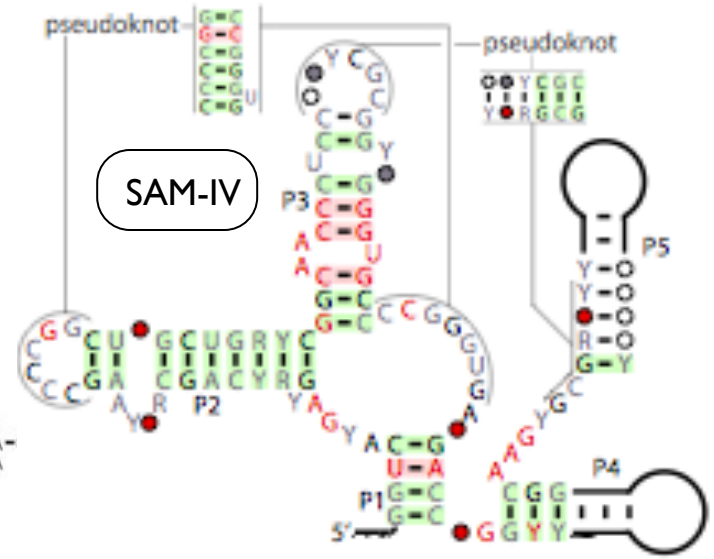



Grundy, Epshtein, Winkler et al., 1998, 2003



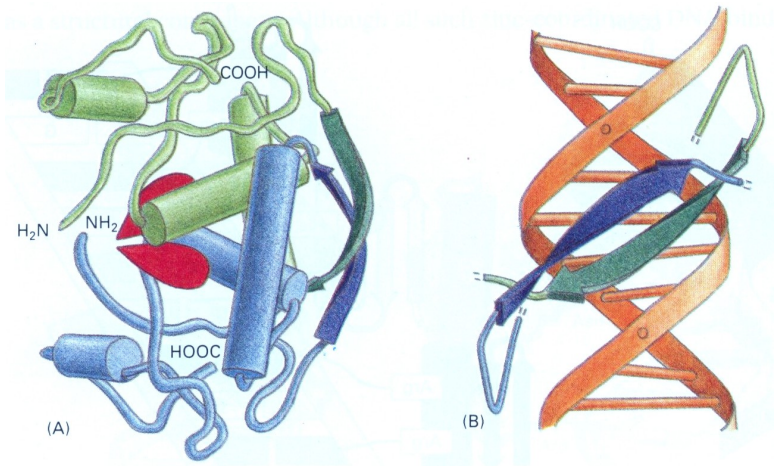

Corbino et al., Genome Biol. 2005


Fuchs et al., NSMB 2006




Weinberg et al., RNA 2008

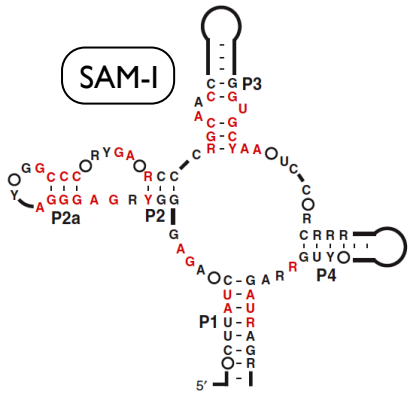
# Not the only way!



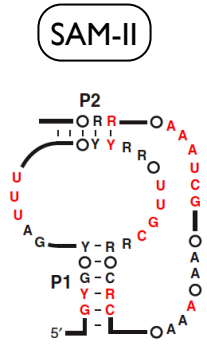

Alberts, et al, 3e.

← Protein way

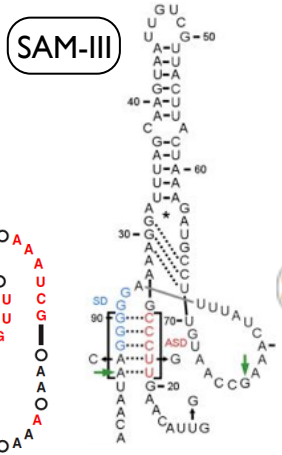
Riboswitch alternatives ↓



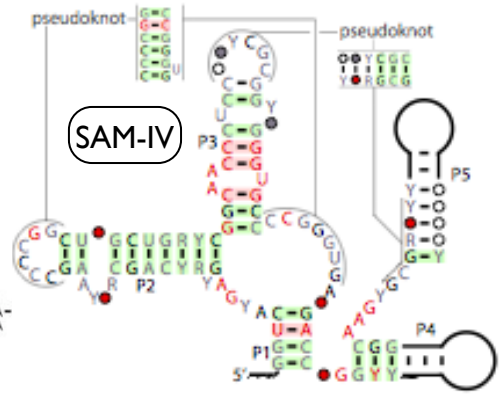

Grundy, Epshtein, Winkler et al., 1998, 2003



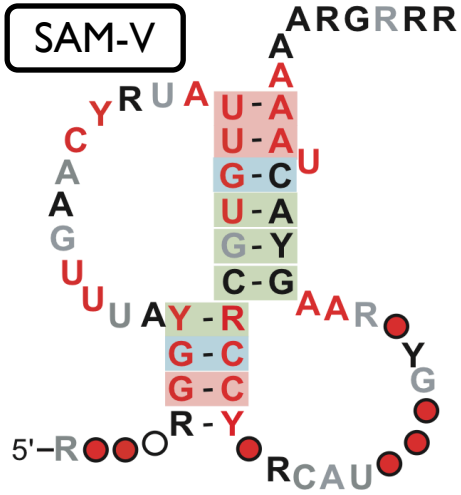

Corbino et al., Genome Biol. 2005



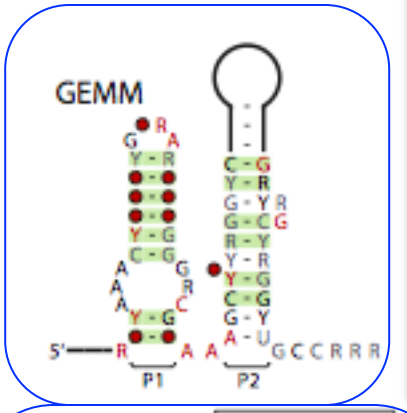

Fuchs et al., NSMB 2006



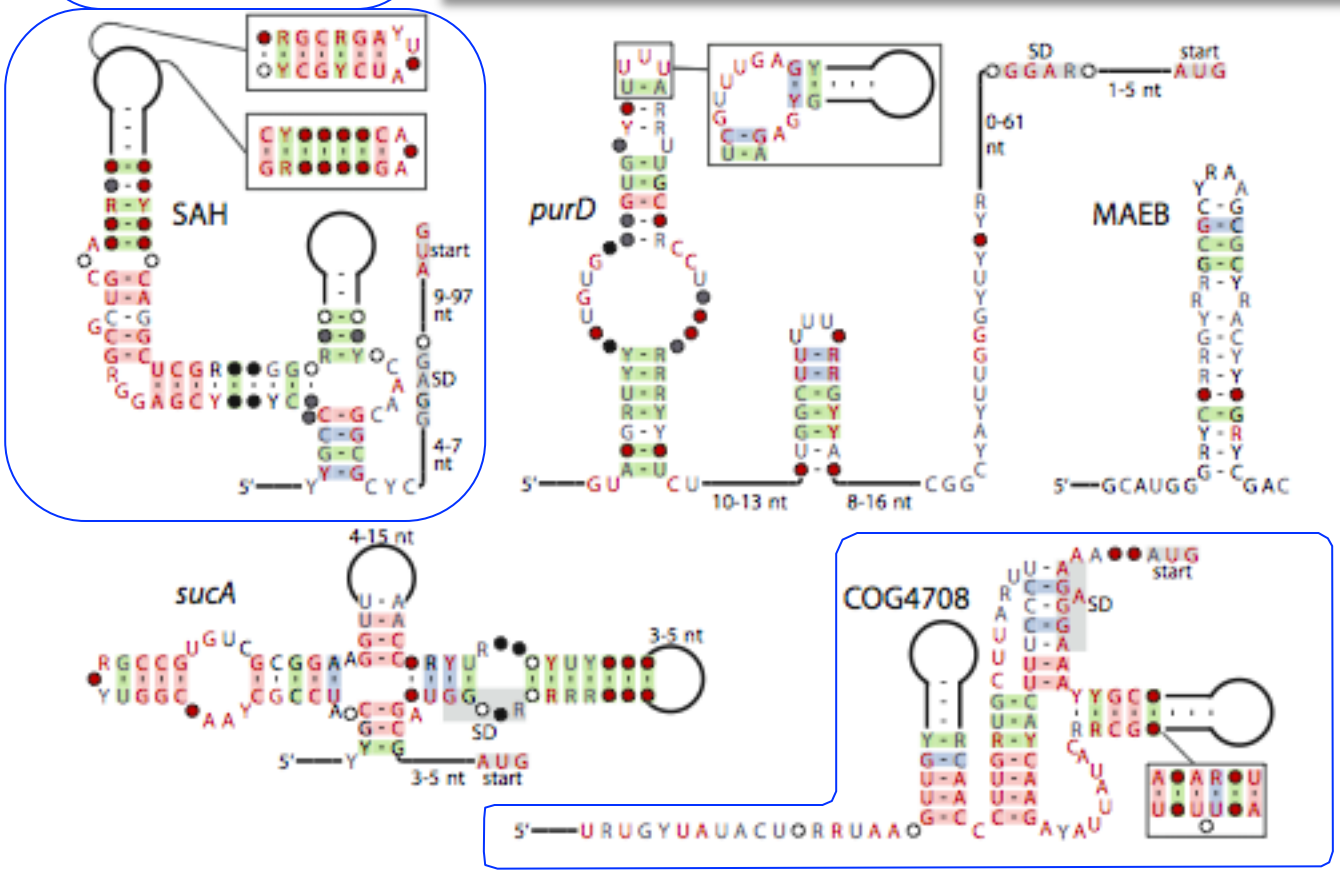

Weinberg et al., RNA 2008




Meyer, etal., BMC Genomics 2009



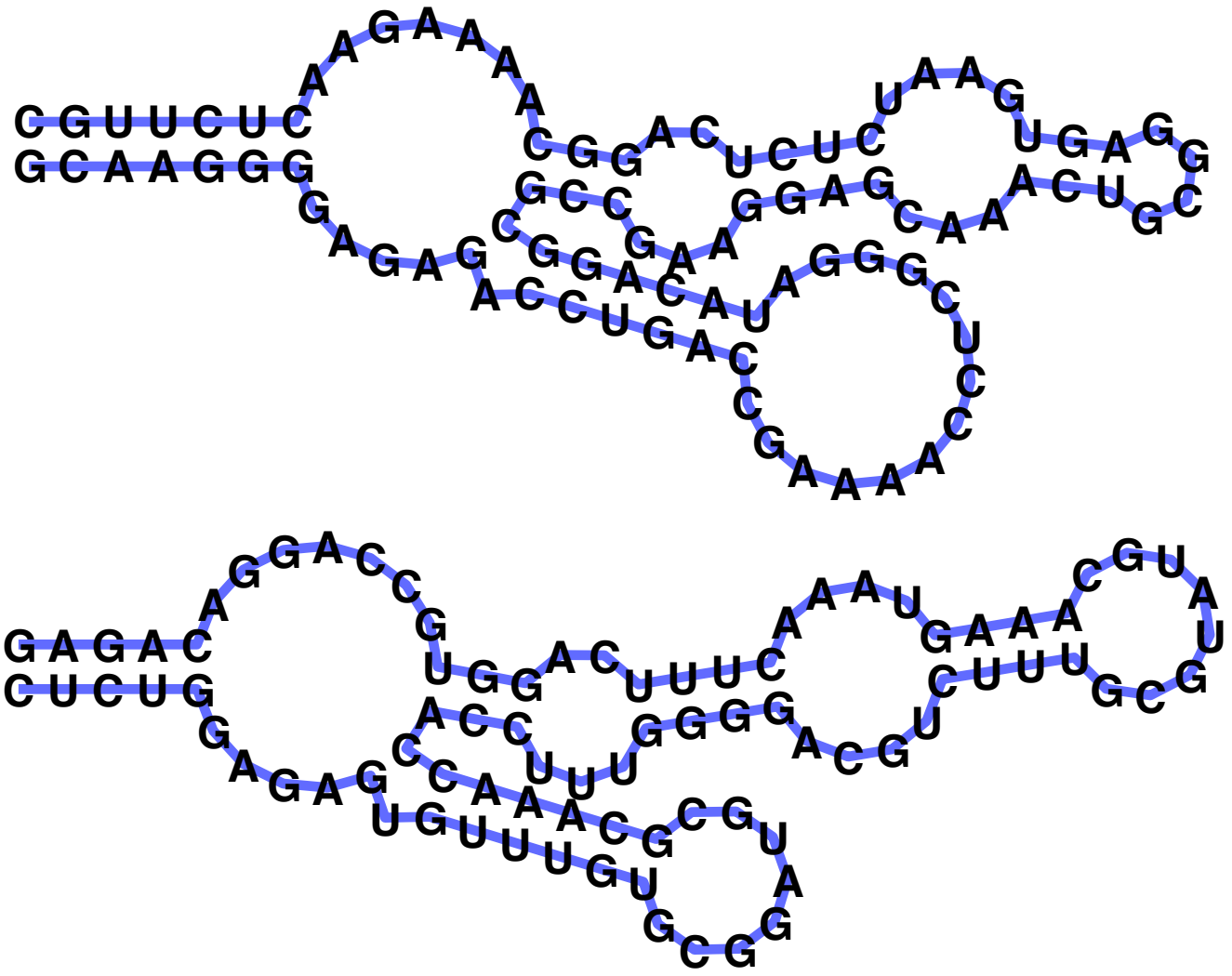

And many other examples. Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.




boxed = confirmed riboswitch (+2 more)

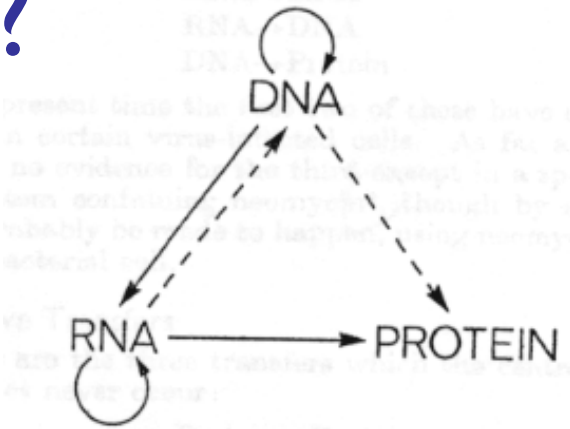
Weinberg, et al. Nucl. Acids Res., July 2007 35: 4809-4819.

# Why is RNA hard to deal with?



A: Structure often more important than sequence

# 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-directed RNA polymerase

RNA can form complex structures

RNA enzymes exist (ribozymes)

RNA can control, do logic (riboswitches)

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

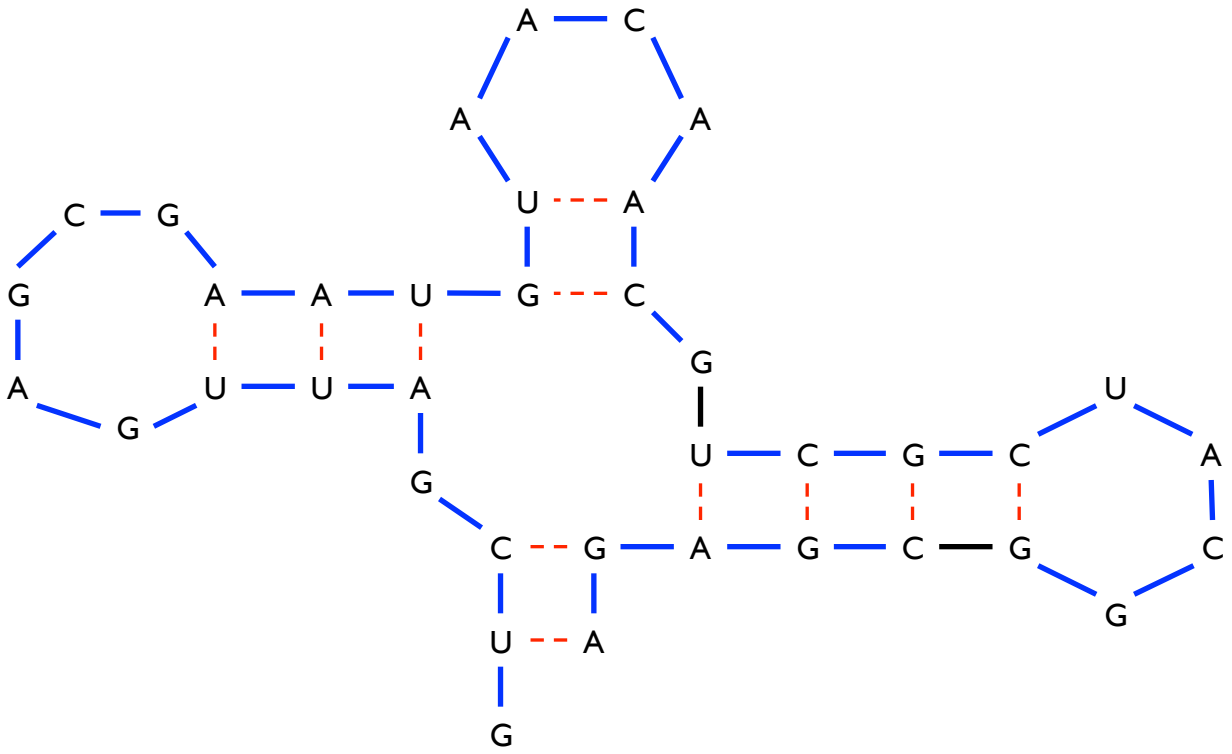
# 6.5 RNA Secondary Structure

Nussinov's Algorithm – core technology
for RNA structure prediction

# RNA Secondary Structure

RNA. String B = $b_1 b_2 \ldots b_n$ over alphabet { A, C, G, U }.

Secondary structure. RNA is usually single-stranded, and tends to loop back and form base pairs with itself. This structure is essential for understanding molecular behavior.




backbone

pairing

complementary base pairs: A-U, C-G

Ex: **GUCGAUUGAGCGAAUGUAACAACGUGGCUACGGCGAGA**

# 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*.)

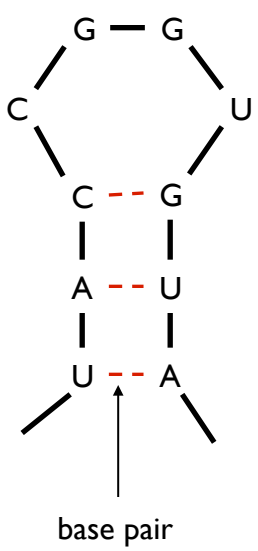
Free energy. Usual hypothesis is that an RNA molecule will form the structure with the optimum total free energy.

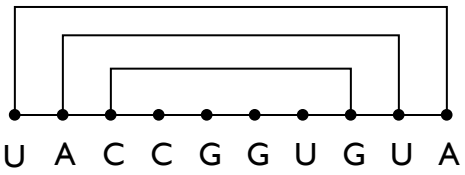
approximated by number of base pairs

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.

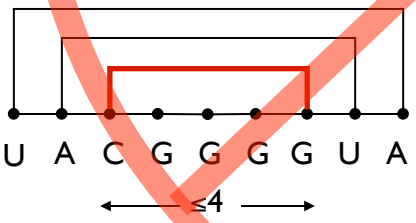
# RNA Secondary Structure: Examples

Examples.

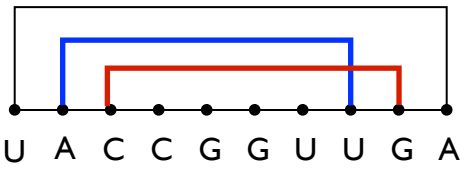


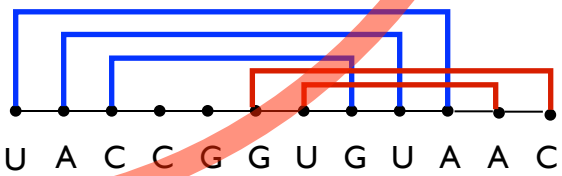




ok




sharp turn

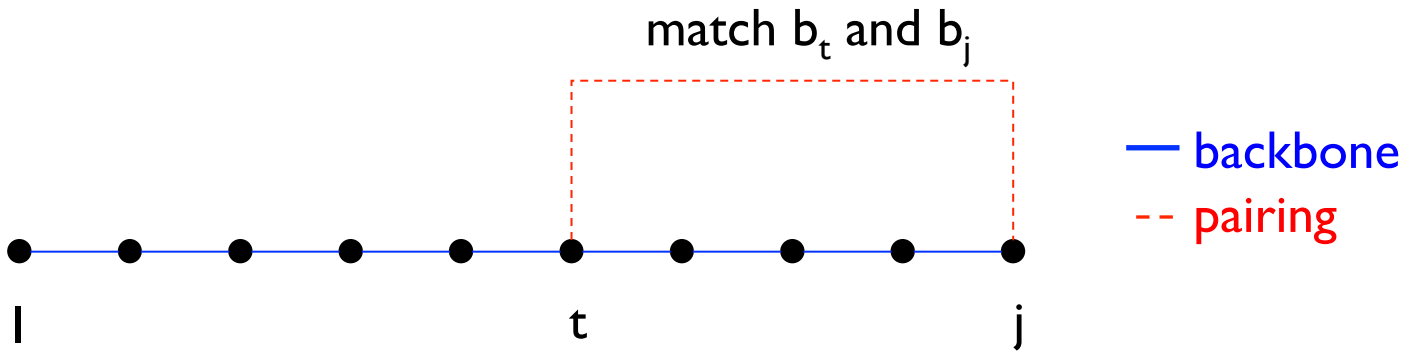



crossing

# 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 flexible set of 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

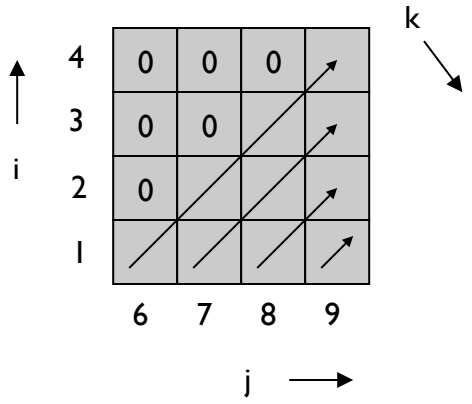
Key point: Either last base is unpaired (case 1,2) or paired (case 3)

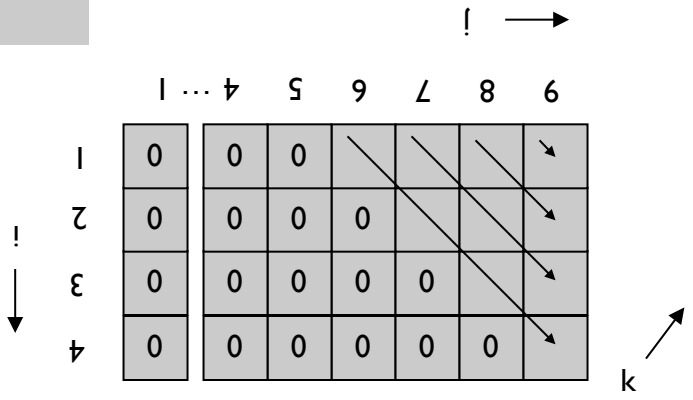
Remark. Core idea in CKY algorithm for context-free parsing

# Bottom Up Dynamic Programming Over Intervals

Q. What order to solve the sub-problems?
A. One way–do shortest intervals first:

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




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

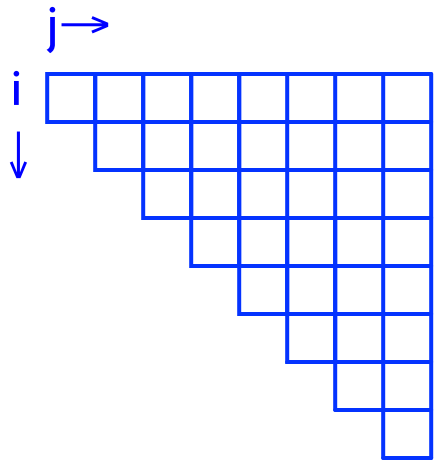
# Nussinov: Max Pairing

opt(i,j) = # pairs in optimal pairing of $b_i$ ... $b_j$

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

opt(i,j) = max of:

$$\begin{cases} \text{opt}(i,j-1) \\ \max \{ \text{opt}(i,t-1)+1+\text{opt}(t+1,j-1) \mid \\ \qquad i \le t < j-4 \text{ and } b_t\text{-}b_j \text{ may pair} \} \end{cases}$$



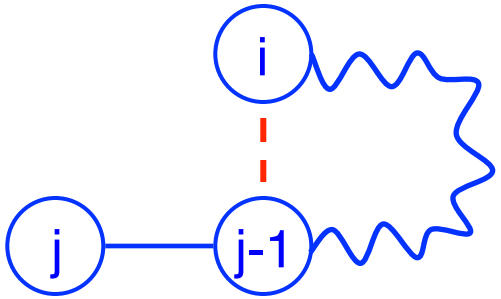

R Nussinov, AB Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA." PNAS 1980

# "Optimal pairing of $b_i$ ... $b_j$"

Two possibilities:

j Unpaired:

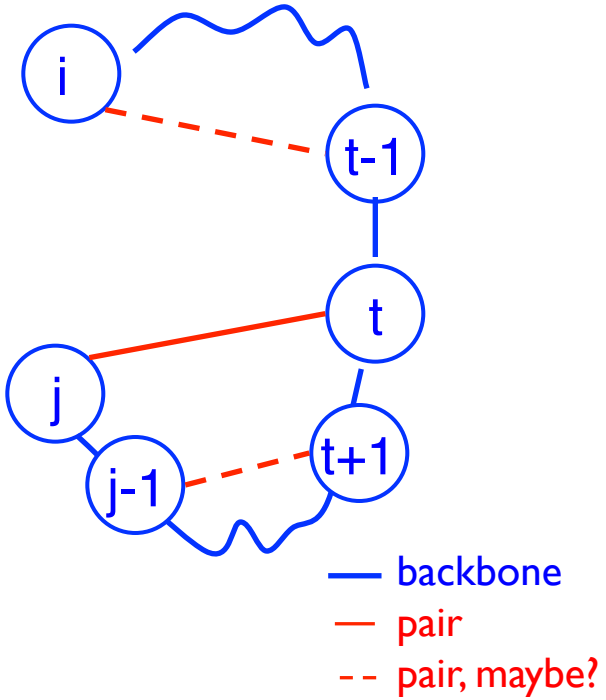
Find best pairing of $b_i$ ... $b_{j-1}$




j Paired (with some t):

Find best $b_i$ ... $b_{t-1}$ +

best $b_{t+1}$ ... $b_{j-1}$ plus 1




Why is it slow?

Why do pseudoknots matter?

# Another Computation Order

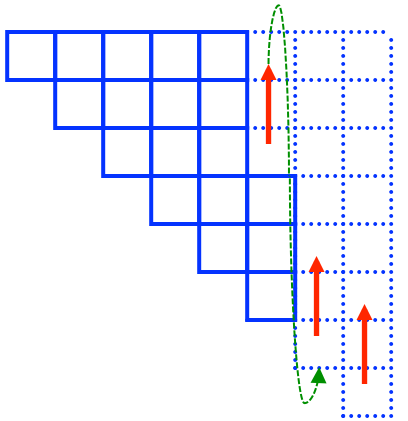
opt(i,j) = # pairs in optimal pairing of $b_i$ ... $b_j$

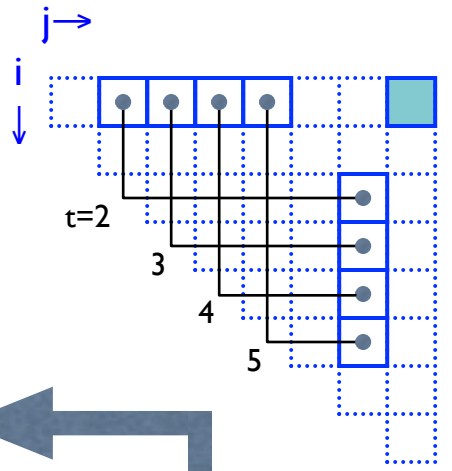
opt(i,j) = 0 for all i, j with $i \geq j-4$;

otherwise:

opt(i,j) = max of:

$$\begin{cases} \text{opt}(i,j-1) \\ \max \{ \text{opt}(i,t-1)+1+\text{opt}(t+1,j-1) \mid \\ \qquad i \leq t < j-4 \text{ and } b_t\text{-}b_j \text{ may pair} \} \end{cases}$$

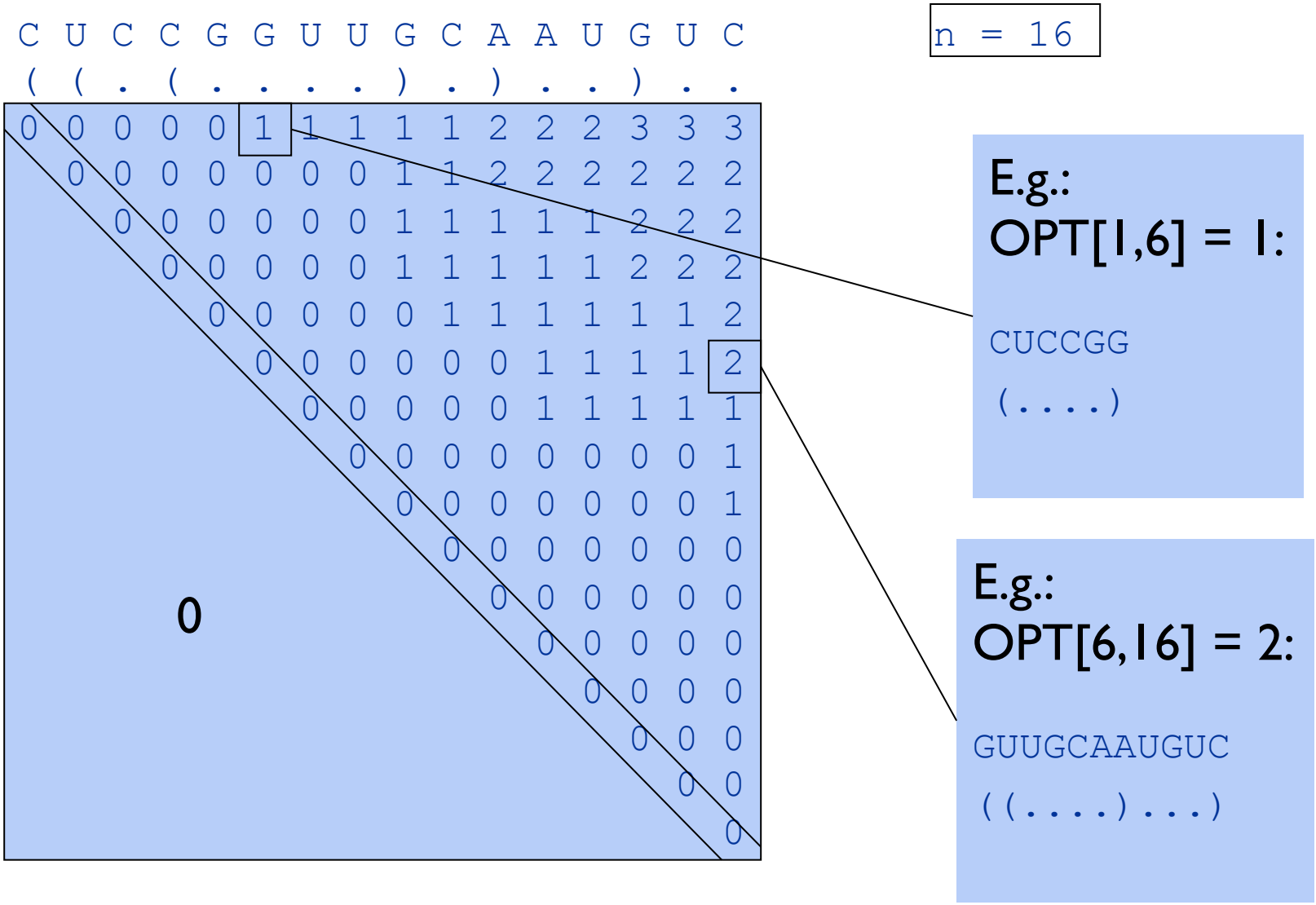





Time: $O(n^3)$

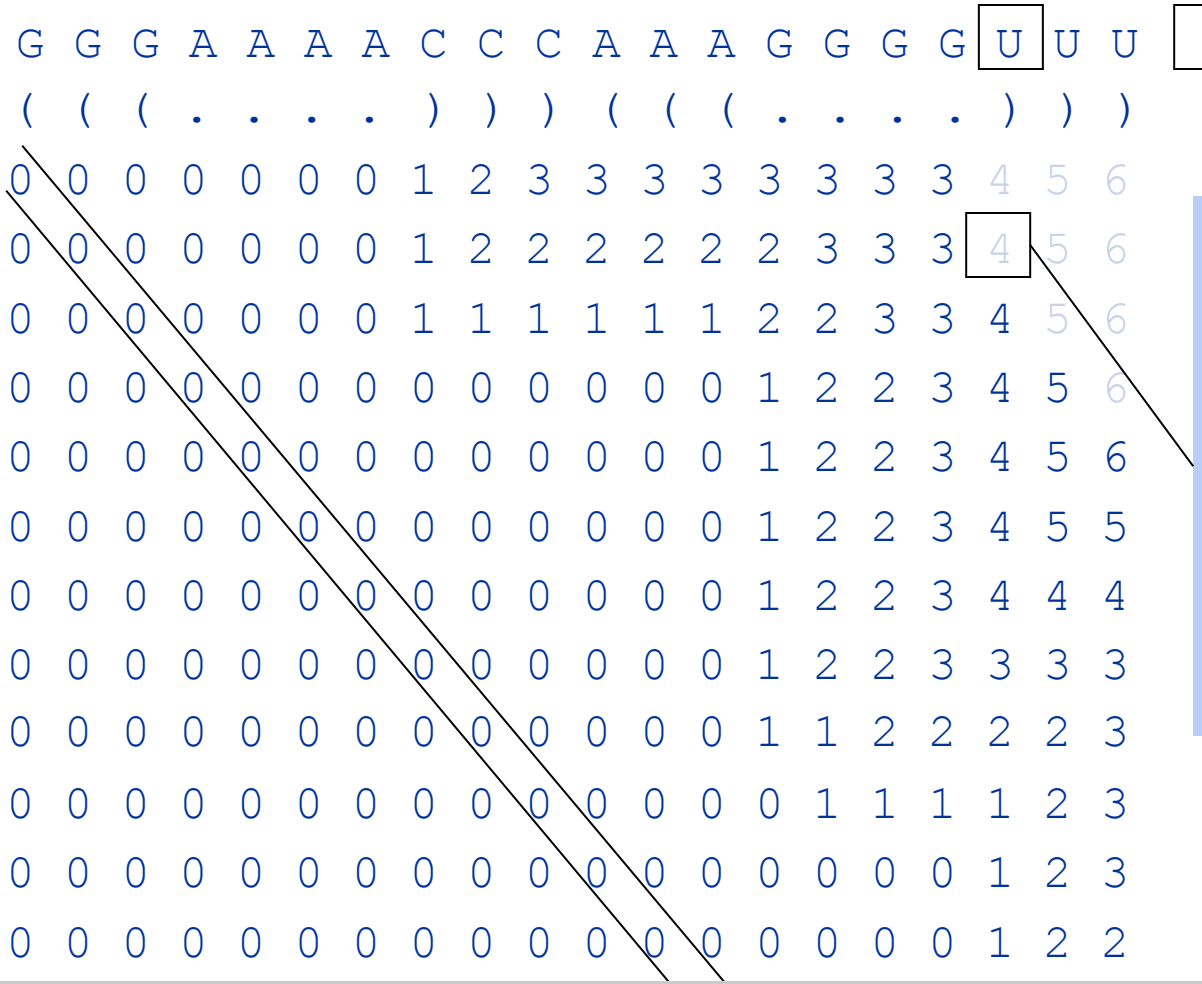
# Which Pairs?

Usual dynamic programming "trace-back" tells you *which* base pairs are in the optimal solution, not just how many




(Examples here and below assume 1-based indexing)

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



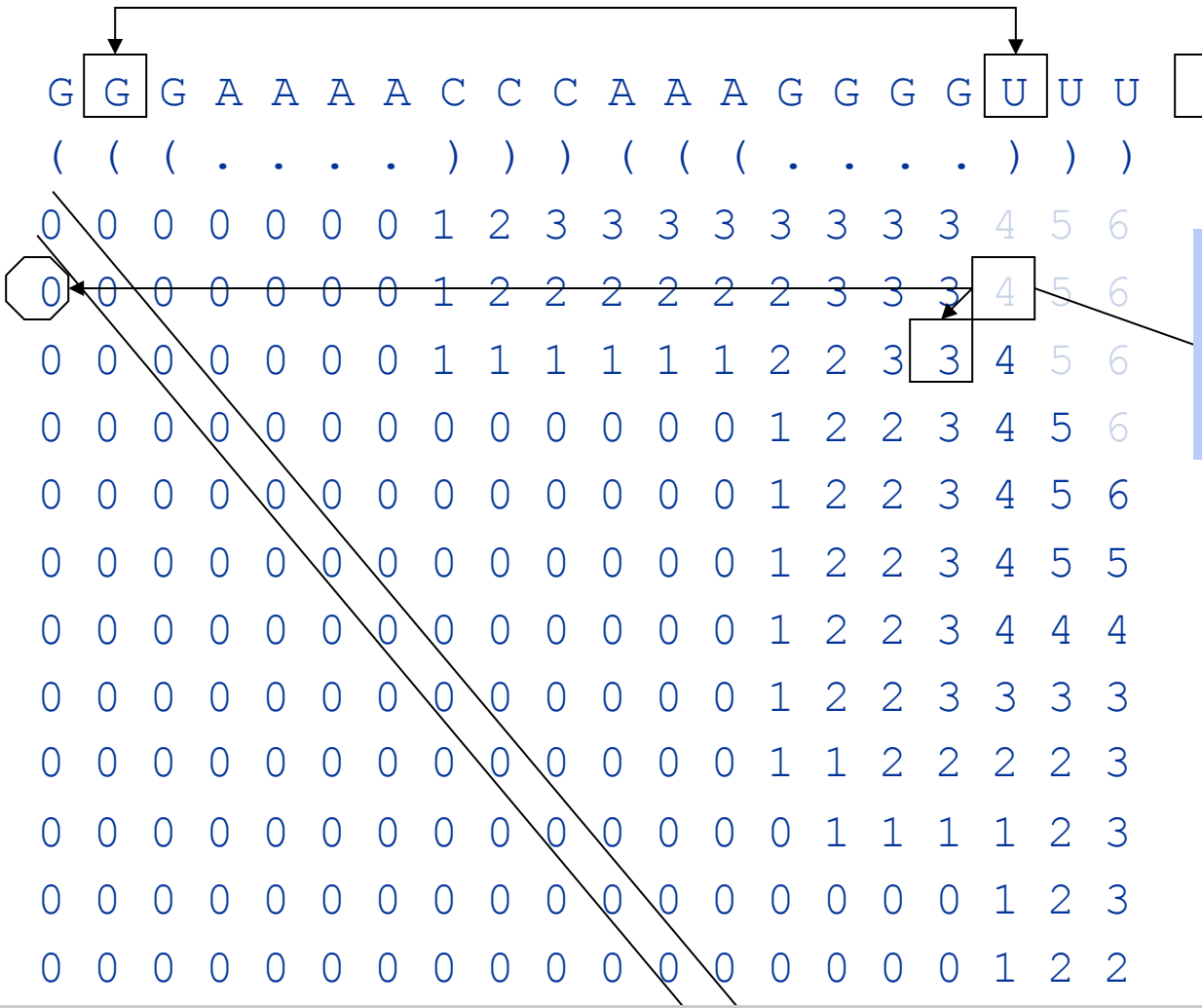

Case 1:
2 ≥ 18-4? no.

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

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

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

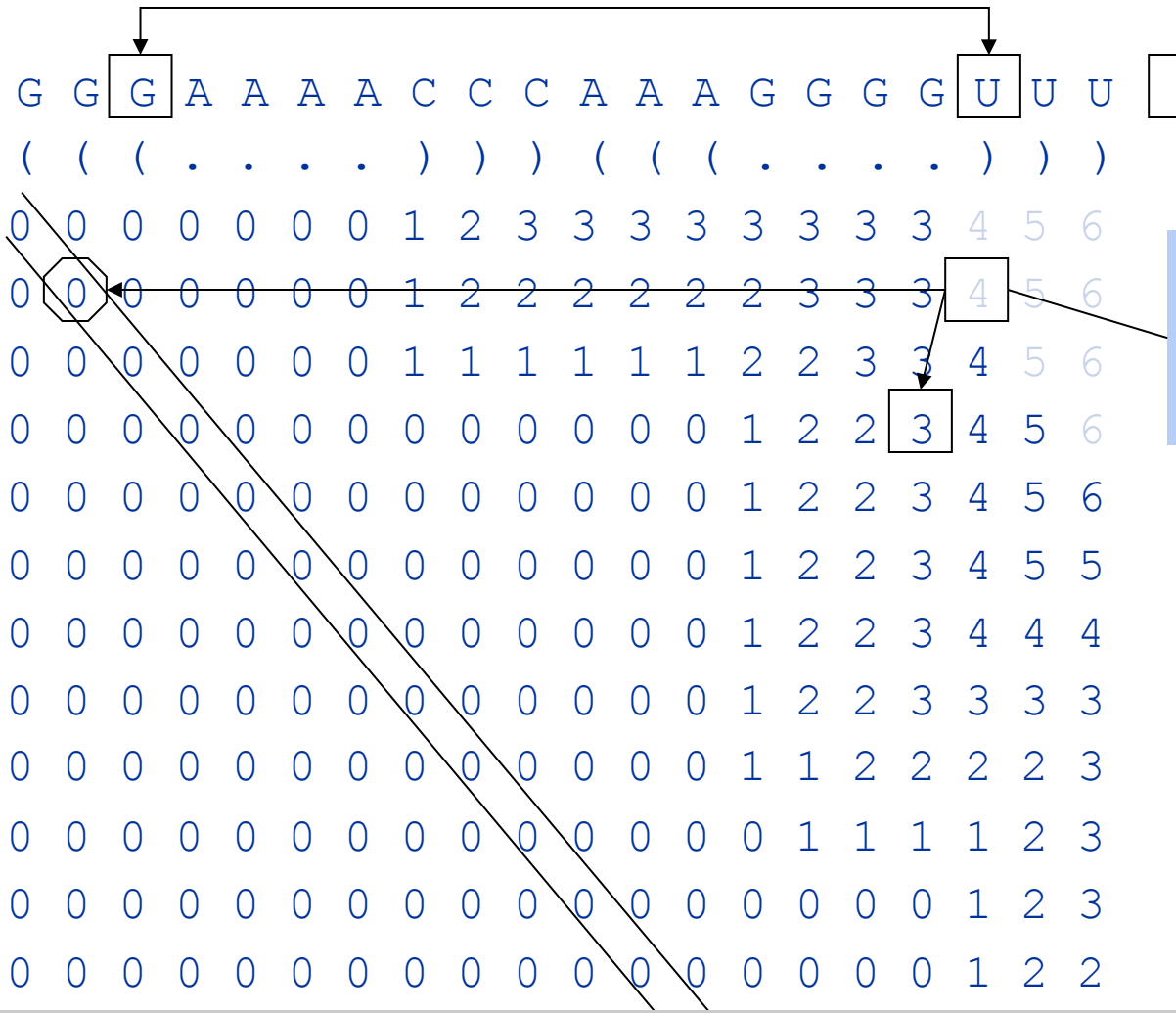
# Computing one cell: OPT[2,18] = ?




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

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

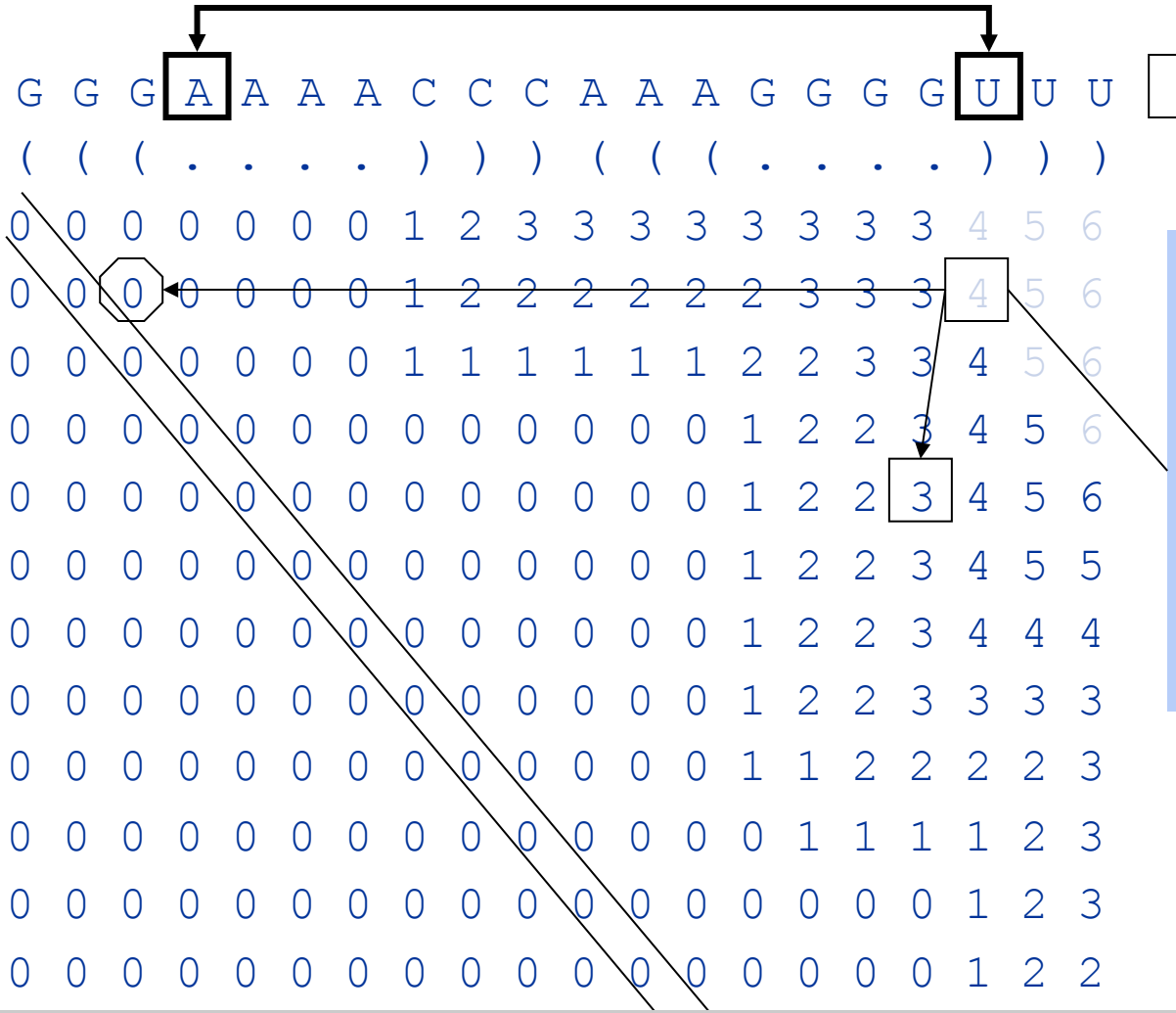
# Computing one cell: OPT[2,18] = ?




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] = ?

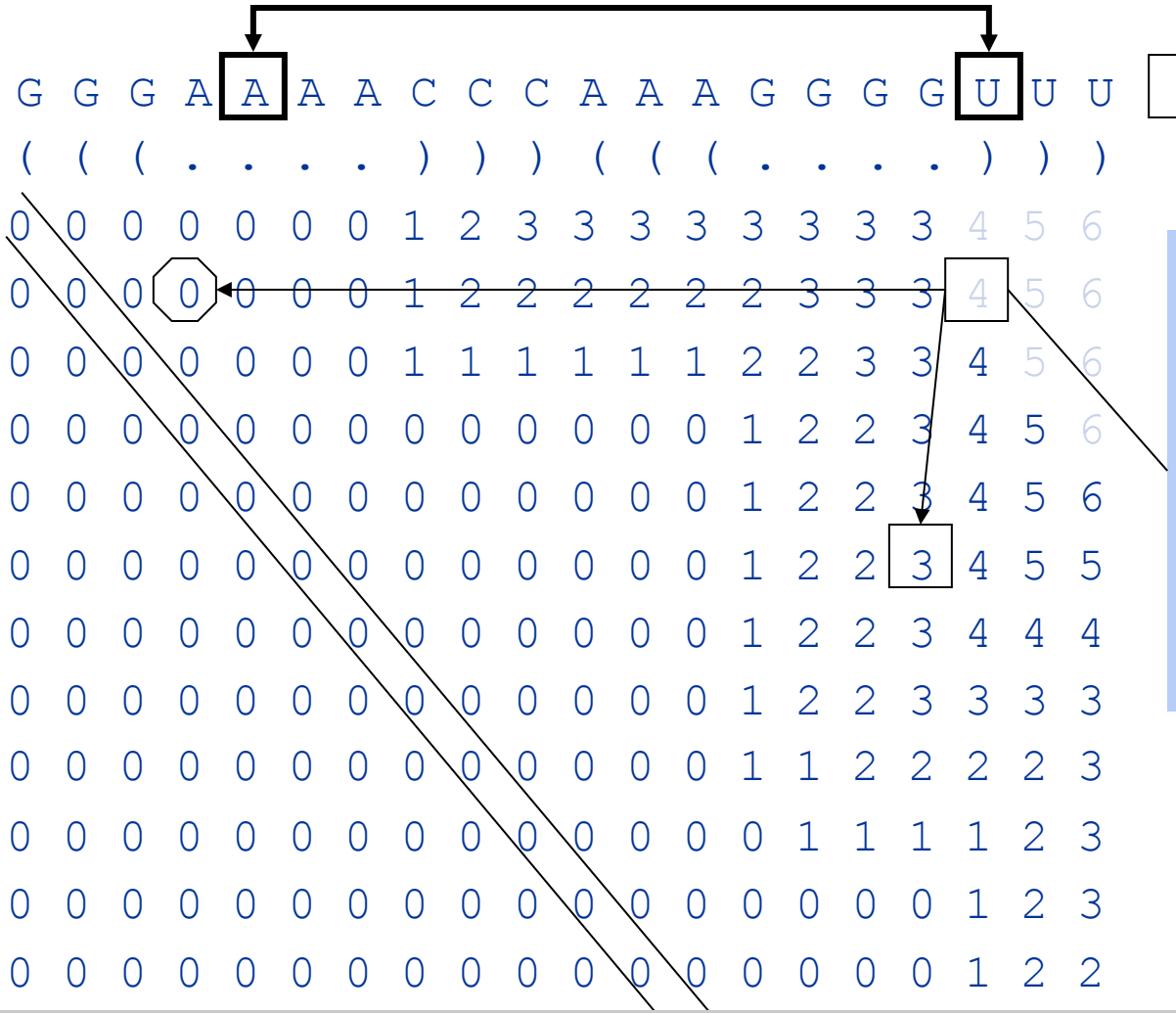



**Case 3, 2 ≤ t <18-4:**
**t = 4: yes pair**
**OPT[2,18]≥1+0+3**

GG**A**AAACCCAAAGGGG**U**
..**(**...**(**((....))))

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

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

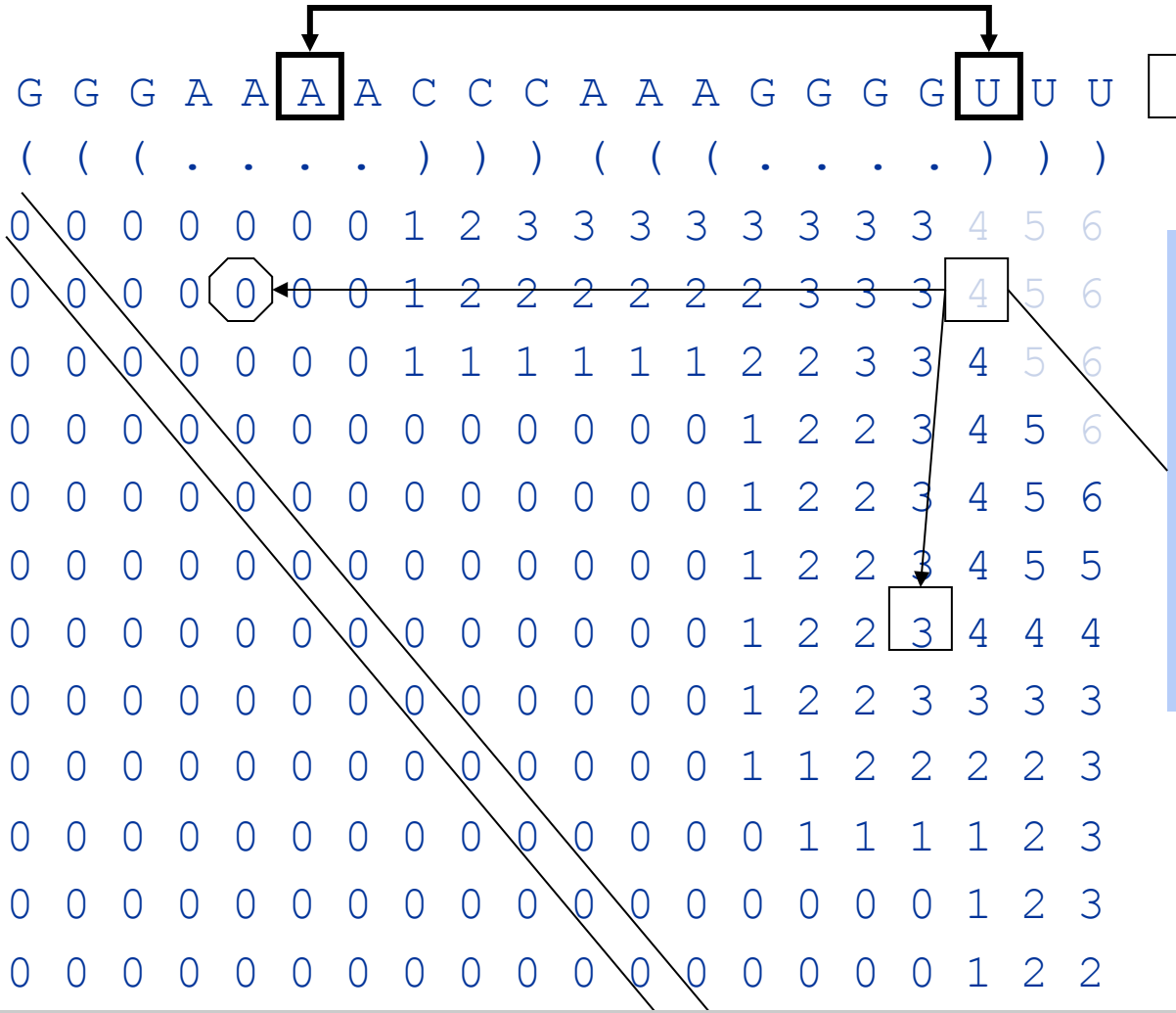



**Case 3, $2 \le t < 18-4$:**
**$t = 5$: yes pair**
**$\mathrm{OPT}[2,18] \ge 1+0+3$**

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] = ?

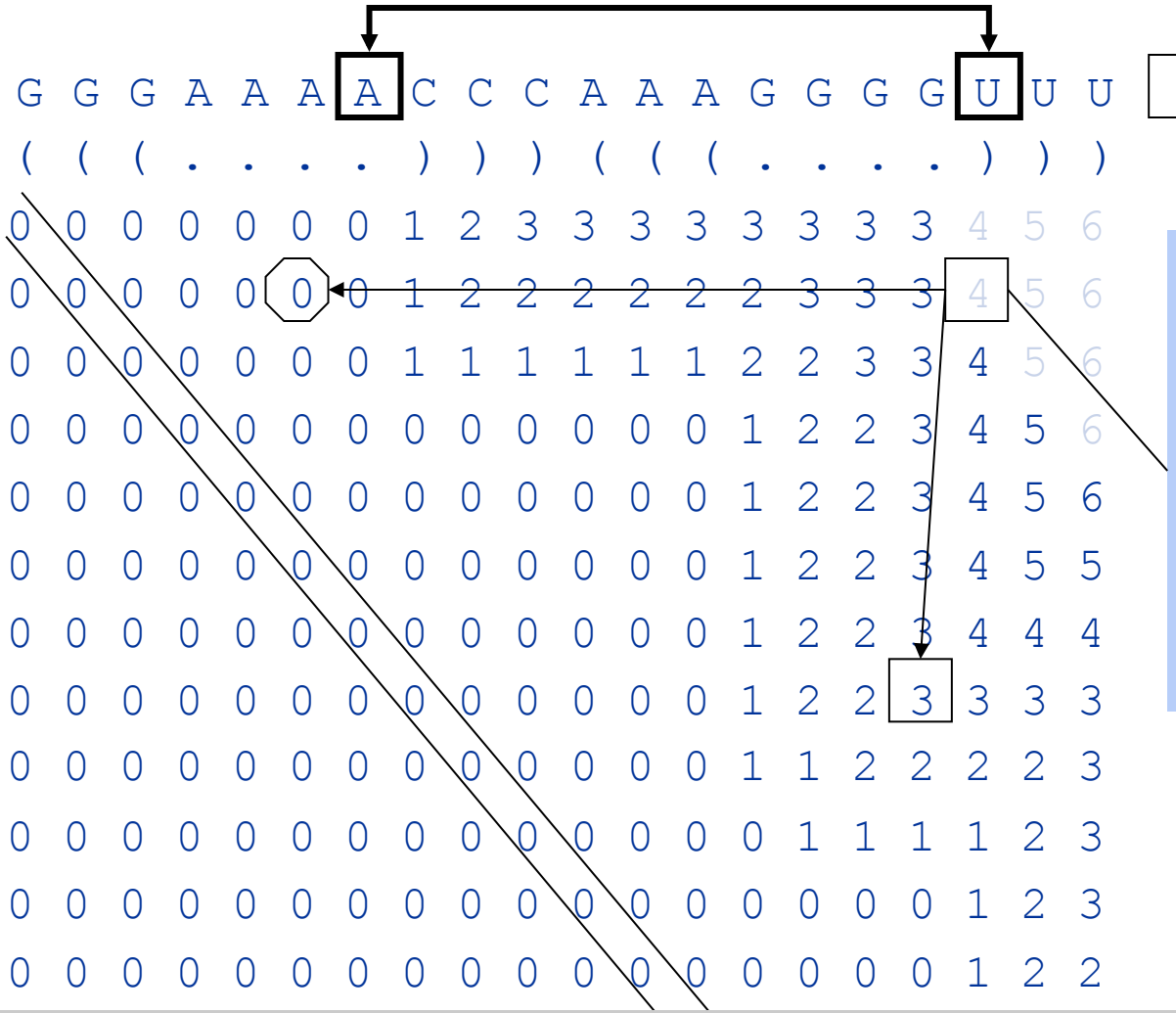



**Case 3, 2 ≤ t <18-4:**
**t = 6: yes pair**
**OPT[2,18]≥1+0+3**

GGAA**A**ACCCAAAGGGG**U**
....**(**.(((....)))**)**

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

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




**Case 3, 2 ≤ t <18-4:**
**t = 7: yes pair**
**OPT[2,18]≥1+0+3**

GGAAA**A**CCCAAAGGGG**U**
.....**(**(((....)))**)**

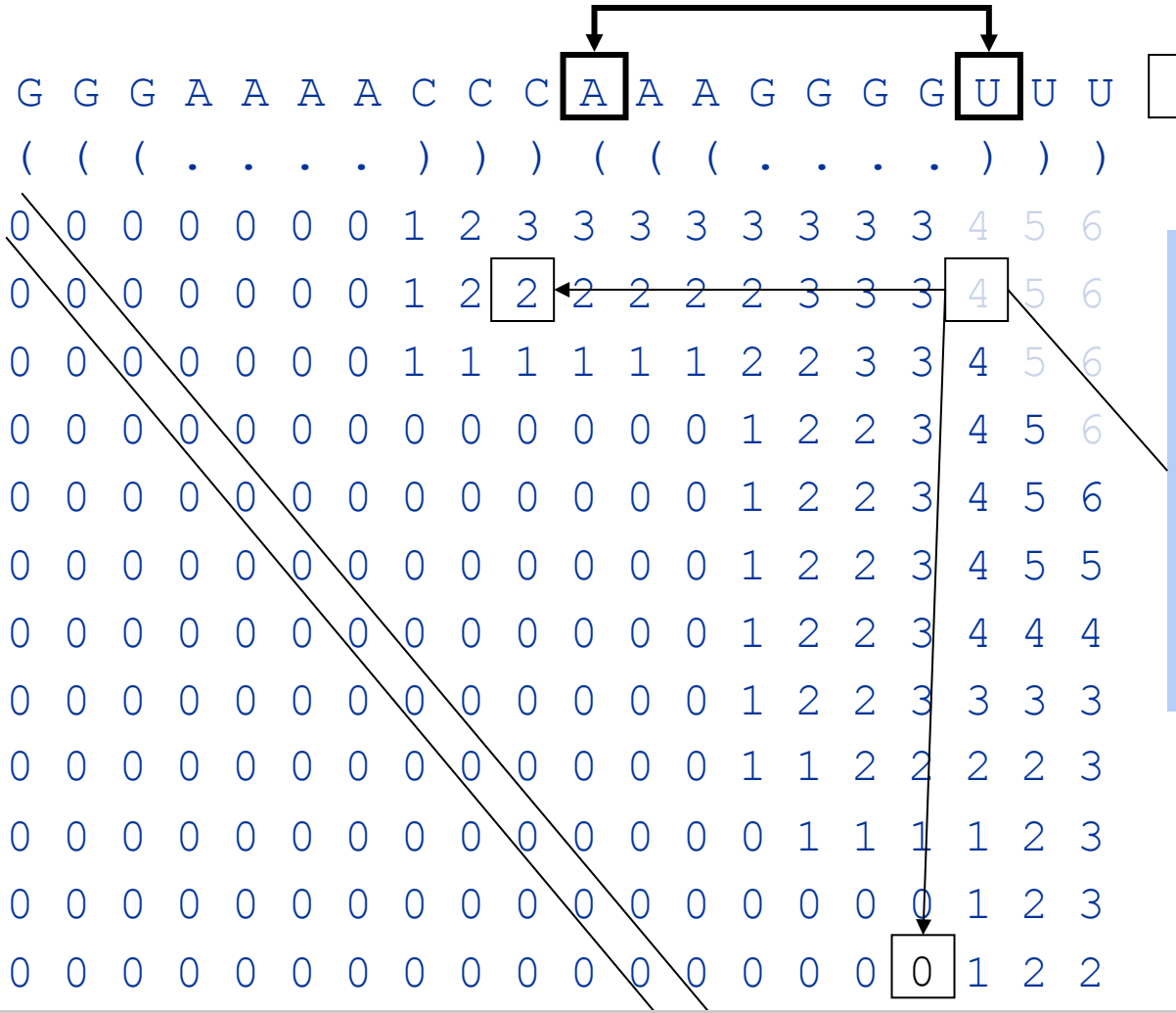
$$\mathrm{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\left\{ \begin{array}{c} \mathrm{OPT}[i,j-1] \\ 1+\max_t(\mathrm{OPT}[i,t-1]+\mathrm{OPT}[t+1,j-1]) \end{array} \right\} & \text{otherwise} \end{cases}$$

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


Case 3, $2 \le t < 18-4$:
$t = 8$: 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] = ?

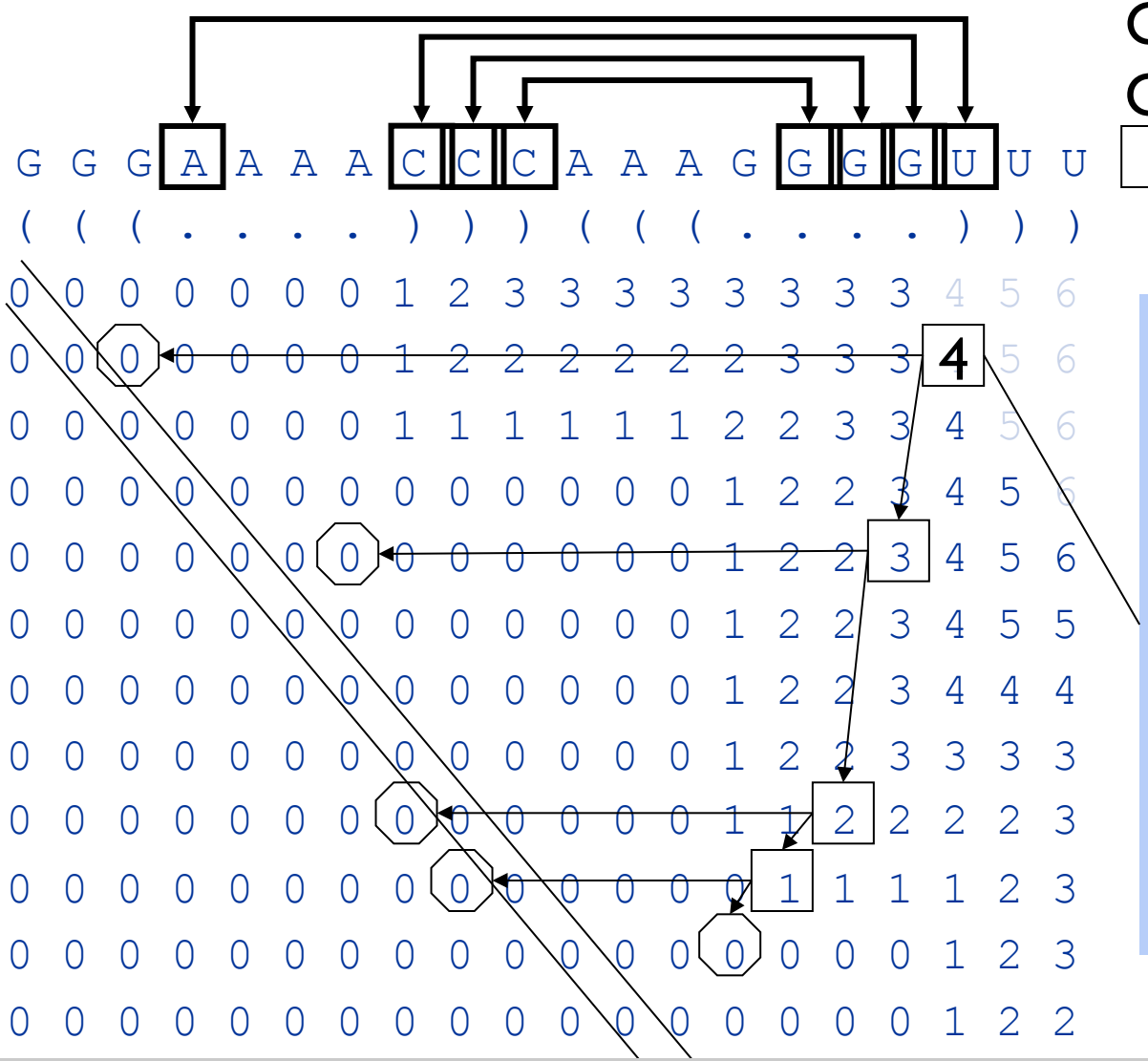



**Case 3, 2 ≤ t <18-4:**
**t = 11: yes pair**
**OPT[2,18]≥1+2+0**

GGAAAACCC**A**AAGGGG**U**
((.....))**(.....)**

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

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




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

n= 20

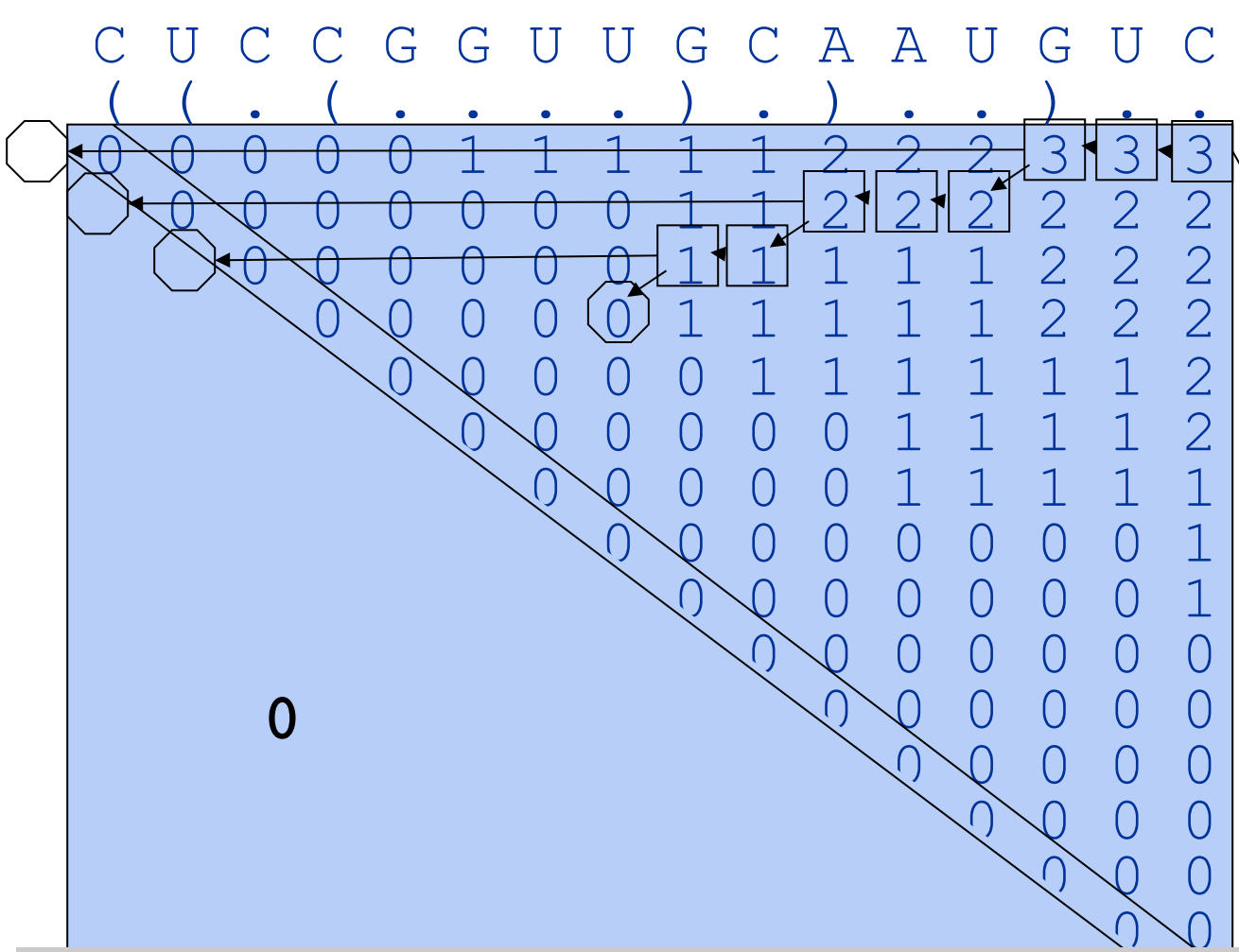
**Overall, Max = 4**
**several ways, e.g.:**

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

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

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

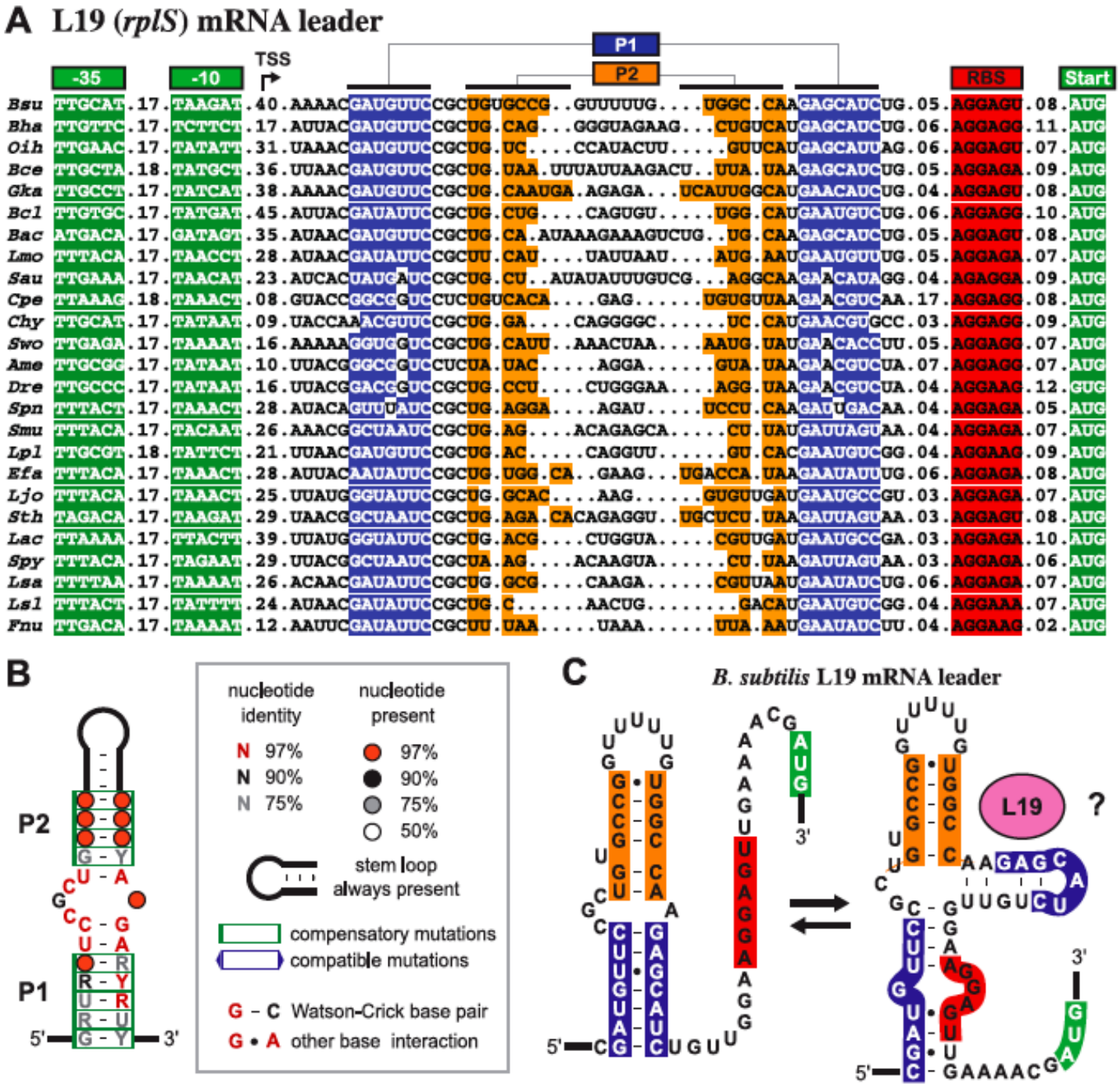



**Figure 3.** Putative Autoregulatory Structure in L19 mRNA Leaders

# Summary

- RNA has important roles
  - Beyond mRNA; many unexpected recent discoveries
- Structure is critical to function
  - True of other molecules, too
- RNA secondary structure prediction is a key tool
- Dynamic programming–useful accuracy, $O(n^3)$ time:
  - Binary choice again: last base is paired or not
  - Optimal substructure again: given last pair, optimally fold inside & outside separately
  - Tabulate again: best folding of all substrings.