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
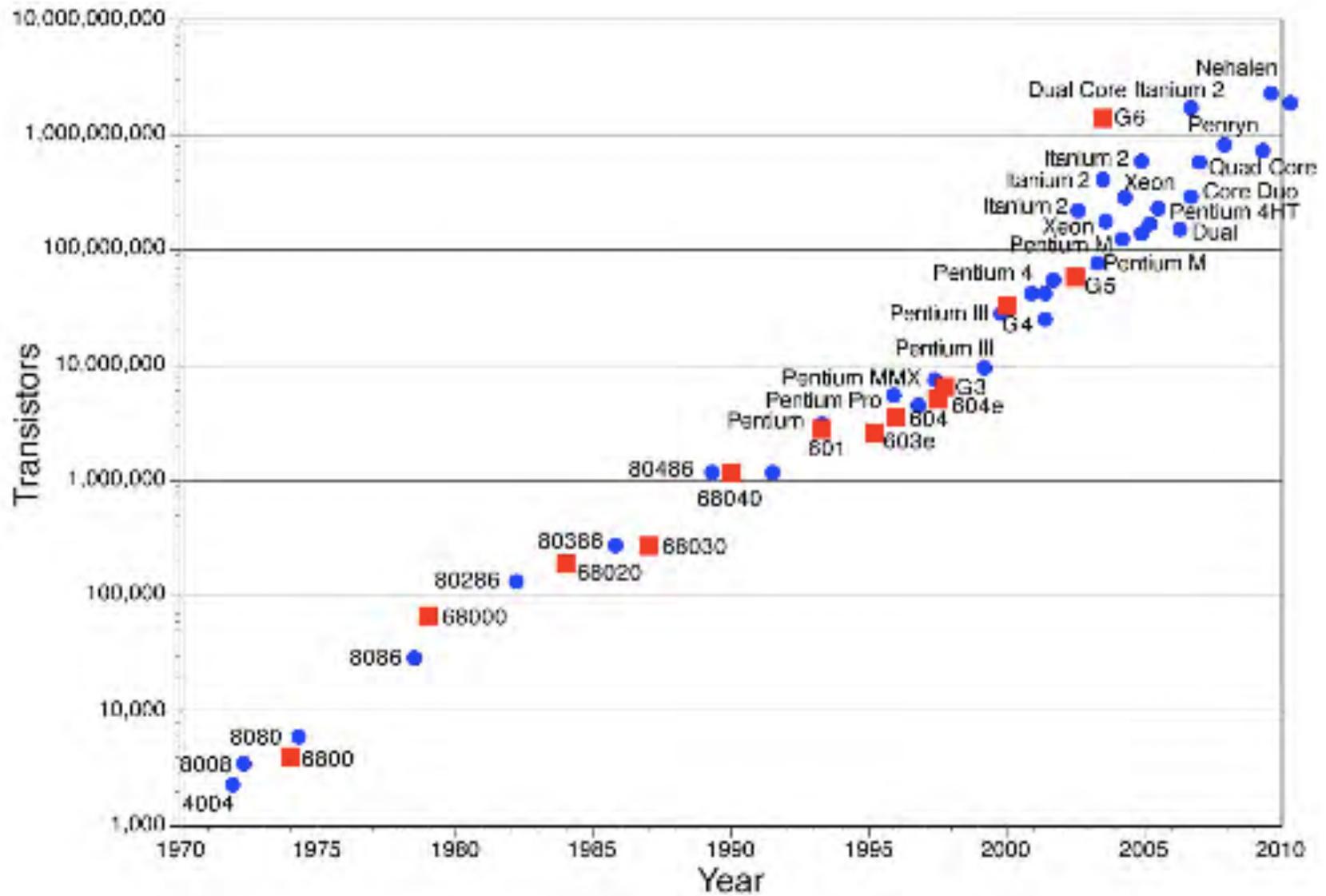
Growth of GenBank (Base Pairs)

Excludes "short-read archive," > 7 terabases by mid-2009

Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html   5

SRA database growth

2,492,489,806,780,758 total bases
1,174,010,897,059,076 open access bases

Size, terabases

1000

100

10

0

2009   2010   2011   2012   2013   2014
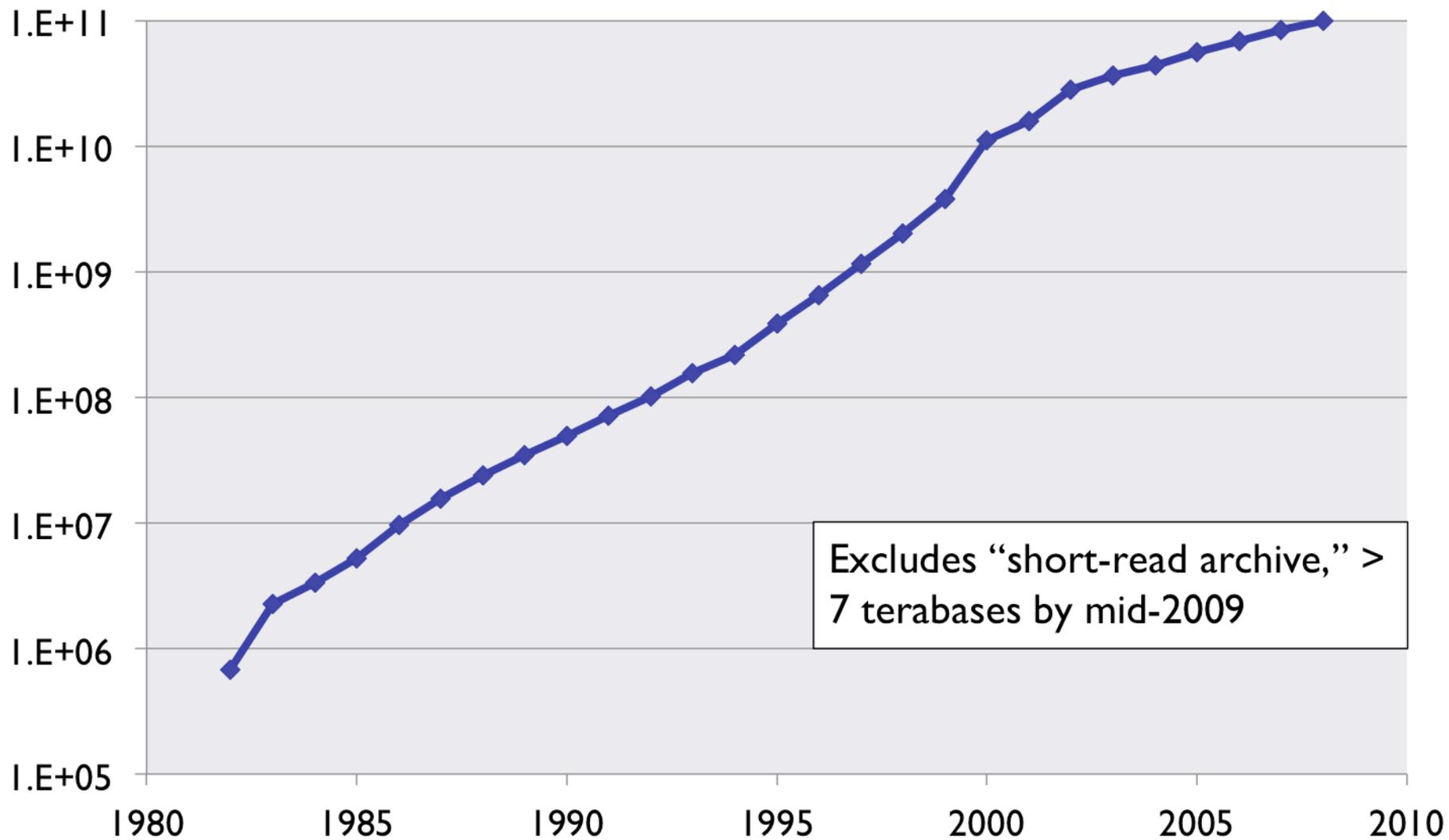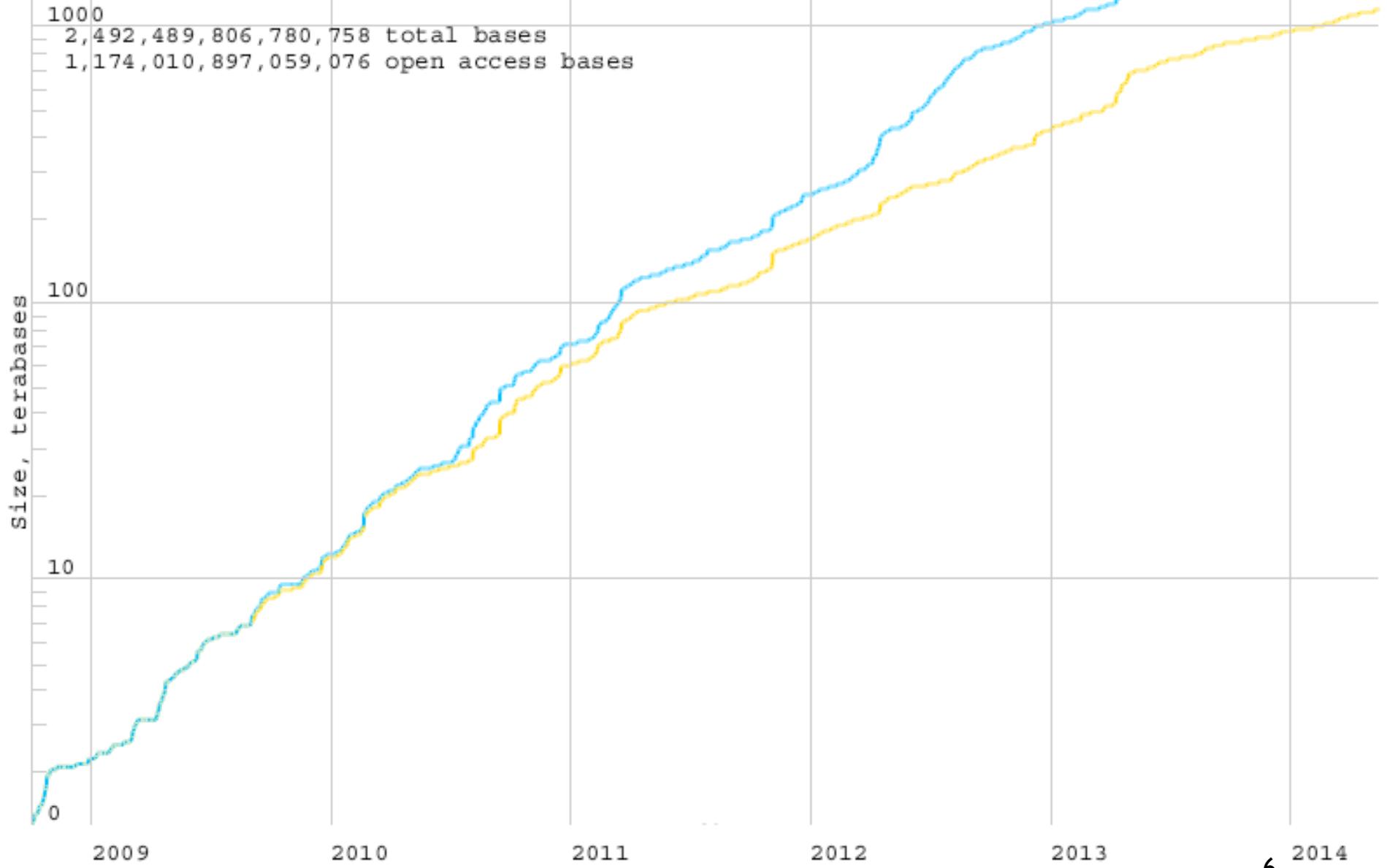
Total bases
Open access bases

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

05/14/2014 12:59pm

Cost per Genome
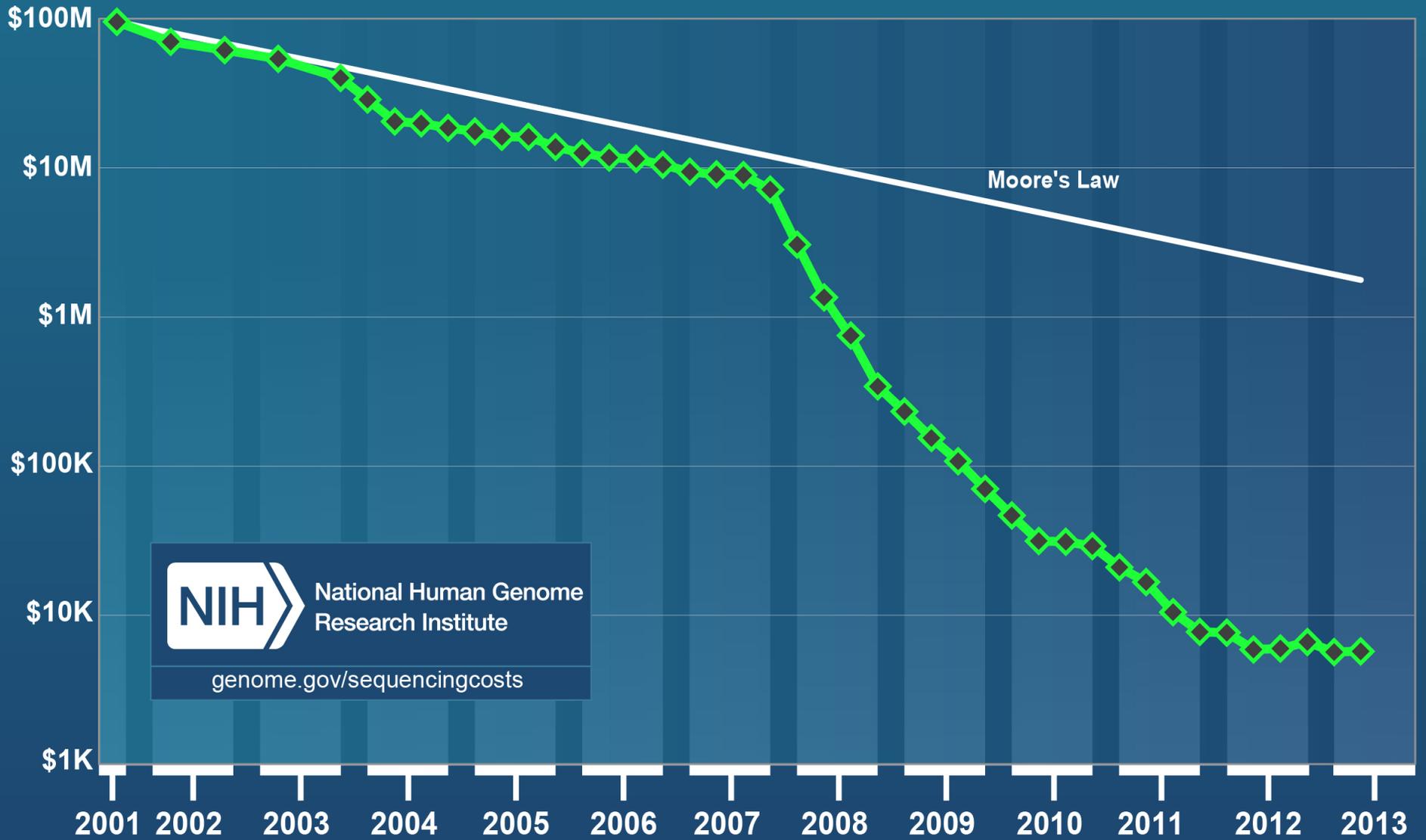
Sequencing Costs Outpace Moore's Law

Chart showing Cost per Genome dropping from $100,000,000 (Sep 01) to about $1,000 (Mar 14), compared against Moore's Law line.

Y-axis: $100,000,000 / $10,000,000 / $1,000,000 / $100,000 / $10,000 / $1,000 / $100

X-axis: Sep 01, Jul 02, May 03, Mar 04, Jan 05, Nov 05, Sep 06, Jul 07, May 08, Mar 09, Jan 10, Nov 10, Sep 11, Jul 12, May 13, Mar 14

Legend:
— Cost per Genome
— Moore's Law
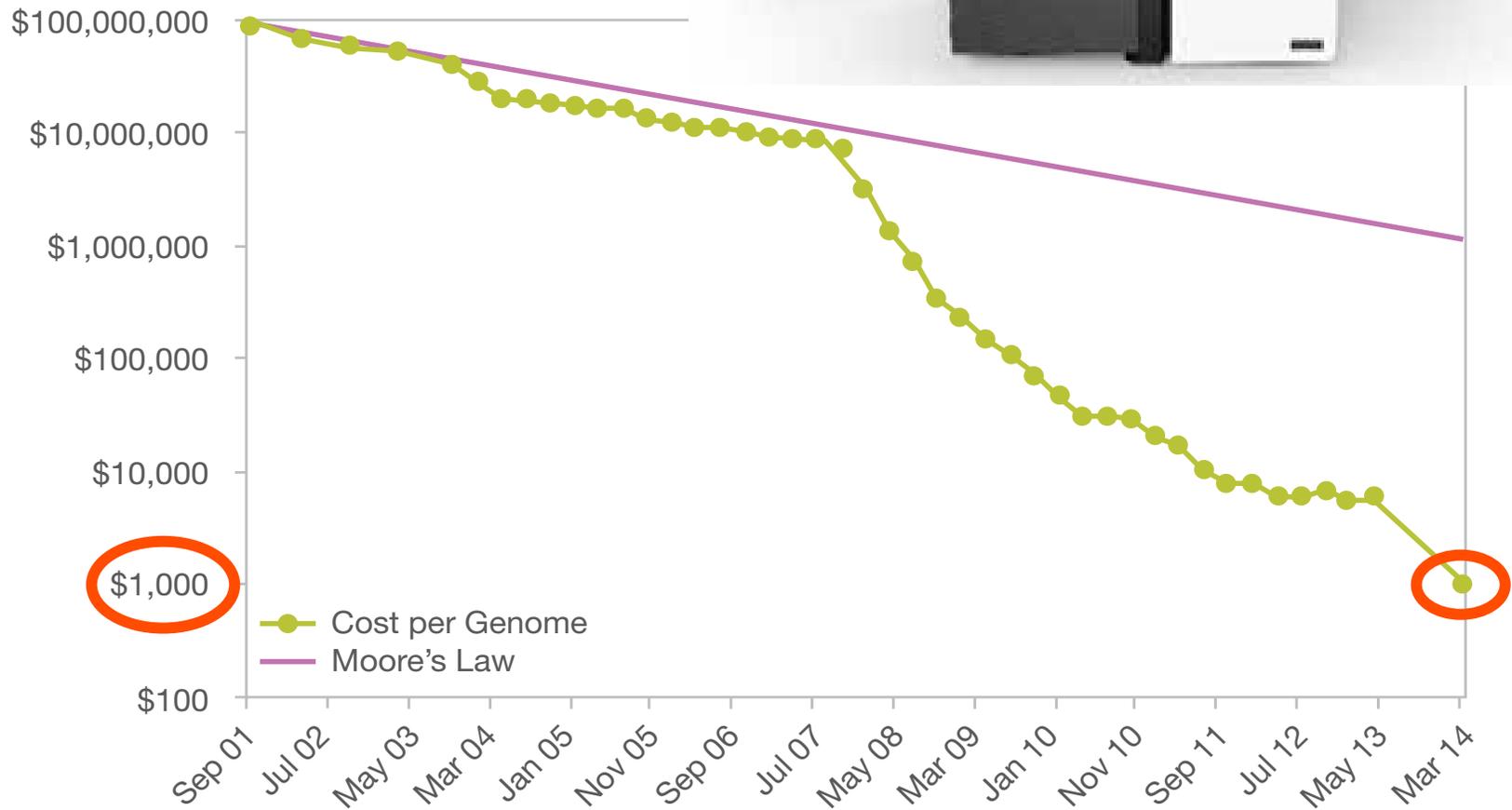
A Database Search

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		frog	MyoD protein (Xenopus tropicalis)

10

…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
```

11

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	4.0×10^{-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	5.0×10^{-30}	GSTENG00034775001
	F1NFS6	F1NFS6_CHICK	☆	Uncharacterized protein	Gallus gallus (Chicken)	237	46.0%	343	5.0×10^{-30}	Gga.378
	Q91151	Q91151_NOTVI	☆	Myogenic regulatory factor; transcription fac…	Notophthalmus viridescens (Eastern newt) (Triturus viridescens)	219	44.0%	342	6.0×10^{-30}	MRF-4
	Q29BN7	Q29BN7_DROPS	☆	GA10192	Drosophila pseudoobscura pseudoobscura (Fruit fly)	330	42.0%	342	6.0×10^{-30}	GA10192 Dpse\GA10192 Dpse_GA10192
	B4GP81	B4GP81_DROPE	☆	GL13832	Drosophila persimilis (Fruit fly)	330	42.0%	342	6.0×10^{-30}	GL13832 Dper\GL13832 Dper_GL13832
	Q92020	MYF6_XENLA	★	Myogenic factor 6	Xenopus laevis (African clawed frog)	240	44.0%	340	1.0×10^{-29}	myf6 mrf4
	B7ZQB0	B7ZQB0_XENLA	☆	MRF4a	Xenopus laevis (African clawed frog)	240	44.0%	340	1.0×10^{-29}	MRF4
	A7UCI1	A7UCI1_XENLA	☆	MRF4a	Xenopus laevis (African clawed frog)	240	44.0%	340	1.0×10^{-29}	MRF4A MRF4
	F7FIX8	F7FIX8_MONDO	☆	Uncharacterized protein	Monodelphis domestica (Gray short-tailed opossum)	243	47.0%	339	1.0×10^{-29}	MYF6
	D9IV56	D9IV56_EPICO	☆	Myogenin	Epinephelus coioides (Orange-spotted grouper) (Epinephelus nebulosus)	250	47.0%	338	2.0×10^{-29}	
	Q6SYV5	MYF6_TAKRU	★	Myogenic factor 6	Takifugu rubripes (Japanese pufferfish) (Fugu rubripes)	239	46.0%	337	2.0×10^{-29}	myf6 mrf4
	G3WJP0	G3WJP0_SARHA	☆	Uncharacterized protein	Sarcophilus harrisii (Tasmanian devil) (Sarcophilus laniarius)	243	47.0%	337	2.0×10^{-29}	MYF6
	G7Y452	G7Y452_CLOSI	☆	Transcription factor SUM-1	Clonorchis sinensis (Chinese liver fluke)	946	64.0%	337	2.0×10^{-29}	CLF_100742
	G4LXJ1	G4LXJ1_SCHMA	☆	Myogenic factor, putative	Schistosoma mansoni (Blood fluke)	864	68.0%	337	2.0×10^{-29}	Smp_167400
	G1EN33	G1EN33_DUGJA	☆	Myogenic determinant factor	Dugesia japonica (Planarian)	498	42.0%	336	3.0×10^{-29}	MDF
	Q8MWP6	Q8MWP6_SCHMD	☆	MyoD-like protein	Schmidtea mediterranea (Freshwater planarian flatworm)	466	69.0%	335	4.0×10^{-29}	
	Q6PUV5	MYF6_TETNG	★	Myogenic factor 6	Tetraodon nigroviridis (Spotted green pufferfish) (Chelonodon nigroviridis)	239	45.0%	333	7.0×10^{-29}	myf6 mrf4 GSTENG00021536001
	Q2PZ12	Q2PZ12_SALSA	☆	Myogenin	Salmo salar (Atlantic salmon)	254	55.0%	333	7.0×10^{-29}	
	Q91207	Q91207_ONCMY	☆	TMyogenin protein	Oncorhynchus mykiss (Rainbow trout) (Salmo gairdneri)	254	45.0%	332	9.0×10^{-29}	TMyogenin

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

12

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

(a) Computer-generated Image of DNA (by Mel Prueitt)

(b) Uncoiled DNA Fragment

Deoxyribose residue

Phosphate group

to 3' carbon of sugar residue

$$O^- - P = O$$

to 3' carbon of sugar residue

Base

Nucleotide

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 strα a polymer made up of four re called deoxyribonucleotides

Los Alamos Science

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

15

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.

16
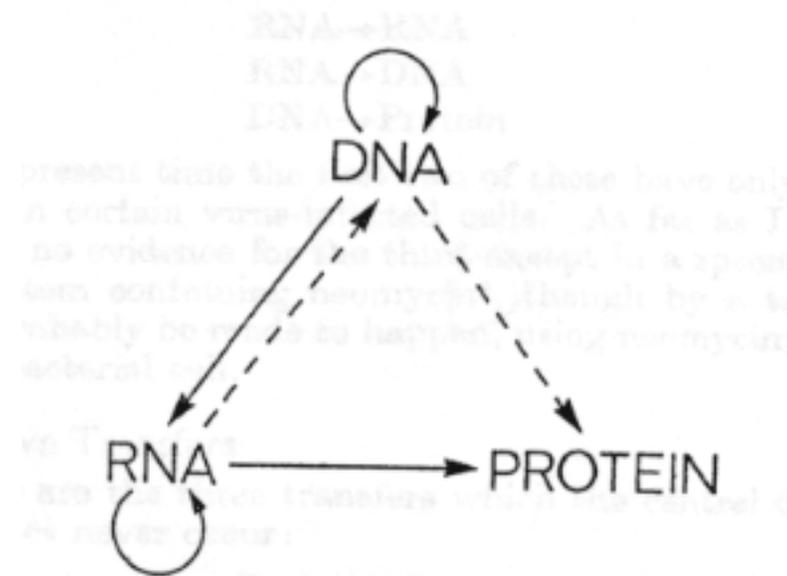
Non-coding RNA

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'
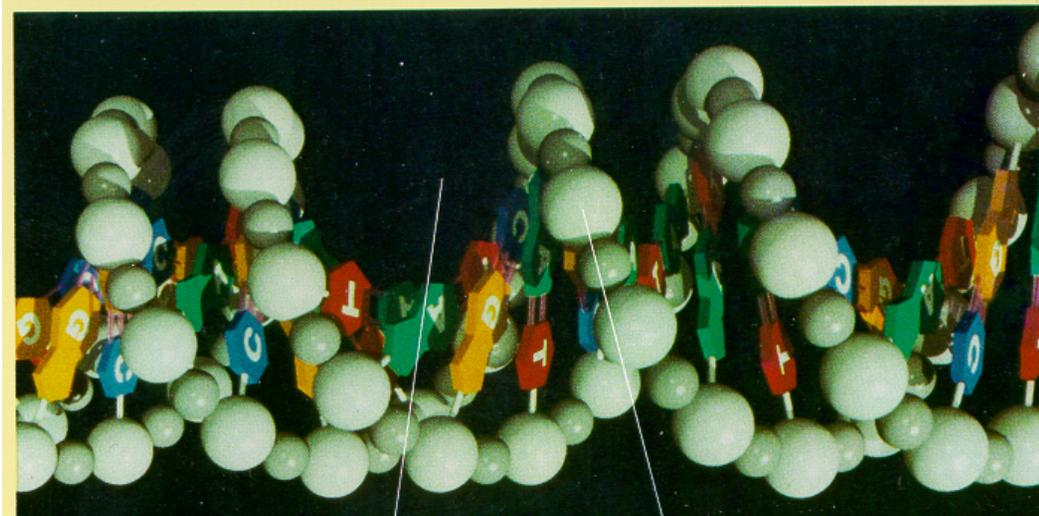
RNA Secondary Structure:
RNA makes helices too

Base pairs

A–U
C–G

U C
A   A
G–C
C–G
A
G–C
U–A
C–G
A–U
G–C
5´   C A A   A A U   3´
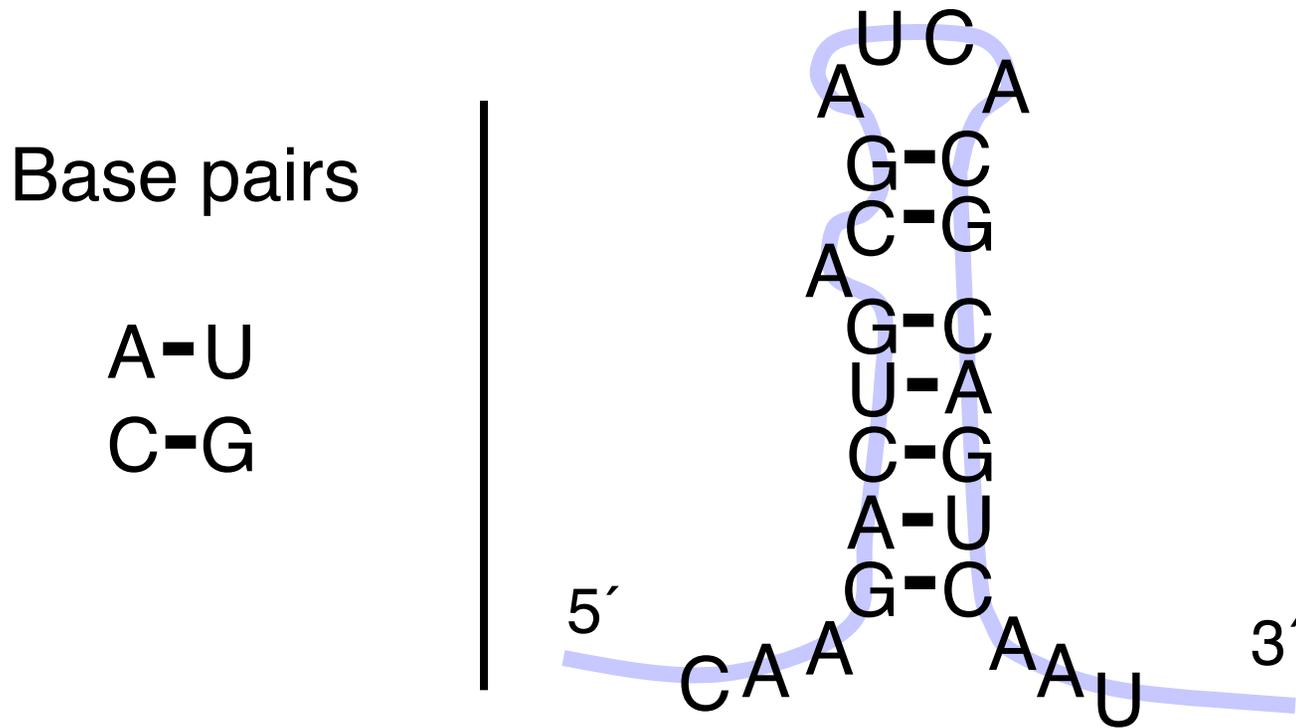
Usually *single* stranded

RNA Secondary Structure:

Not everything, but important, easier than 3d
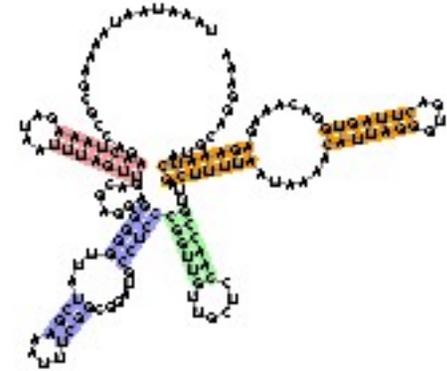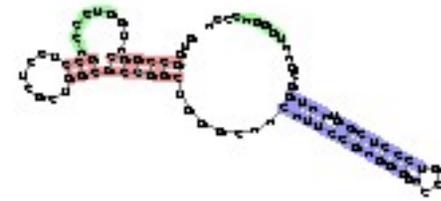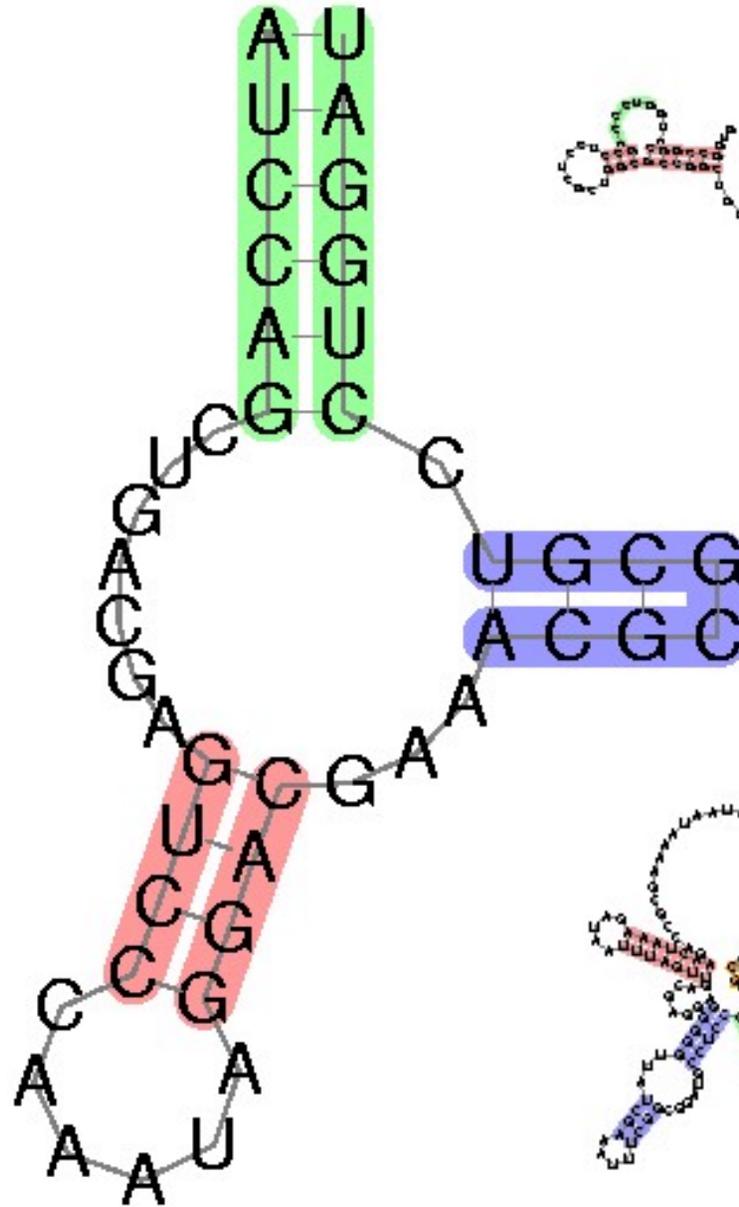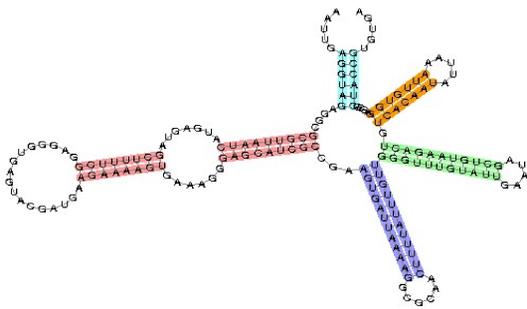
Why is structure important?

- 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 non-coding RNA, many *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
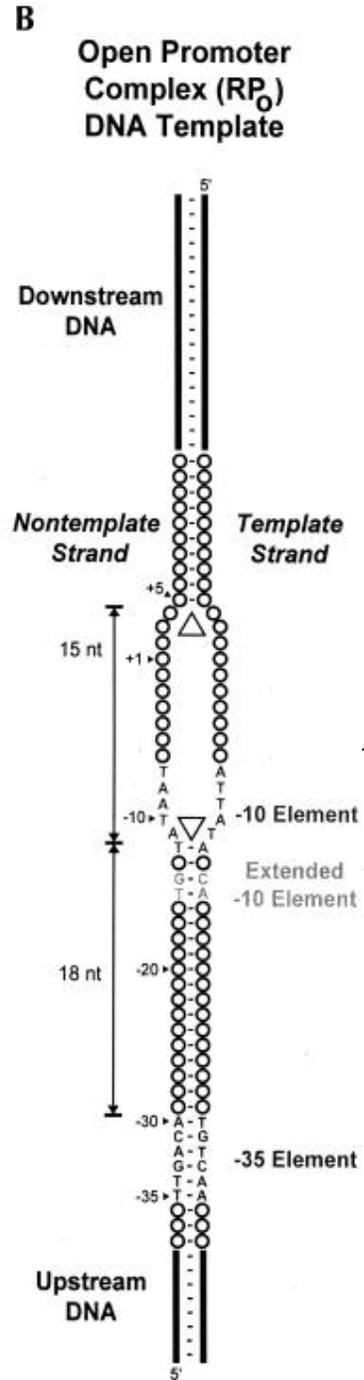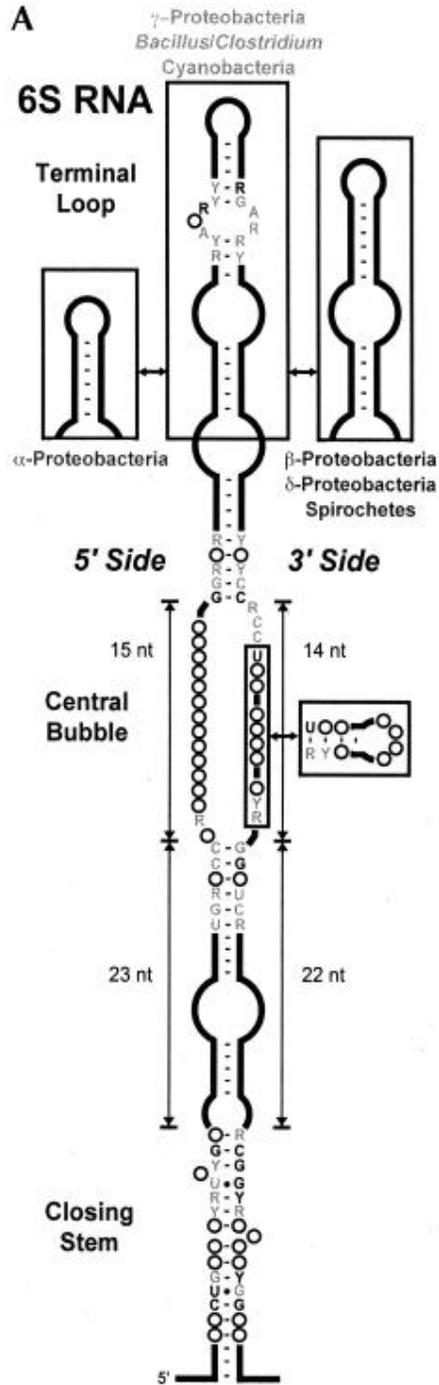
6S mimics an open promoter
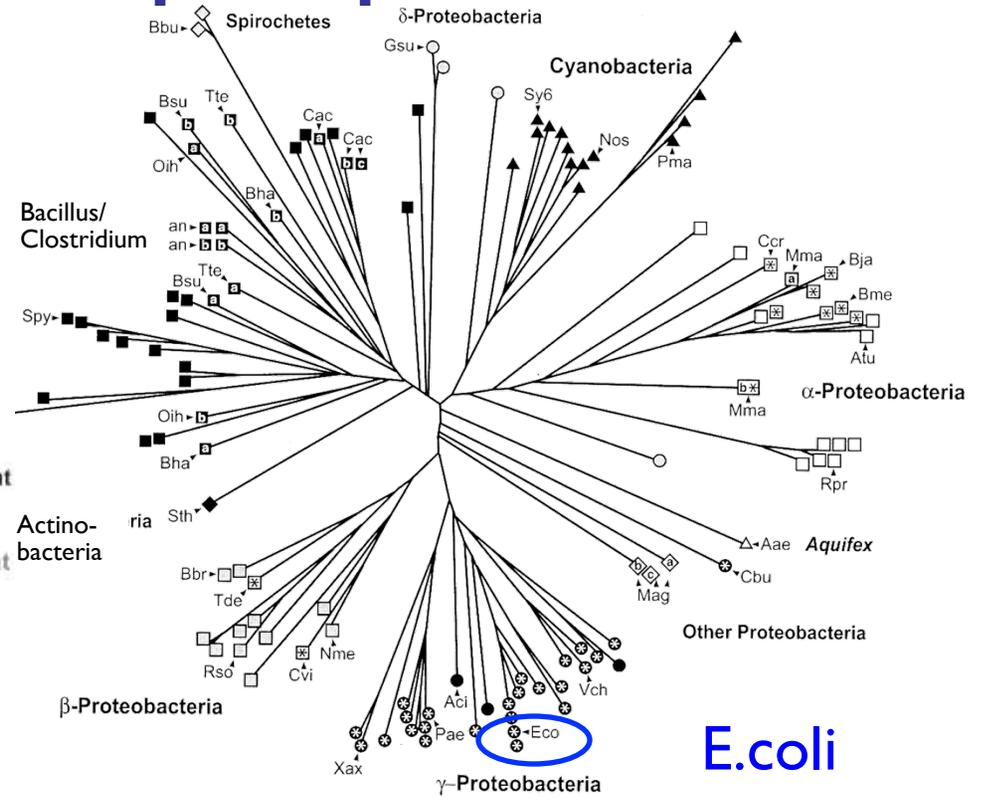
E.coli

Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005 [22]

Chloroflexi

δ -Proteobacteria

Chloroflexus aurantiacus

Geobacter metallireducens
Geobacter sulphurreducens

A — *Salmonella enterica*
Salmonella typhimurium
Escherichia coli ← (E. coli)
Yersinia pestis
Haemophilus influenzae
Pasteurella multocida
Vibrio cholerae
Buchnera aphidicola
Pseudomonas aeruginosa
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis

γ-Proteobacteria

B
Neisseria meningitidis
Ralstonia solanacearum

β-Proteobacteria

C
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

I
Mycobacterium leprae
Mycobacterium tuberculosis
Corynebacterium glutamicum
Streptomyces coelicolor

Actinobacteria
(high GC)

F
Deinococcus radiodurans

E
D
Tsc. elongatus
Nostoc sp.
Synechocystis sp.

Cyanobacteria

H
Fusobacterium nucleatum
Clostridium acetobutylicum
Clostridium perfringens
Tab. tengcongensis
G
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)

J
K
Thermotoga maritima
Aquifex aeolicus

4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0

Billion years ago

23

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

24
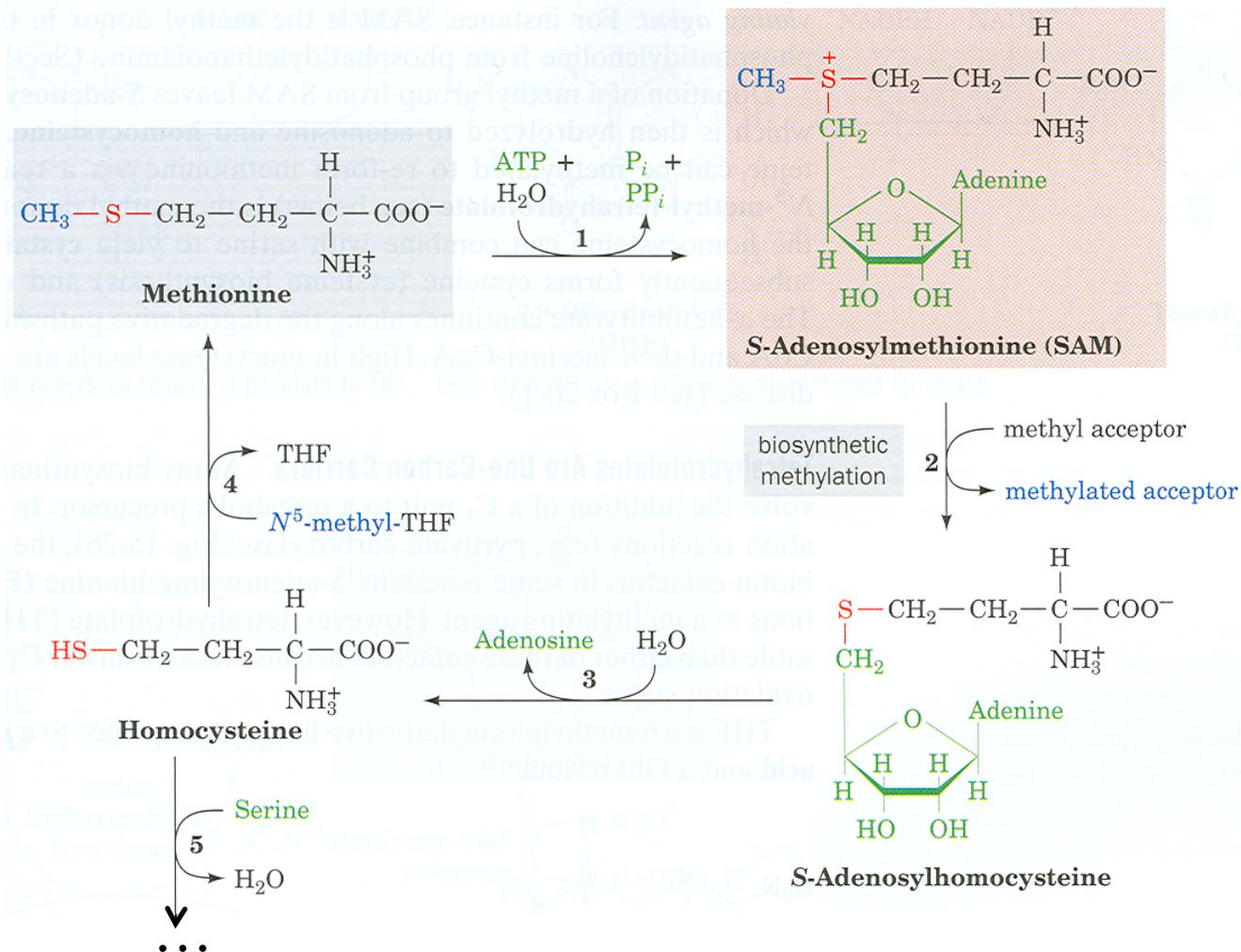
Gene Regulation: The MET Repressor

SAM

(A)

COOH

HOOC

(B)
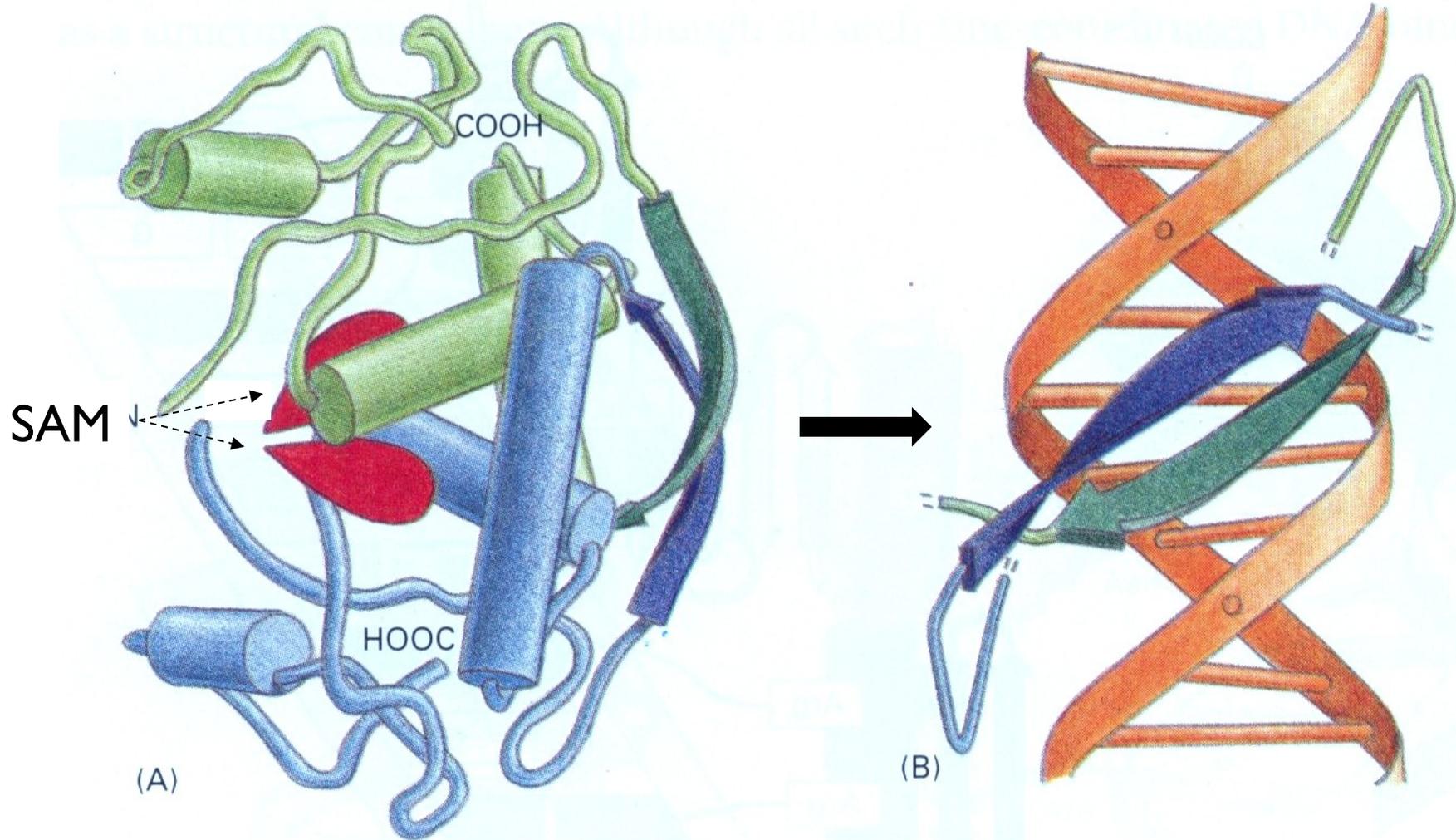
Protein

Alberts, et al, 3e.

DNA

25

Not the only way!

Protein way

Riboswitch alternative

SAM
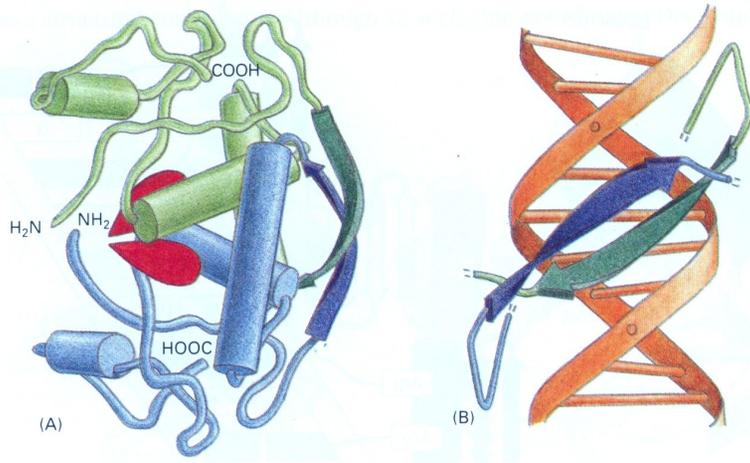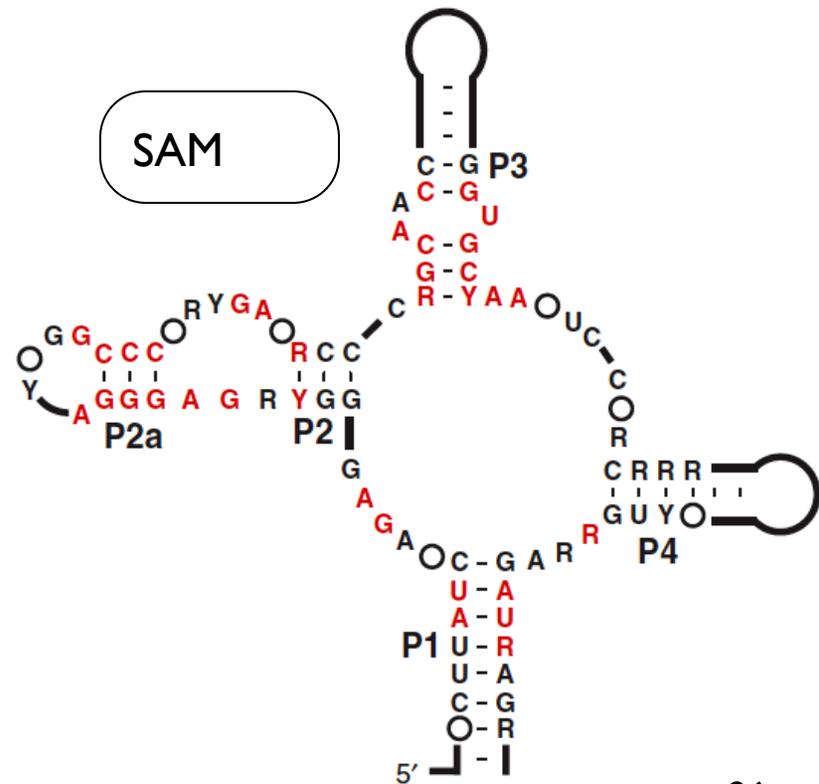
Grundy & Henkin, Mol. Microbiol 1998
Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003
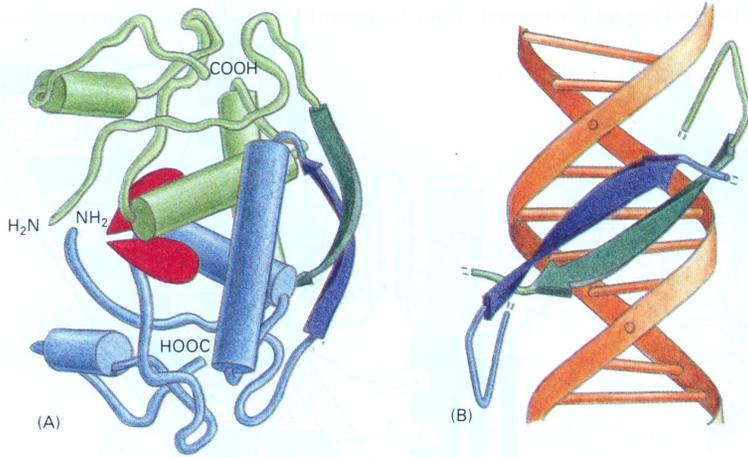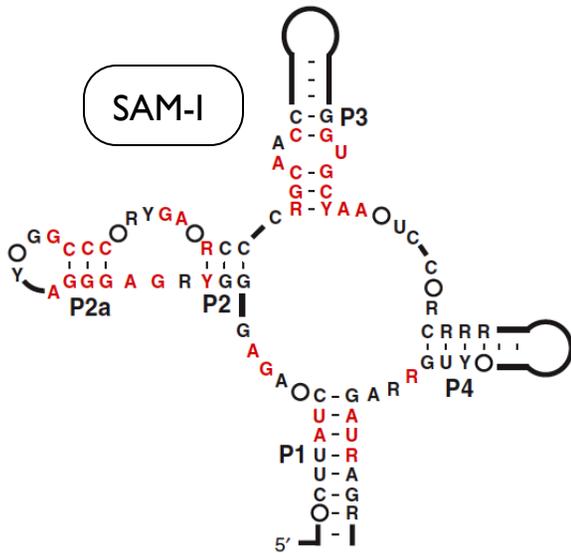
Not the only way!

Protein way

Riboswitch alternatives

SAM-II

P2

SAM-I

P3

P2a   P2

P4

P1

5′

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

Corbino et al.,
Genome Biol. 2005

P2

P1

5′
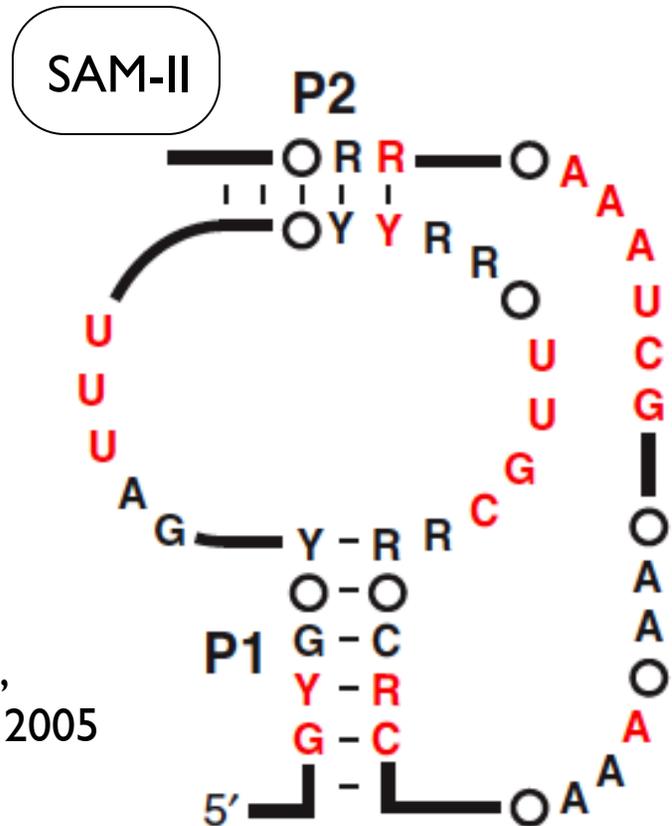
Not the only way!
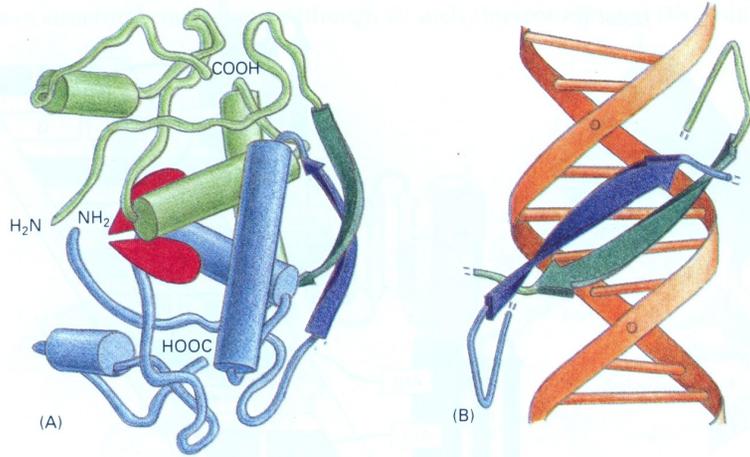
Protein way

Riboswitch alternatives

Alberts, et al, 3e.

SAM-III

SAM-I

SAM-II
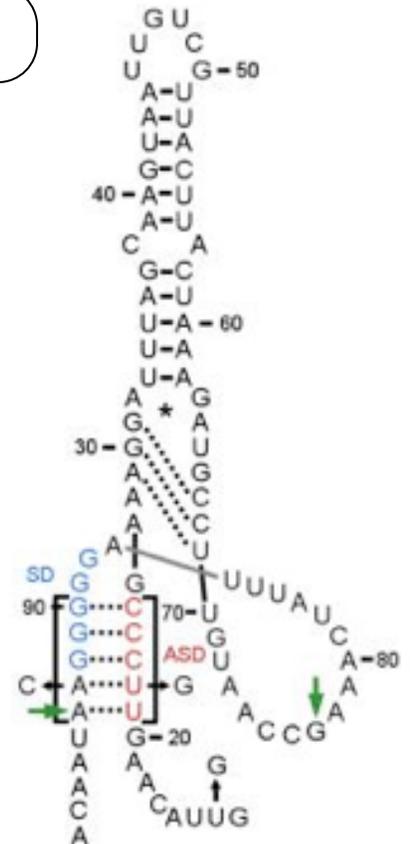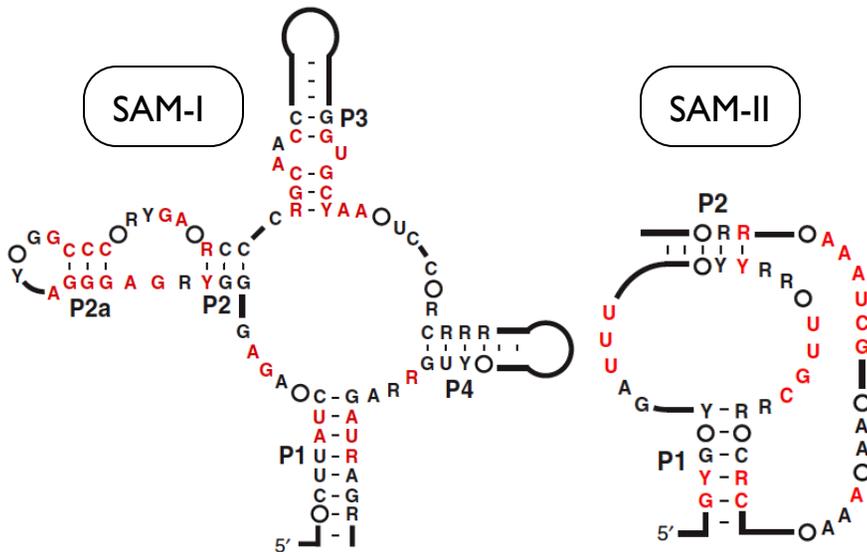
Fuchs et al.,
NSMB 2006

Grundy, Epshtein, Winkler
et al., 1998, 2003
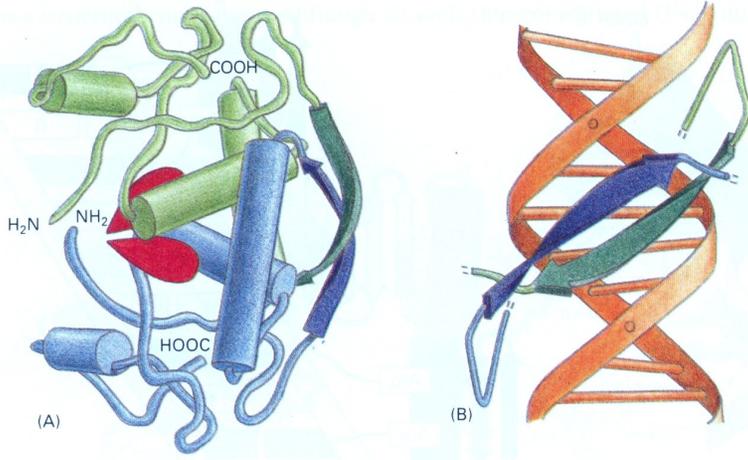
Corbino et al.,
Genome Biol. 2005

Not the only way!

Alberts, et al, 3e.

Protein way ← Riboswitch alternatives

SAM-I

SAM-II

SAM-III

SAM-IV
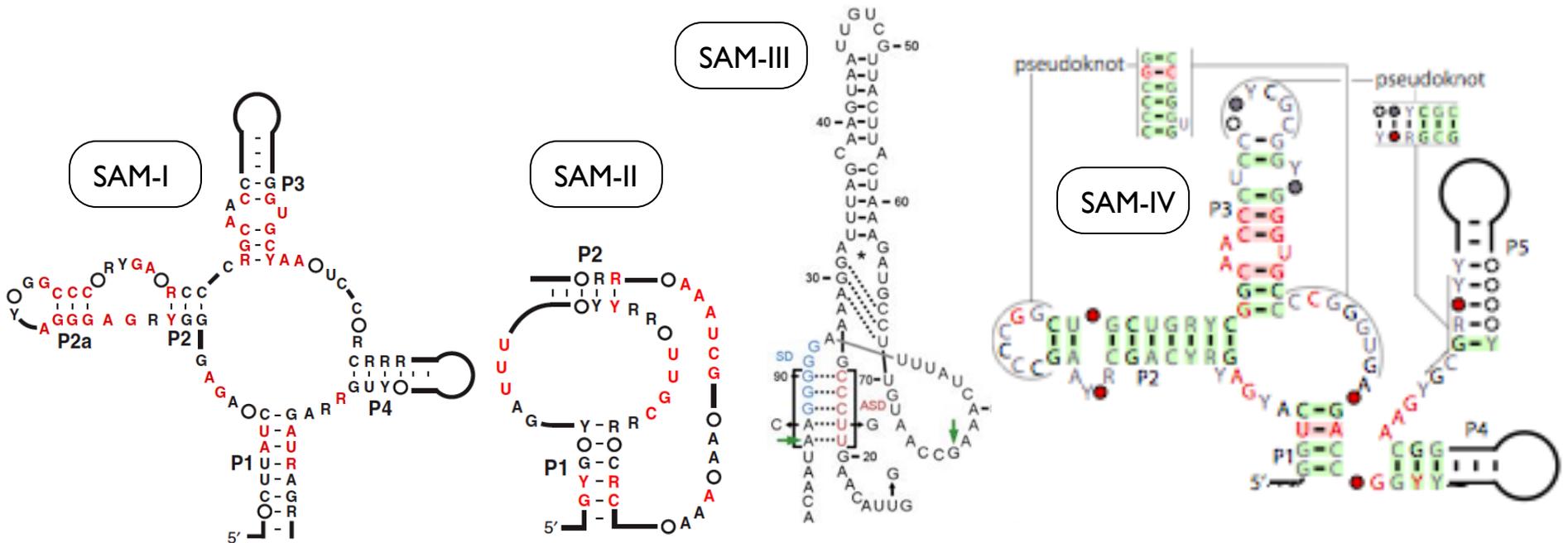
Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008

29

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

SAM-IV

Weinberg et al., RNA 2008

SAM-V

Meyer, etal., BMC Genomics 2009

30

And many other examples. Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.
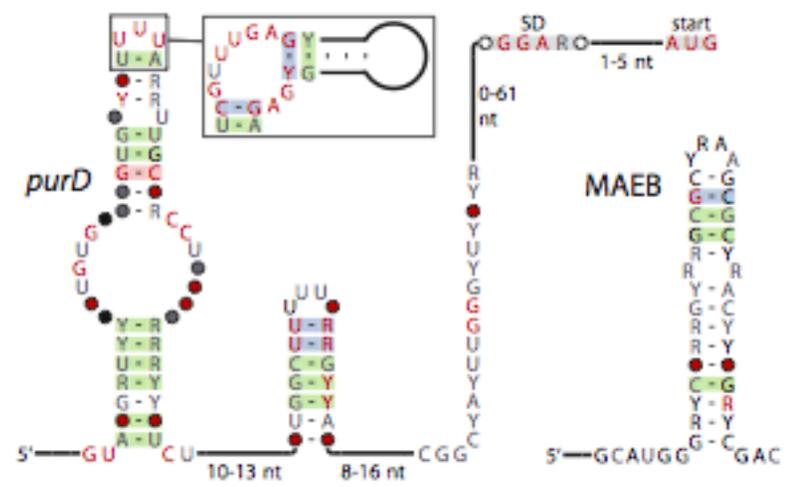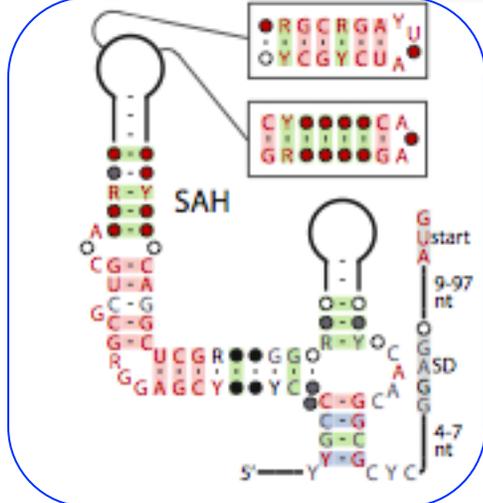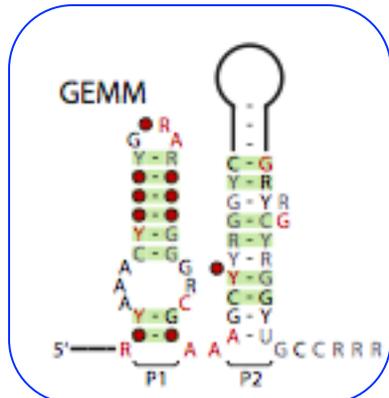
boxed = confirmed riboswitch (+2 more)

Weinberg, et al. Nucl. Acids Res., July 2007 35: 4809-4819.
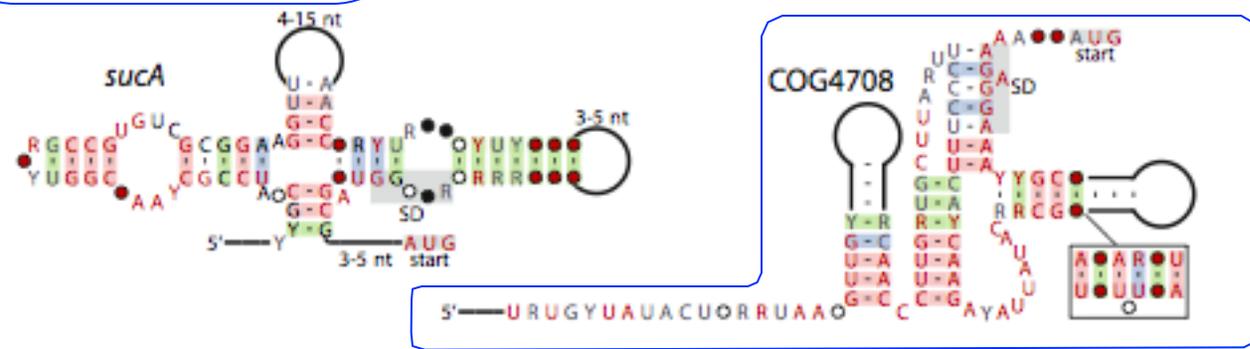
Why is RNA hard to deal with?
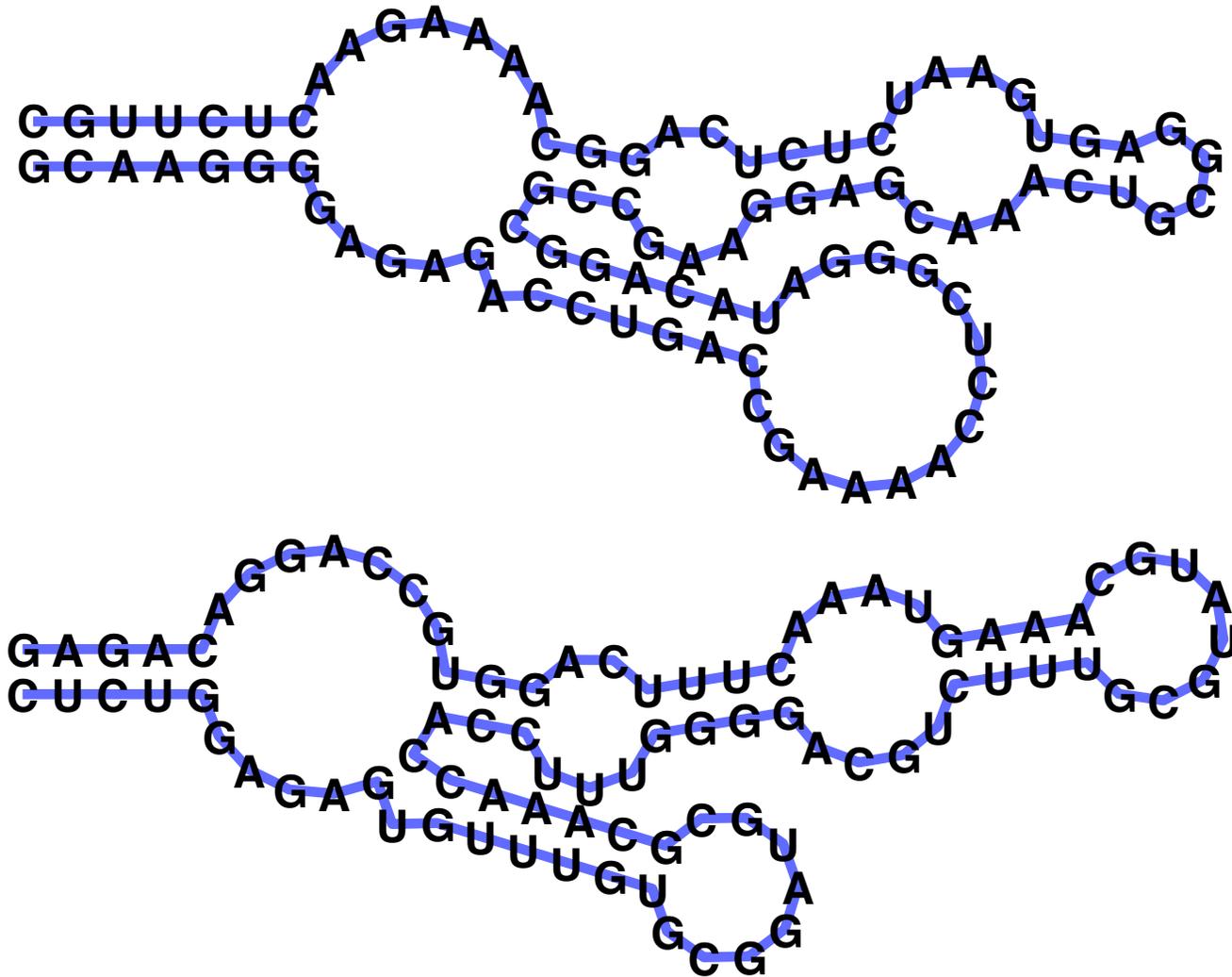
A: Structure often more important than sequence [32]

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?
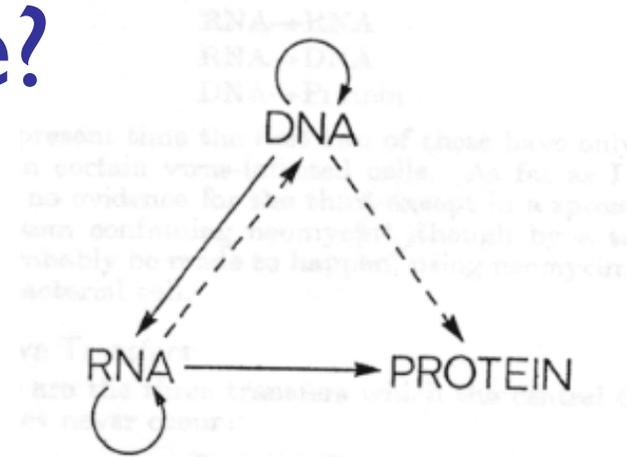
Origin of Life?

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
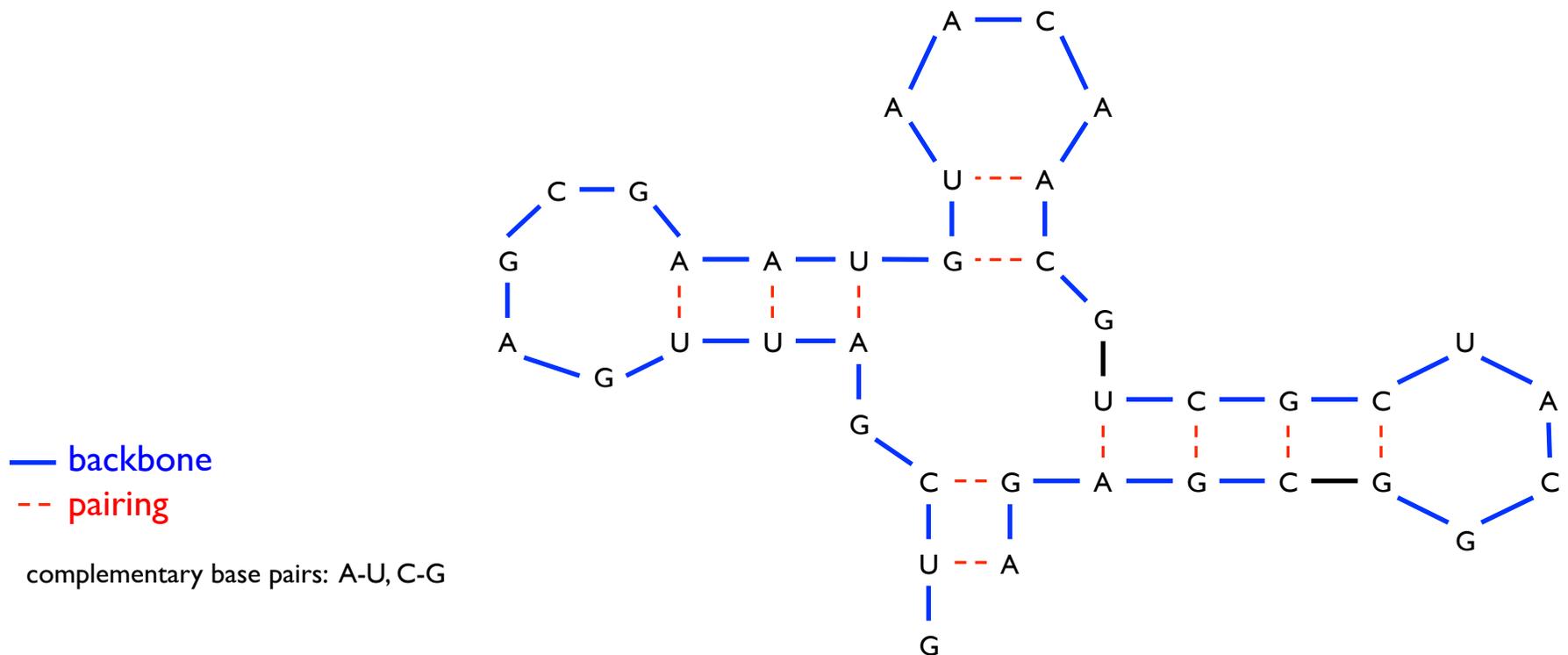
Ex: GUCGAUUGAGCGAAUGUAACAACGUGGCUACGGCGAGA

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.

base pair

ok

sharp turn

crossing

U A C C G G U G U A

U A C G G G U A

≤4

U A C C G G U U G A

U A C C G G U G U A A C
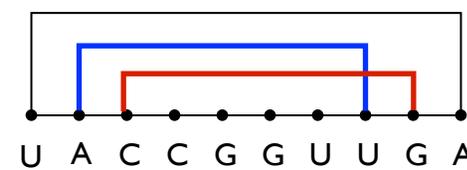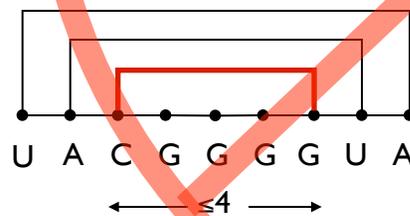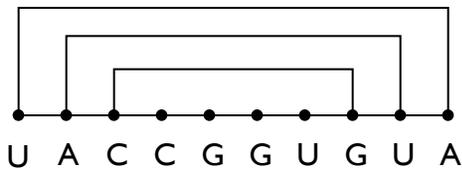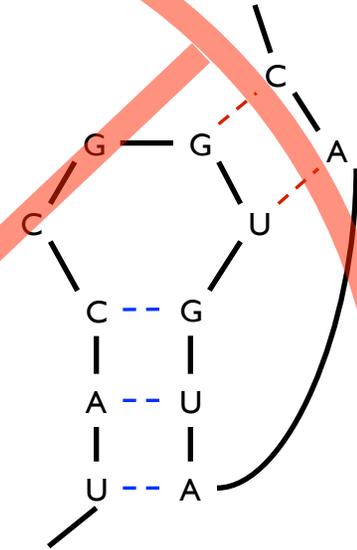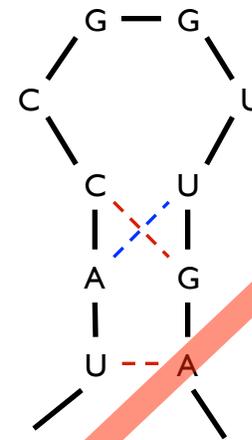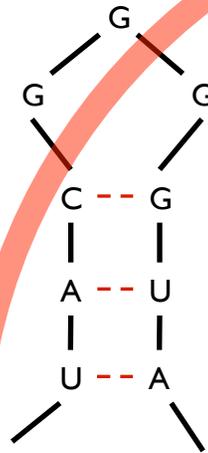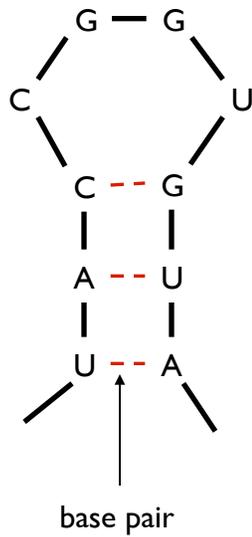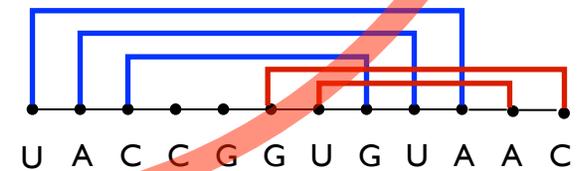
RNA Secondary Structure: Subproblems

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

match b_t and b_j

— backbone
-- pairing

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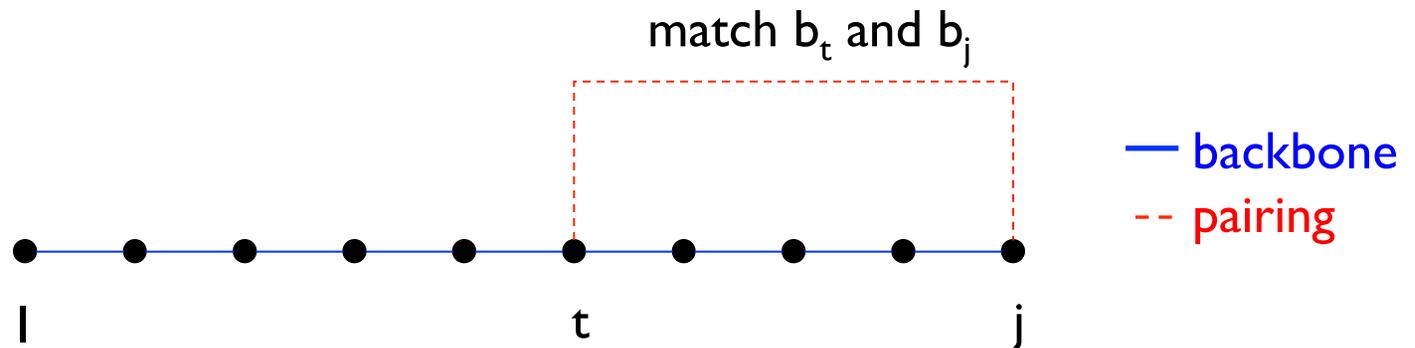
Difficulty. Results in two sub-problems.
- Finding secondary structure in: $b_1 b_2 \ldots b_{t-1}$.   ← OPT(t-1)
- Finding secondary structure in: $b_{t+1} b_{t+2} \ldots b_{j-1}$. ← not "OPT" of anything; need more flexible set of sub-problems

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.

  > $OPT[i, j] = OPT[i, j-1]$

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

- 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

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
        for i = 1, 2, …, n-k
            j = i + k
            Compute OPT[i, j]

    return OPT[1, n]

}
```

Interval length → `for k = 5, 6, …, n-1`

Start position → `for i = 1, 2, …, n-k`

End position → `j = i + k`

using recurrence
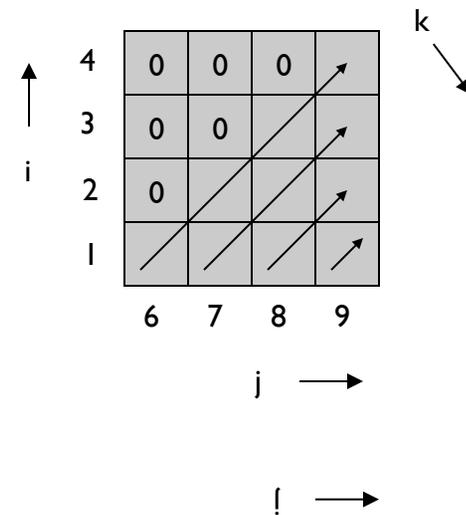
Running time. $O(n^3)$.
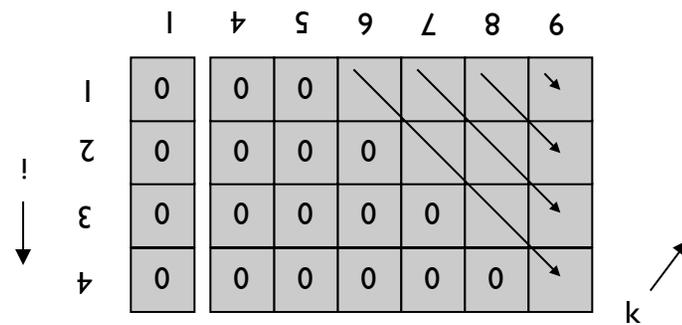
Nussinov: Max Pairing

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:

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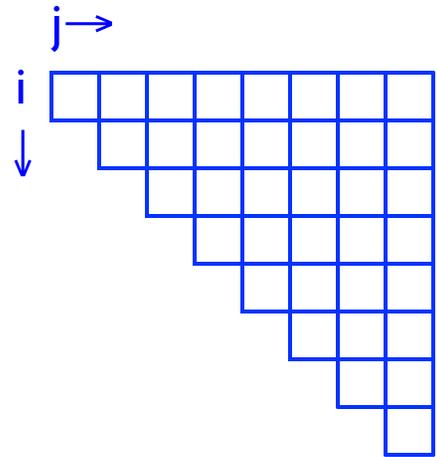
"Optimal pairing of b_i ... b_j"

Two possibilities:

j Unpaired:
Find best pairing of b_i ... b_{j-1}
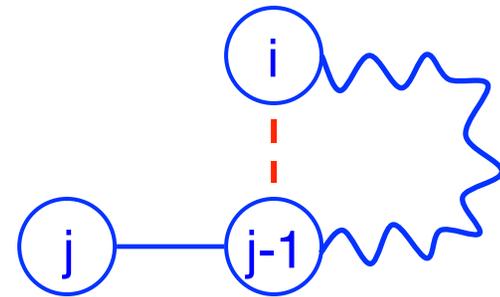
j Paired (with some t):
Find best b_i ... b_{t-1} +
best b_{t+1} ... b_{j-1} plus 1
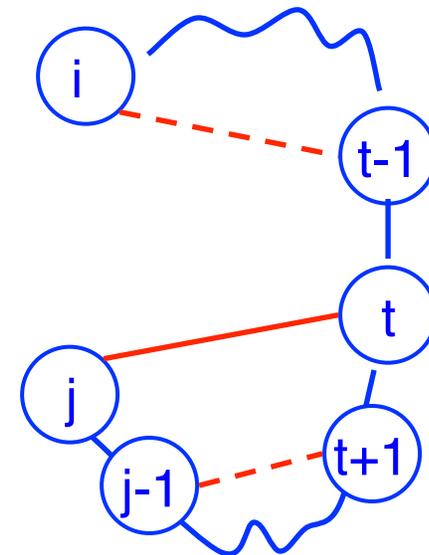
Why is it slow?
Why do pseudoknots matter?

— backbone
— pairing

43

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 ≥ j-4;
    otherwise:

opt(i,j) = max of:

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

j→

i↓

t=2

3

4

5
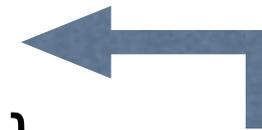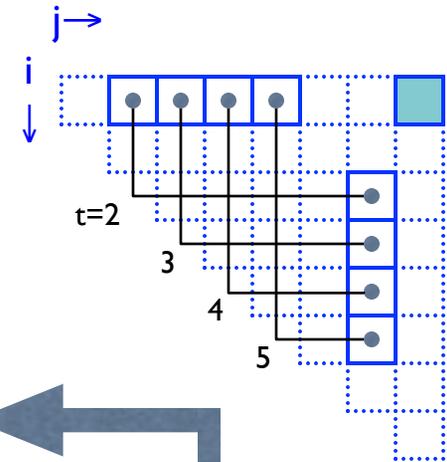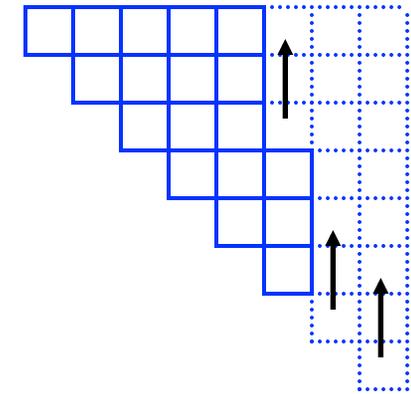
Time: $O(n^3)$

Which Pairs?

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

C U C C G G U U G C A A U G U C

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

n = 16

```
0 0 0 0 0 1 1 1 1 1 2 2 2 3 3 3
  0 0 0 0 0 0 1 1 2 2 2 2 2 2
    0 0 0 0 0 1 1 1 1 1 2 2 2
      0 0 0 0 1 1 1 1 1 2 2 2
        0 0 0 0 1 1 1 1 1 1 2
          0 0 0 0 0 1 1 1 1 2
            0 0 0 0 0 1 1 1 1
              0 0 0 0 0 0 0 1
                0 0 0 0 0 0 1
                  0 0 0 0 0 0
                    0 0 0 0 0
                      0 0 0 0
                        0 0 0
                          0 0
                            0
```

0

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

CUCCGG

(. . . .)

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

GUUGCAAUGUC

((. . . .) . . .)

(Examples here and below assume 1-based indexing)

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

```
G G G A A A A C C C A A A G G G G U U U      n= 20
( ( ( . . . . ) ) ) ( ( ( . . . . ) ) )
0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6
0 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 4 5 6
0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3
0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2
```

Case 1:
  $2 \geq 18-4$? no.
Case 2:
  B_{18} unpaired?
  Always a possibility;
  then OPT[2,18] $\geq$ 3

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

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

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

```
G G G A A A A C C C A A A G G G G U U U     n= 20
( ( ( . . . . ) ) ) ( ( ( . . . . ) ) )
0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6
0 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 4 5 6
0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3
0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2
```

Case 3, $2 \leq t < 18\text{-}4$:
t = 2: no pair

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

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

```
G  G  G  A  A  A  A  C  C  C  A  A  A  G  G  G  G  U  U  U     n=  20
(  (  (  .  .  .  .  )  )  )  (  (  (  .  .  .  .  )  )  )
0  0  0  0  0  0  0  1  2  3  3  3  3  3  3  3  3  4  5  6
0  0  0  0  0  0  0  1  2  2  2  2  2  2  3  3  3  4  5  6
0  0  0  0  0  0  0  1  1  1  1  1  1  2  2  3  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  5
0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  4  4
0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  3  3  3
0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  2  2  2  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2
```

Case 3, $2 \leq t < 18\text{-}4$:
   $t = 3$: no pair

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

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

G G G [A] A A A C C C A A A G G G [U] U U   n= 20

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

0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6

0 0 ⬡0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 [4] 5 6

0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 [3] 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2

Case 3, $2 \leq t < 18-4$:
t = 4: yes pair
OPT[2,18]$\geq$1+0+3

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

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

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

G G G A A A A C C C A A A G G G G U U U   n= 20

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

```
0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6
0 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 4 5 6
0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3
0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2
```

Case 3, $2 \le t < 18-4$:
t = 5: yes pair
OPT[2,18]$\geq$1+0+3

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

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

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

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

G G G A A [A] A C C C A A A G G G [U] U U | n= 20 |

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

0	0	0	0	0	0	0	1	2	3	3	3	3	3	3	3	3	4	5	6
0	0	0	0	0	0	0	1	2	2	2	2	2	2	3	3	3	4	5	6
0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	3	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	5
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	4	4
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	3	3
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	2	2	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2

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

GGAA**A**CCCAAAGGGG**U**
. . . . (. (((. . . .))))

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

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

```
G  G  G  A  A  A  A  C  C  C  A  A  A  G  G  G  U  U  U      n= 20
(  (  (  .  .  .  .  )  )  )  (  (  (  .  .  .  .  )  )  )
0  0  0  0  0  0  0  1  2  3  3  3  3  3  3  3  3  4  5  6
0  0  0  0  0  0  0  0  1  2  2  2  2  2  2  3  3  3  4  5  6
0  0  0  0  0  0  0  0  1  1  1  1  1  1  2  2  3  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  5
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  4  4
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  3  3  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  2  2  2  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2
```
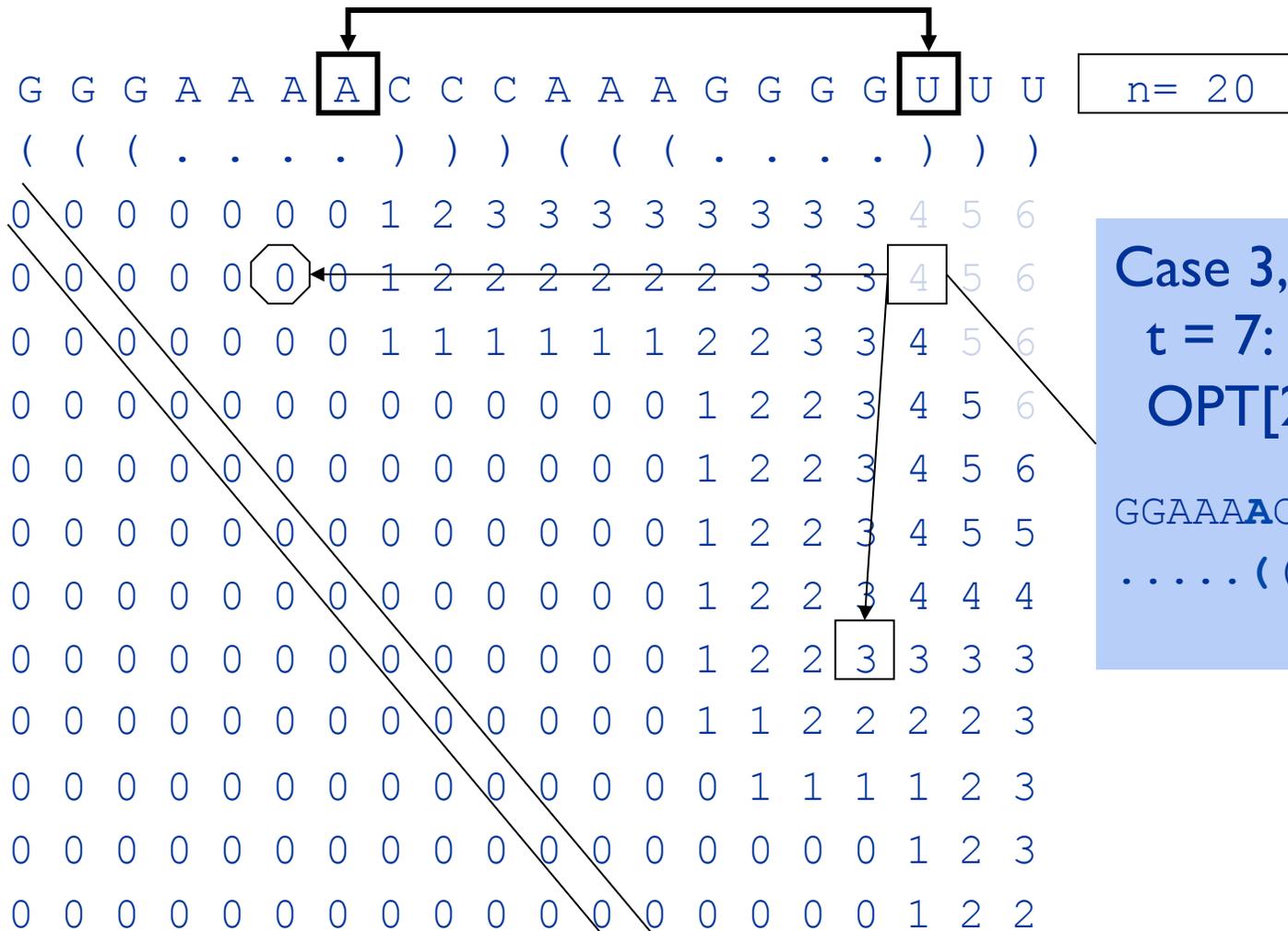
Case 3, $2 \le t < 18-4$:
$t = 7$: yes pair
$OPT[2,18] \ge 1+0+3$

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

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

53

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

```
G  G  G  A  A  A  A  C  C  C  A  A  A  G  G  G  U  U  U      n= 20
(  (  (  .  .  .  .  )  )  )  (  (  (  .  .  .  .  )  )  )
0  0  0  0  0  0  0  1  2  3  3  3  3  3  3  3  3  4  5  6
0  0  0  0  0  0  0  0  1  2  2  2  2  2  2  3  3  3  4  5  6
0  0  0  0  0  0  0  1  1  1  1  1  1  2  2  3  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  6
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  5  5
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  4  4  4
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2  3  3  3  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  2  2  2  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  3
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  2
```
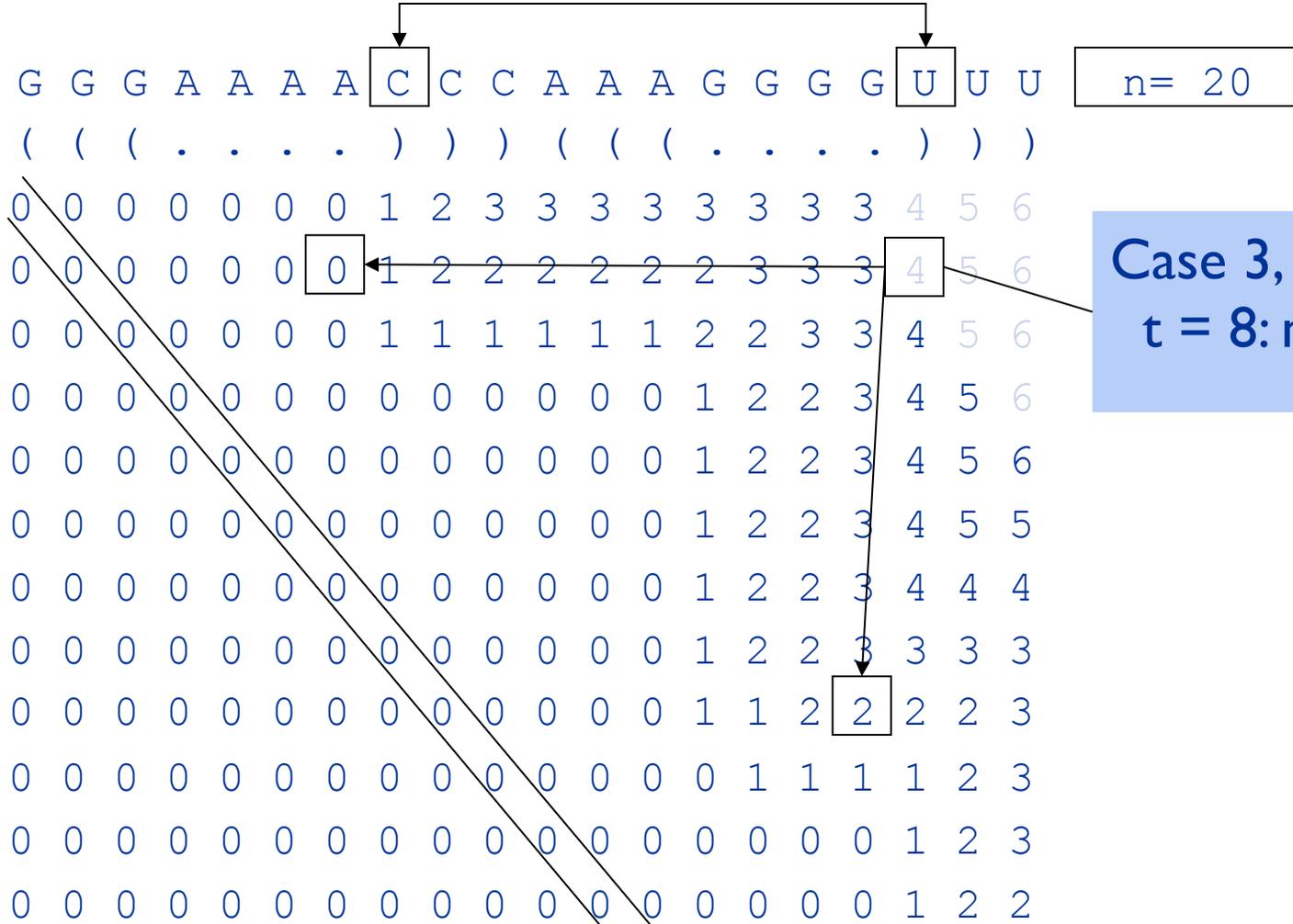
Case 3, $2 \le t < 18\text{-}4$: $t = 8$: no pair

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

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

G G G A A A A C C C A A A G G G G U U U   n= 20

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

0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6
0 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 4 5 6
0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4
0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3
0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2

Case 3, $2 \le t < 18-4$:
t = 11: yes pair
OPT[2,18] $\geq$ 1+2+0

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

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

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

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

55

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

n= 20

G G G A A A A C C C A A A G G G U U U

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

0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 4 5 6

0 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 **4** 5 6

0 0 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4

0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 3 3 3

0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 3

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2
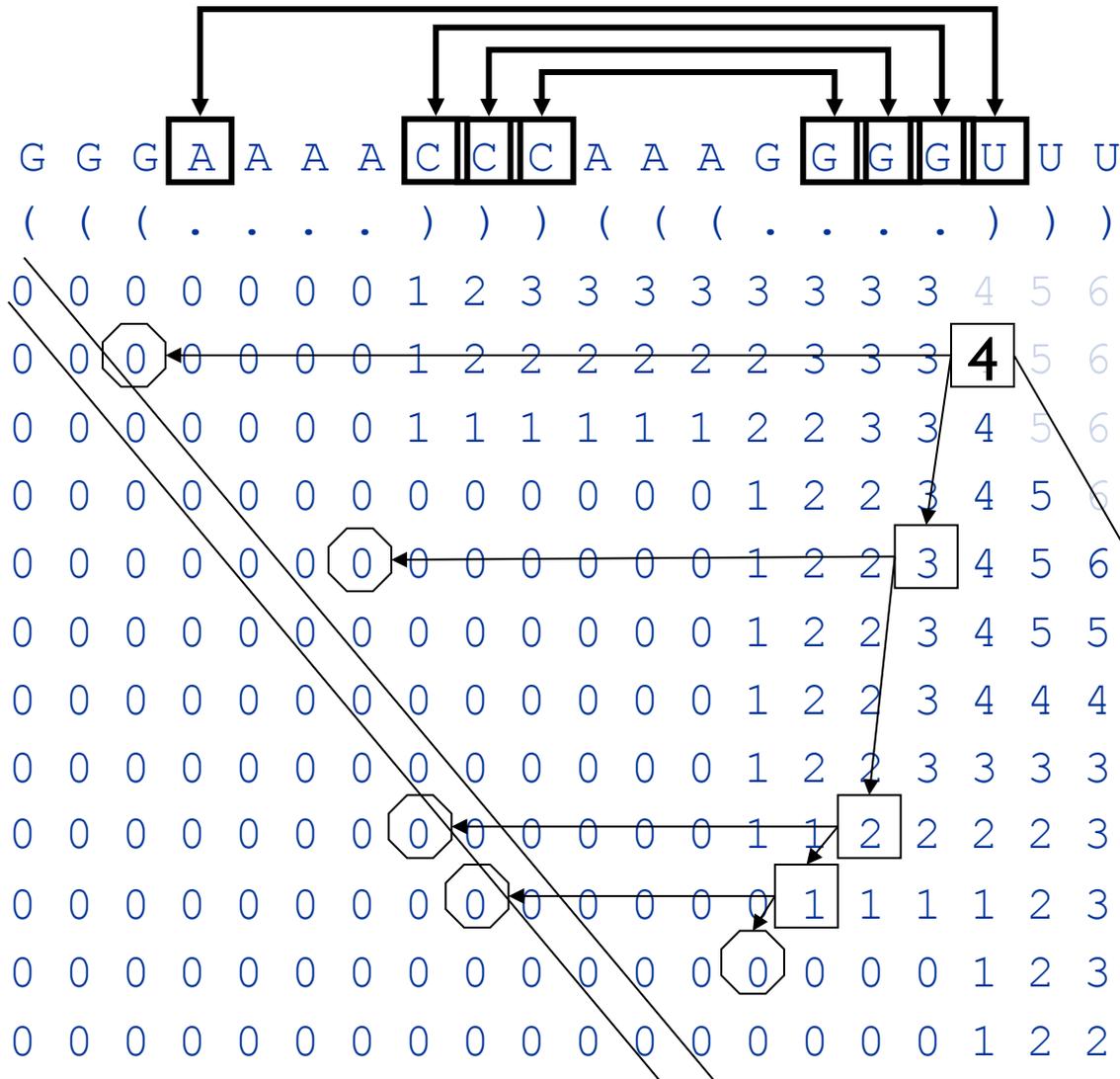
Overall, Max = 4
several ways, e.g.:

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

tree shows trace back:
square = case 3
octagon = case 1

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

56

Another Trace Back Example

C U C C G G U U G C A A U G U C $n = 16$

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

0	0	0	0	0	1	1	1	1	1	2	2	2	3	3	3
	0	0	0	0	0	0	1	1	2	2	2	2	2	2	2
		0	0	0	0	0	1	1	1	1	1	2	2	2	2
			0	0	0	0	1	1	1	1	1	2	2	2	2
				0	0	0	0	1	1	1	1	1	1	1	2
					0	0	0	0	1	1	1	1	1	1	2
						0	0	0	0	1	1	1	1	1	1
							0	0	0	0	0	0	0	0	1
								0	0	0	0	0	0	0	1
									0	0	0	0	0	0	0
										0	0	0	0	0	0
											0	0	0	0	0
												0	0	0	0
													0	0	0
														0	0
															0

0
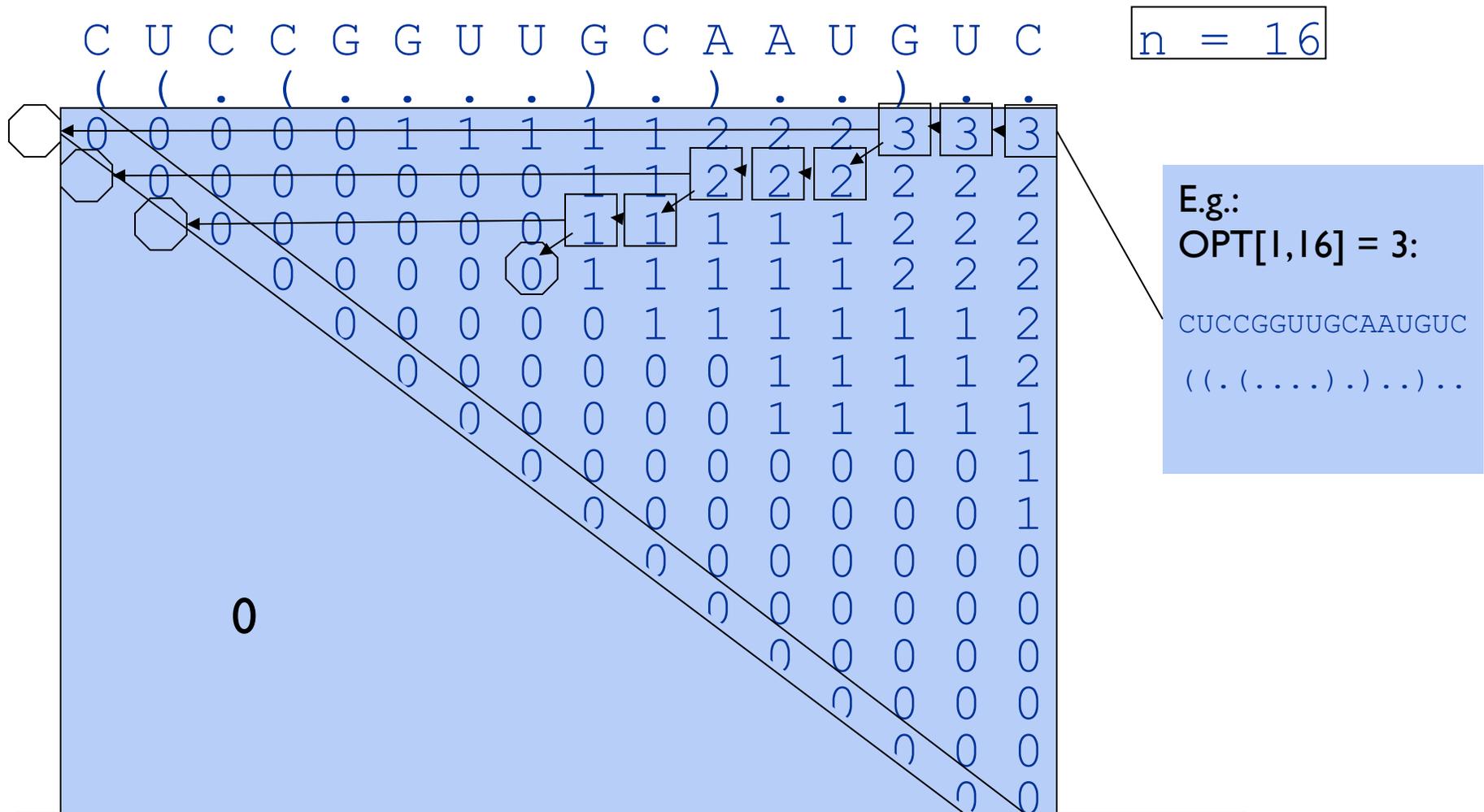
E.g.:
OPT[1,16] = 3:

CUCCGGUUGCAAUGUC

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

$$\text{OPT}(i,j) = \begin{cases} 0 & \text{if } i \geq j-4 \\ \max\begin{cases} \text{OPT}[i,j-1] \\ 1 + \max_t(\text{OPT}[i,t-1] + \text{OPT}[t+1,j-1]) \end{cases} & \text{otherwise} \end{cases}$$
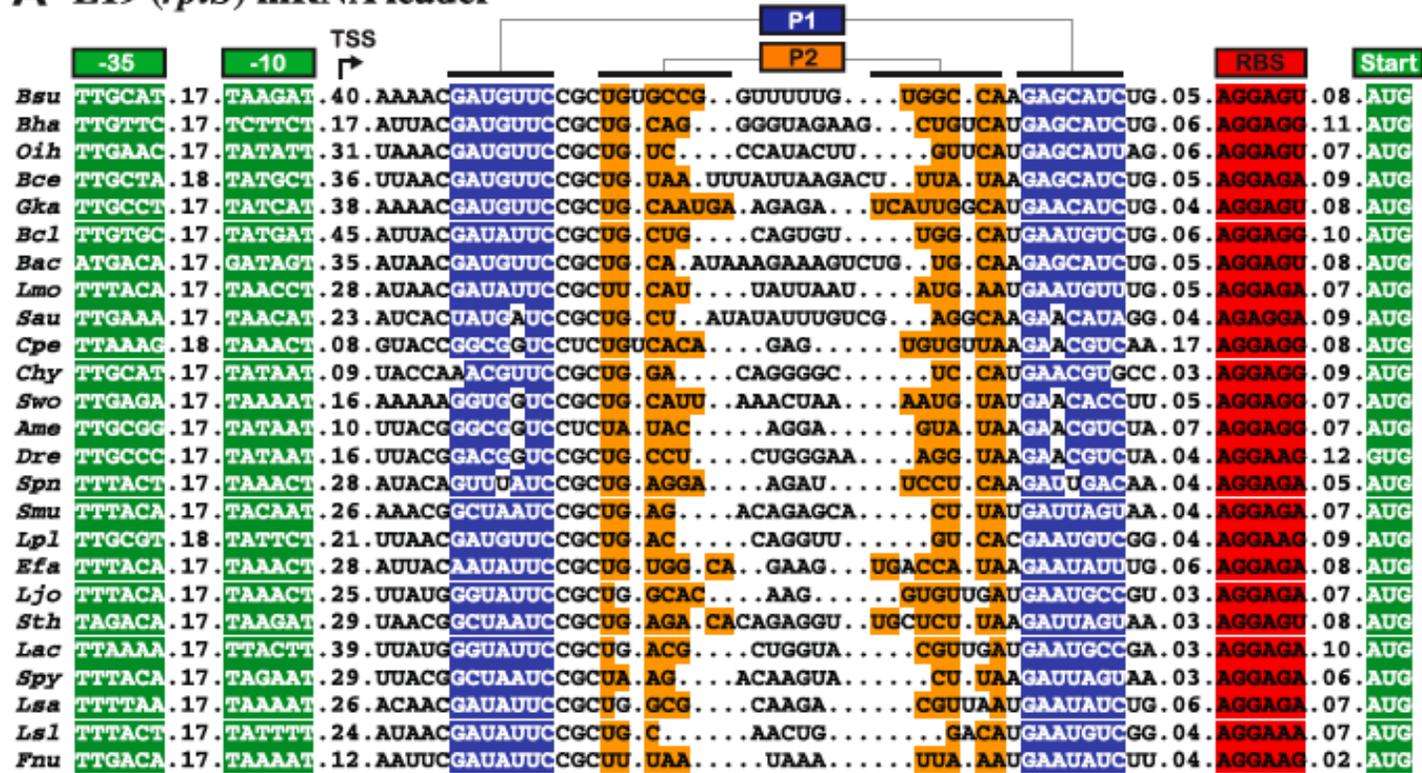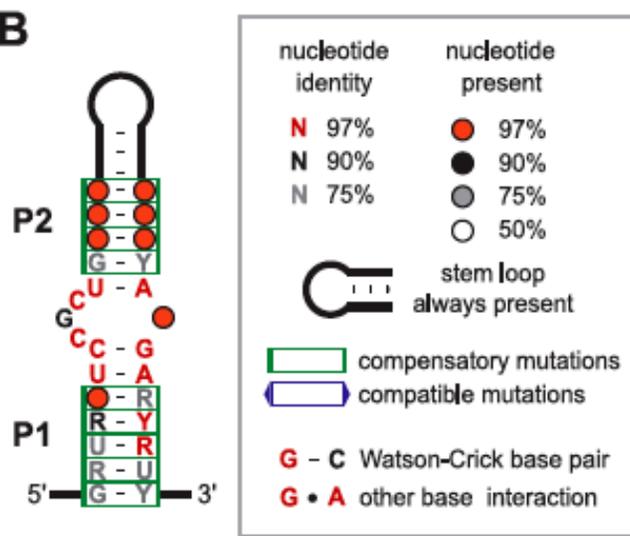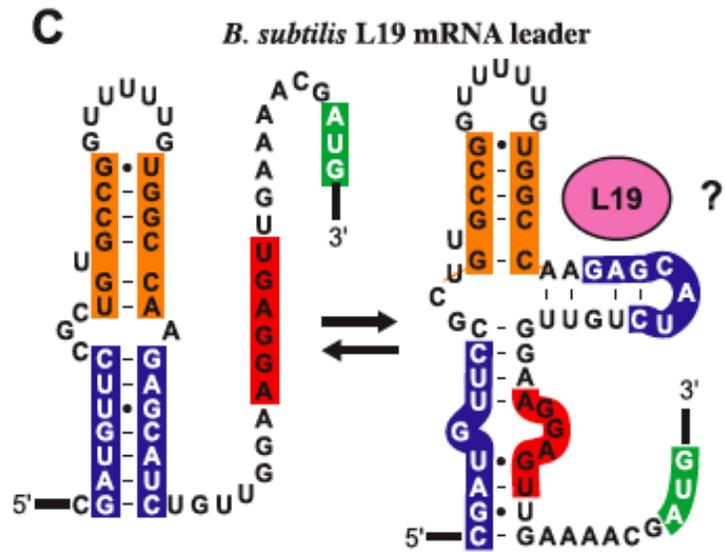
Figure 3. Putative Autoregulatory Structure in L19 mRNA Leaders

58