

CSE 427

Computational Biology

RNA: Function, Secondary Structure
Prediction, Search, Discovery

The Message

Cells make lots of ~~RNA~~ *noncoding* RNA

Functionally important, functionally diverse

Structurally complex

New tools required

alignment, discovery, search, scoring, etc.

Rough Outline

Today

Noncoding RNA Examples

RNA structure prediction

RNA

DNA: DeoxyriboNucleic Acid

RNA: RiboNucleic Acid

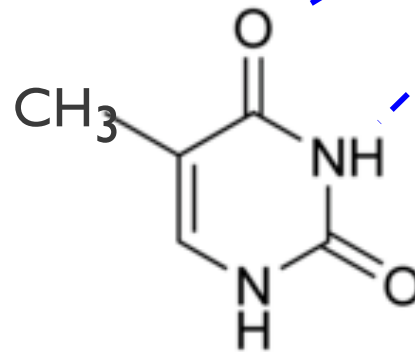
Like DNA, except:

Adds an OH on ribose (backbone sugar)

Uracil (U) in place of thymine (T)

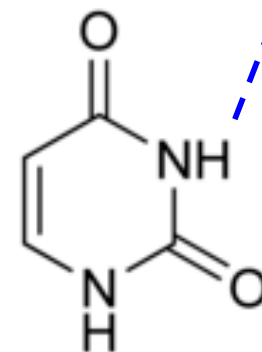
A, G, C as before

thymine

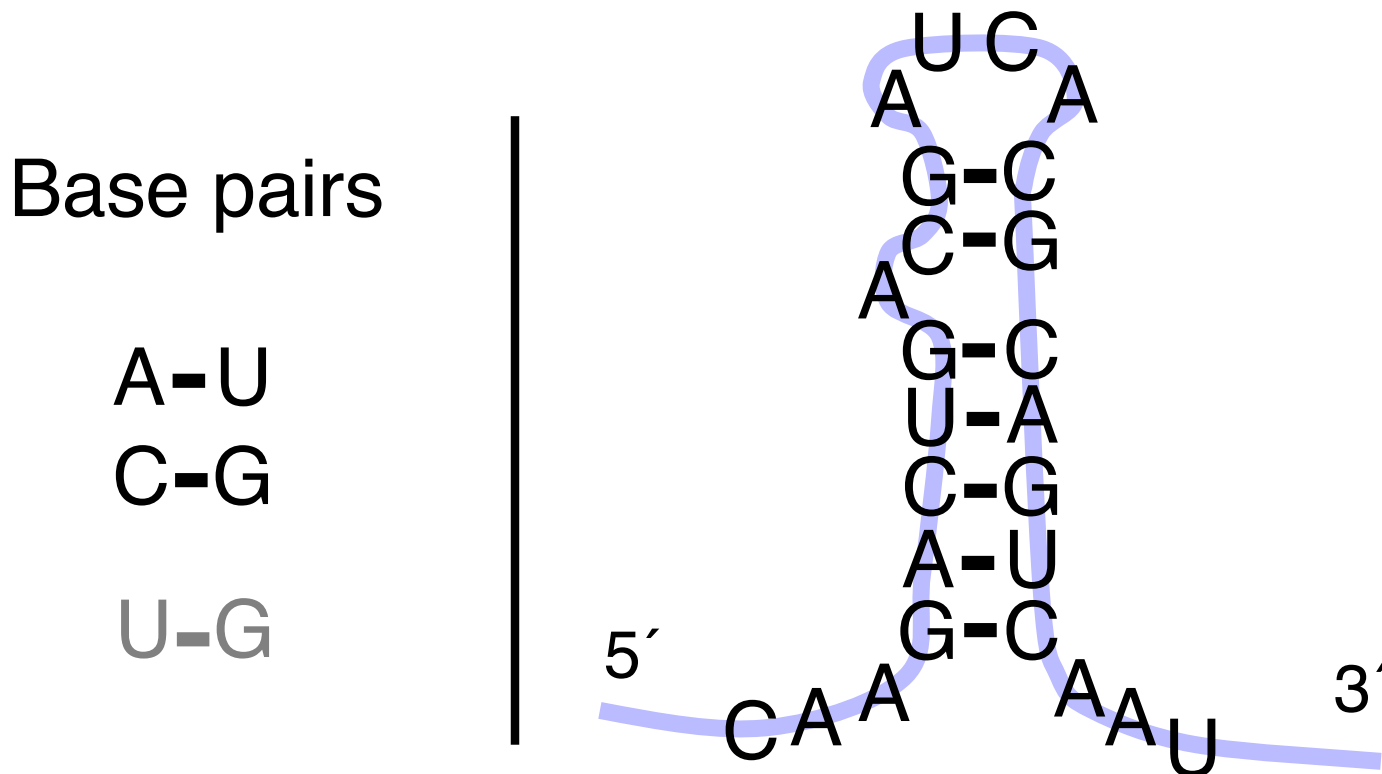


pairs
with A

uracil



RNA Secondary Structure: RNA makes helices too



Usually *single* stranded

Central Dogma of Molecular Biology

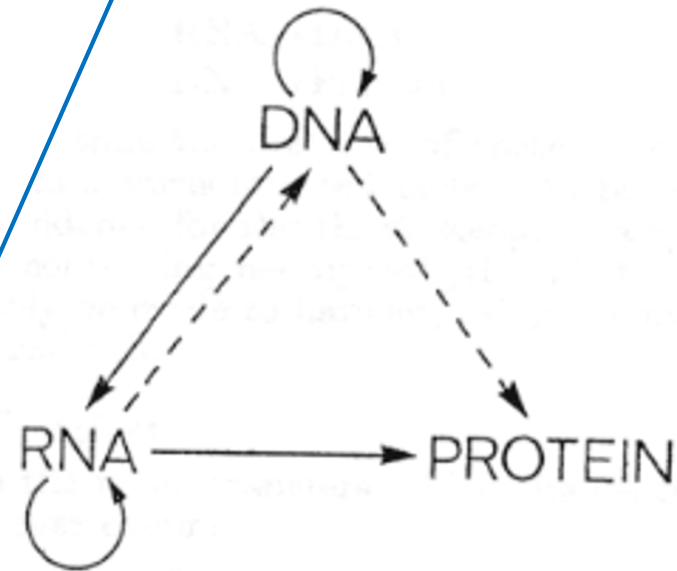
by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

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 nucleic acid.

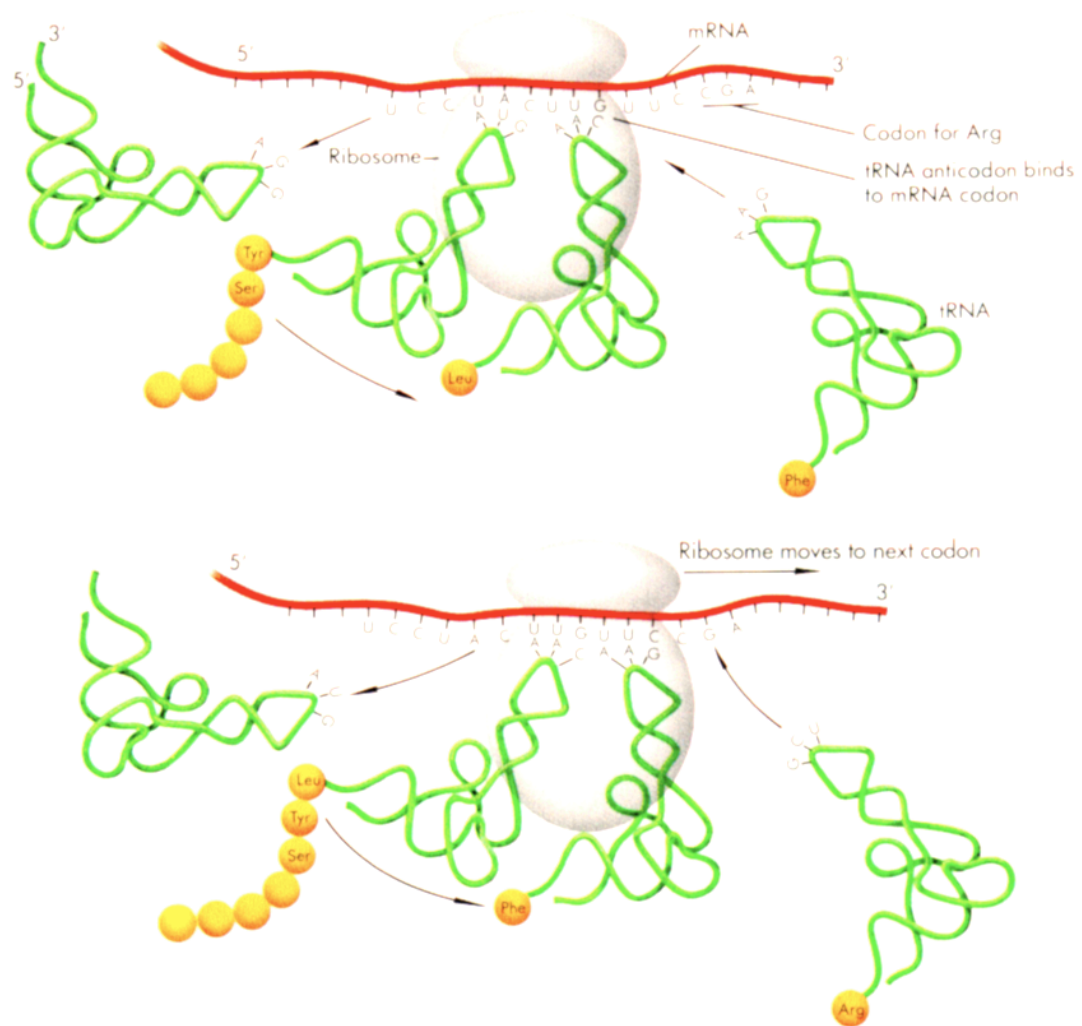
"The central dogma, the keystone of molecular biology, has been the subject of considerable over-simplification."

*Re: Origins -
Chicken & Egg
Problem*

Fig. 2. The arrows show the situations that 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.



Ribosomes



Ribosomes

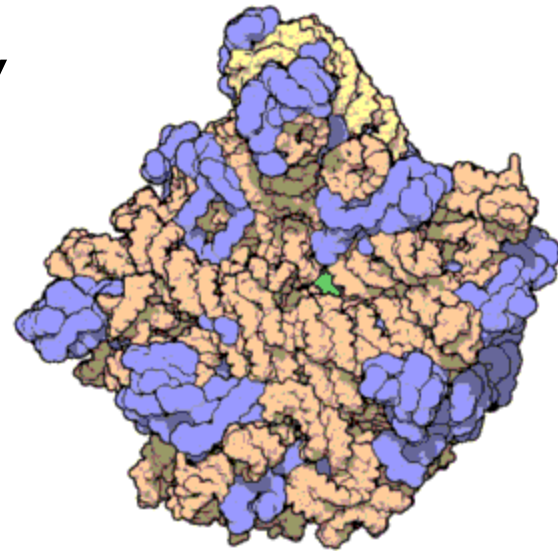
1974 Nobel prize to Romanian biologist George Palade (1912-2008) for discovery in mid 50's

50-80 proteins

3-4 RNAs (half the mass)

Catalytic core is RNA

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



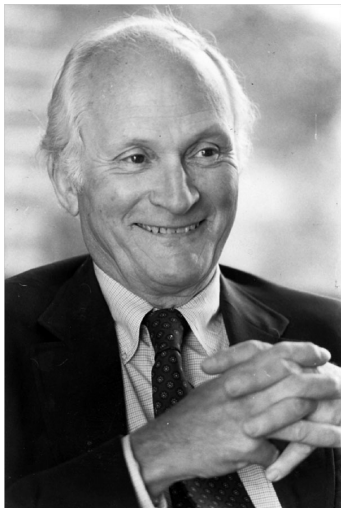
Atomic structure of the 50S Subunit from *Haloarcula marismortui*. Proteins are shown in blue and the two RNA strands in orange and yellow. The small patch of green in the center of the subunit is the active site.

- Wikipedia

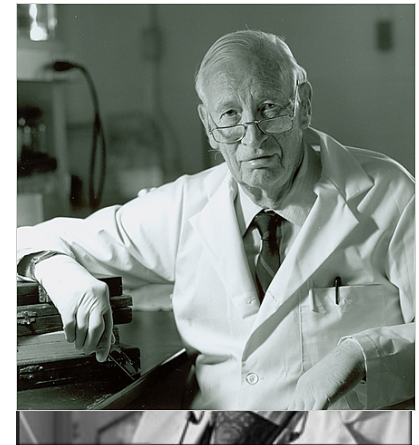
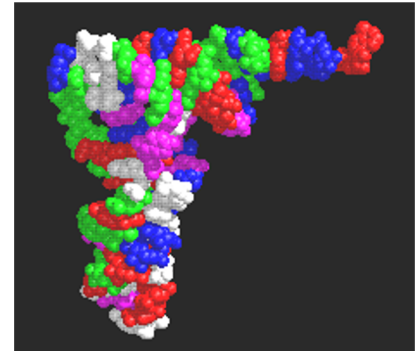
Transfer RNA

The “adapter” coupling mRNA to protein synthesis.

Discovered in the mid-1950s by



Mahlon Hoagland (1921-2009, left), Mary Stephenson, and Paul Zamecnik (1912-2009; Lasker award winner, right).



Bacteria

Triumph of proteins

50-80% of genome is coding DNA

Functionally diverse

receptors

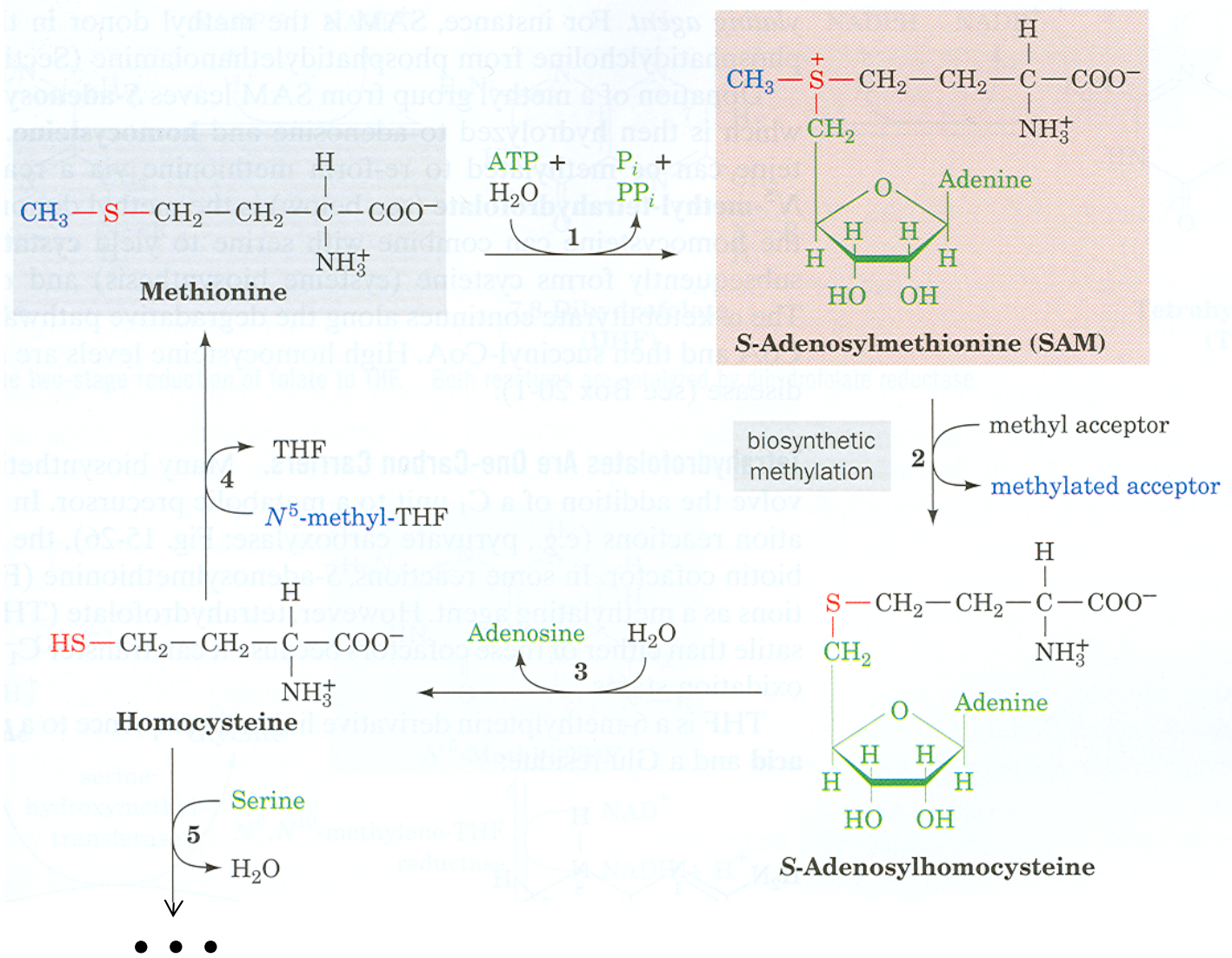
motors

catalysts

regulators (Monod & Jakob, Nobel prize 1965)

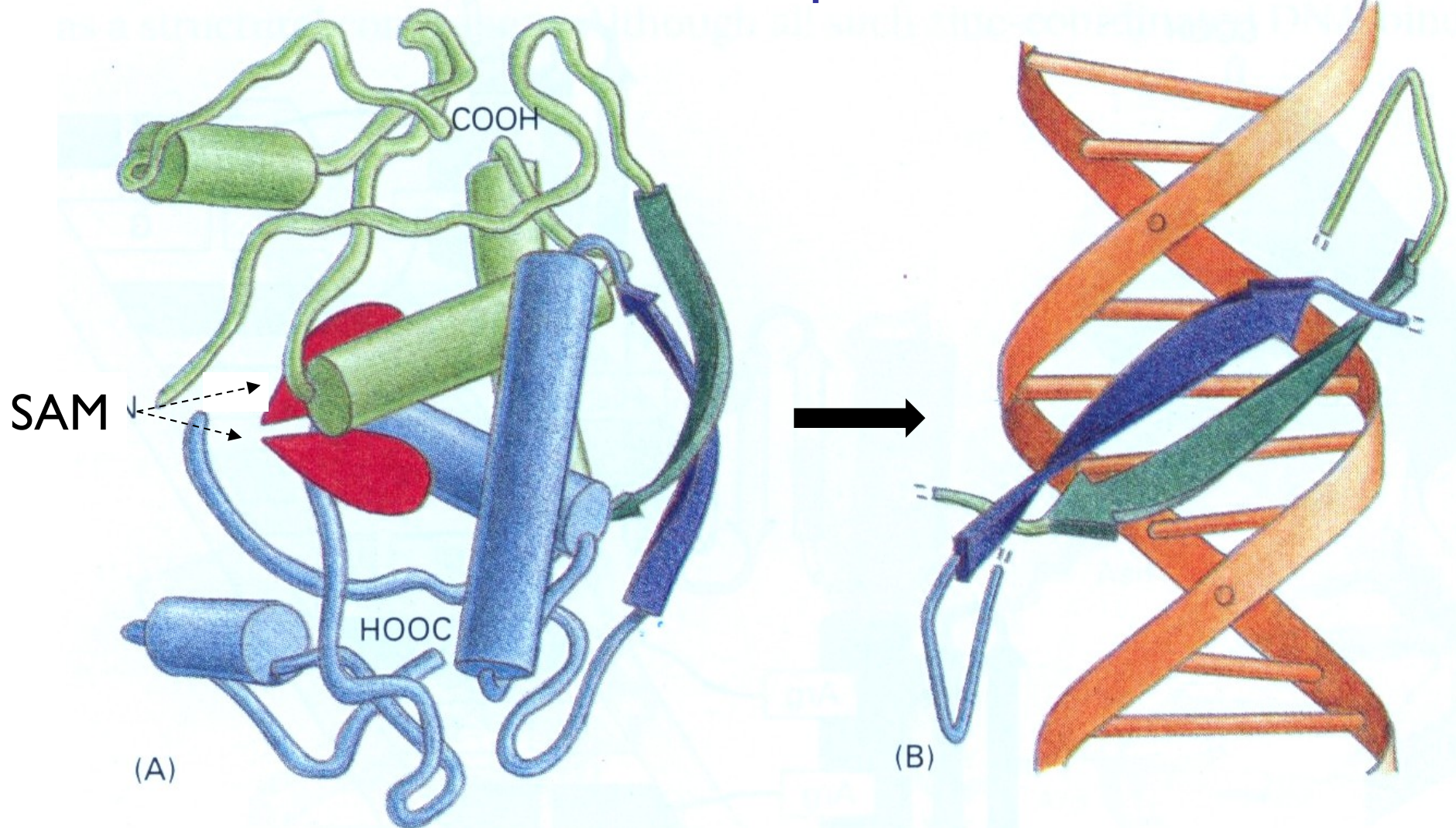
...

Proteins Catalyze Biochemistry: Met Pathways



Proteins Regulate Biochemistry:

The MET Repressor

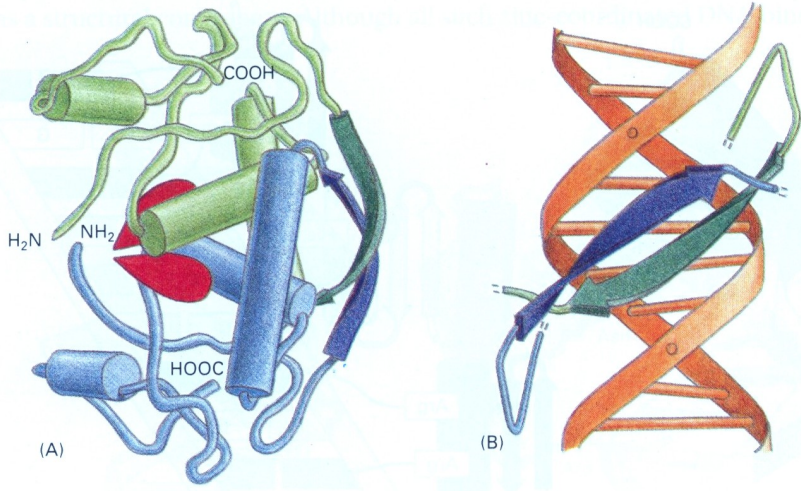


Protein

Alberts, et al, 3e.

DNA

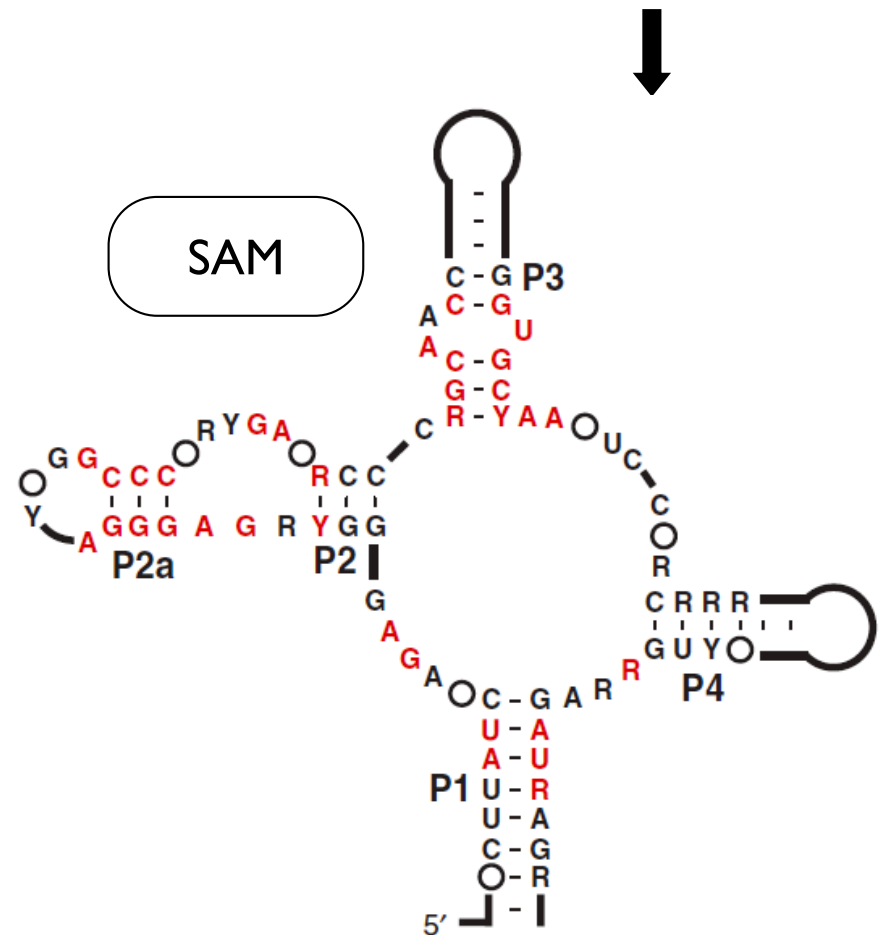
Alberts, et al, 3e.



Not the only way!

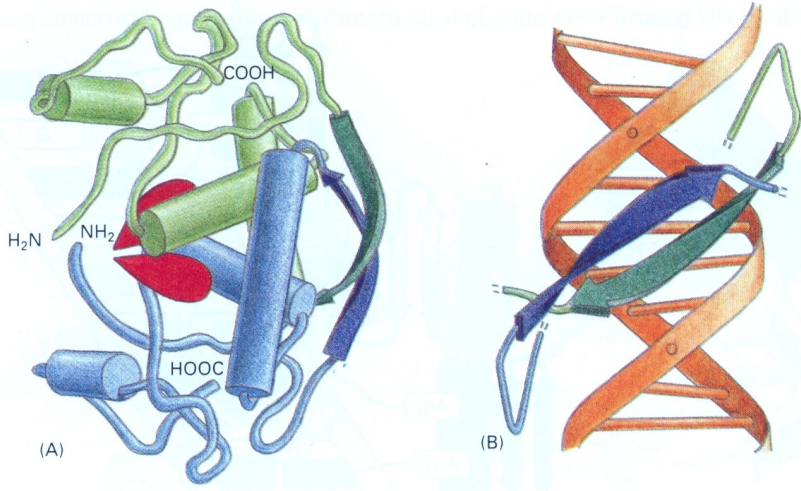
← Protein way

Riboswitch alternative



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

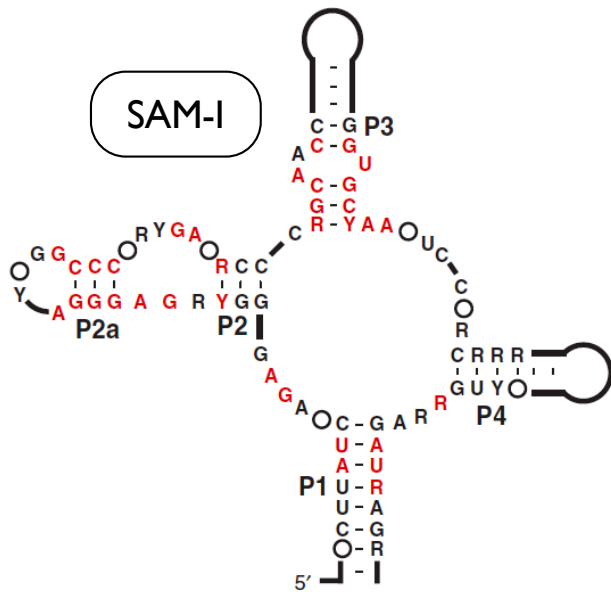
Alberts, et al, 3e.



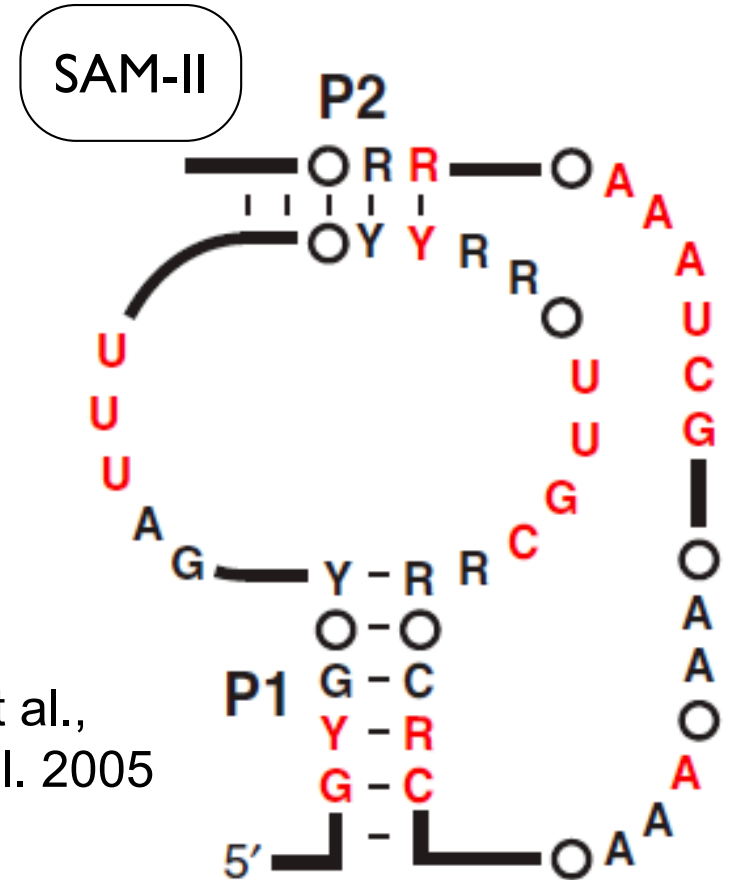
Not the only way!

Protein way

Riboswitch alternatives

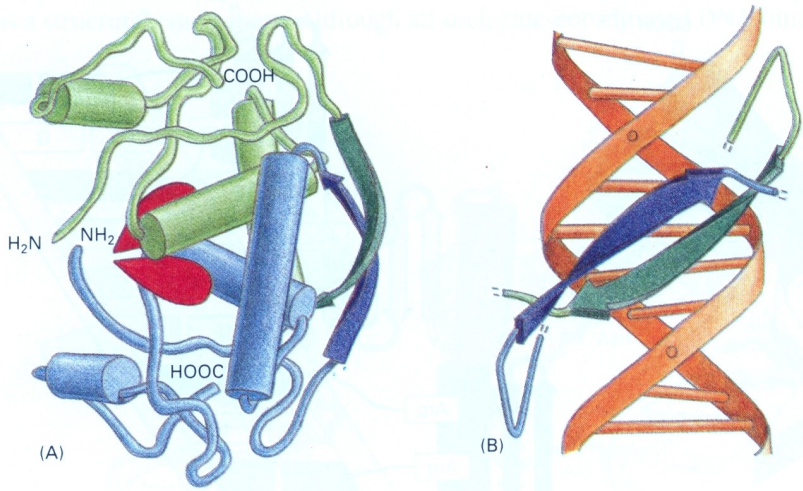


Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

Alberts, et al, 3e.



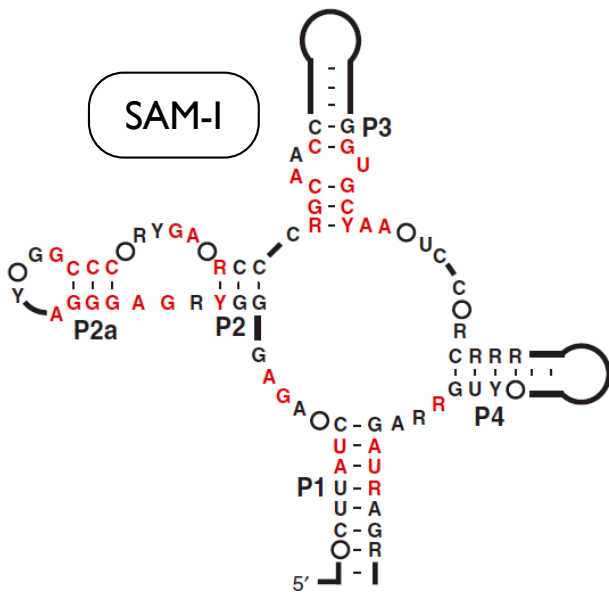
Not the only way!

Protein way

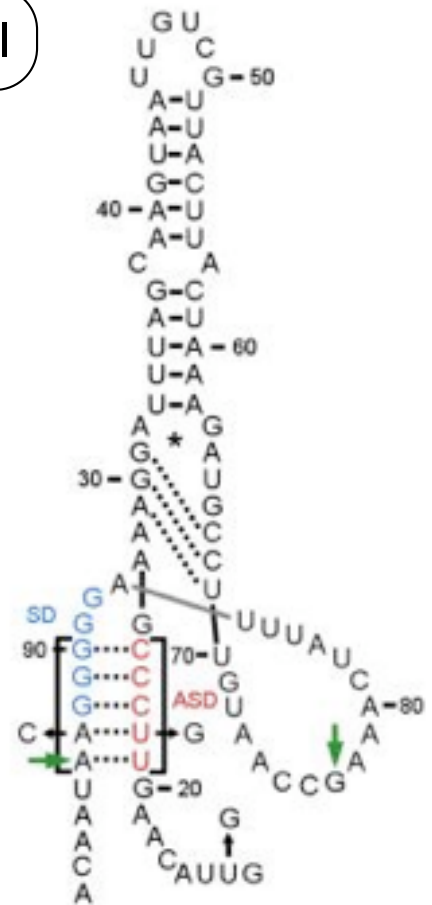
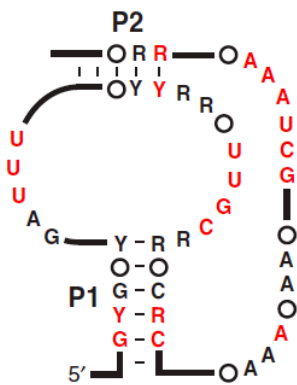
Riboswitch alternatives



SAM-III



SAM-II

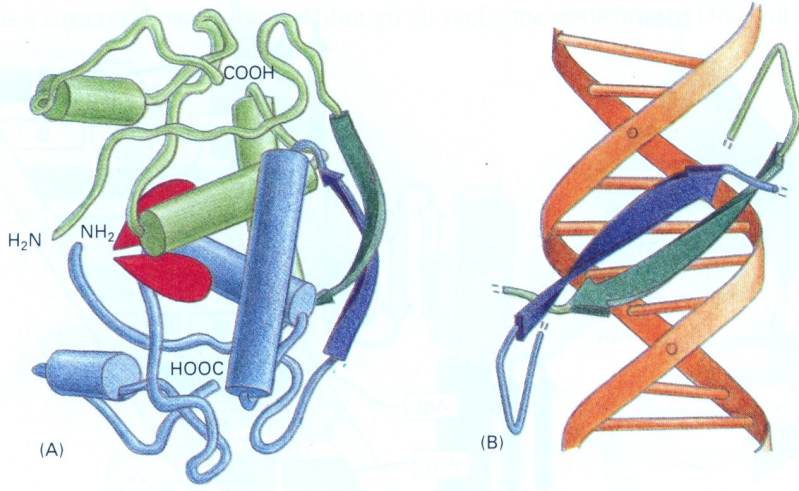


Fuchs et al., NSMB 2006

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

Corbino et al., Genome Biol. 2005

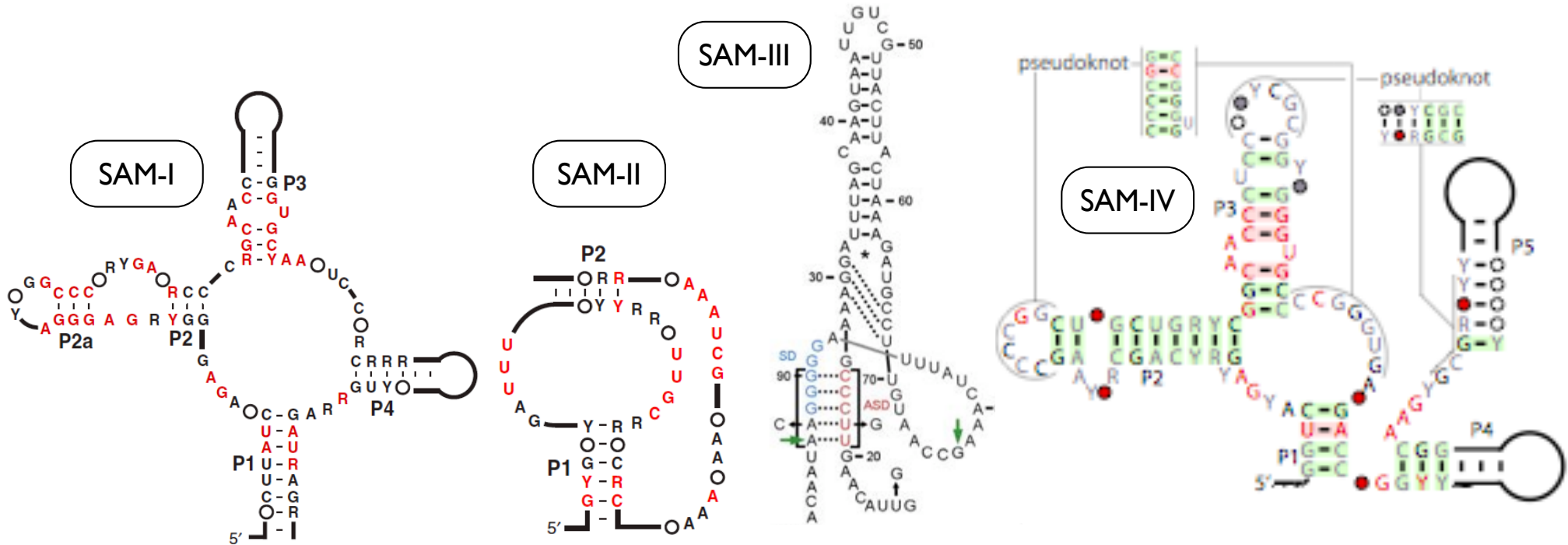
Alberts, et al, 3e.



Not the only way!

Protein way

Riboswitch alternatives



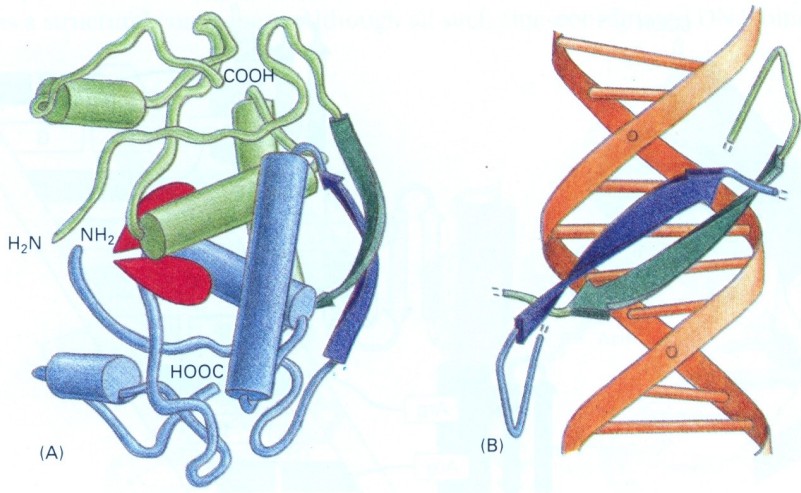
Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008

Alberts, et al, 3e.



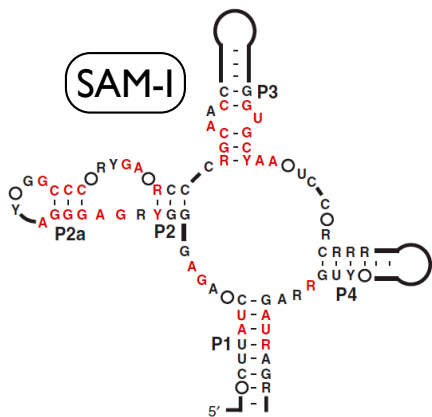
Not the only way!

Protein way

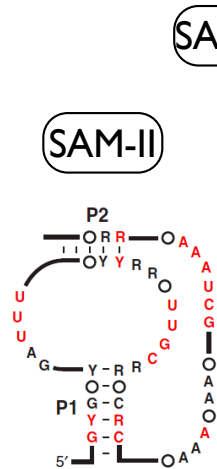
Riboswitch alternatives

SAM-VI

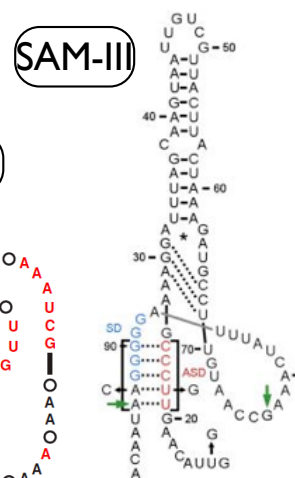
...



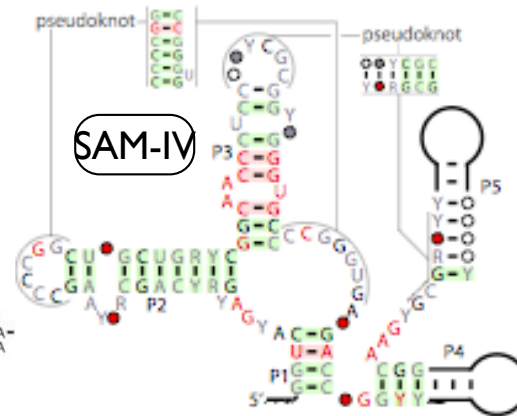
Grundy, Epshtein, Winkler et al., 1998, 2003



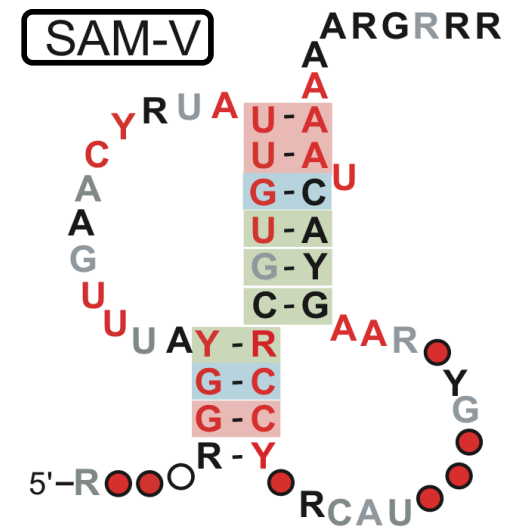
Corbino et al., Genome Biol. 2005



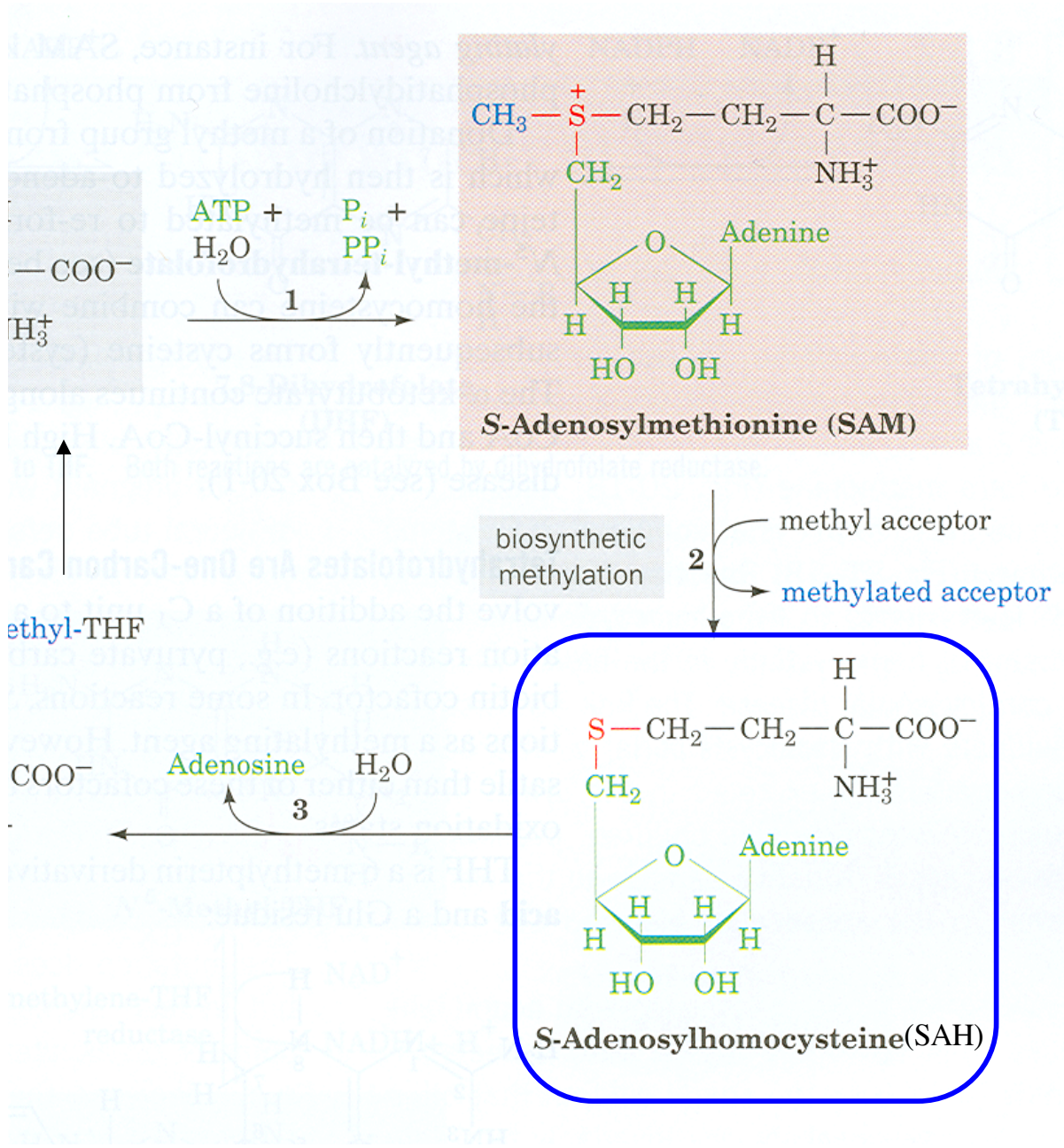
Fuchs et al., NSMB 2006



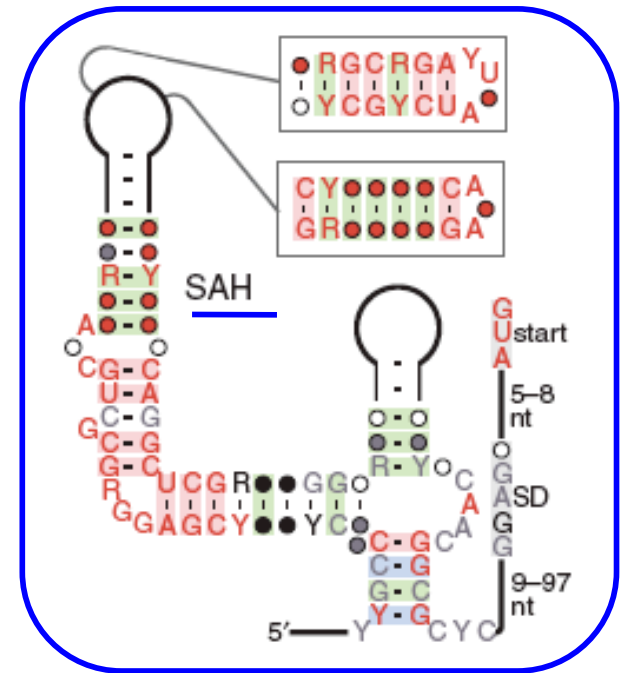
Weinberg et al., RNA 2008



Meyer, et al., BMC Genomics 2009



And in other bacteria, a riboswitch senses SAH



ncRNA Example: Riboswitches

UTR structure that directly senses/binds small molecules & regulates mRNA

widespread in prokaryotes

some in eukaryotes & archaea, one in a phage

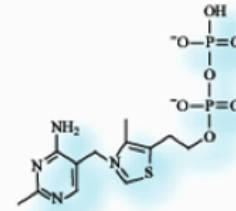
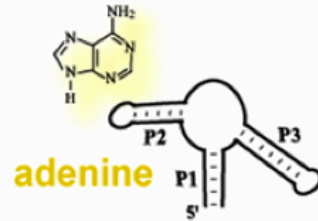
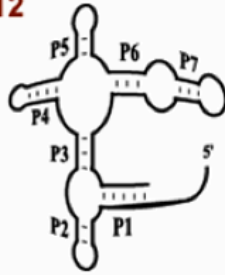
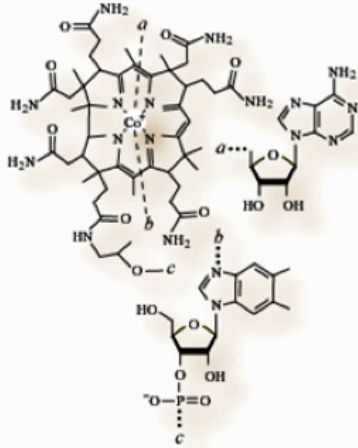
~ 20 ligands known; multiple nonhomologous solutions for some

dozens to hundreds of instances of each

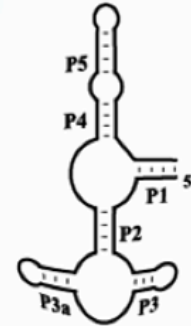
on/off; transcription/translation; splicing; combinatorial control

all found since ~2003; most via bioinformatics

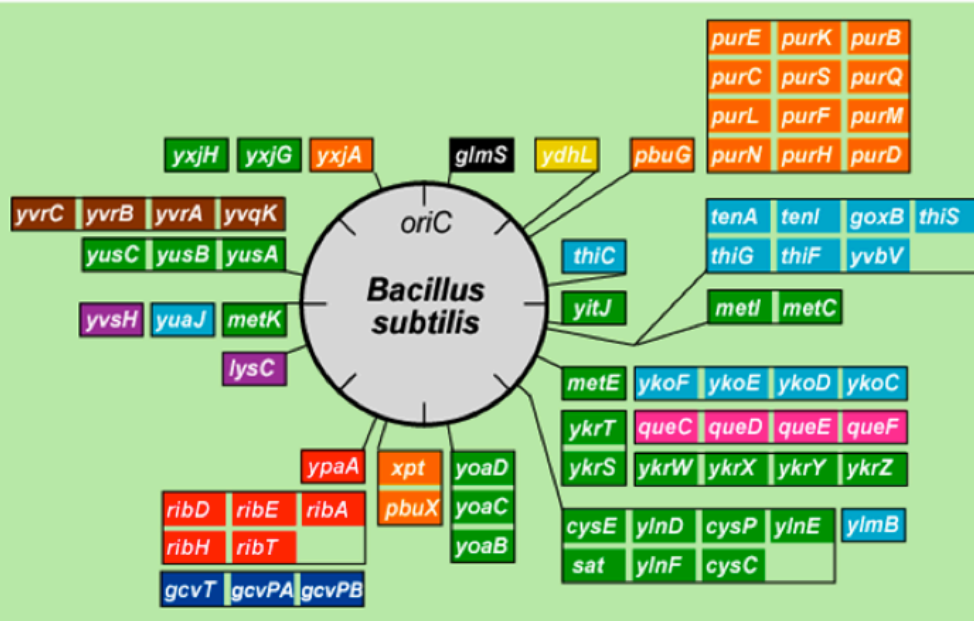
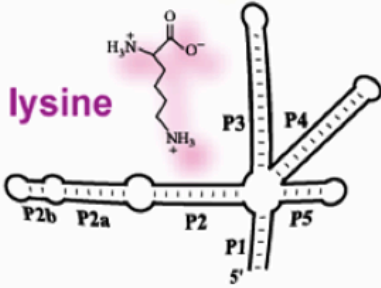
coenzyme B₁₂



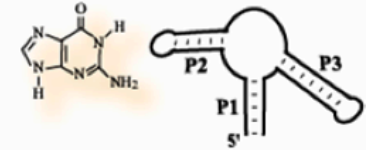
thiamine pyrophosphate



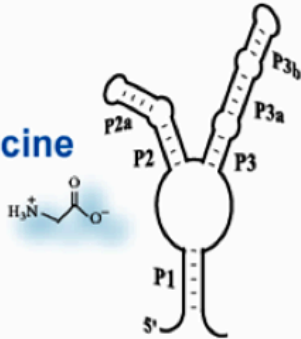
lysine



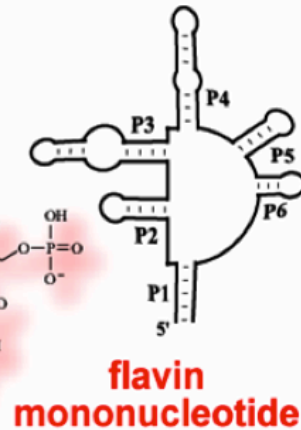
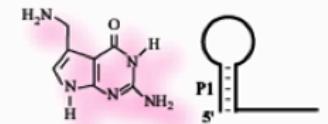
guanine



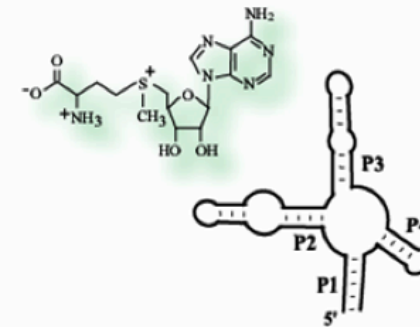
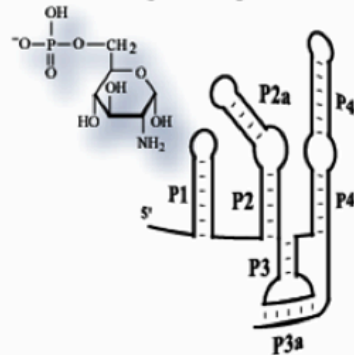
glycine



pre-queosine₁



glucosamine-6-phosphate



S-adenosyl-methionine

New Antibiotic Targets?

Old drugs, new understanding:

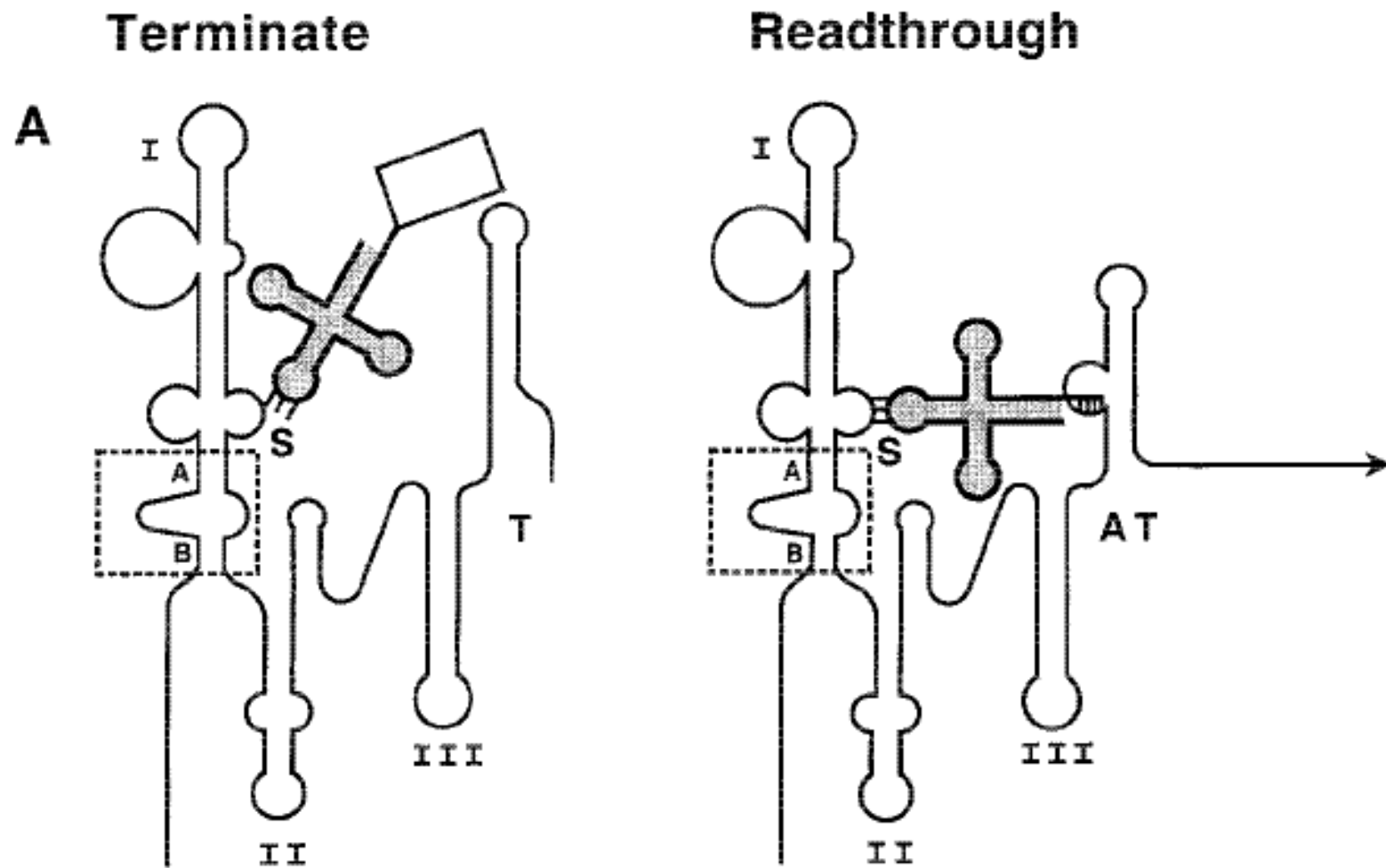
TPP riboswitch ~ pyrithiamine

lysine riboswitch ~ L-aminoethylcysteine, DL-4-oxalysine

FMN riboswitch ~ roseoflavin

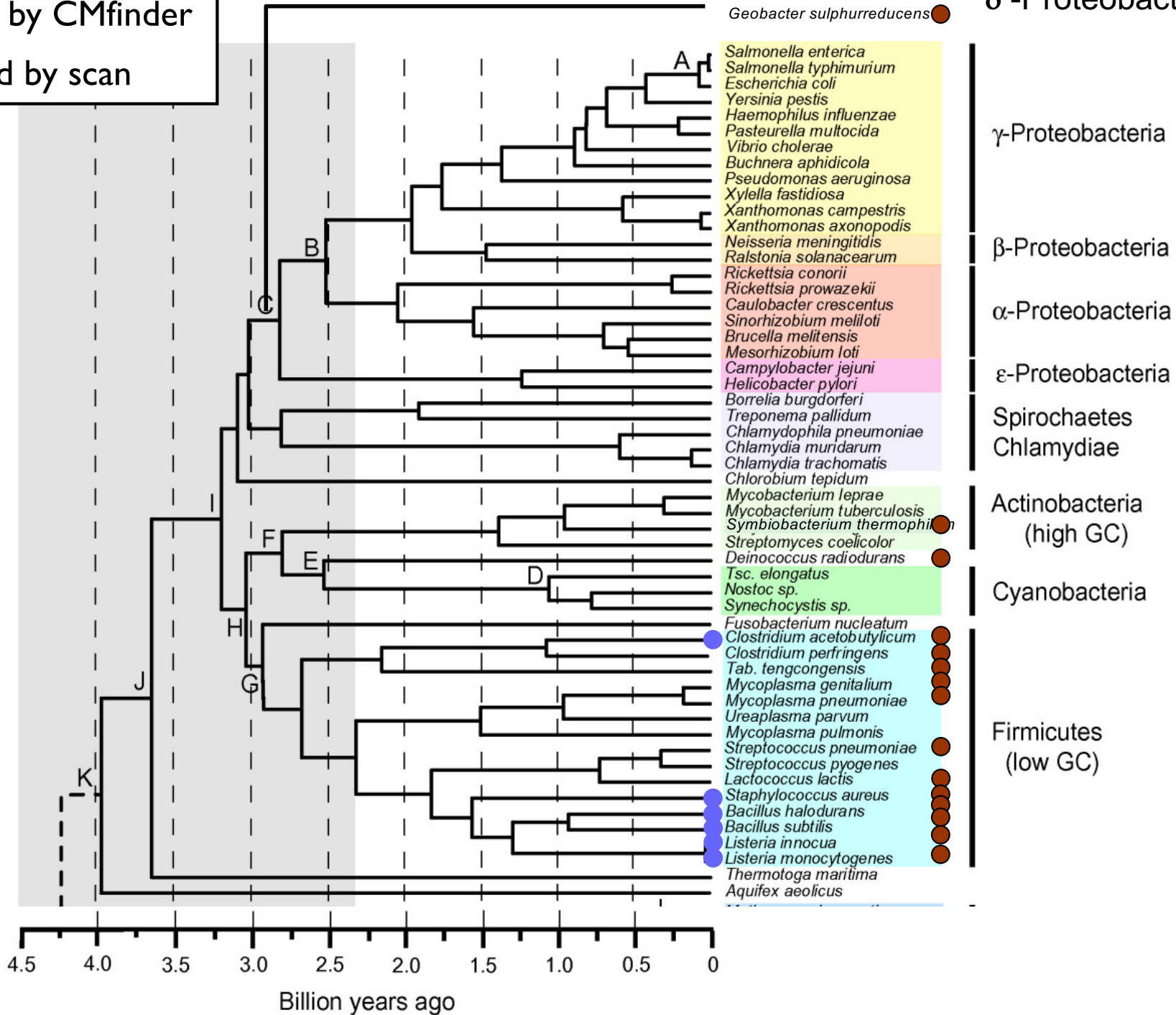
Potential advantages - no (known) human riboswitches, but often multiple copies in bacteria, so potentially efficacious with few side effects?

ncRNA Example: T-boxes



● Used by CMfinder
● Found by scan

Chloroflexus aurantiacus ● Chloroflexi
Geobacter metallireducens ● δ -Proteobacteria
Geobacter sulphurreducens ●



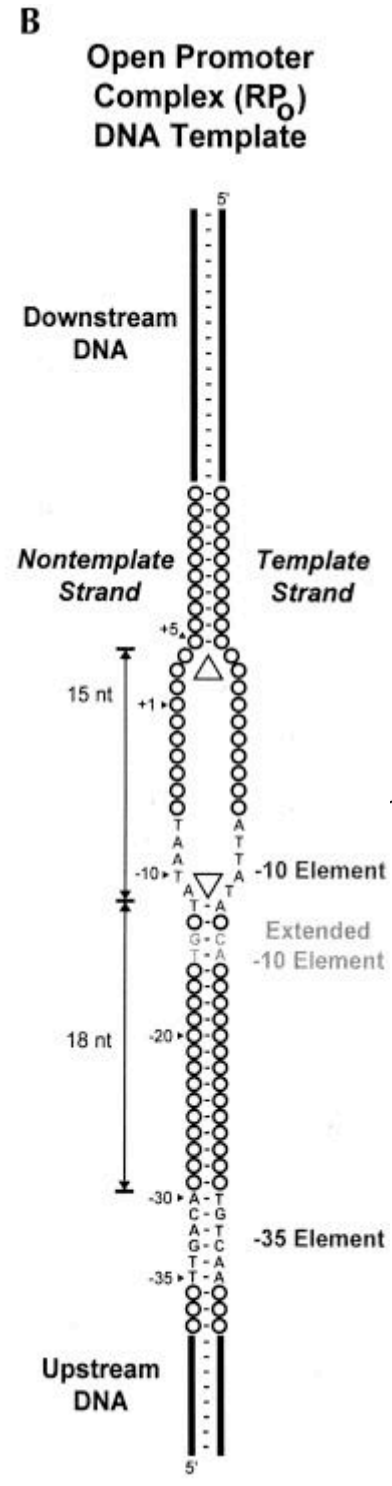
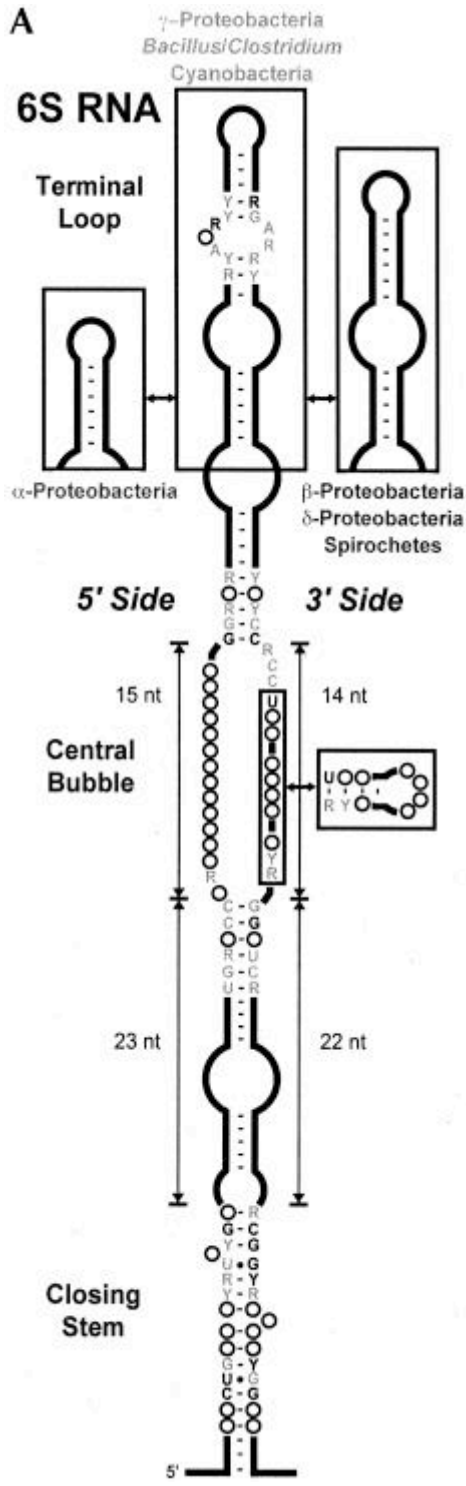
ncRNA Example: 6S

medium size (175nt)

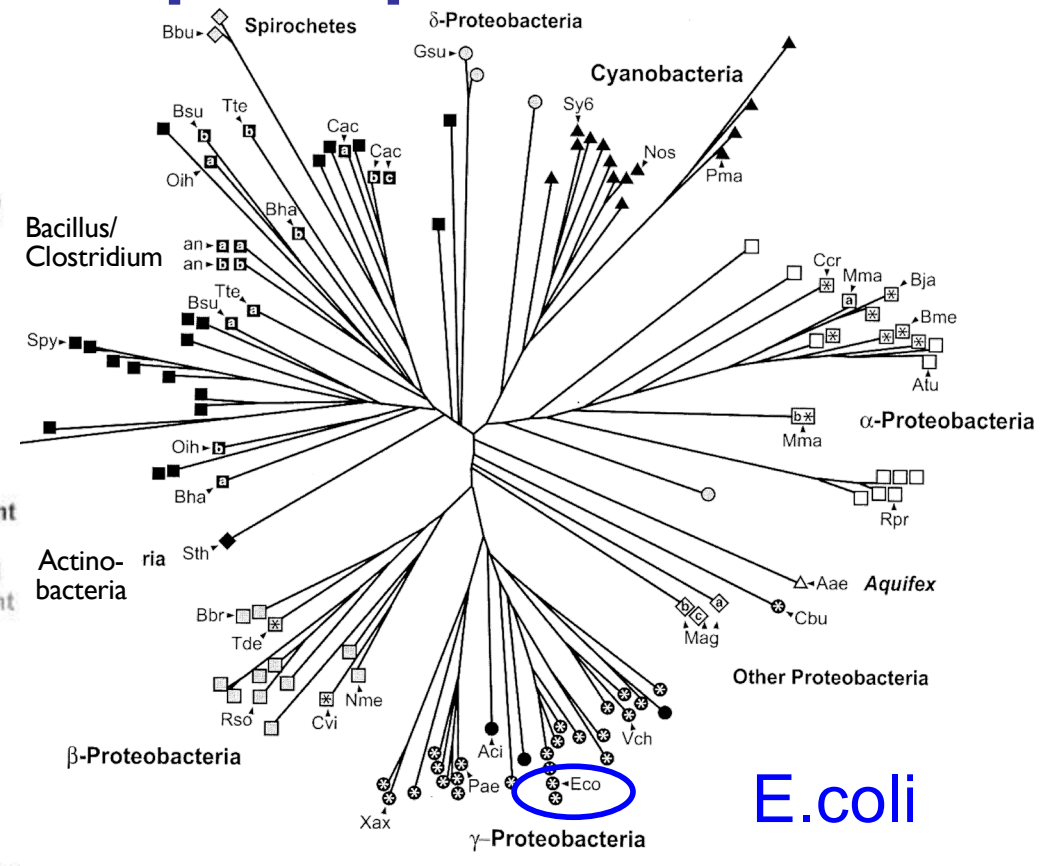
structured

highly expressed in *E. coli* in certain growth conditions

sequenced in 1971; function unknown for 30 years



6S mimics an open promoter



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

Summary: RNA in Bacteria

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

Regulation of MANY genes involves RNA

In some species, we know identities of more riboregulators than protein regulators

Dozens of classes & thousands of new examples in just the last ~10-15 years

Vertebrate ncRNAs

mRNA, tRNA, rRNA, ... of course

PLUS:

snRNA, spliceosome, snoRNA, telomerase,
microRNA, RNAi, SECIS, IRE, piwi-RNA,
XIST (X-inactivation), ribozymes, ...

MicroRNA

1st discovered 1992 in *C. elegans*

2nd discovered 2000, also *C. elegans*

and human, fly, everything between – basically all
multi-celled plants & animals

21-23 nucleotides

literally fell off ends of gels

100s – 1000s now known in human

may regulate 1/3-1/2 of all genes

development, stem cells, cancer, infectious
disease,...

siRNA

2006 Nobel Prize
Fire & Mello

“Short Interfering RNA”

Also discovered in *C. elegans*

Possibly an antiviral defense, shares
machinery with miRNA pathways

Allows artificial repression of most genes in
most higher organisms

Huge tool for biology & biotech

ncRNA Example: Xist

large (\approx 12kb)

largely unstructured RNA

required for X-inactivation in mammals

(Remember calico cats?)

One of many thousands of “Long NonCoding RNAs” (lncRNAs) now recognized, tho most others are of completely unknown significance

Human Predictions

EvoFold

S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, ES Lander, J Kent, W Miller, D Haussler, "Identification and classification of conserved RNA secondary structures in the human genome." [PLoS Comput. Biol., 2, #4 \(2006\) e33.](#)

48,479 candidates (~70% FDR?)

FOLDALIGN

E Torarinsson, M Sawera, JH Havgaard, M Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure." [Genome Res., 16, #7 \(2006\) 885-9.](#)

1800 candidates from 36970 (of 100,000) pairs

RNAz

S Washietl, IL Hofacker, M Lukasser, A Huttenhofer, RF Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." [Nat. Biotechnol., 23, #11 \(2005\) 1307-90.](#)

30,000 structured RNA elements

1,000 conserved across all vertebrates.

~1/3 in introns of known genes, ~1/6 in UTRs

~1/2 located far from any known gene

CMfinder

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. [Genome Research, Feb 2008, 18\(2\):242-251 PMID: 18096747](#)

Seemann, Mirza, Hansen, Bang-Berthelsen, Garde, Christensen-Dalsgaard, Torarinsson, Yao, Workman, Pociot, Nielsen, Tommerup, Ruzzo, Gorodkin. The identification and functional annotation of RNA structures conserved in vertebrates. [Genome Res, Aug 2017, 27\(8\):1371-1383 PMID: 28487280.](#)

Thousands of Predictions

Bottom line?

A significant number of “one-off” examples

Extremely wide-spread ncRNA expression

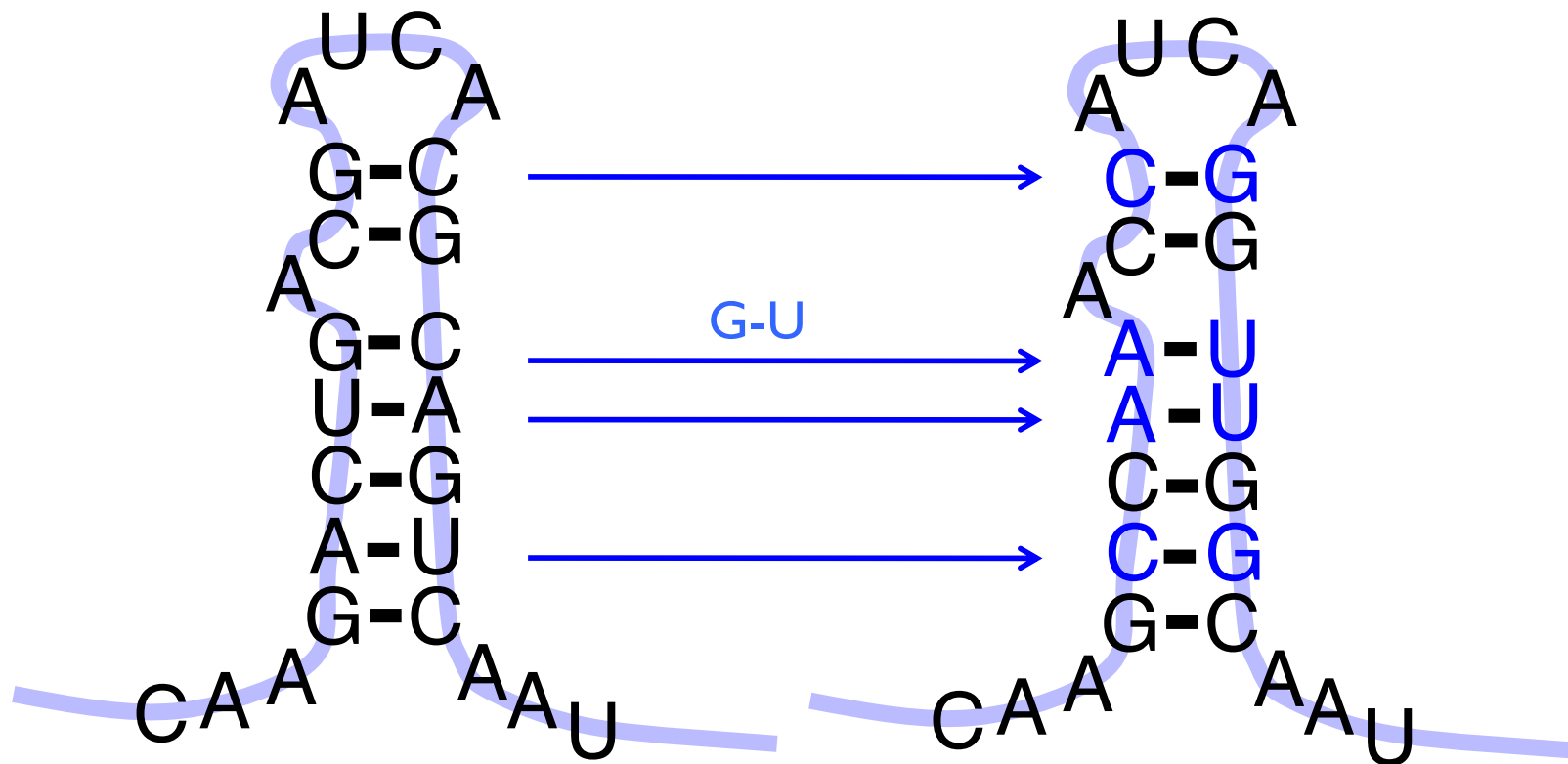
At a minimum, a vast evolutionary substrate

New technology (e.g., RNAseq) exposing
more

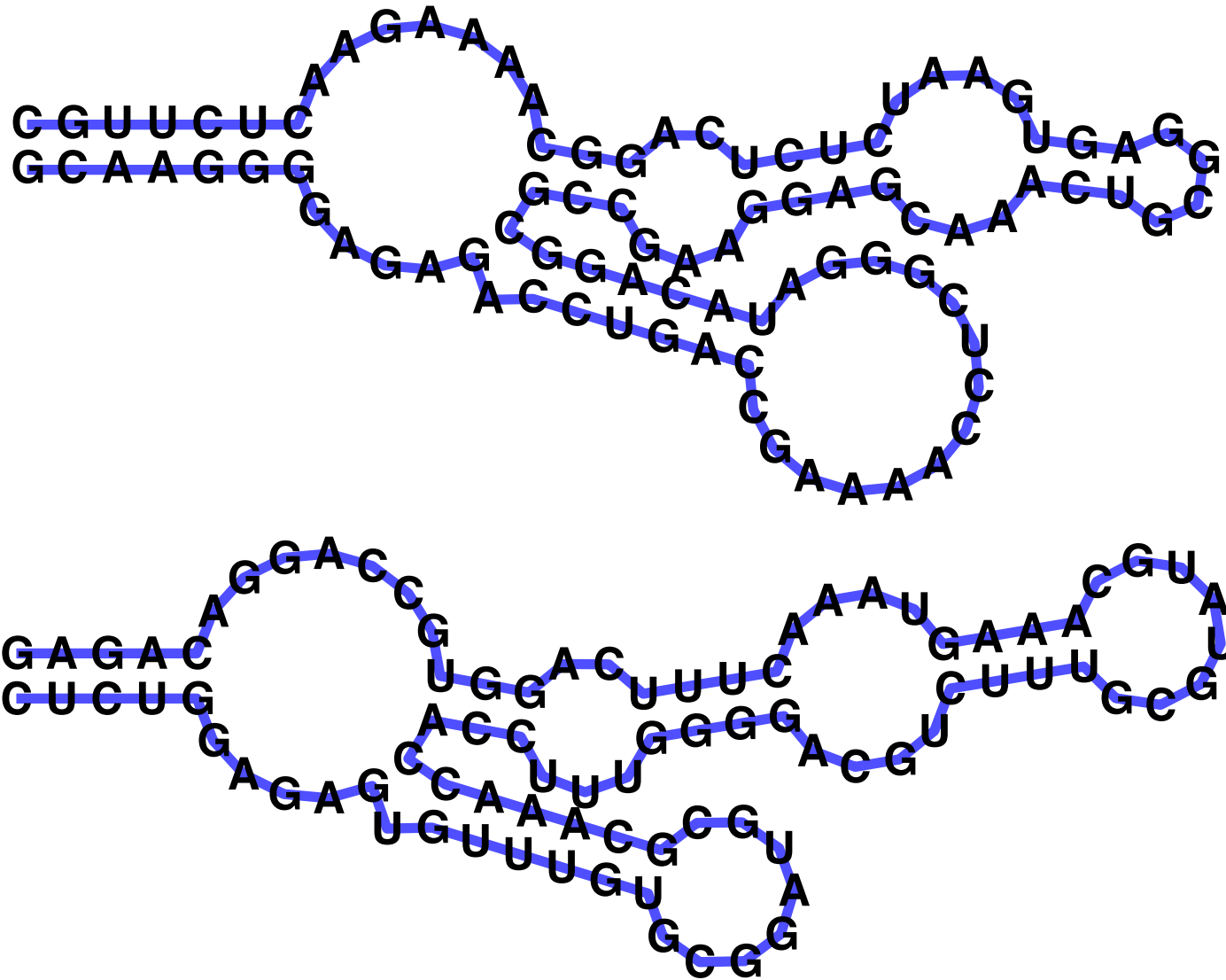
How do you recognize an interesting one?

A Clue: Conserved secondary structure

RNA Secondary Structure: can be fixed while sequence evolves



Why is RNA hard to deal with?



A: *Structure* often more important than *sequence*₃₅

Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

RNA Pairing

Watson-Crick Pairing

C - G

~ 3 kcal/mole

A - U

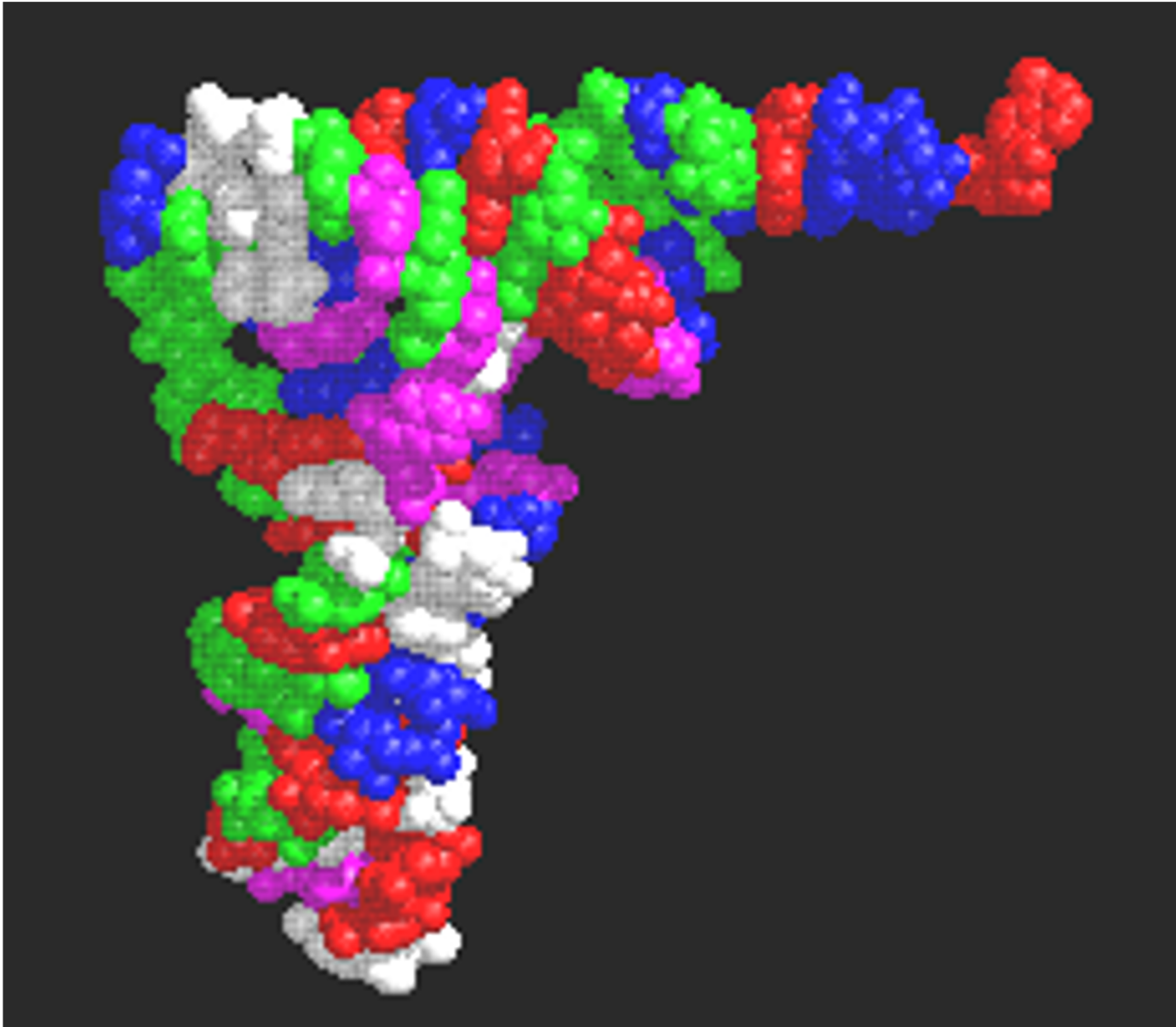
~ 2 kcal/mole

“Wobble Pair” G - U

~1 kcal/mole

Non-canonical Pairs (esp. if modified)

tRNA 3d Structure



tRNA - Alt. Representations

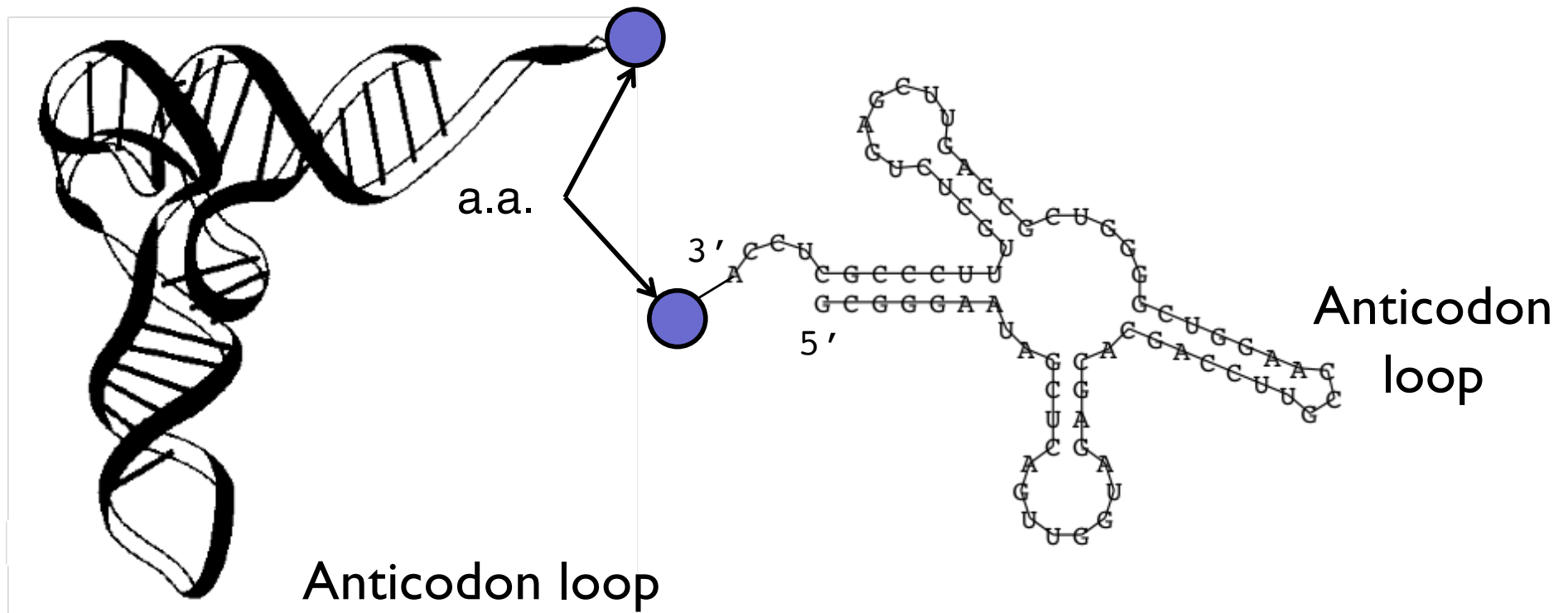
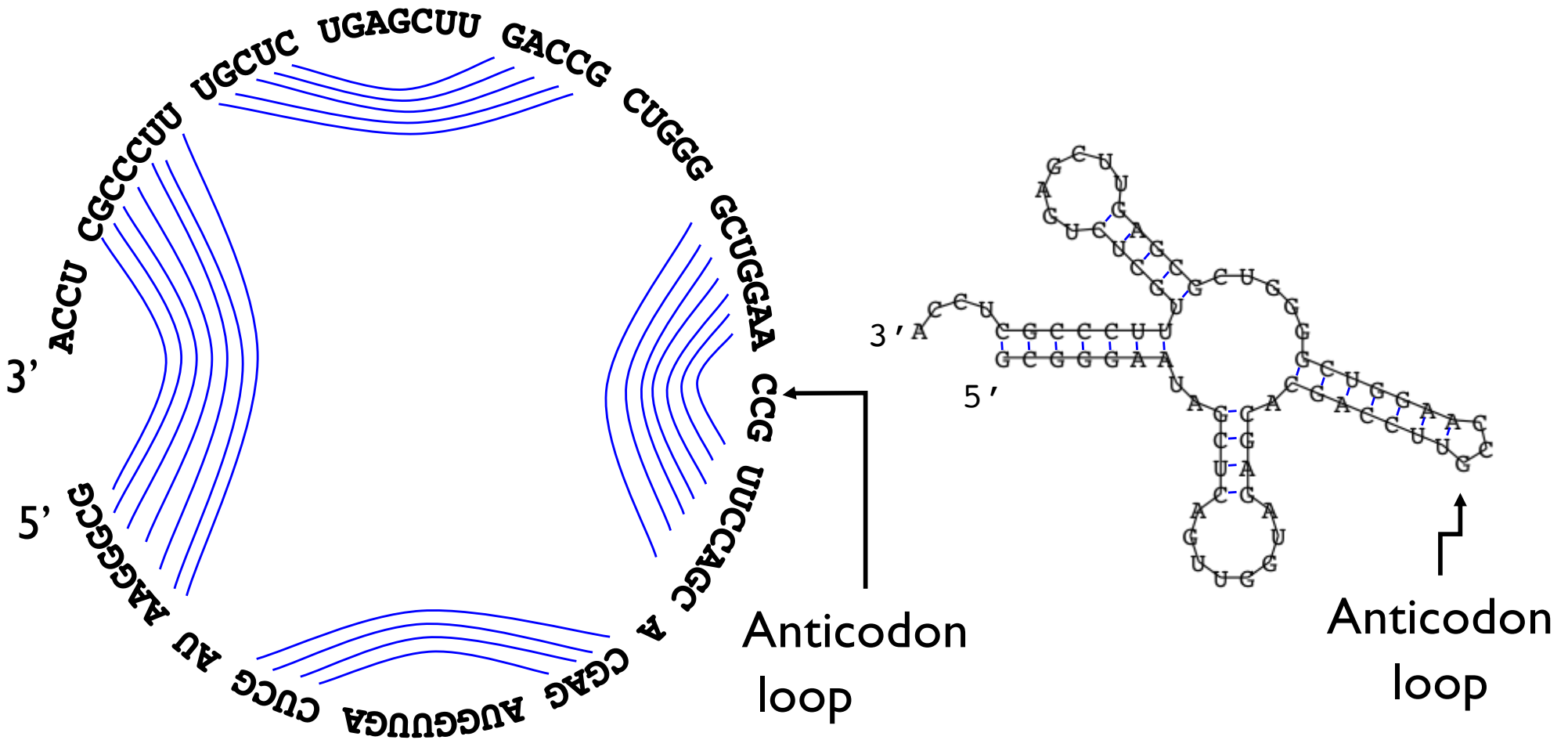


Figure 1: a) The spatial structure of the phenylalanine tRNA from yeast

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

tRNA - Alt. Representations



Definitions

Sequence $5' r_1 r_2 r_3 \dots r_n 3'$ in $\{A, C, G, T/U\}$

A **Secondary Structure** is a set of pairs $i \bullet j$ s.t.

$i < j-4$, and } no sharp turns

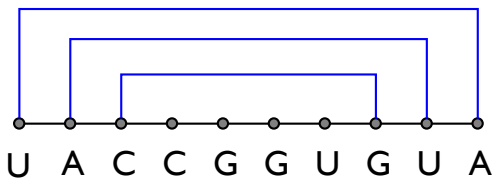
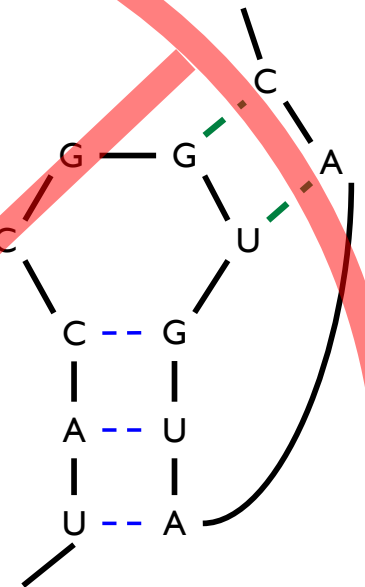
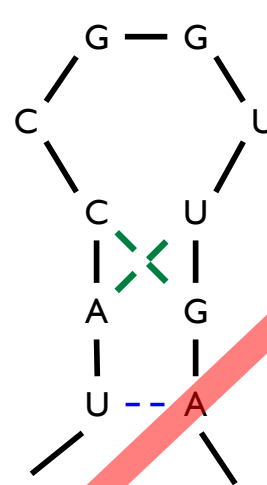
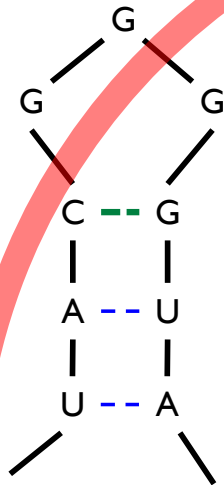
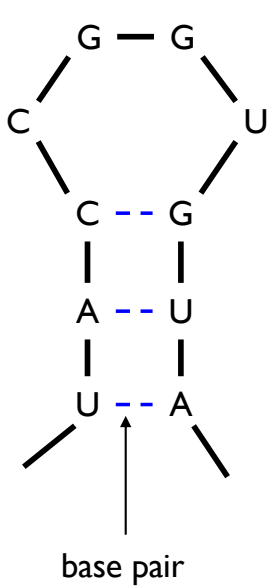
if $i \bullet j$ & $i' \bullet j'$ are two different pairs with $i \leq i'$, then

$j < i'$, or

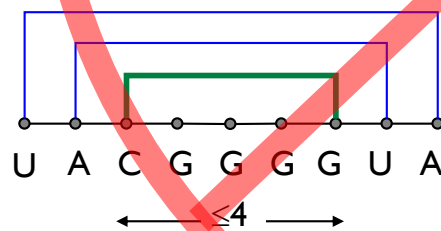
$i < i' < j' < j$

} 2nd pair follows 1st, or is nested within it;
no “pseudoknots”
And pairs, not triples, etc.

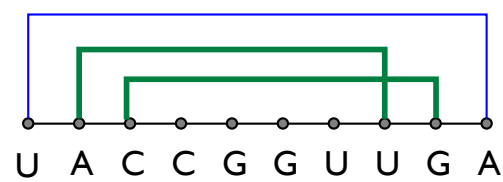
RNA Secondary Structure: Examples



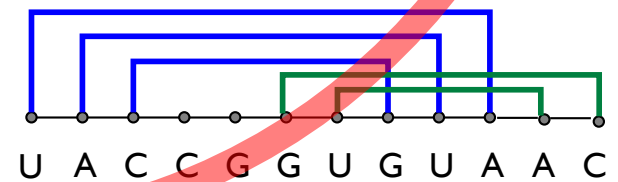
ok



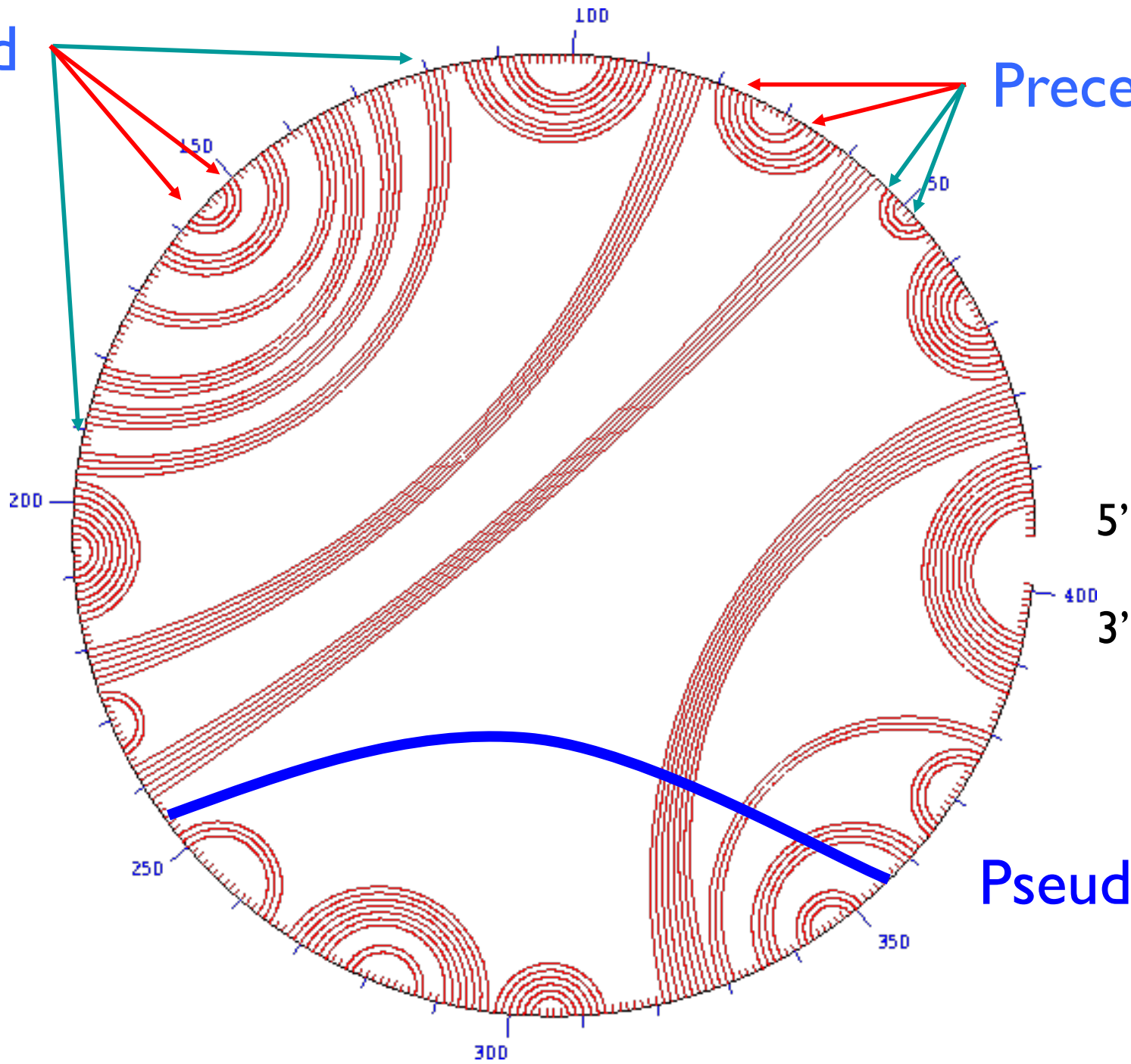
sharp turn



crossing



Nested



Precedes

5'
3'

Pseudoknot

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

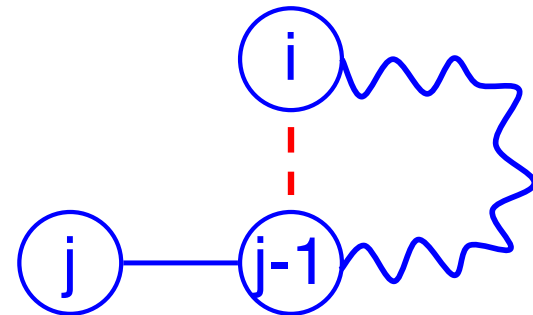
- + finds all folds
- ignores pseudoknots

“Optimal pairing of $r_i \dots r_j$ ”

Two possibilities

j Unpaired:

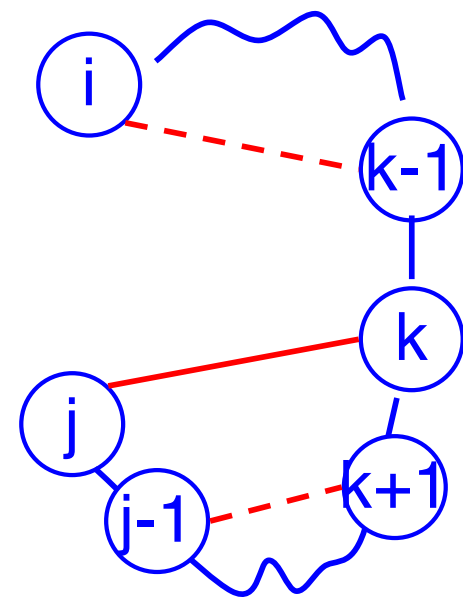
Find best pairing of $r_i \dots r_{j-1}$



j Paired (with some k):

Find best $r_i \dots r_{k-1}$ +

best $r_{k+1} \dots r_{j-1}$ **plus 1**



Why is it slow?

Why do pseudoknots matter?

Nussinov: Max Pairing

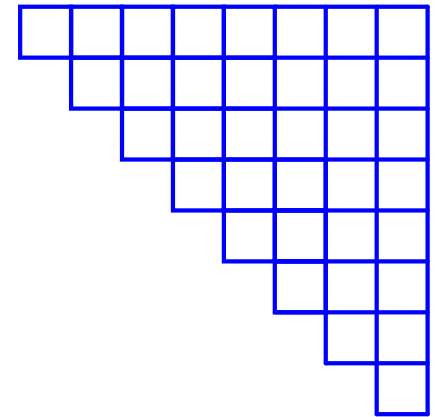
$B(i,j) = \#$ pairs in optimal pairing of $r_i \dots r_j$

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

Otherwise

$B(i,j) = \max$ of:

$$\left\{ \begin{array}{l} B(i,j-1) \\ \max \{ B(i,k-1) + 1 + B(k+1,j-1) \mid \\ \quad i \leq k < j-4 \text{ and } r_k - r_j \text{ may pair} \} \end{array} \right.$$



Nussinov:

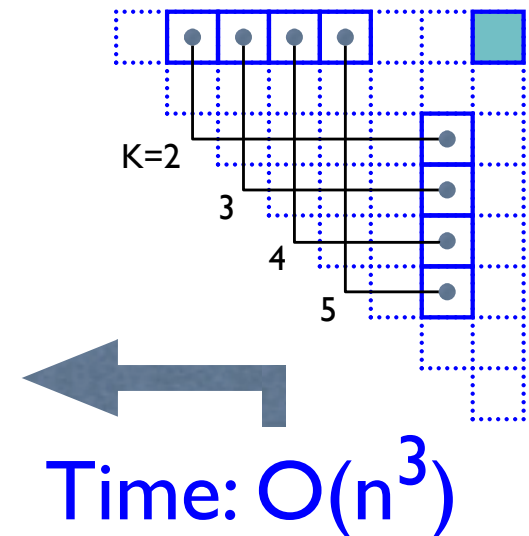
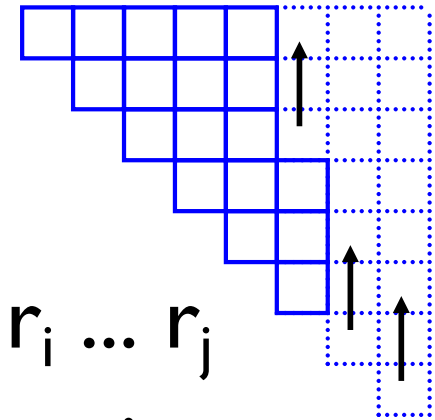
A Computation Order

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

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

$B(i,j) = \max$ of:

$$\left\{ \begin{array}{l} B(i,j-1) \\ \max \{ B(i,k-1) + 1 + B(k+1,j-1) \mid \\ \quad i \leq k < j-4 \text{ and } r_k - r_j \text{ may pair} \} \end{array} \right.$$



Which Pairs?

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

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	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	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
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

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

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

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

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	0	0	0	0	0	0	0	0	0	0	0

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



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

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

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



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

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

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

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

GGAAAACCCAAAGGGGU
 $\dots(\dots(((\dots))))$

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

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

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

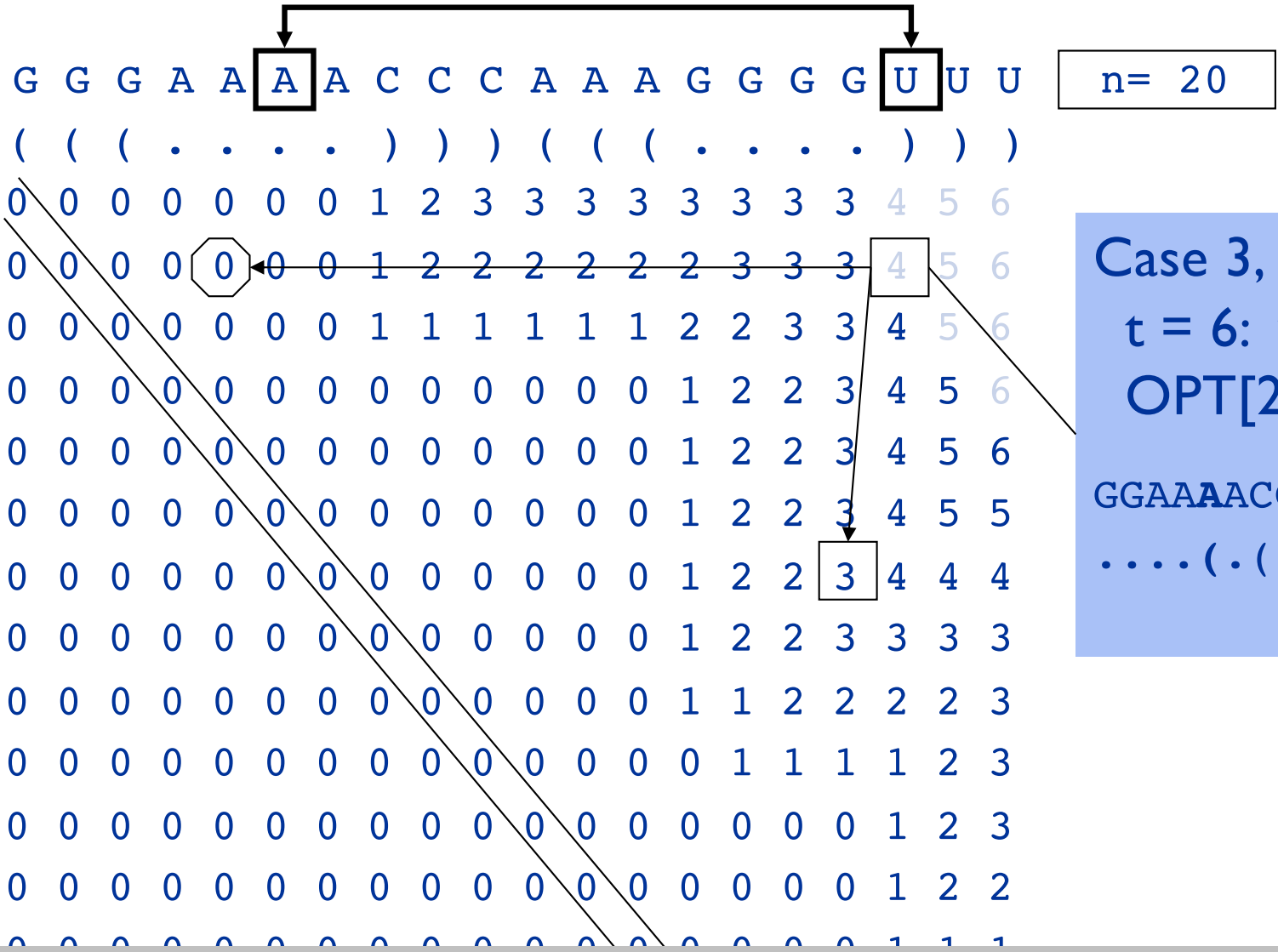
Case 3, $2 \leq t < 18-4$:
 $t = 5$: yes pair
 $OPT[2,18] \geq 1+0+3$

GGAAACCCAAAGGGGU
 $\dots(\dots(((\dots))))$

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

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

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



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

GGAAACCCAAAGGGGU
 $\dots \cdot (\cdot ((\cdot \dots))))$

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

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	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	1	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	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	

Case 3, $2 \leq t < 18-4$:
 $t = 7$: yes pair
 $OPT[2,18] \geq 1+0+3$

GGAAA**ACC**CAAAGGGGU
((((.....))))

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

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

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

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

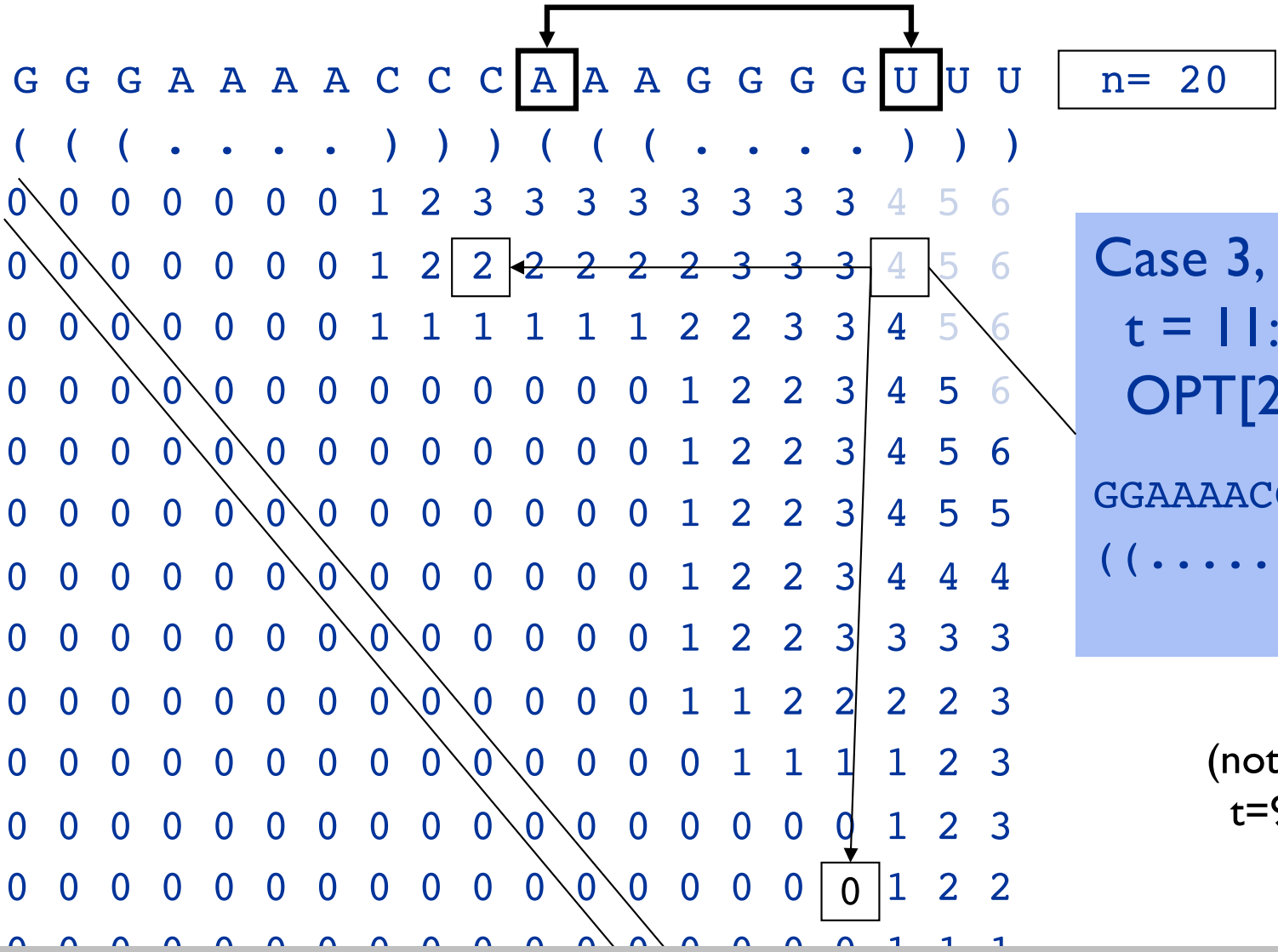
0	0	0	0	0	0	0	1	2	3	3	3	3	3	3	3	3	4	5	6
0	0	0	0	0	0	0	1	2	2	2	2	2	2	3	3	3	4	5	6
0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	3	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	6
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	5	5
0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	4	4	4
0	0	0	0	0	0	0	0	0	0	0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

Case 3, $2 \leq t < 18-4$:
 $t = 8$: no pair

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

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

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



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

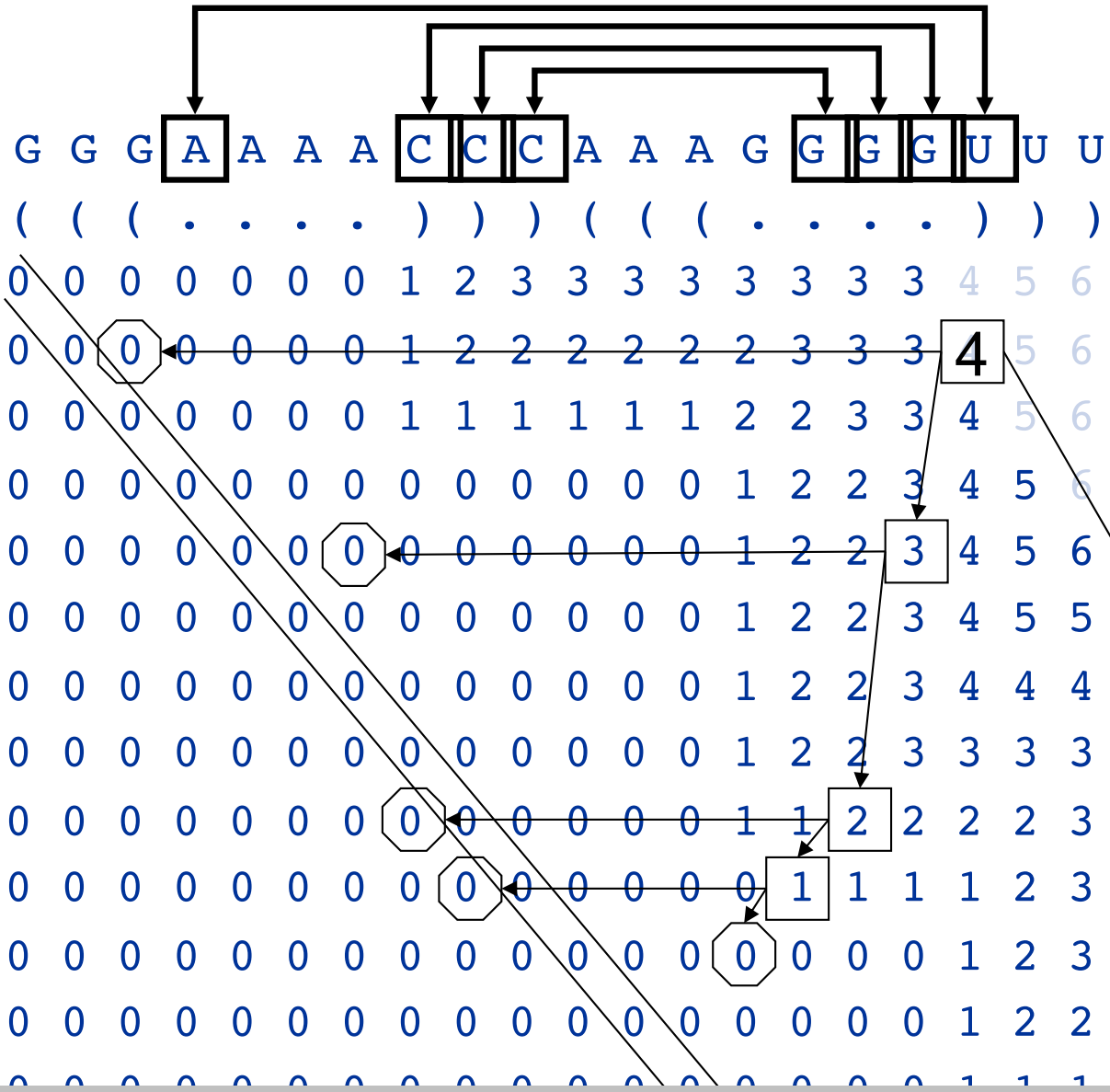
GGAAAACCC**AA**AGGGGU
 ((.....))(.....)

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

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

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

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

All 5 optimal structures on the above example

GGGAAAACCCAAAGGGGUUU

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

```
 0 0 0 0 0 0 0 1 2 3 3 3 3 3 3 3 3 3 4 5 6
-7 0 0 0 0 0 0 1 2 2 2 2 2 2 3 3 3 3 4 5 6
-7-7 0 0 0 0 0 1 1 1 1 1 1 2 2 3 3 4 5 6
-7-7-7 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
-7-7-7-7 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 6
-7-7-7-7-7 0 0 0 0 0 0 0 0 0 1 2 2 3 4 5 5
-7-7-7-7-7-7 0 0 0 0 0 0 0 0 1 2 2 3 4 4 4
-7-7-7-7-7-7-7 0 0 0 0 0 0 0 1 2 2 3 3 3 3
-7-7-7-7-7-7-7-7 0 0 0 0 0 1 1 2 2 2 2 3
-7-7-7-7-7-7-7-7-7 0 0 0 0 0 1 1 1 1 2 3
-7-7-7-7-7-7-7-7-7-7 0 0 0 0 0 0 0 1 2 3
-7-7-7-7-7-7-7-7-7-7-7 0 0 0 0 0 0 1 2 2
-7-7-7-7-7-7-7-7-7-7-7-7 0 0 0 0 0 1 1 1
-7-7-7-7-7-7-7-7-7-7-7-7-7 0 0 0 0 0 0
-7-7-7-7-7-7-7-7-7-7-7-7-7-7 0 0 0 0 0
-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7 0 0 0 0
-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7 0 0 0
-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7 0 0
-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7-7 0
n= 20 Pairs= 6 AltStructs= 5 0.000117 (sec. total)
```

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


Pair-based Energy Minimization


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

$E(i,j) = \infty$ 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 $k-j$ pair

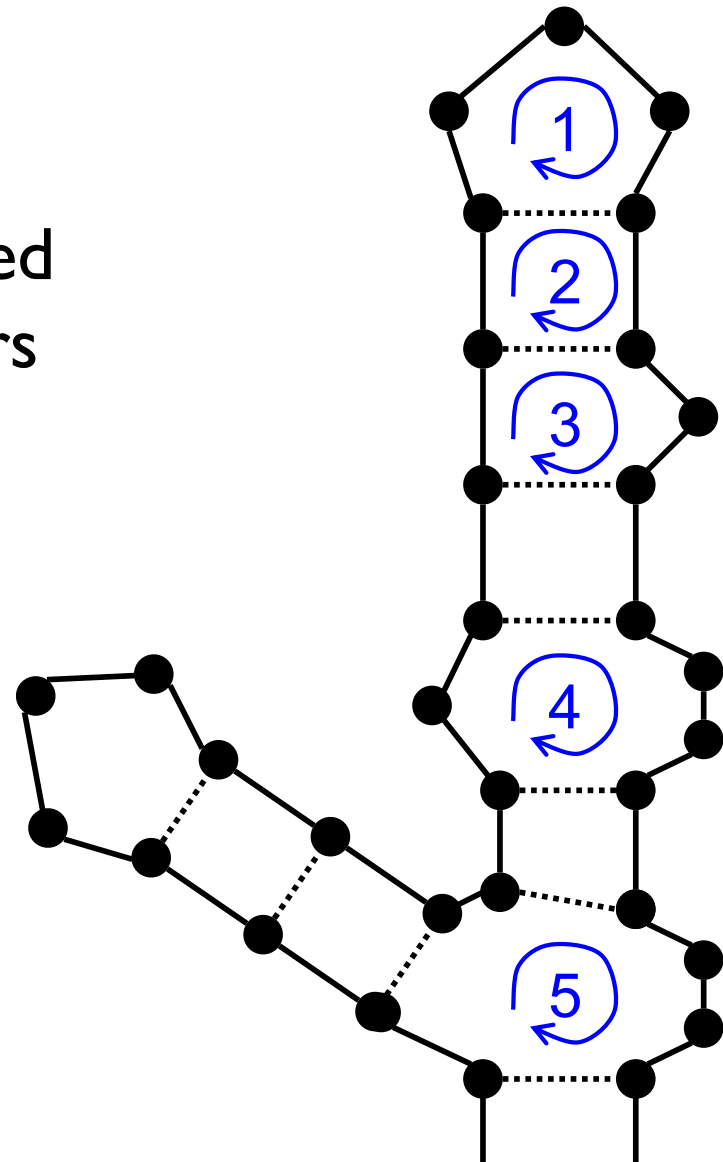
Time: $O(n^3)$ 

Loop-based Energy Minimization

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

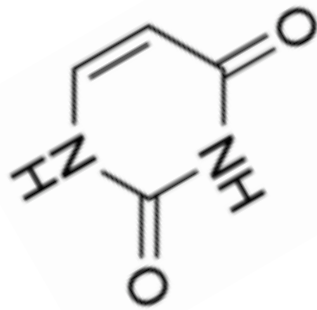
Loop types

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

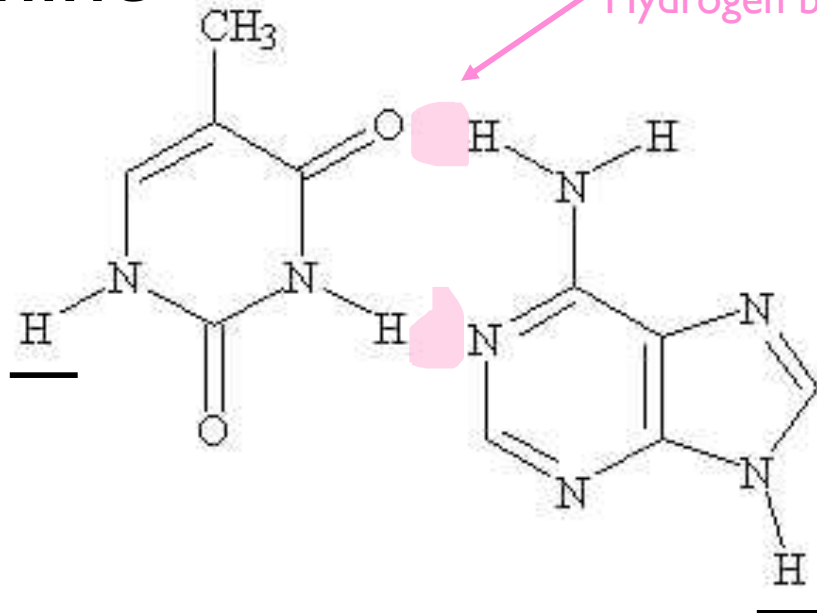


Base Pairs and Stacking

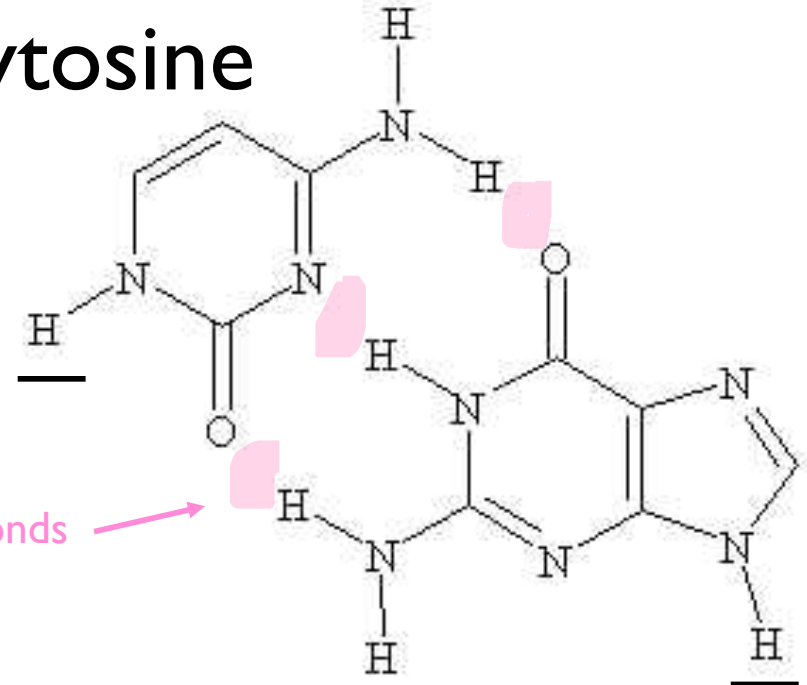
uracil



thymine



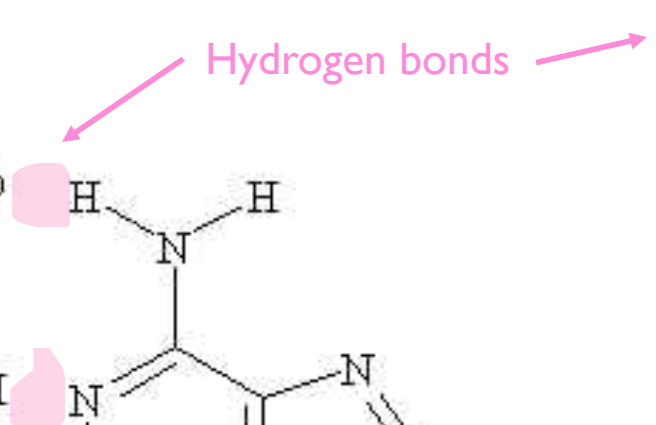
cytosine



guanine

adenine

Hydrogen bonds



Zuker: Loop-based Energy, I

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

$V(i,j)$ = as above, but forcing pair $i \bullet j$

$W(i,j) = V(i,j) = \infty$ for all i, j with $i \geq j-4$

$W(i,j) = \min(W(i,j-1),$
 $\min \{ W(i,k-1) + V(k,j) \mid i \leq k < j-4 \}$
 $)$

Zuker: Loop-based Energy, II

hairpin stack bulge/ multi-
interior loop

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

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

$$\text{VBI}(i,j) = \min \{ \text{ebi}(i,j,i',j') + V(i',j') \mid$$

$$i < i' < j' < j \ \& \ i'-i+j-j' > 2 \}$$

bulge/
interior



Time: $O(n^4)$



$O(n^3)$ possible if $\text{ebi}(\cdot)$ is “nice”

Energy Parameters

Q. Where do they come from?

A1. Experiments with carefully selected synthetic RNAs

A2. Learned algorithmically from trusted alignments/structures [Andronescu et al., 2007]

Single Seq Prediction Accuracy

Mfold, Vienna,... [Nussinov, Zuker, Hofacker, McCaskill]

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

Definitely useful, but obviously imperfect

Approaches, II

Comparative sequence analysis

- + handles all pairings (potentially 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 crystallography, 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 from sequence alone due, e.g., to codon structure, which RNAs lack

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

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