CSE 428
Computational Biology Capstone

A Quick Tour of RNA: Function & Secondary Structure Prediction

The Message

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

Functionally important, functionally diverse

Structurally complex

New tools required

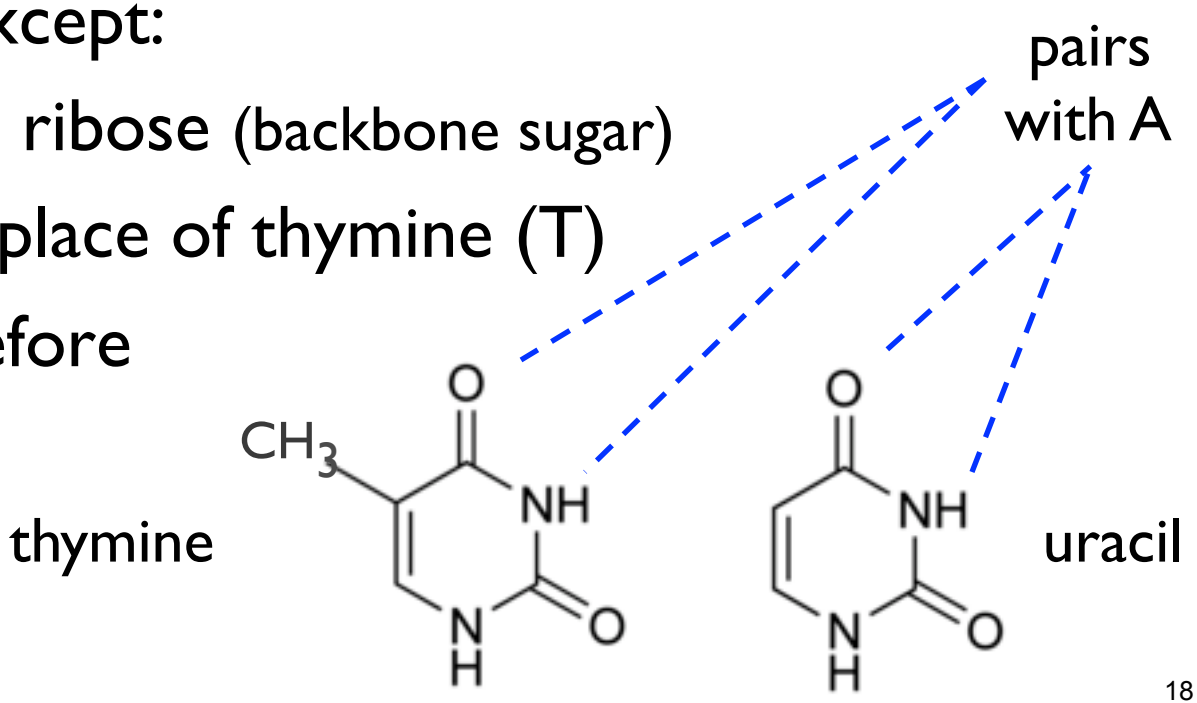
alignment, discovery, search, scoring, etc.

RNA

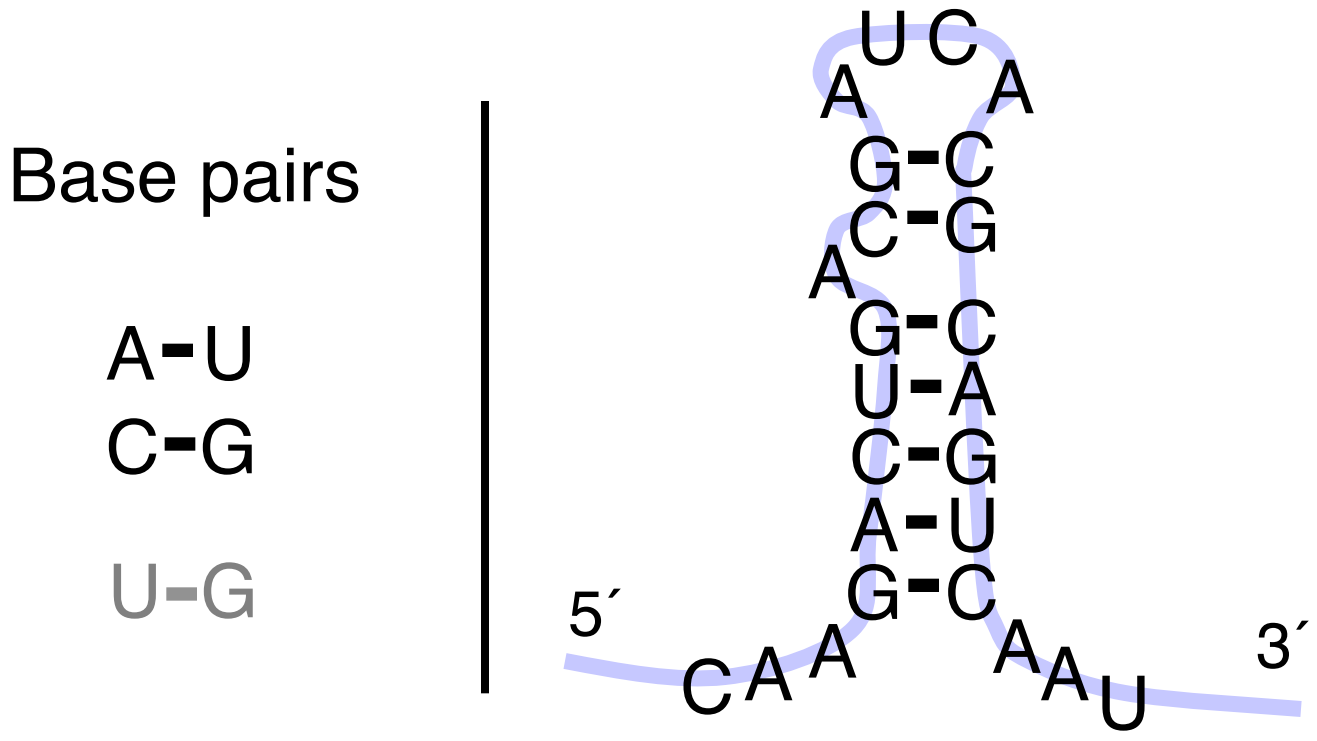
DNA: DeoxyriboNucleic Acid

RNA: RiboNucleic Acid

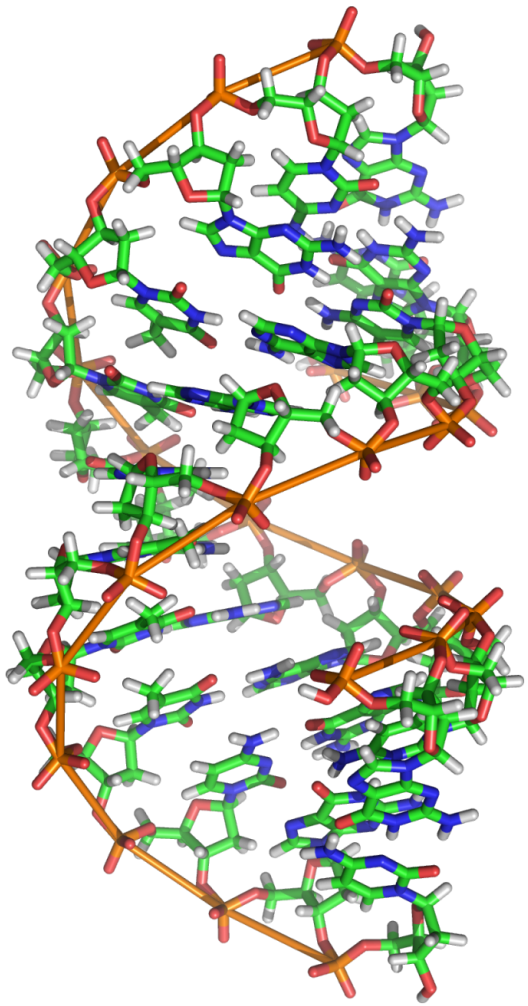
- Like DNA, except:
- Lacks OH on ribose (backbone sugar)
- Uracil (U) in place of thymine (T)
- A, G, C as before



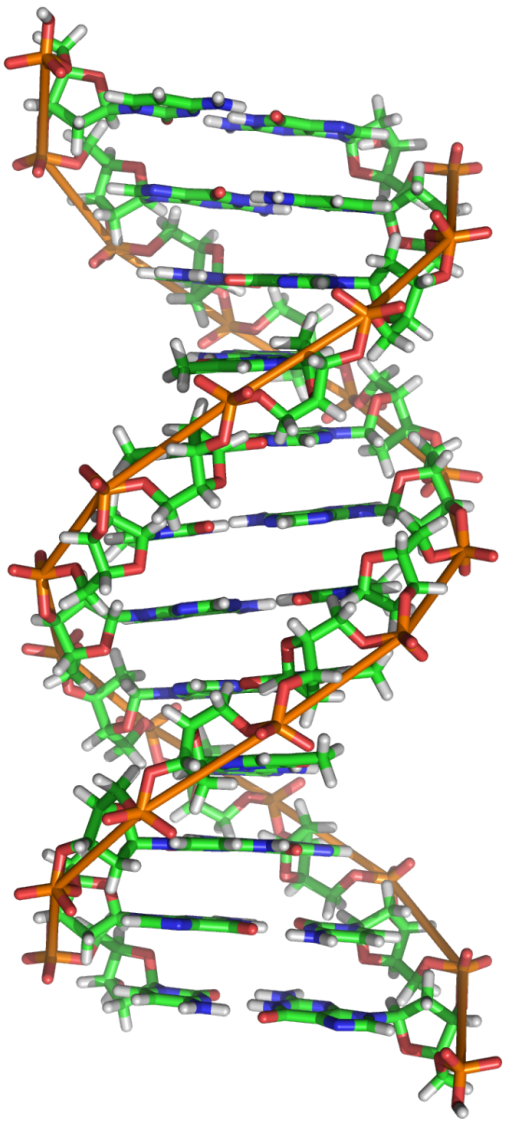
RNA Secondary Structure:
RNA makes helices too



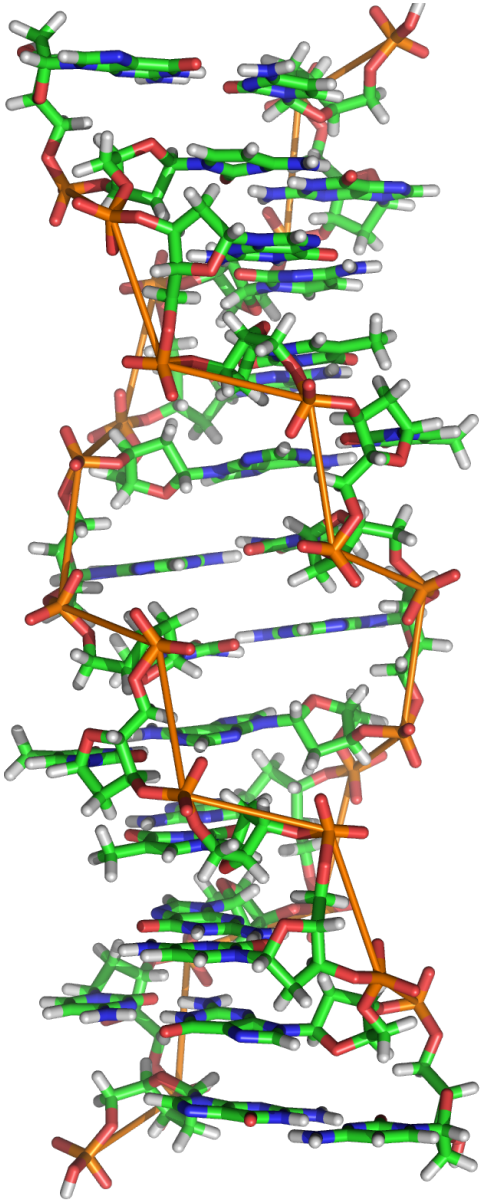
Usually *single* stranded



A
(norm for RNA)



B
(norm for DNA)



Z

http://en.wikipedia.org/wiki/File:A-DNA,_B-DNA_and_Z-DNA.png

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."

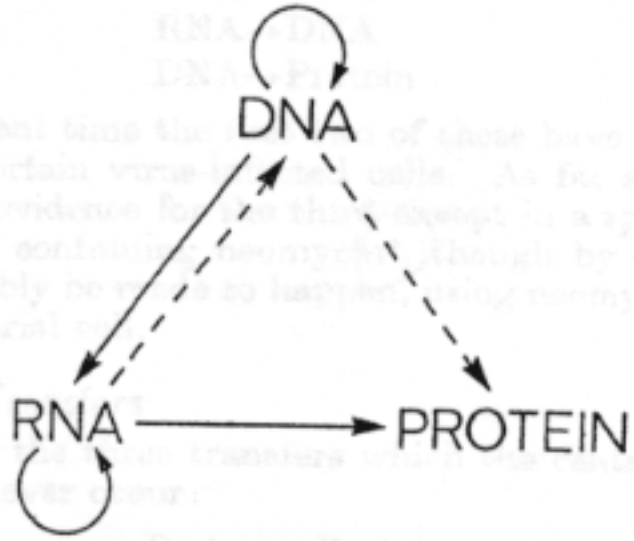


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

Ribosomes

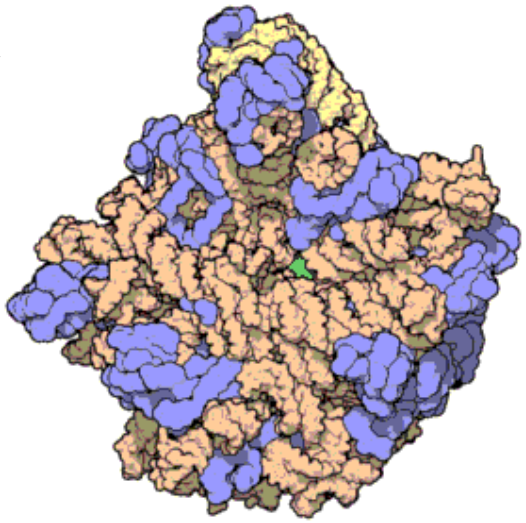
1974 Nobel prize to Romanian biologist George Palade (1912-2008) for discovery in mid 50's

50-80 proteins

3-4 RNAs (half the mass)

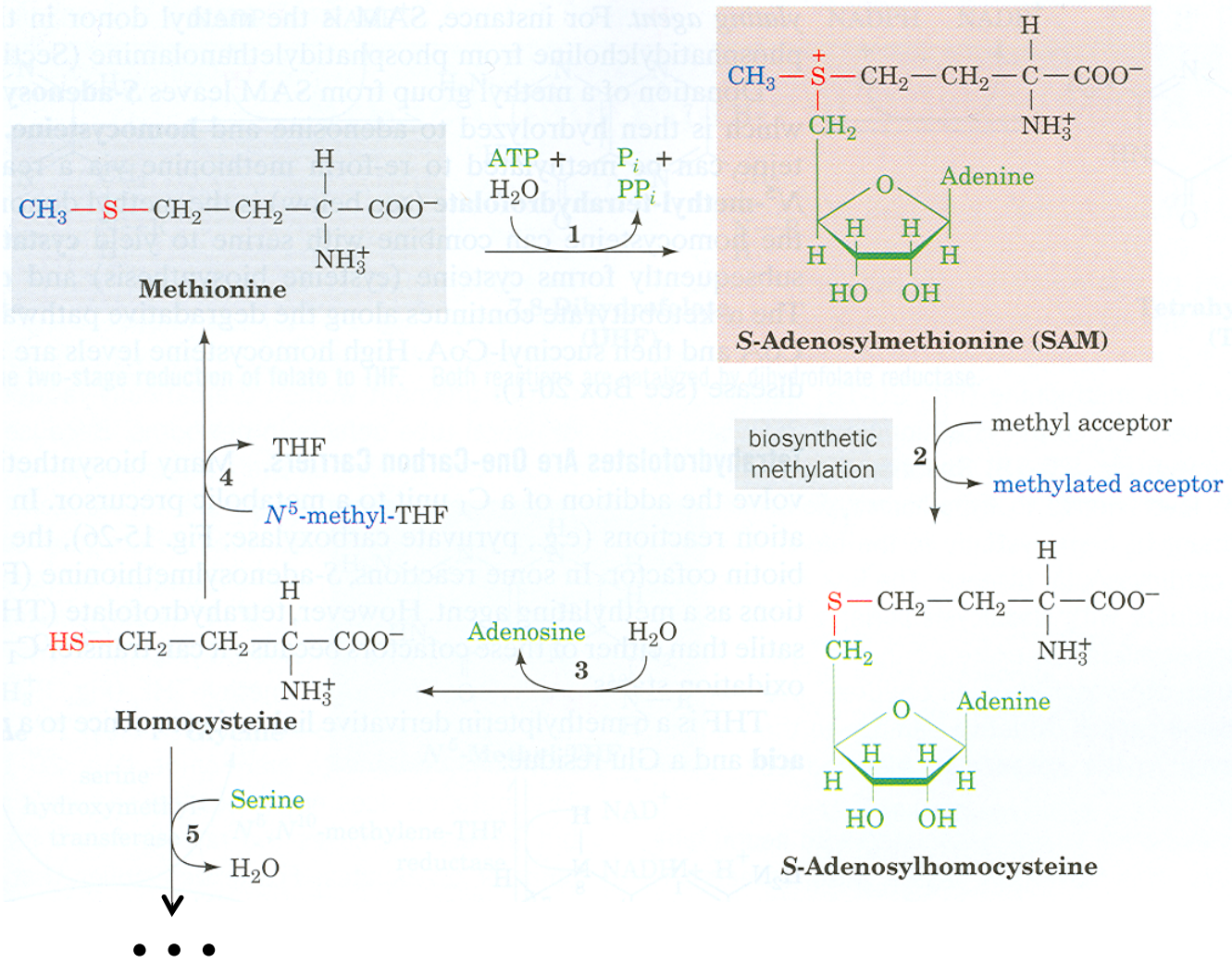
Catalytic core is RNA

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



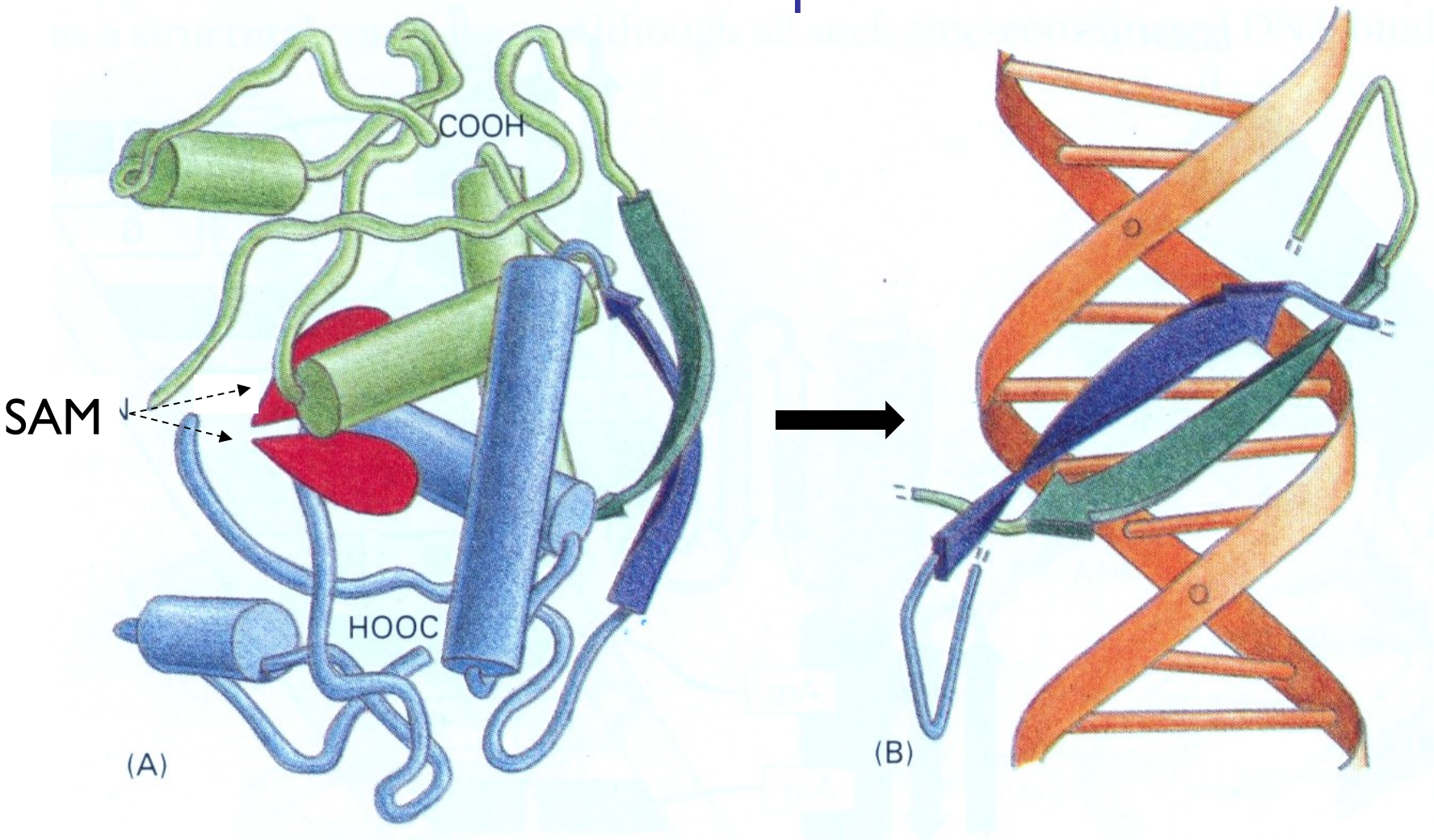
Atomic structure of the 50S Subunit from *Haloarcula marismortui*. Proteins are shown in blue and the two RNA strands in orange and yellow. The small patch of green in the center of the subunit is the active site.
- Wikipedia

Proteins Catalyze Biochemistry: Met Pathways



Proteins Regulate Biochemistry:

The MET Repressor



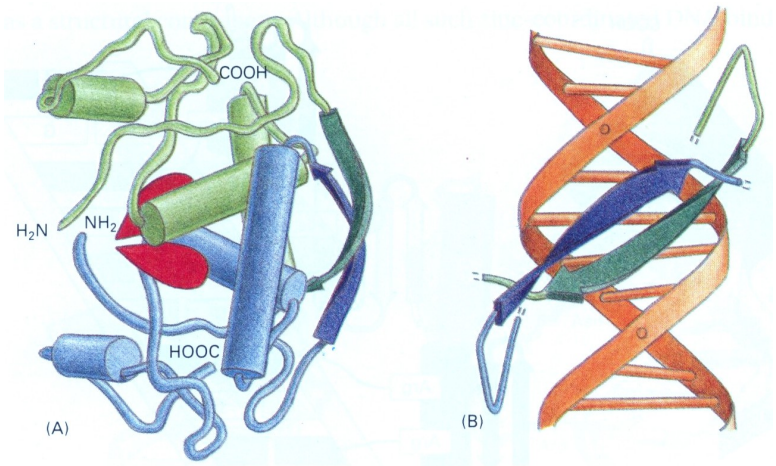
Protein

Alberts, et al, 3e.

DNA

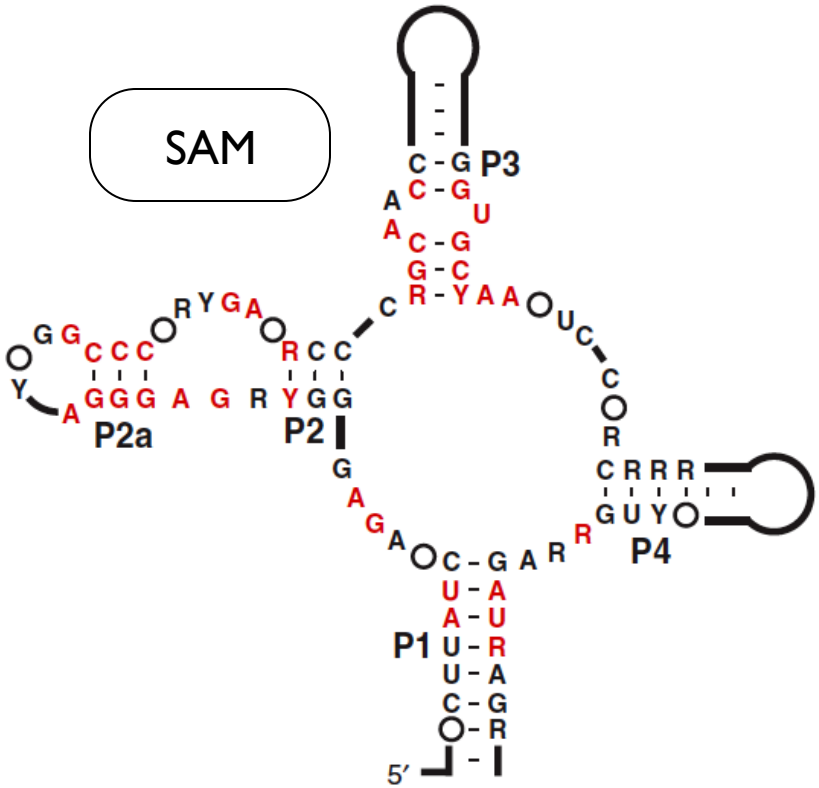
Not the only way!

Alberts, et al, 3e.



Protein way

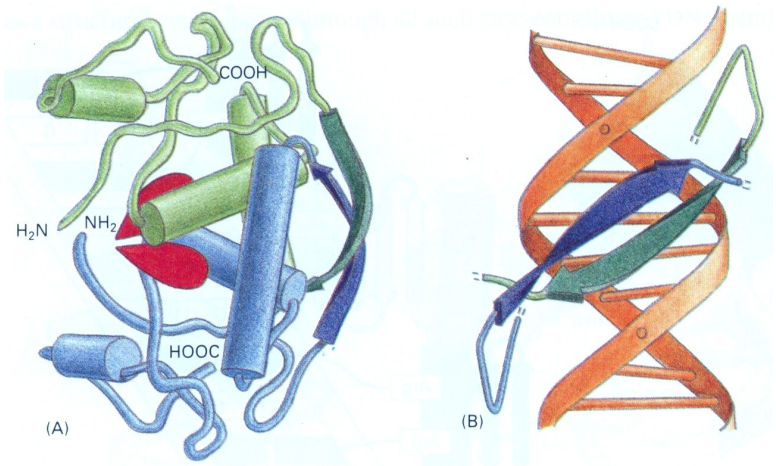
Riboswitch alternative



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

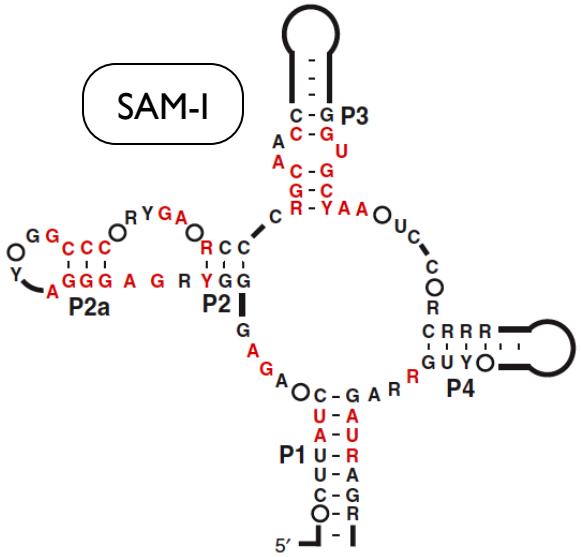
Not the only way!

Alberts, et al, 3e.

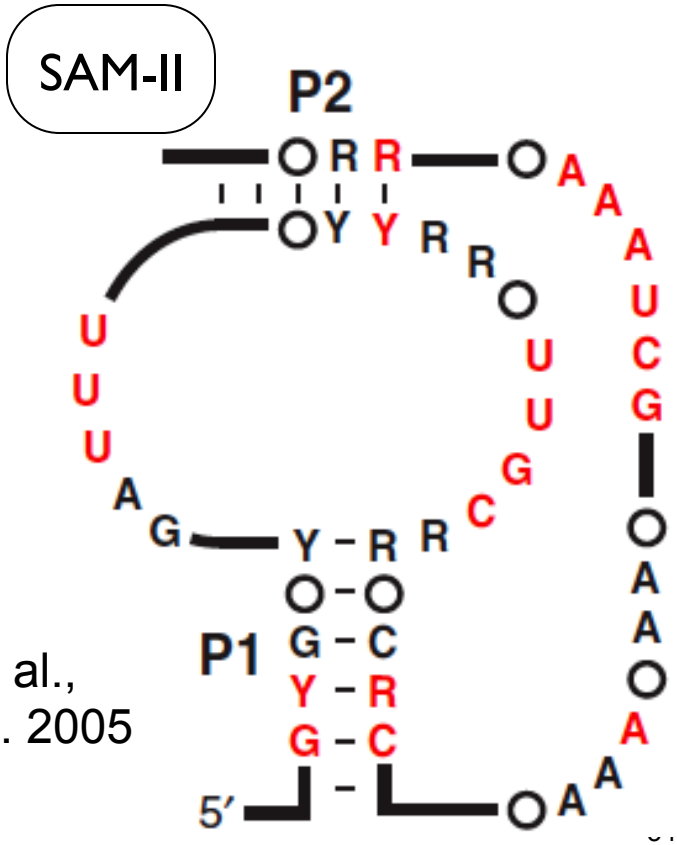


Protein way

Riboswitch alternatives



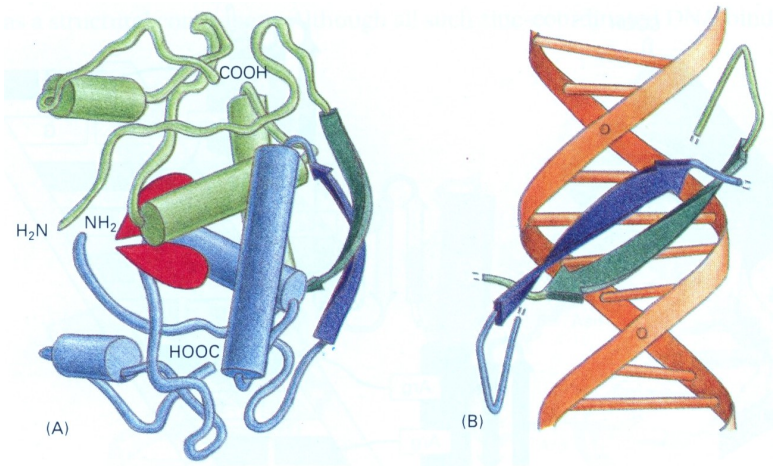
Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

Not the only way!

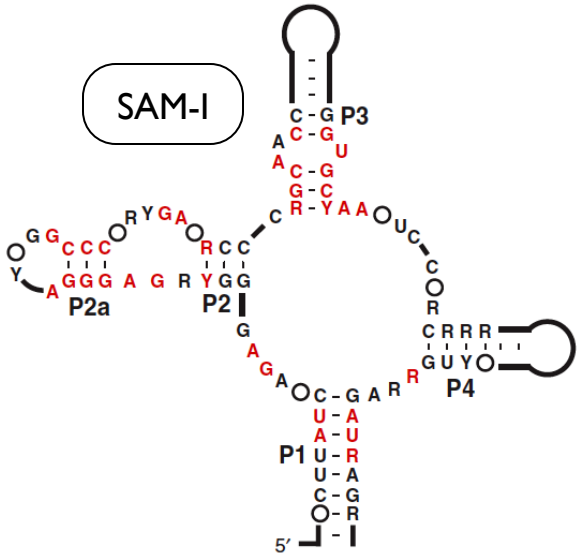
Alberts, et al, 3e.



Protein way

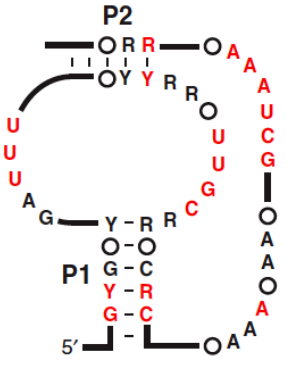
Riboswitch alternatives

SAM-I



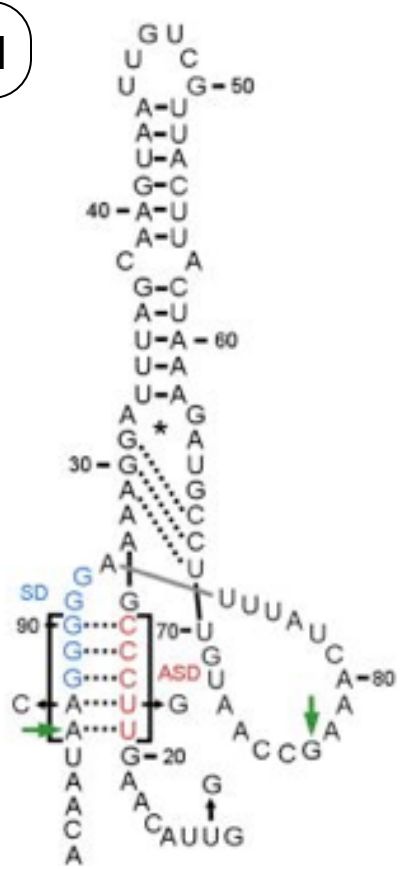
Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



Corbino et al., Genome Biol. 2005

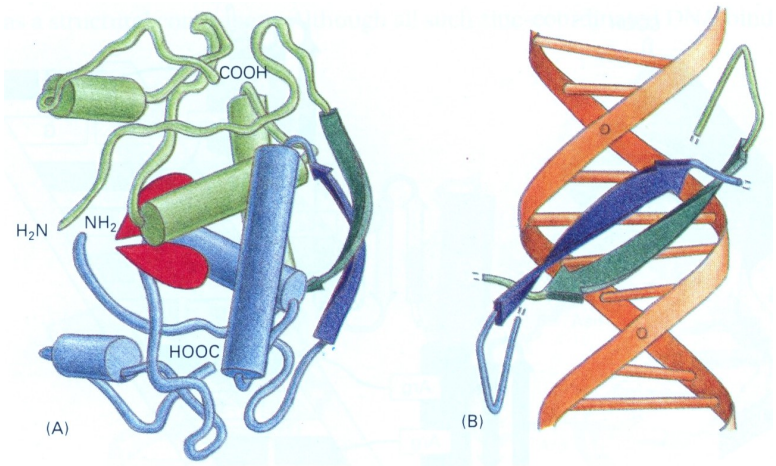
SAM-III



Fuchs et al., NSMB 2006

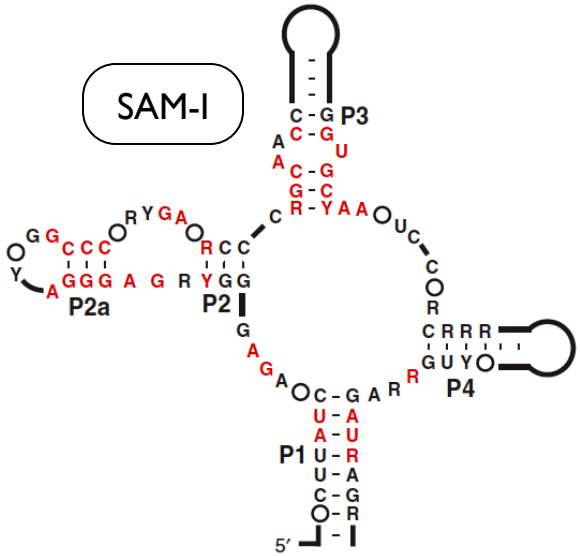
Not the only way!

Alberts, et al, 3e.

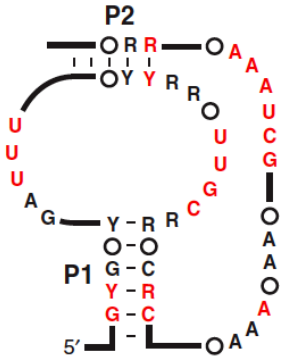


← Protein way

Riboswitch alternatives ↓

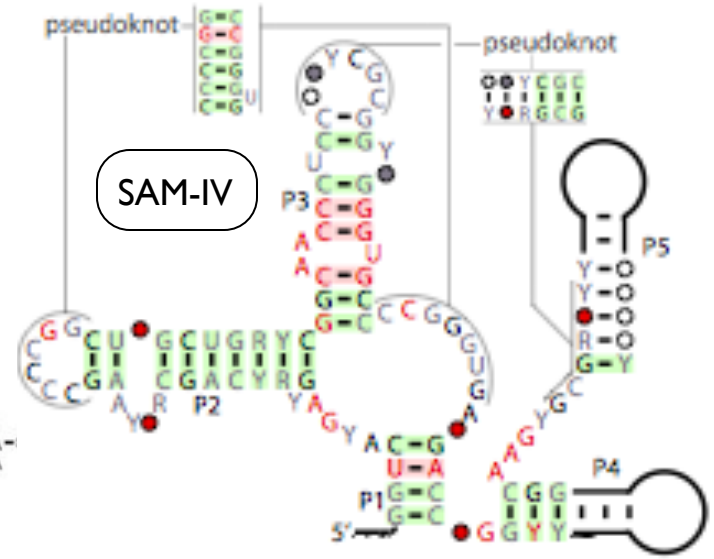


Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

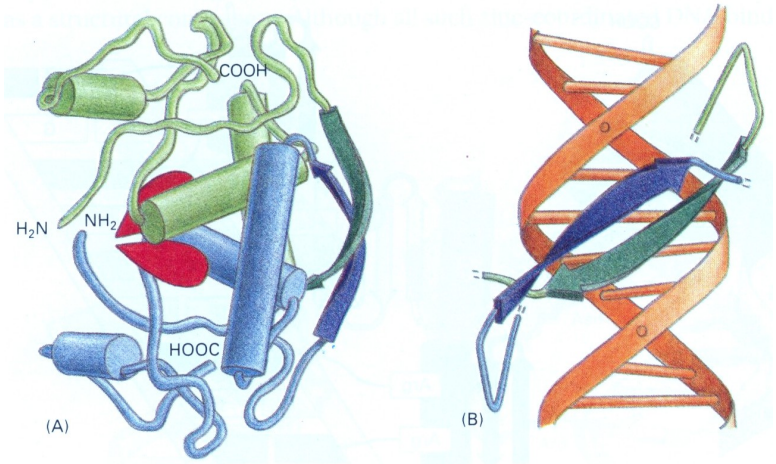
Fuchs et al., NSMB 2006



Weinberg et al., RNA 2008

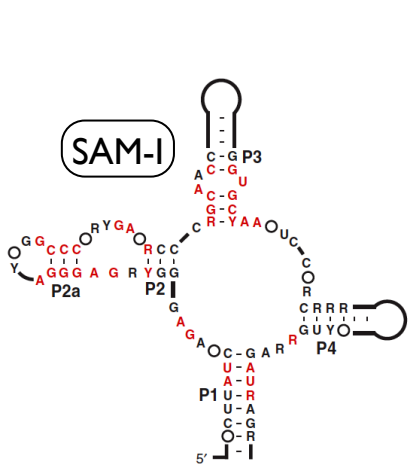
Not the only way!

Alberts, et al, 3e.

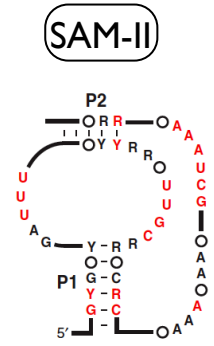


← Protein way

Riboswitch alternatives ↓

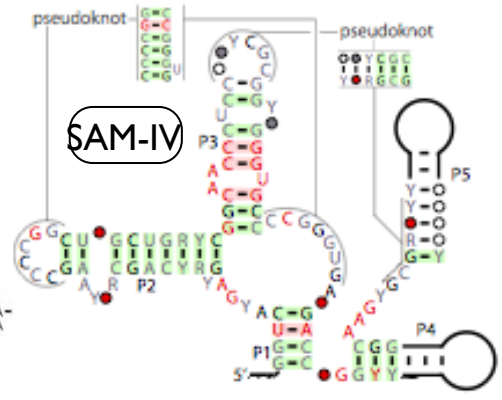


Grundy, Epshtein, Winkler et al., 1998, 2003

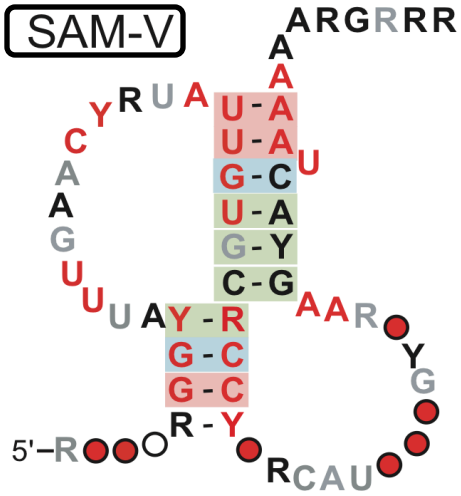


Corbino et al., Genome Biol. 2005

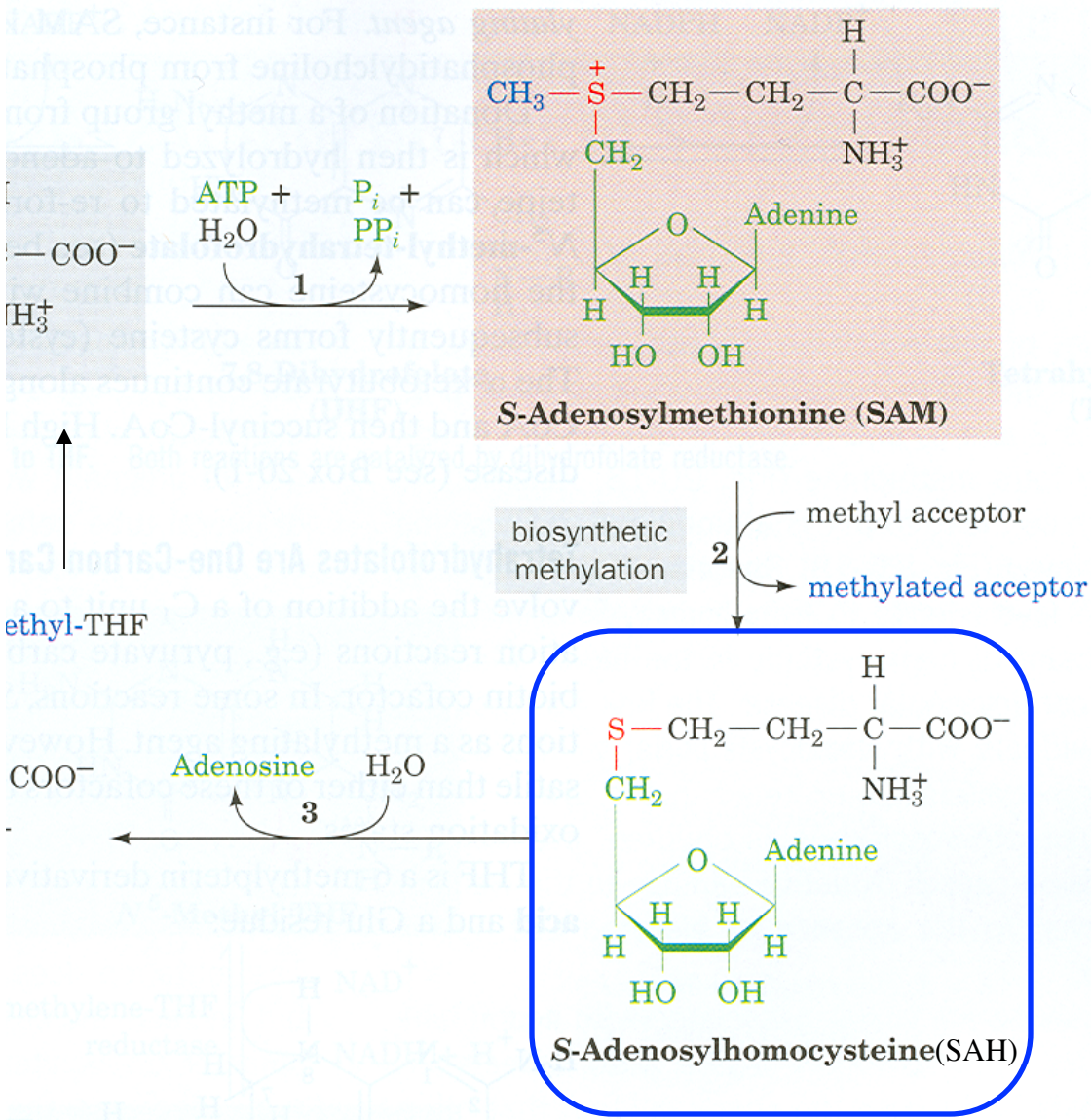
Fuchs et al., NSMB 2006



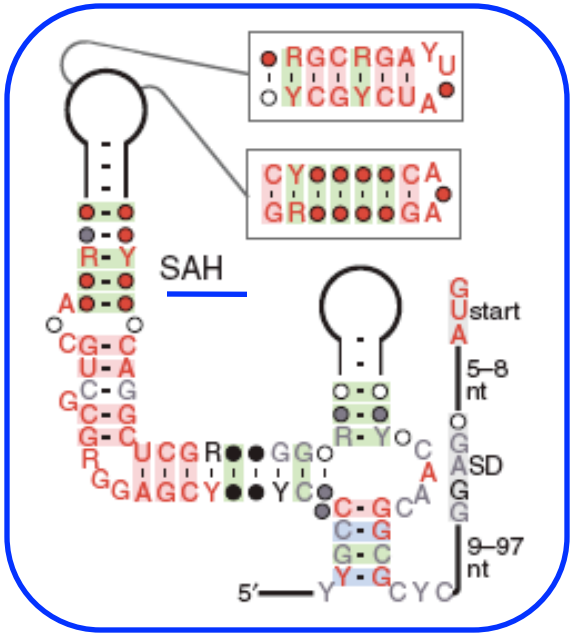
Weinberg et al., RNA 2008

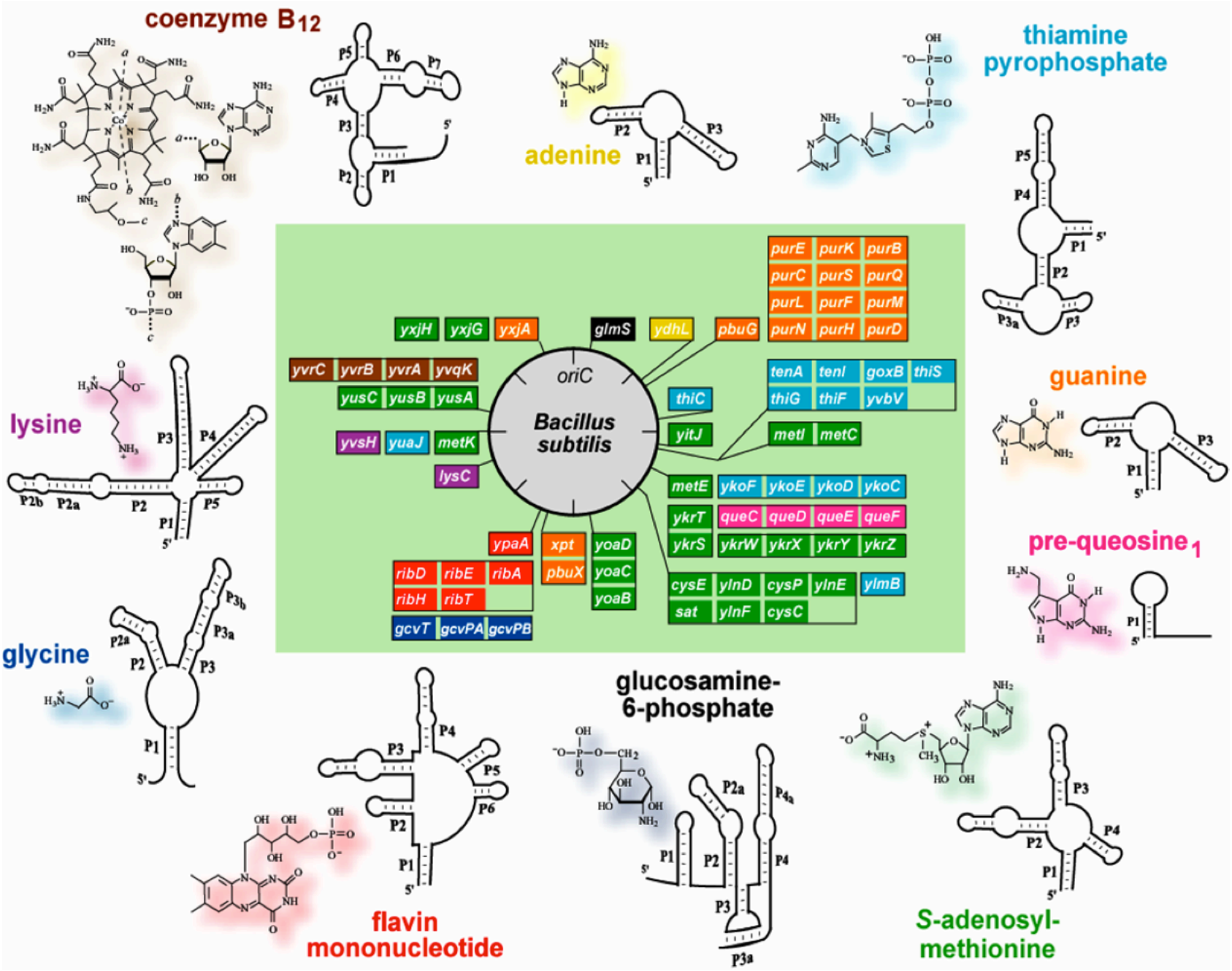


Meyer, etal., BMC Genomics 2009



And in other bacteria, a riboswitch senses SAH




coenzyme B12
adenine
thiamine pyrophosphate
guanine
lysine
glycine
pre-queosine1
glucosamine-6-phosphate
flavin mononucleotide
S-adenosyl-methionine
oriC
Bacillus subtilis
yxjH yxjG yxjA
glmS
ydhL
pbuG
purE purK purB
purC purS purQ
purL purF purM
purN purH purD
yvrC yvrB yvrA yvqK
yusC yusB yusA
yvsH yuaJ metK
lysC
thiC
yitJ
tenA tenI goxB thiS
thiG thiF yvbV
metI metC
metE ykoF ykoE ykoD ykoC
ykrT queC queD queE queF
ykrS ykrW ykrX ykrY ykrZ
ypaA xpt yoaD
pbuX yoaC
yoaB
ribD ribE ribA
ribH ribT
gcvT gcvPA gcvPB
cysE ylnD cysP ylnE ylmB
sat ylnF cysC

Antibiotics?

Old drugs, new understanding:

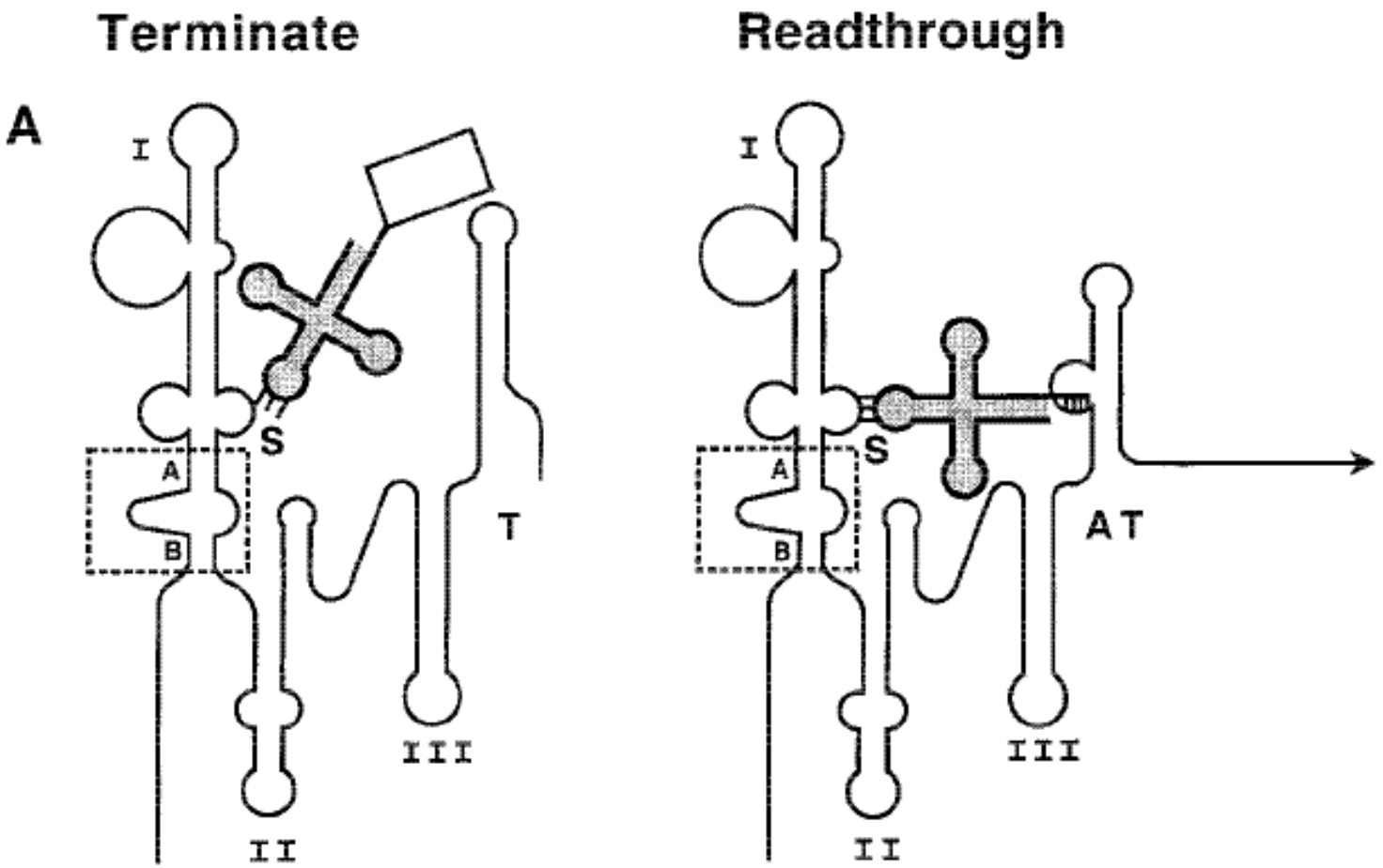
TPP ~ pyrithiamine

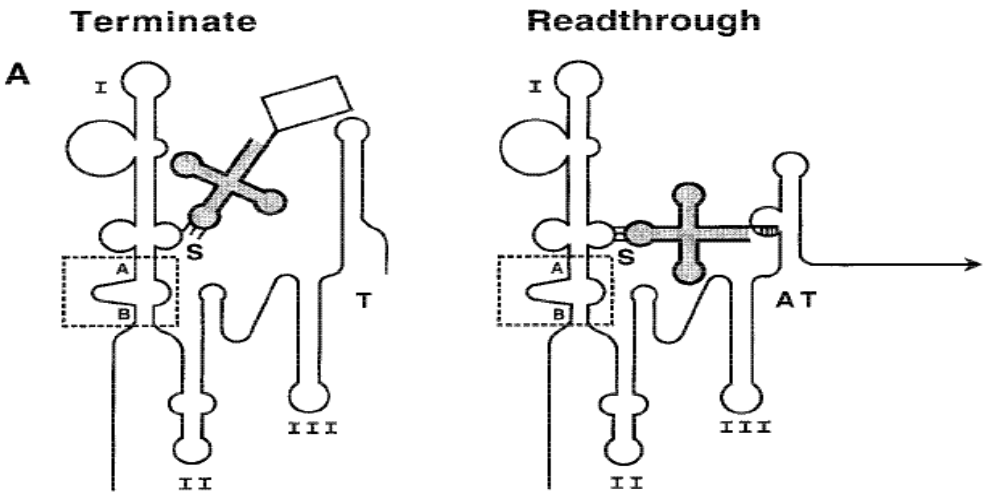
lysine ~ L-aminoethylcysteine, DL-4-oxalysine

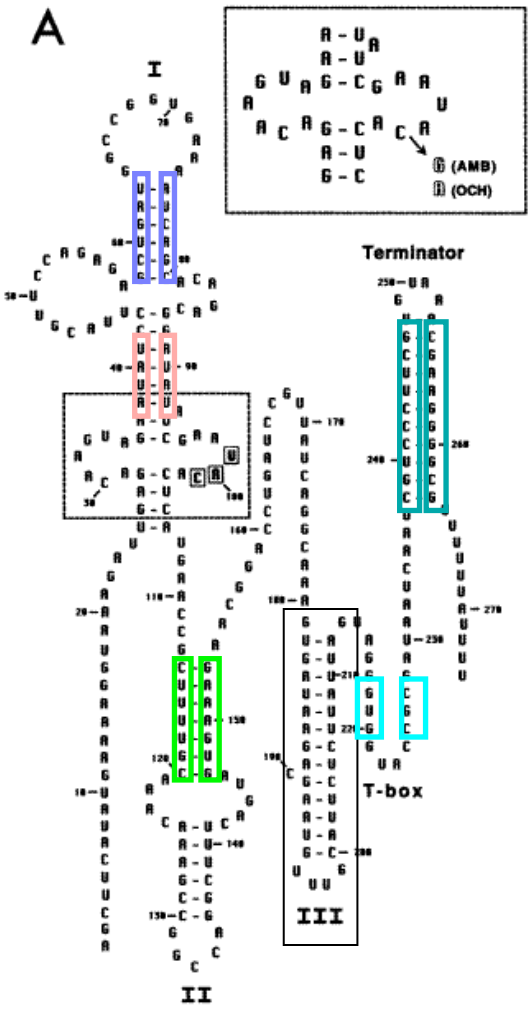
FMN ~roseoflavin

Potential advantages - no (known) human riboswitches, but often multiple copies in bacteria

ncRNA Example: T-boxes








```
NC_000964.1 AUAUC.CUUACGU..UCCAGAGAGCUGAUGGCCGGUGAAA.AUCAGCACAGACGGAUAUAU
NC_004722.1 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG
NC_004193.1 AAAAGUAGAACCG.AUCUAGCGAAUUGAGGAU.GGUGUGAGCUCAGUGC.GGAAAGCUUUU
NC_003997.3 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG

NC_000964.1 CGAA..UACACUCAUGAACCGCUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
NC_004722.1 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUUCC
NC_004193.1 AGAAAAUC.ACUCUUGAGUU.UUCAUUACGAAA..CA............AGUAGUAAUGGA
NC_003997.3 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUUCC

NC_000964.1 aCGGAC.CUGAUCCGUUAUCAGGCAAAGUGGUACCGCGAUAAUCAAUCGUCCCUUCGUGUAAaCGAAGGGGCGUUU
NC_004722.1 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAUaAGGGACGGGAGUU
NC_004193.1 .CGGUUcAUC.UCCGUUAUCGAUCUUAGUGGUACCGCGA.....GUCUUCUCGUCCCUUUU..GGGAUUAGAAGGC
NC_003997.3 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAUaAGGGACGGGAGUU
```

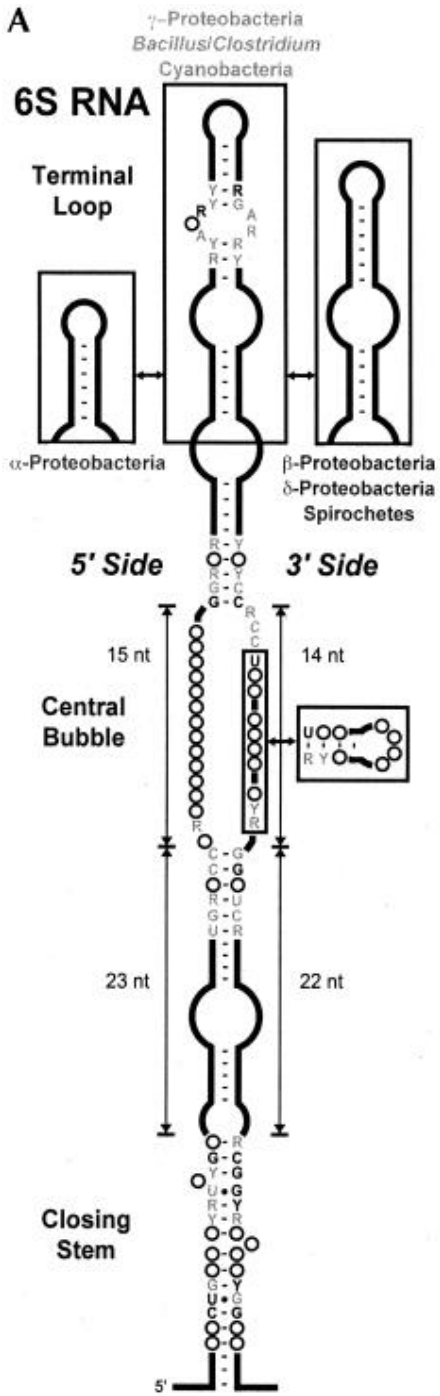
ncRNA Example: 6S

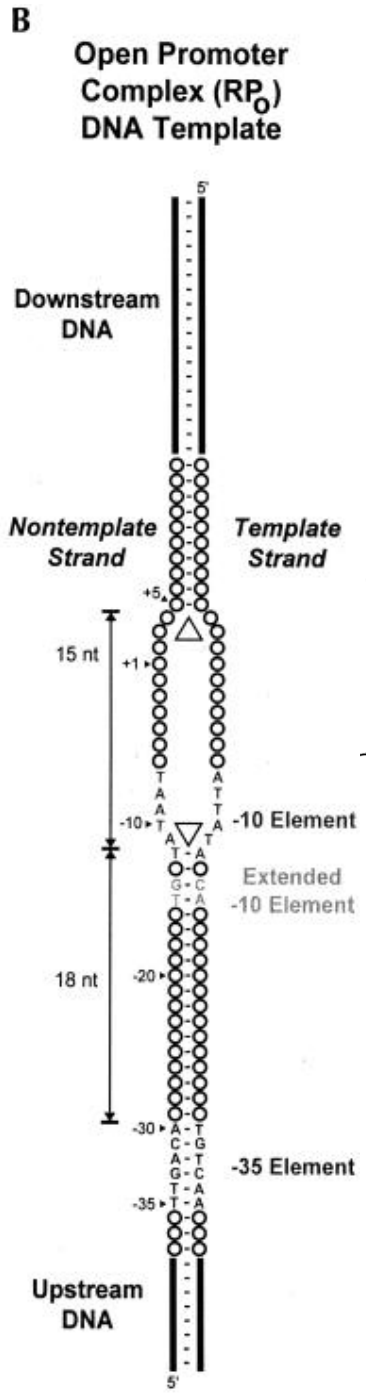
medium size (175nt)

structured

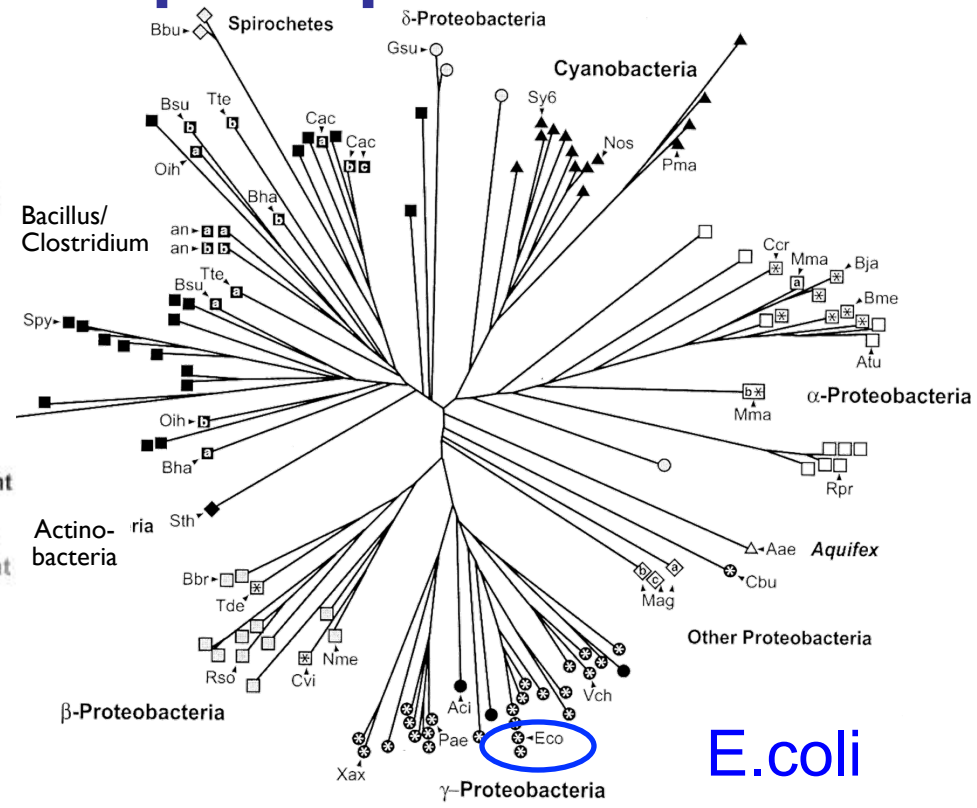
highly expressed in E. coli in certain growth conditions

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

Bottom line?

A significant number of "one-off" examples

Extremely wise-spread ncRNA expression

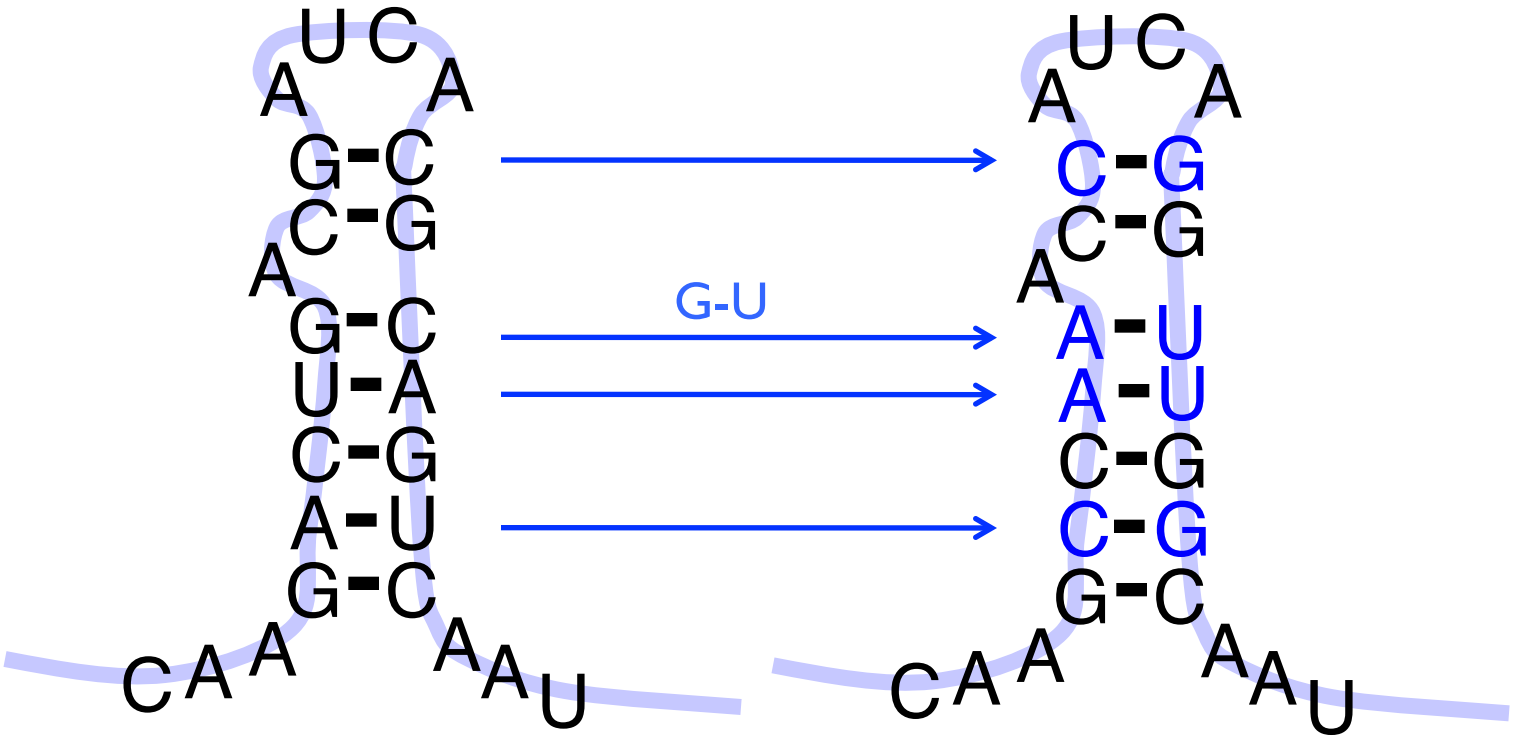
At a minimum, a vast evolutionary substrate

New technology (e.g. RNAseq) exposing more

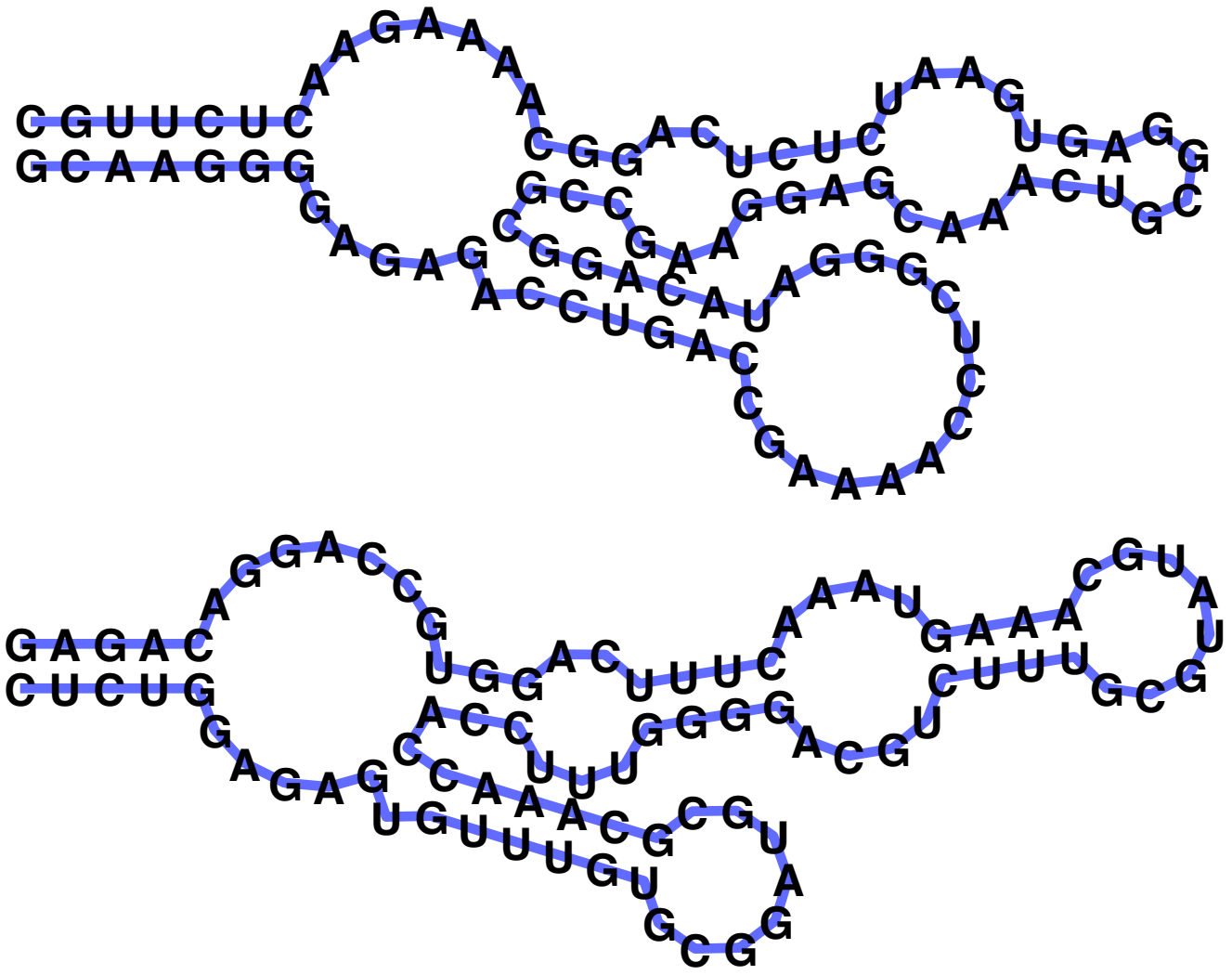
How do you recognize an interesting one?

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*

Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

RNA Pairing

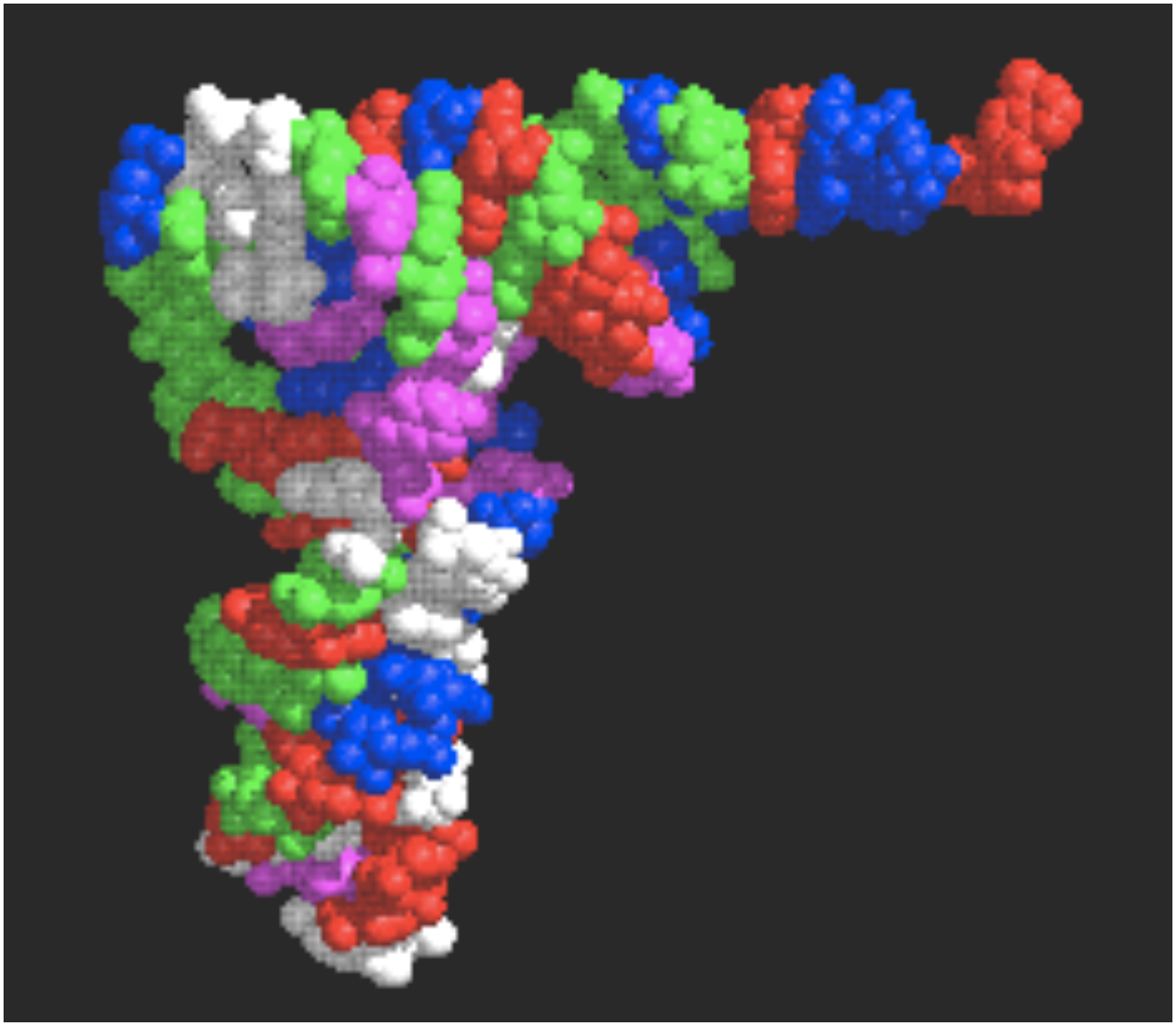
Watson-Crick Pairing

- 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

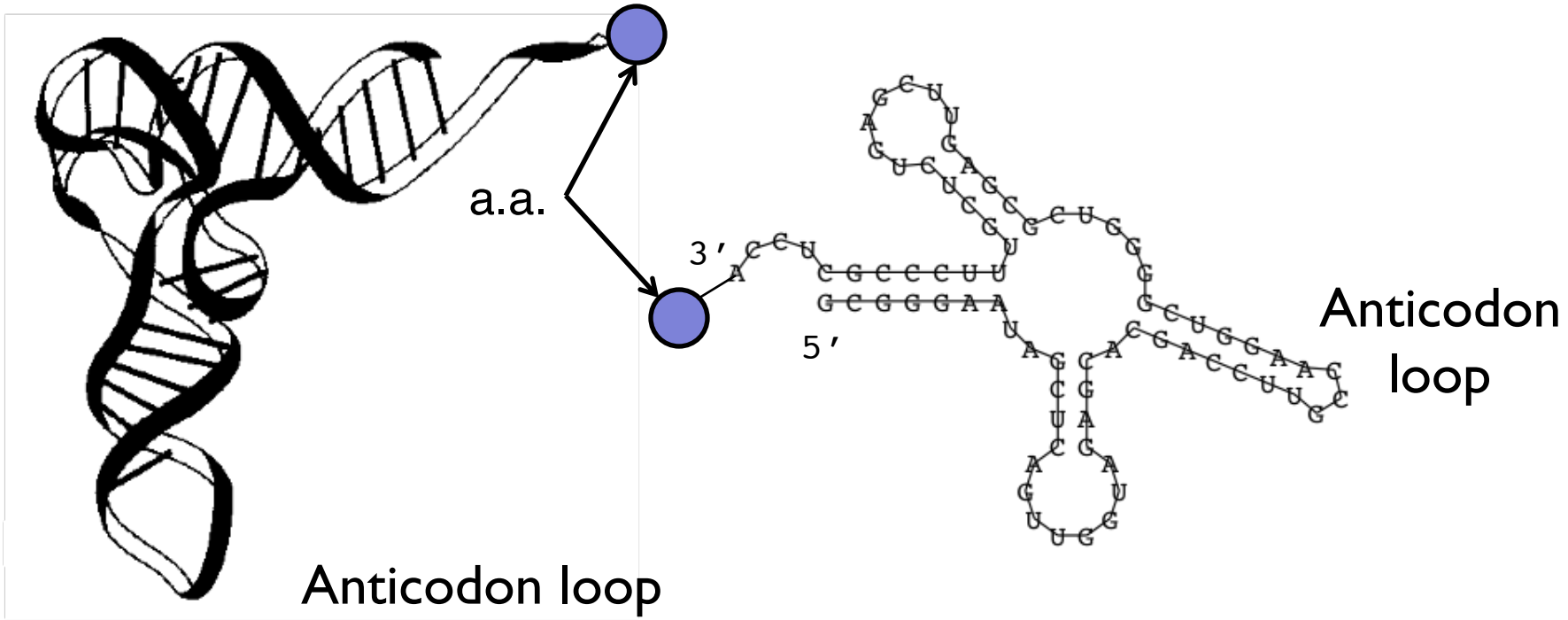
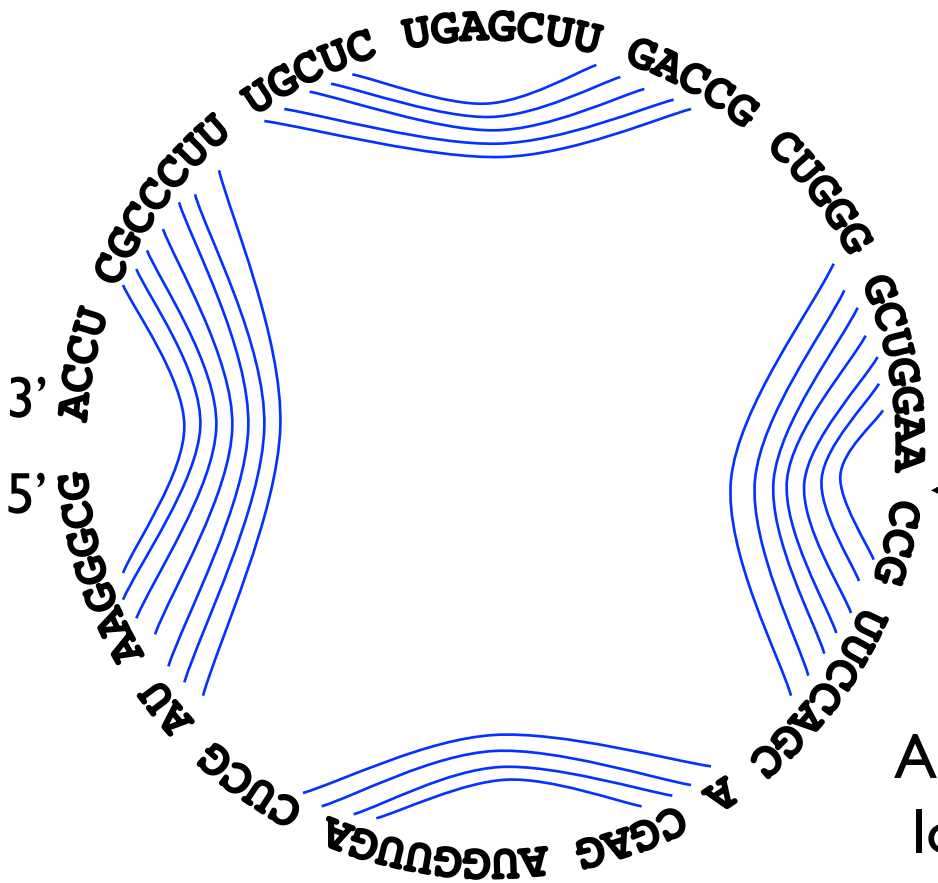


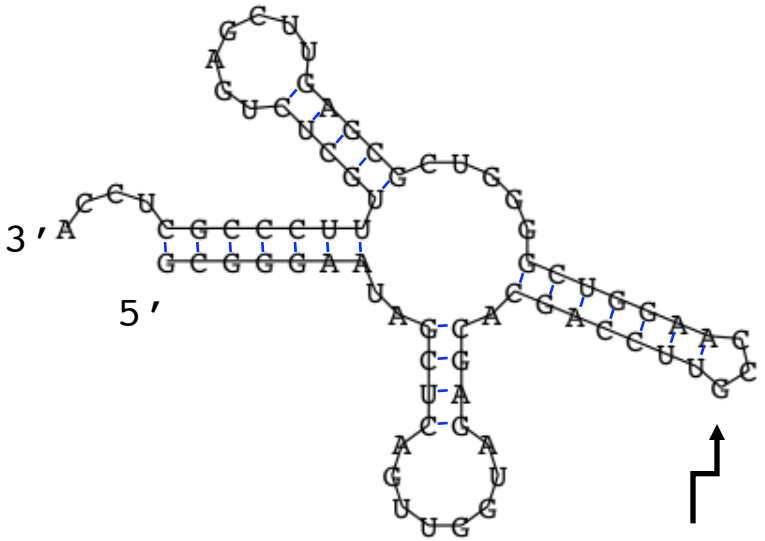
Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

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

tRNA - Alt. Representations



Anticodon loop



Anticodon loop

Definitions

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

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

i < j-4, and } no sharp turns

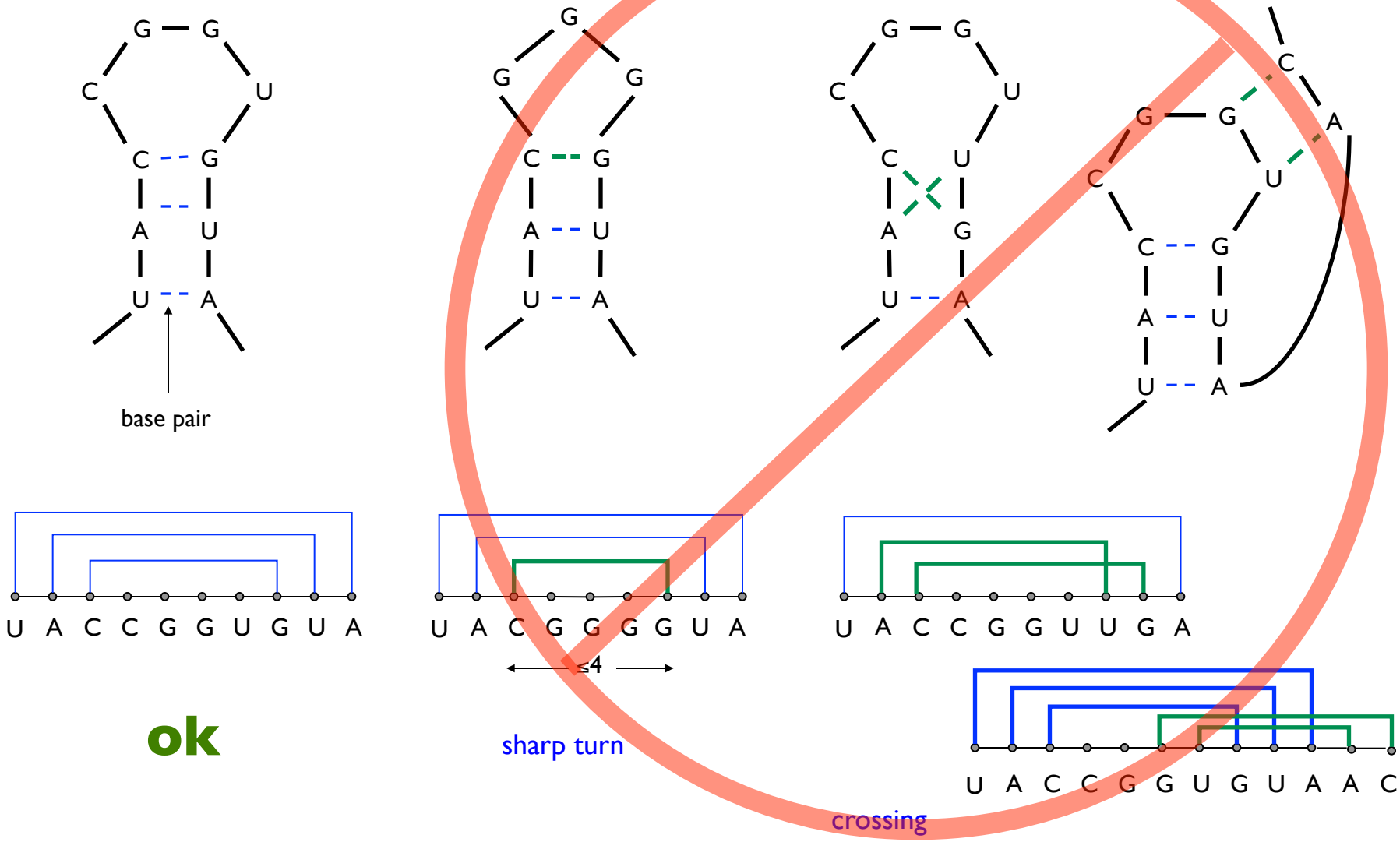
if i•j & i'•j' are two different pairs with i ≤ i', then

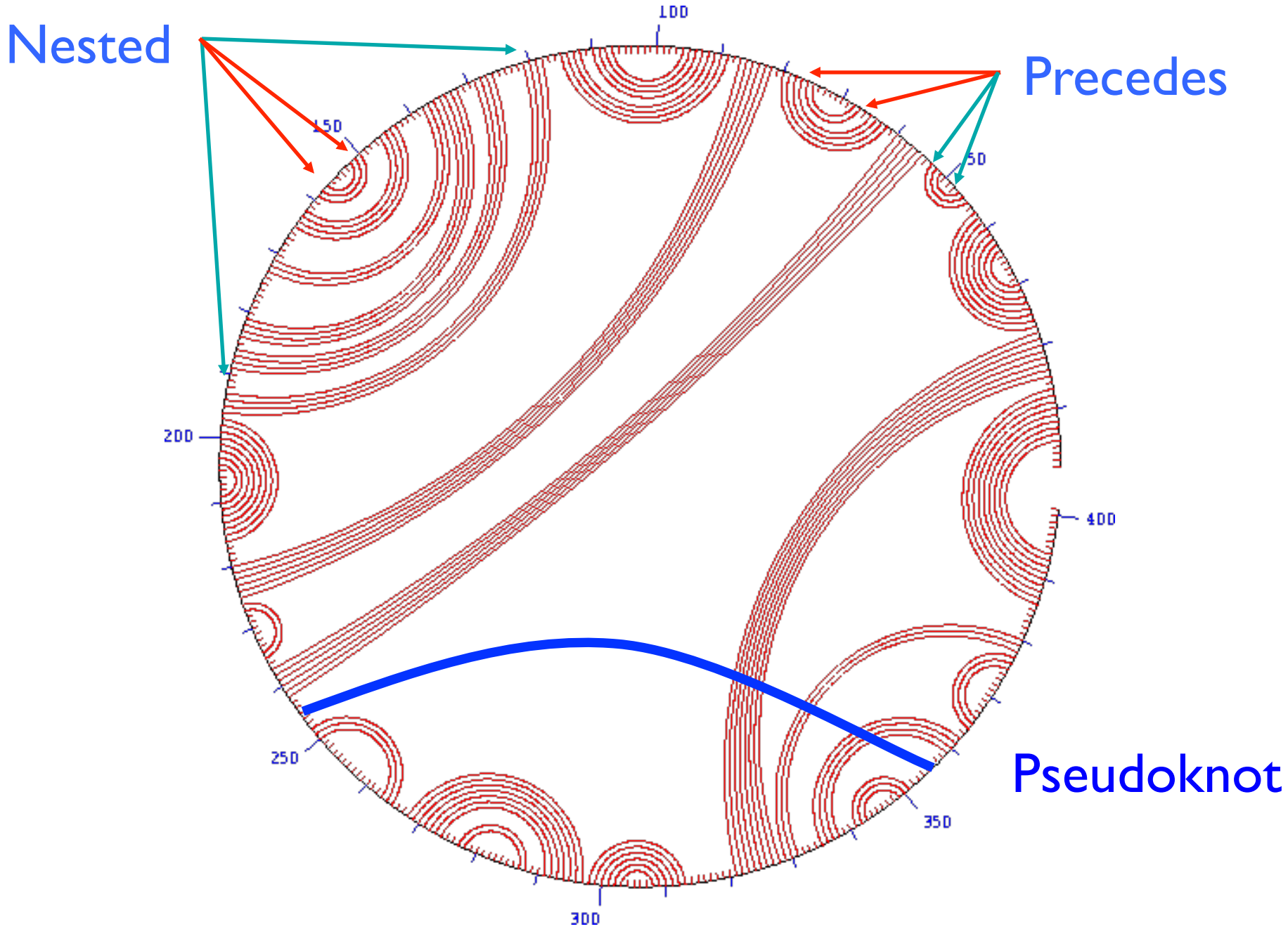
j < i', or

i < i' < j' < j

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

RNA Secondary Structure: Examples




Nested
Precedes
Pseudoknot
100
150
50
200
400
250
350
300

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

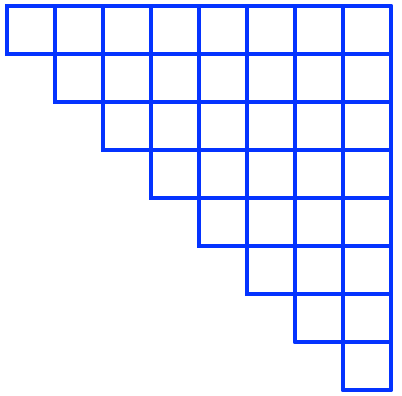
Nussinov: Max Pairing

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

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

B(i,j) = max of:

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

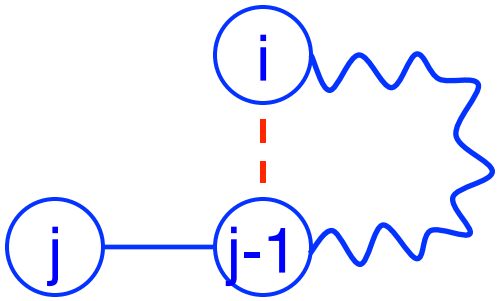


R Nussinov, AB Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA." PNAS 1980.

"Optimal pairing of r_i ... r_j"
Two possibilities

j Unpaired:

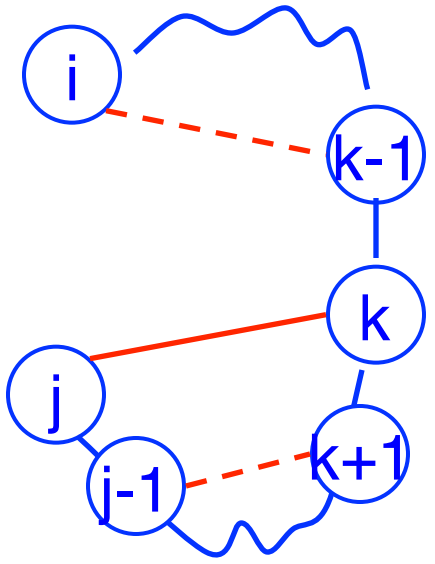
Find best pairing of r_i ... r_{j-1}



j Paired (with some k):

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

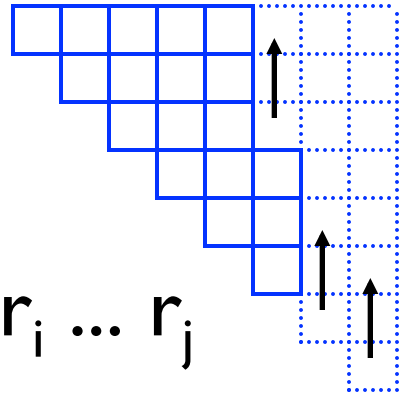
best r_{k+1} ... r_{j-1} plus 1



Why is it slow?

Why do pseudoknots matter?

Nussinov:
A Computation Order

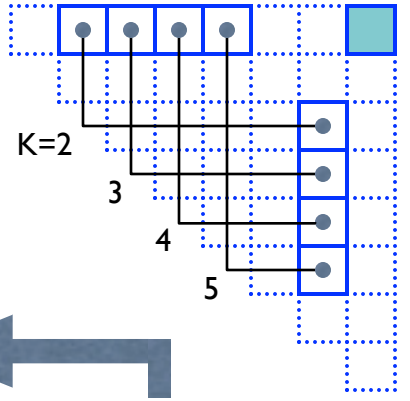


B(i,j) = # pairs in optimal pairing of r_i ... r_j

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

B(i,j) = max of:

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



Time: $O(n^3)$

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"