

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

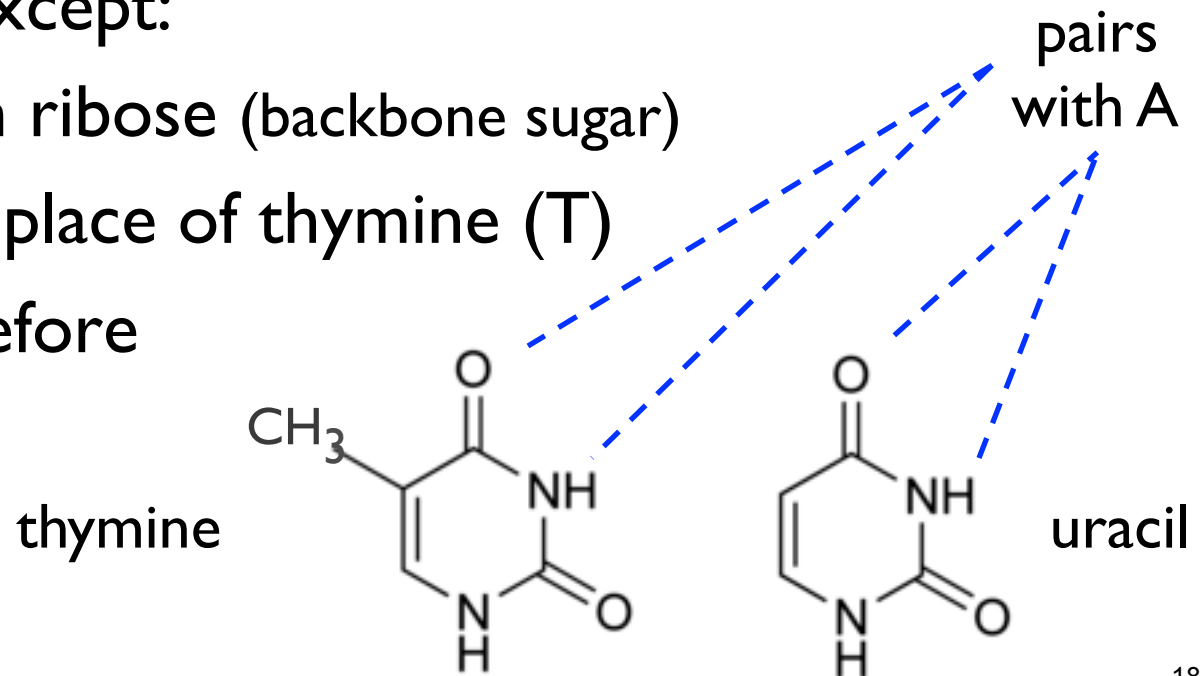
RNA: RiboNucleic Acid

Like DNA, except:

Lacks OH on ribose (backbone sugar)

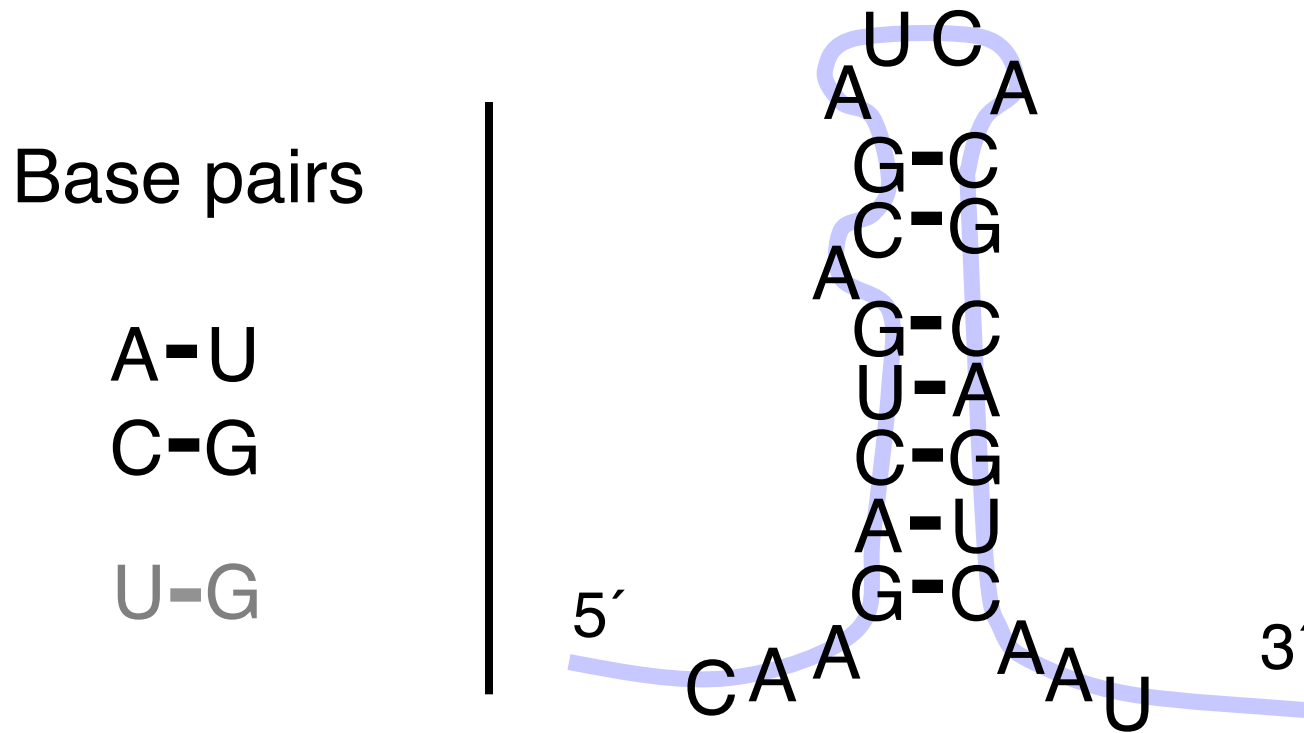
Uracil (U) in place of thymine (T)

A, G, C as before

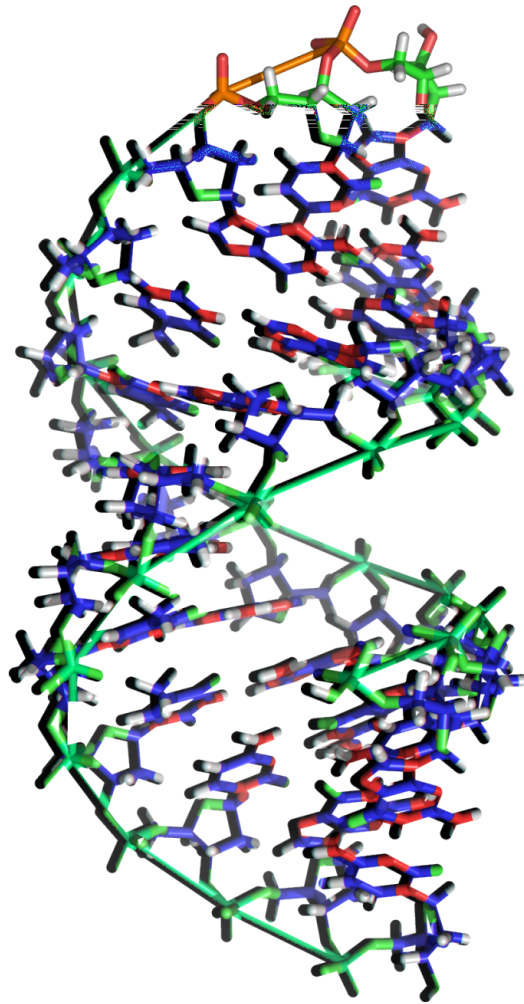


RNA Secondary Structure:

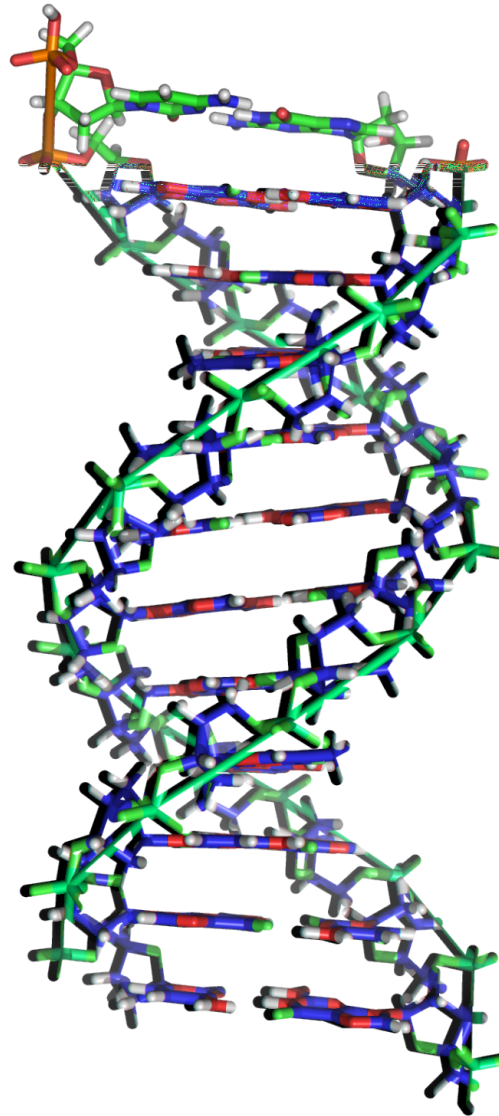
RNA makes helices too



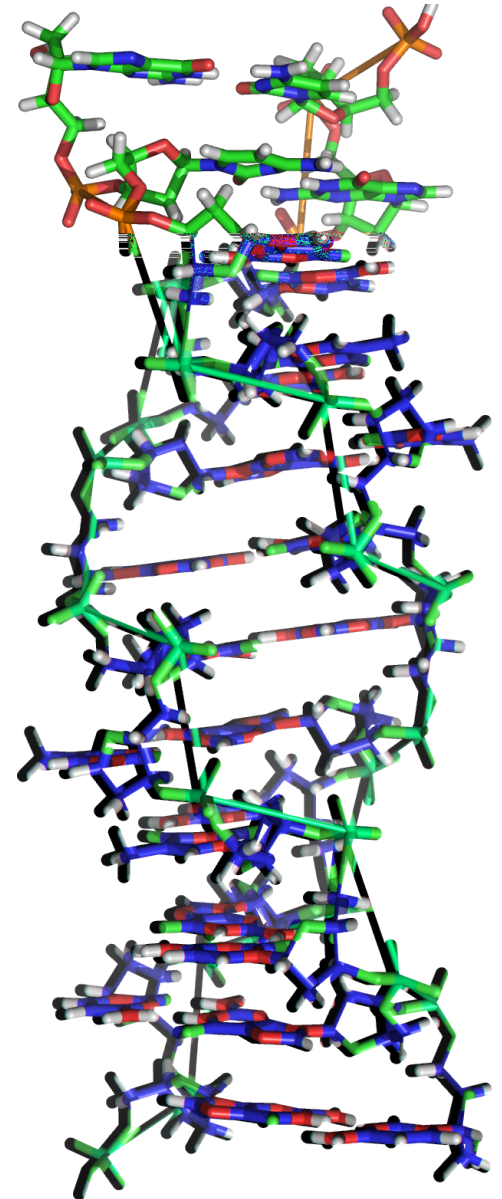
Usually *single* stranded



A
(norm for RNA)



B
(norm for DNA)



Z

NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

by

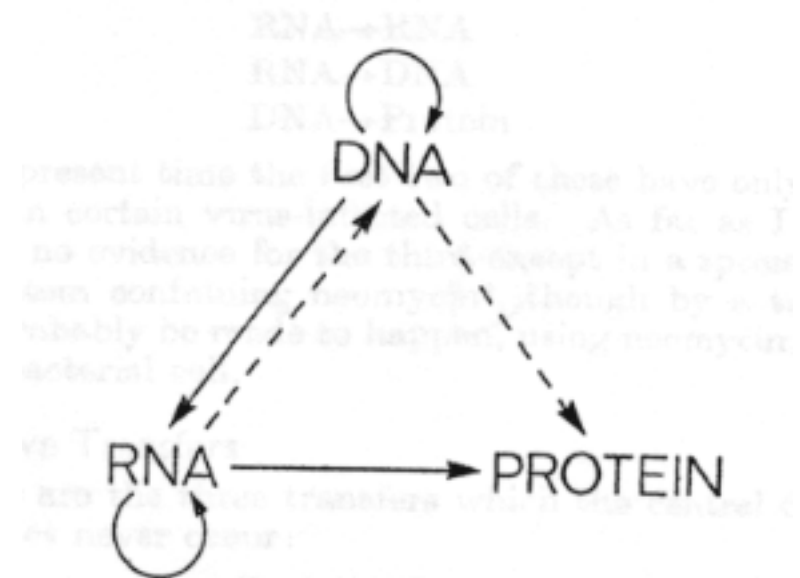
FRANCIS CRICK

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

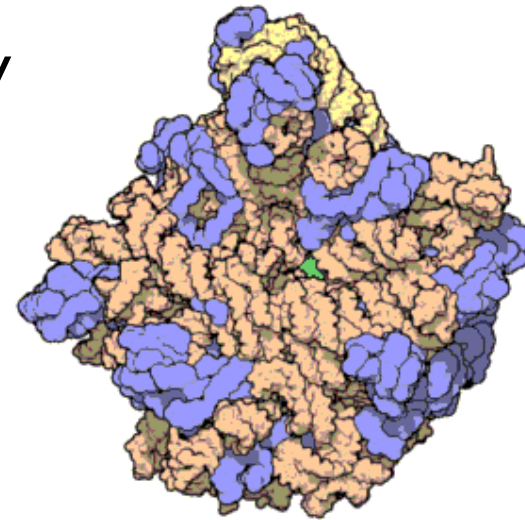
1974 Nobel prize to Romanian biologist George Palade (1912-2008) for discovery in mid 50's

50-80 proteins

3-4 RNAs (half the mass)

Catalytic core is RNA

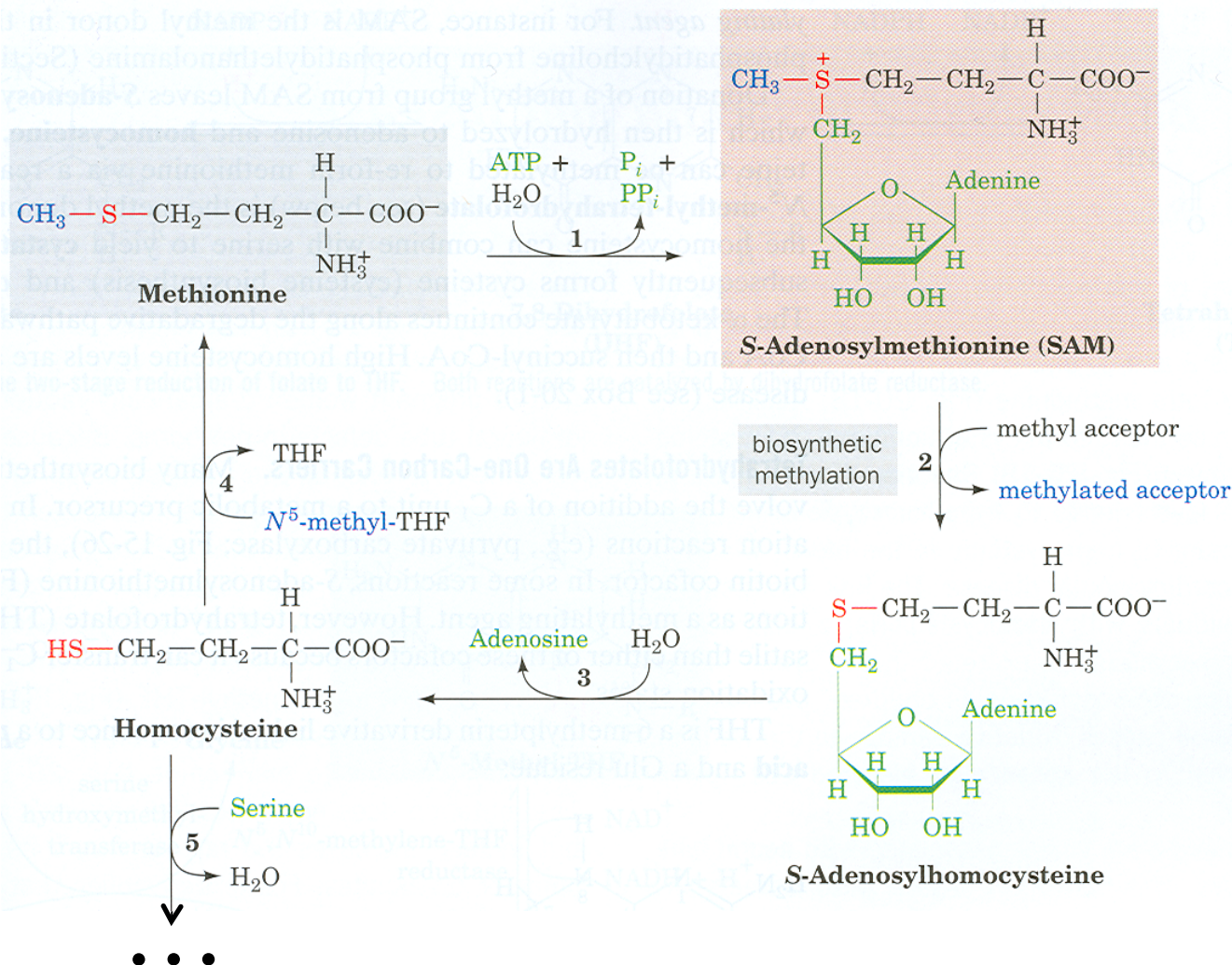
Of course, mRNAs and tRNAs (messenger & transfer RNAs) are critical too



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.

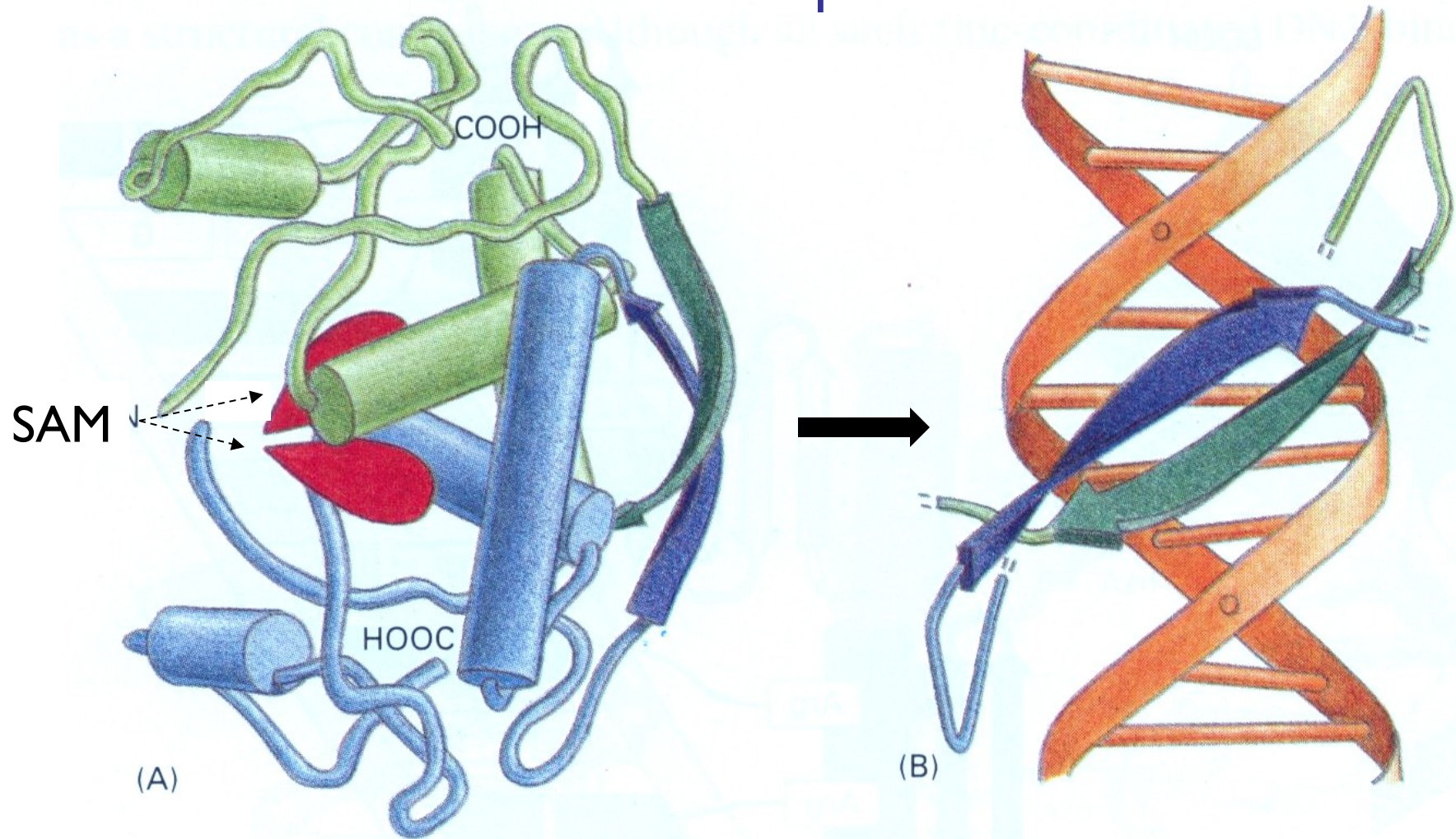
- Wikipedia

Proteins Catalyze Biochemistry: Met Pathways



Proteins Regulate Biochemistry:

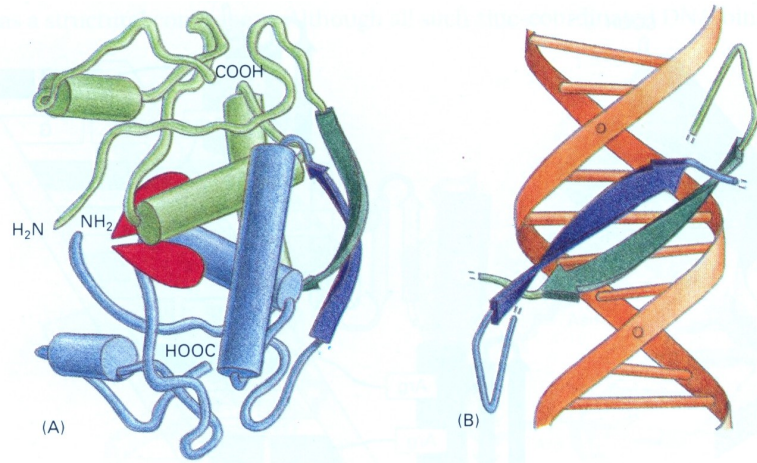
The MET Repressor



Protein

Alberts, et al, 3e.

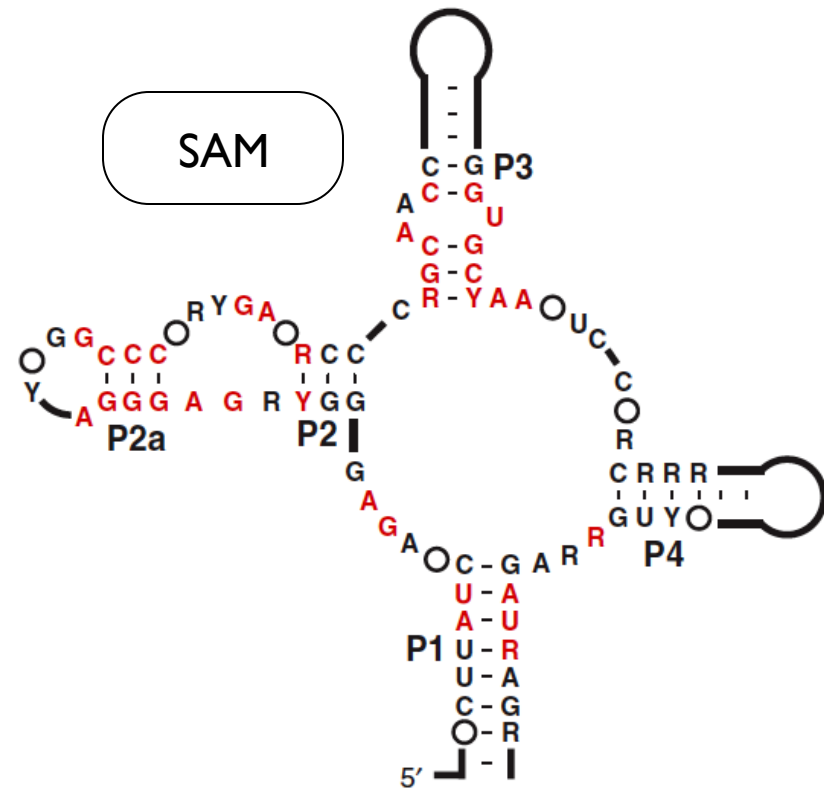
DNA



Not the only way!

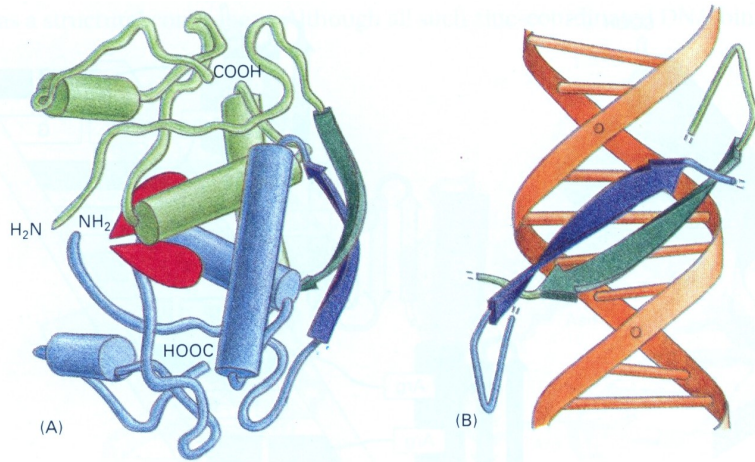
Protein
way

Riboswitch
alternative



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

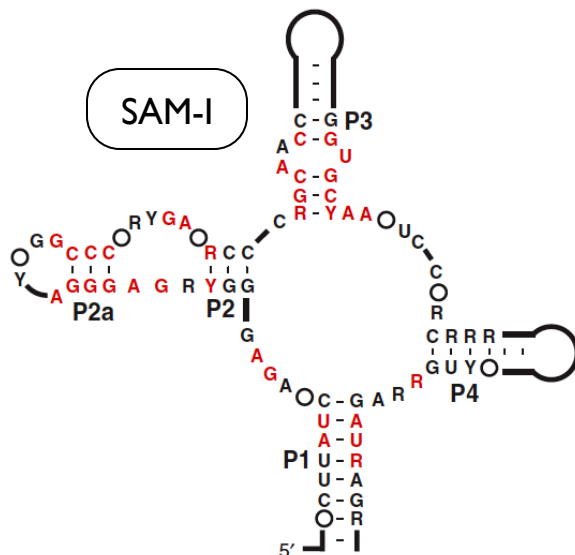
Alberts, et al, 3e.



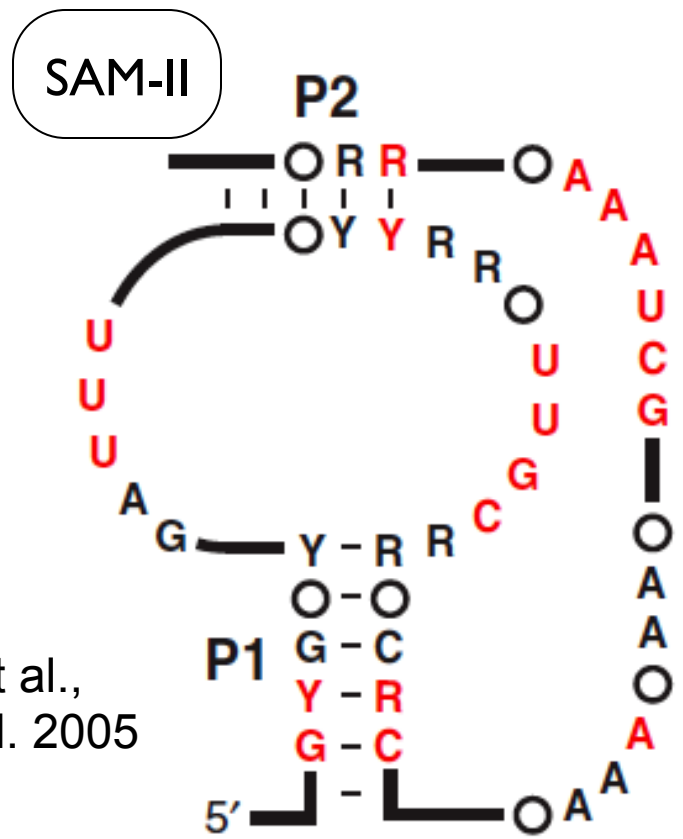
Not the only way!

Protein
way

Riboswitch
alternatives



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



Corbino et al.,
Genome Biol. 2005

Diagram illustrating the structure of a protein and a DNA double helix.

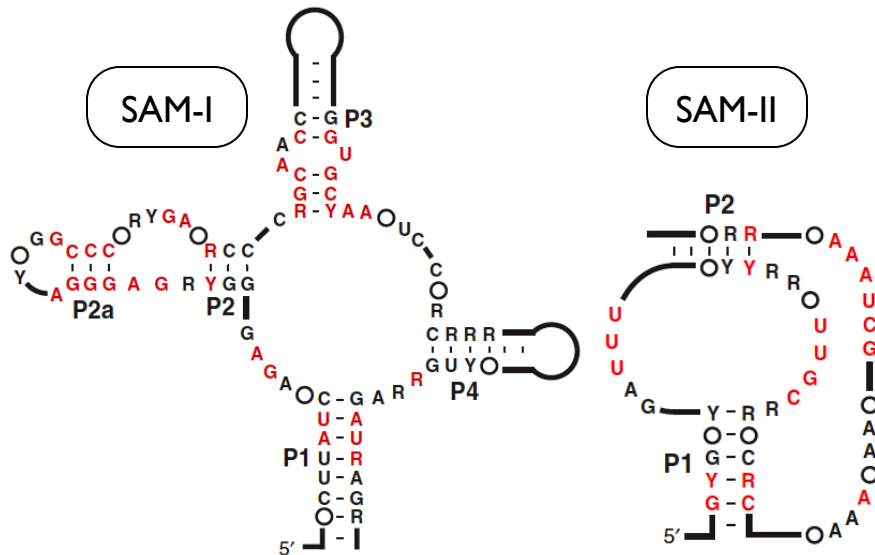
(A) Protein structure: A ribbon diagram of a protein. The structure is colored green and blue. Labels include COOH (carboxyl group), NH_2 (amino group), and HOOC (carboxyl group).

(B) DNA double helix structure: A ribbon diagram of a DNA double helix. The structure is colored orange and blue. Labels include H_2N (amino group) and HOOC (carboxyl group).

Not the only way!

Protein way

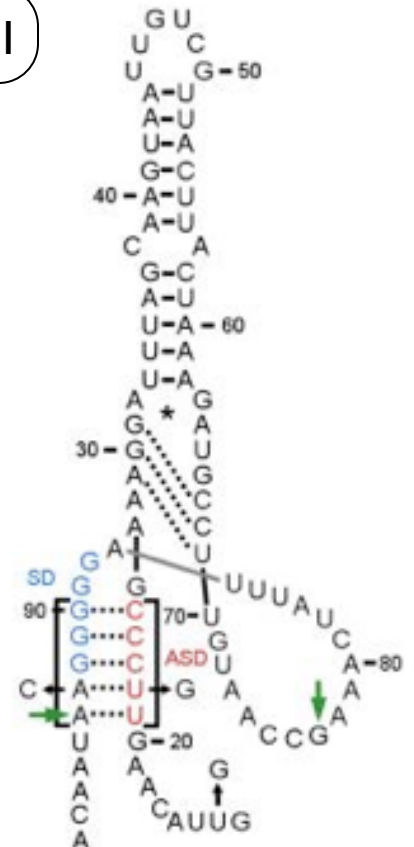
Riboswitch alternatives



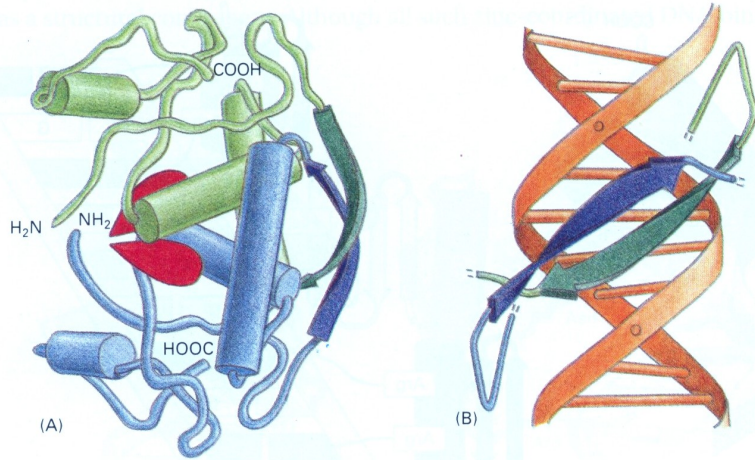
Grundy, Epshtein, Winkler
et al., 1998, 2003

Corbino et al.,
Genome Biol. 2005

SAM-III



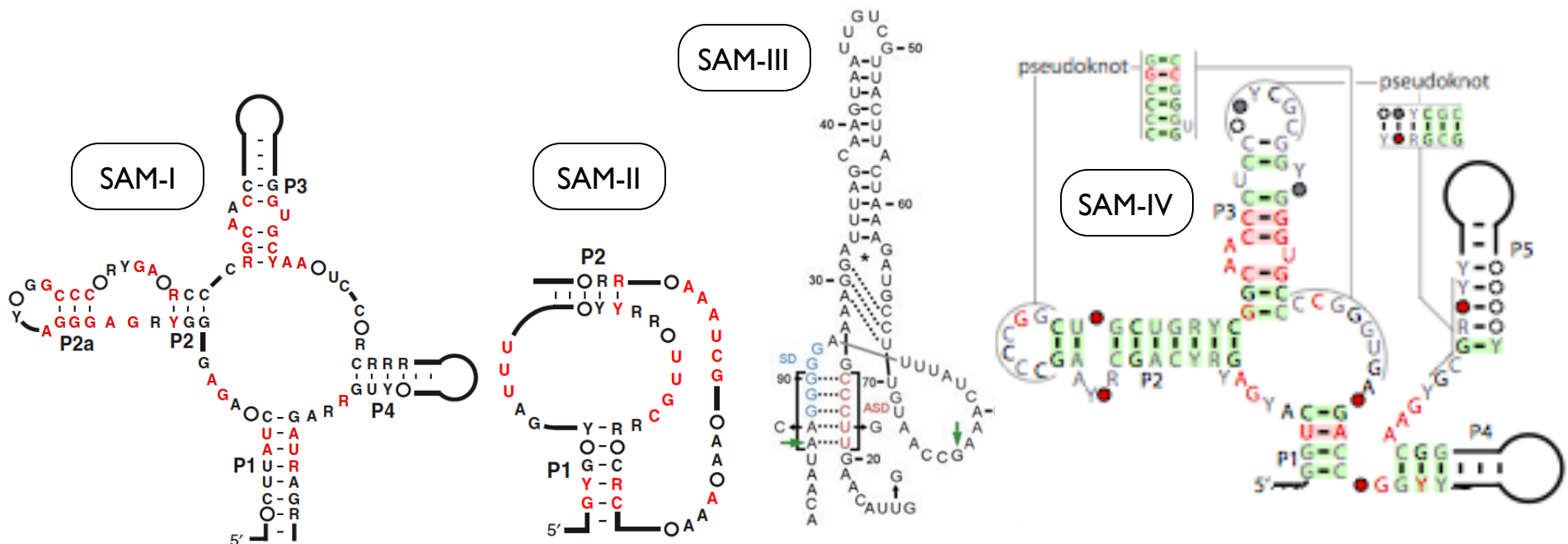
Alberts, et al, 3e.



Not the only way!

Protein
way

Riboswitch
alternatives



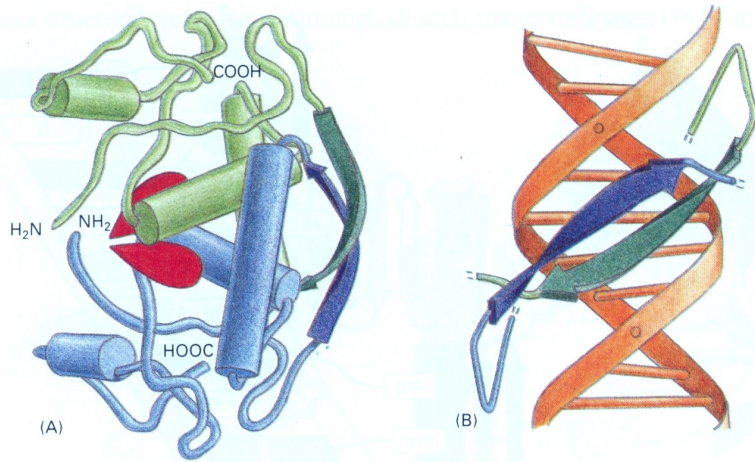
Grundy, Epshtein, Winkler
et al., 1998, 2003

Corbino et al.,
Genome Biol. 2005

Fuchs et al.,
NSMB 2006

Weinberg et al.,
RNA 2008

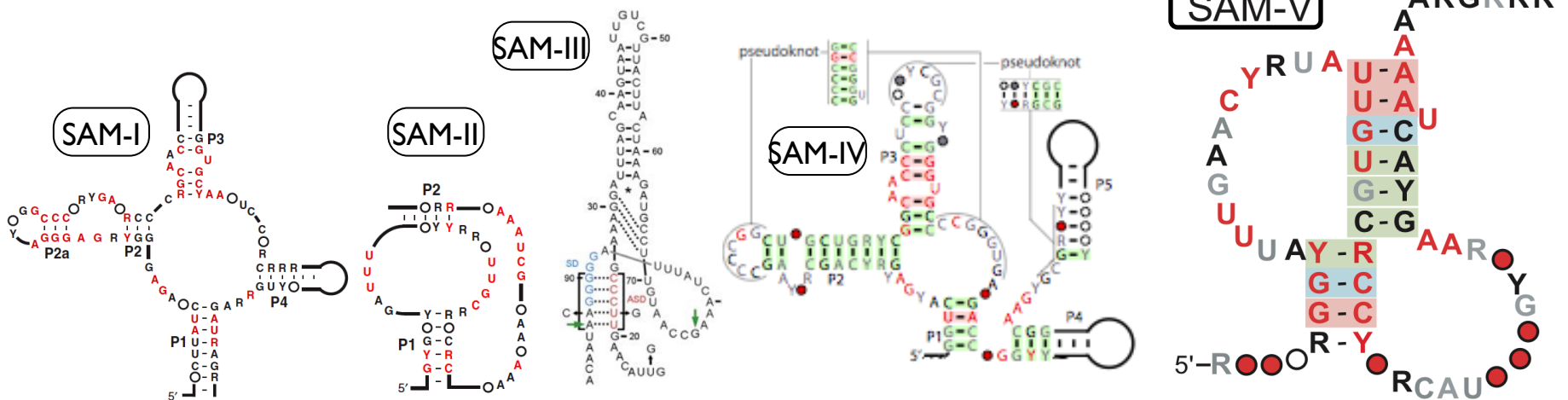
Alberts, et al, 3e.



Not the only way!

Protein
way

Riboswitch
alternatives



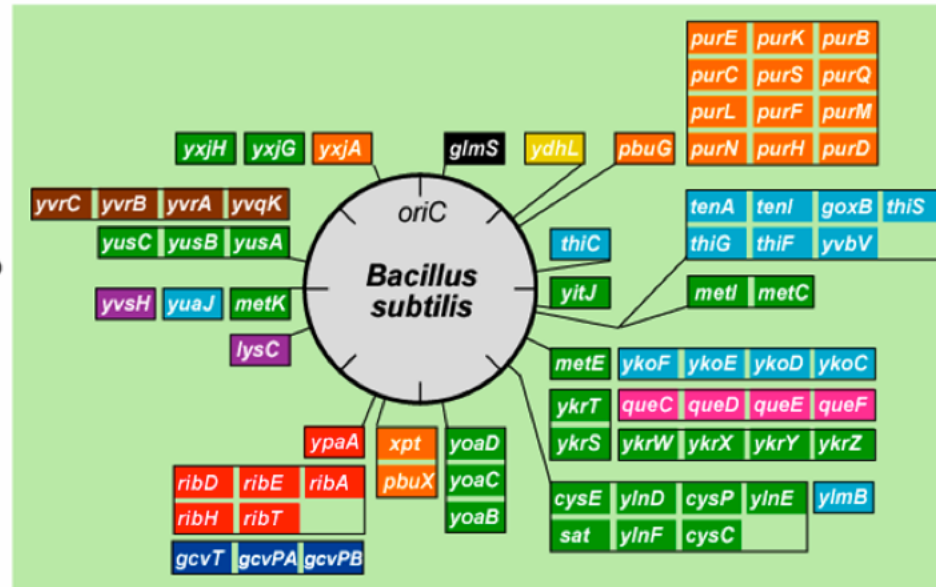
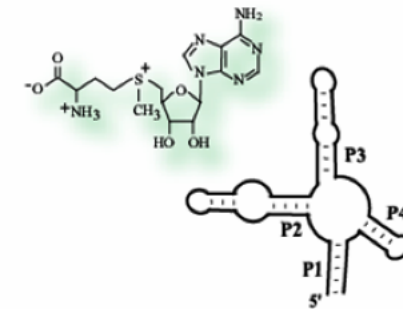
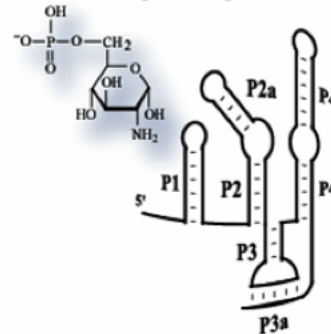
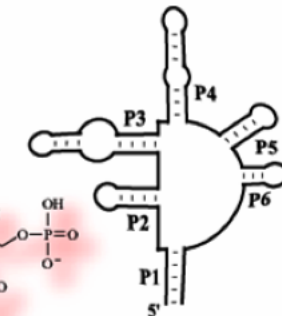
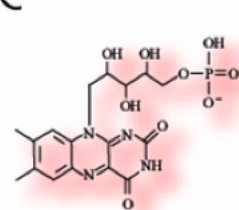
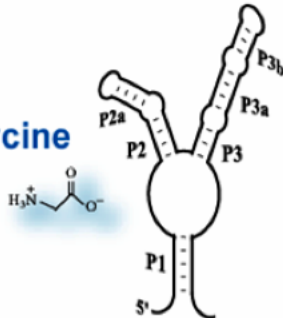
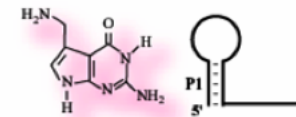
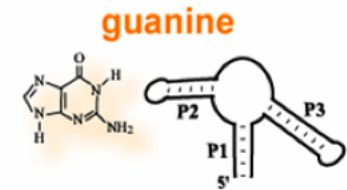
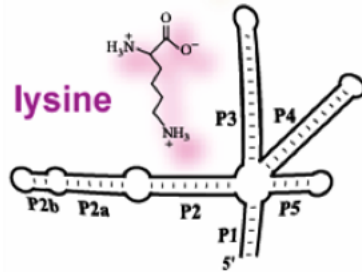
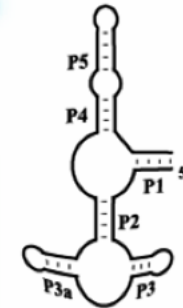
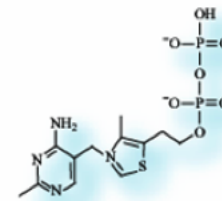
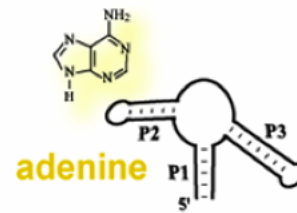
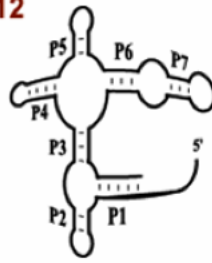
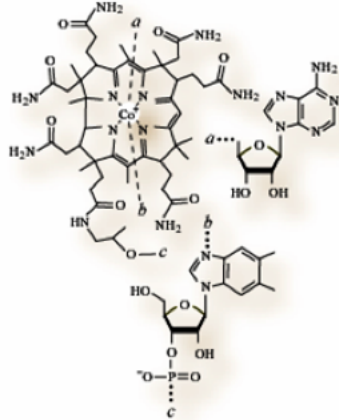
Grundy, Epshtein,
Winkler
et al., 1998, 2003

Corbino et
al.,
Genome
Biol. 2005

Fuchs
et al.,
NSMB
2006

Weinberg
et al.,
RNA 2008

Meyer, et al., BMC
Genomics 2009



Antibiotics?

Old drugs, new understanding:

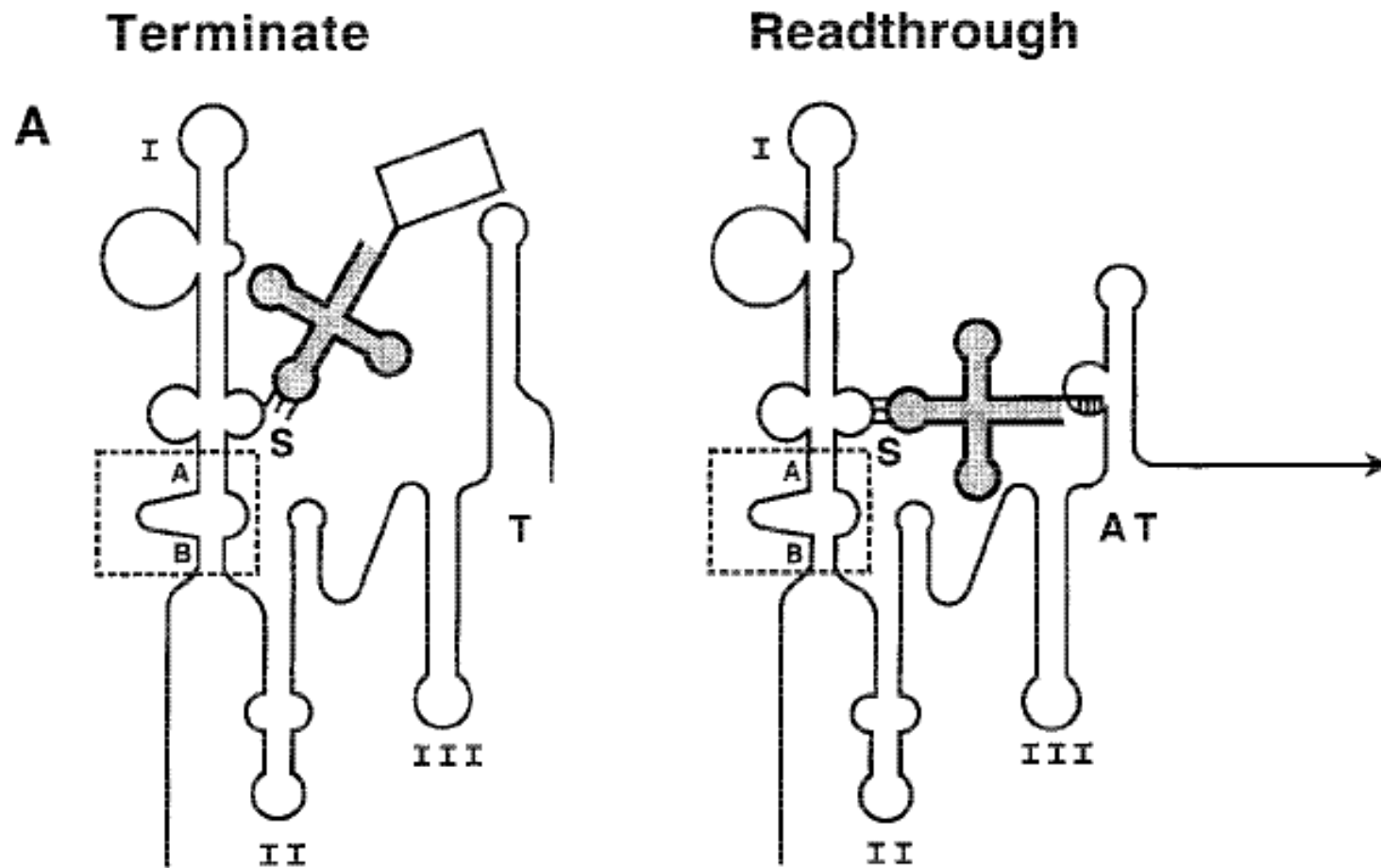
TPP ~ pyriithiamine

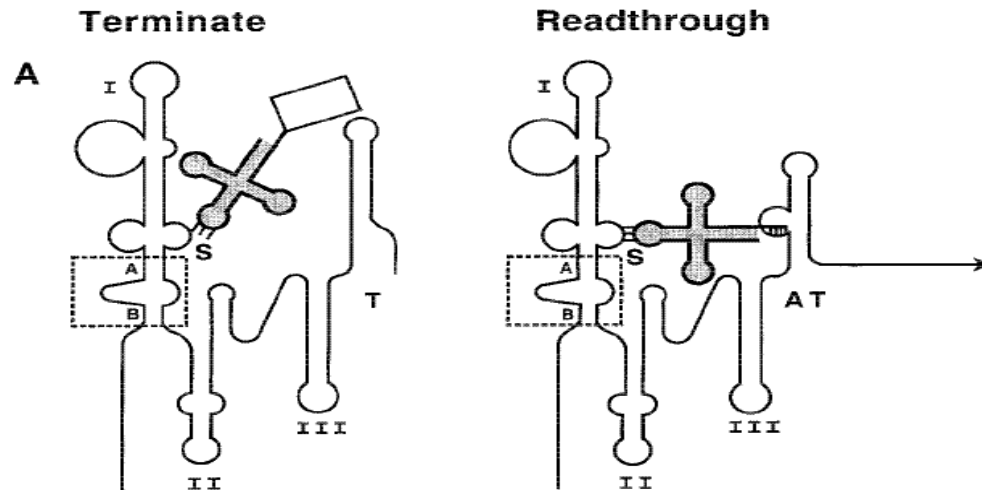
lysine ~ L-aminoethylcysteine, DL-4-oxalysine

FMN ~ roseoflavin

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

ncRNA Example: T-boxes

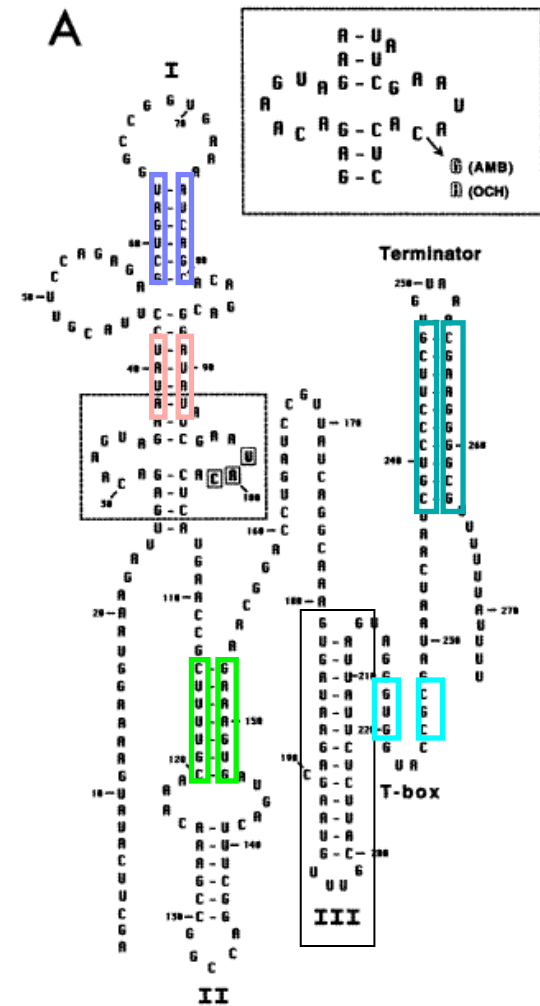




NC_000964.1 **AUAUC**.CUUACGU..UCCAGAGAG**CUGAU**GGCCGGUGAAA..**AUCAGC**ACAGACGGAU**AUAU**
 NC_004722.1 **CAAAU**.GUCGUUUcUUUAGAGAG**UUCGAU**GGUUGGUGGAA..**AUCGAU**AG..AAACA**GUUUG**
 NC_004193.1 **AAAAG**UAGAACCG.AUCUAGCGAA**AUUGAG**GAU..GGUGUGAG**CUCAGU**GC..GGAAAG**CUUUU**
 NC_003997.3 **CAAAU**.GUCGUUUcUUUAGAGAG**UUCGAU**GGUUGGUGGAA..**AUCGAU**AG..AAACA**GUUUG**

NC_000964.1 CGAA..UACACUCAUGAACCG**CUUUUUGC**AAACAAAGccggccaggcuuucAGUA..**GUGAAAG**
 NC_004722.1 UGAA..UCCAUCCUGGAAU..**GGAAUGU**GGAAUAUCUuuuggauu.....AGUAA**GCAUUC**
 NC_004193.1 AGAAAAUC.ACUCUUGAGUU..**UUCAUUAC**GAAA..CA.....AGUA**GUAUUGGA**
 NC_003997.3 UGAA..UCCAUCCUGGAAU..**GGAAUGU**GGAAUAUCUuuuugauu.....AGUAA**ACAUUC**

NC_000964.1 aCGGAC.CUGAUCCGUUAUCAGGCAAAG**GUG**GUAC**CGC**GAUAAUCAA**CGUCCCUUCG**UGUAAa**CGAAGGGGCGUUU**
 NC_004722.1 .CGGUG.AAGAGCCGUUAUU...UCuA**GUG**GCAA**CGC**GG..GUU**AACUCCCGUCCCU**UUUAUa**AGGGACGGGAGUU**
 NC_004193.1 .CGGUUcAUC.UCCGUUAUCGAUCUUAG**GUG**GUAC**CGC**GA.....**GUCUUCU****GUCCCUUUU**..**GGGAUUAGAAGGC**
 NC_003997.3 .CGGUG.AAGAGCCGUUAUU...UCuA**GUG**GCAA**CGC**GG..GUU**AACUCCCGUCCCU**UUUAUa**AGGGACGGGAGUU**



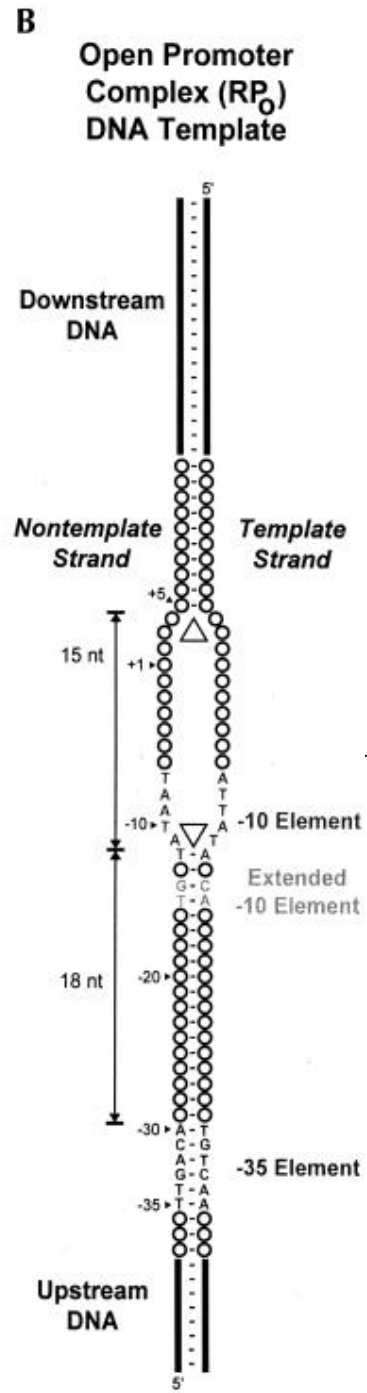
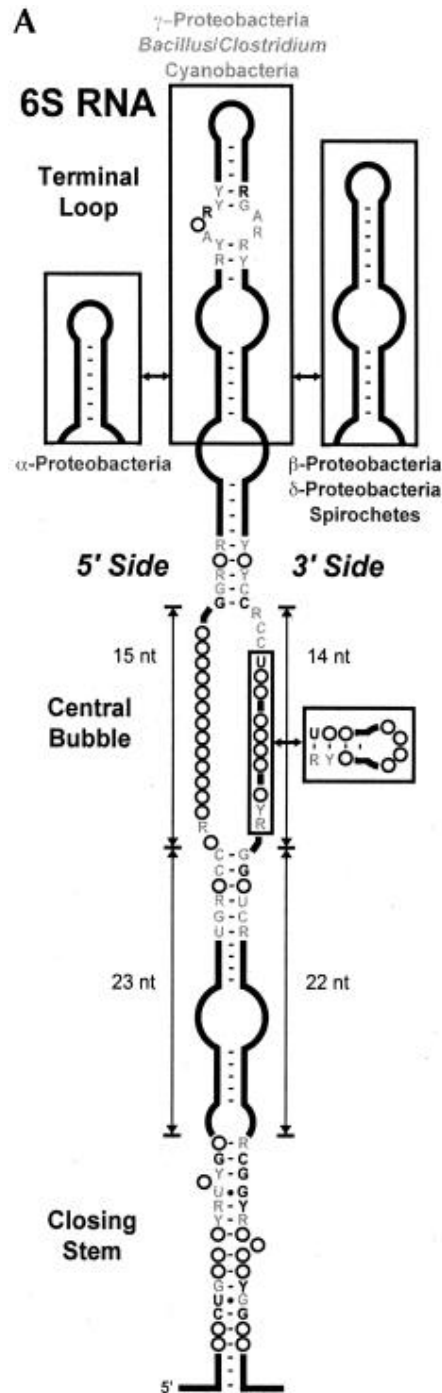
ncRNA Example: 6S

medium size (175nt)

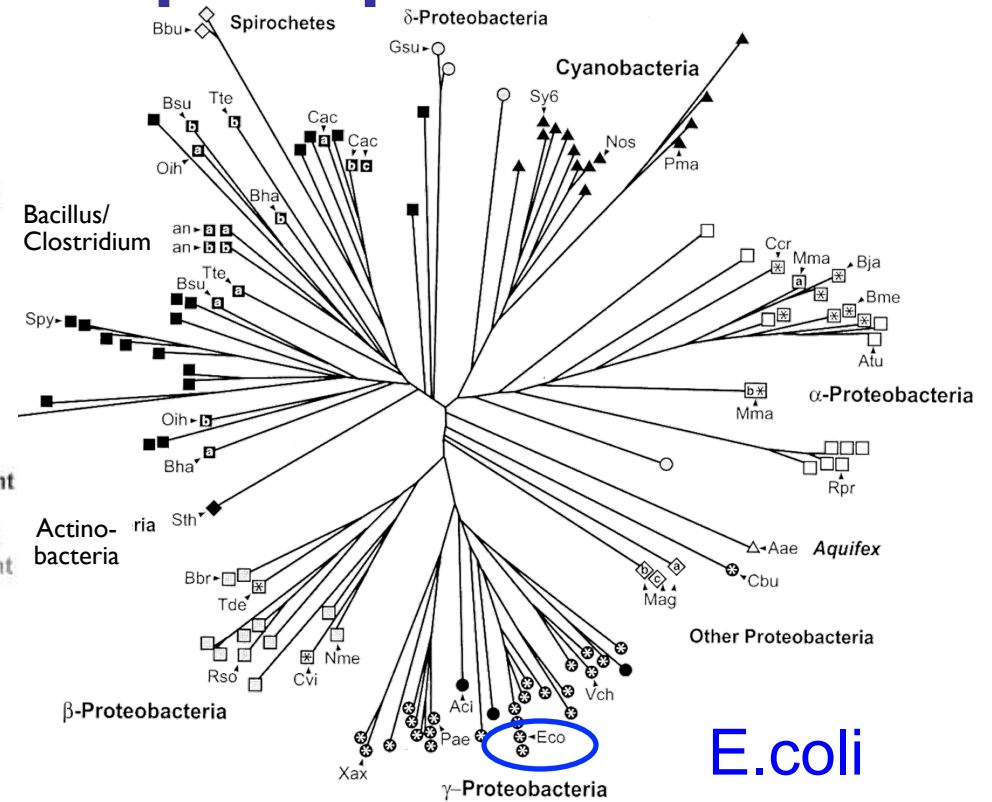
structured

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 wide-spread ncRNA expression

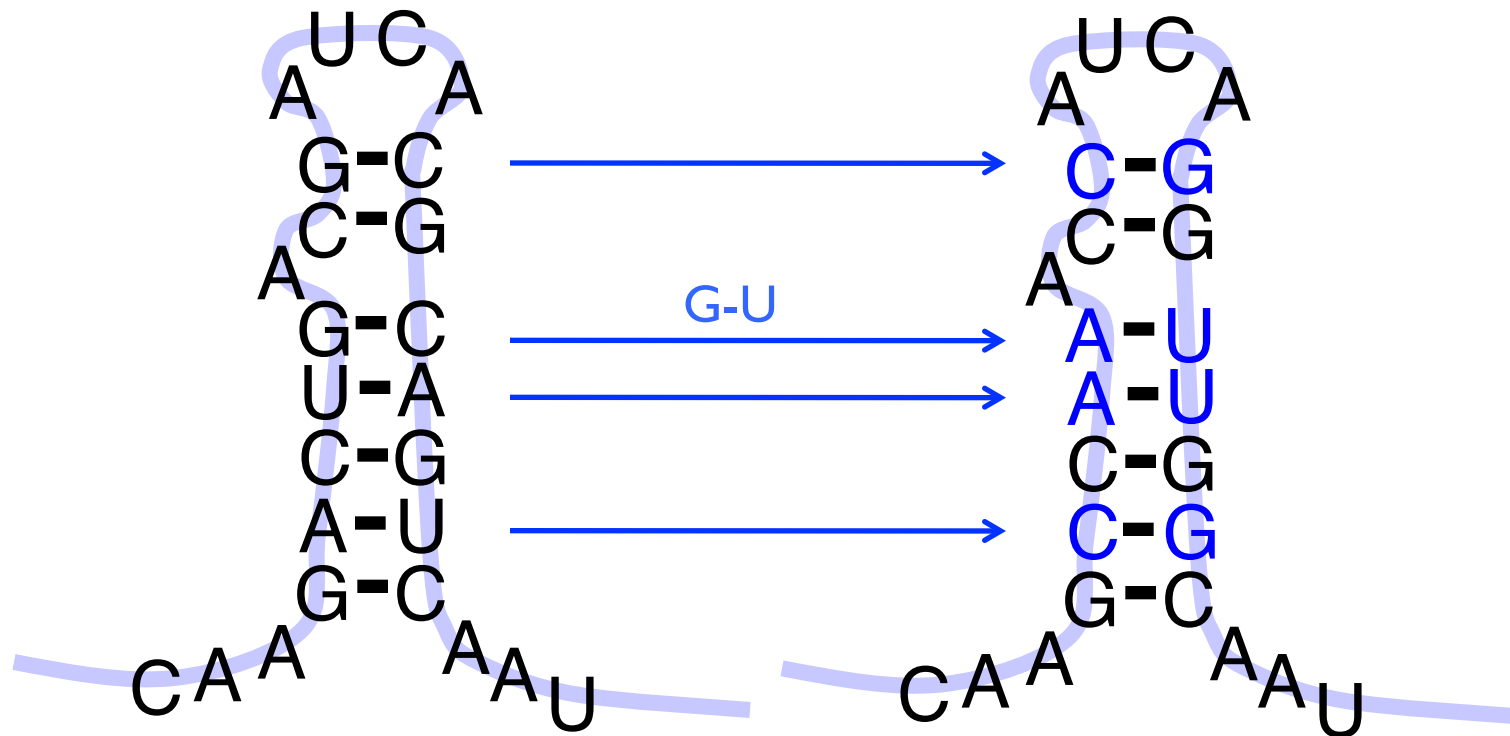
At a minimum, a vast evolutionary substrate

New technology (e.g. RNAseq) exposing
more

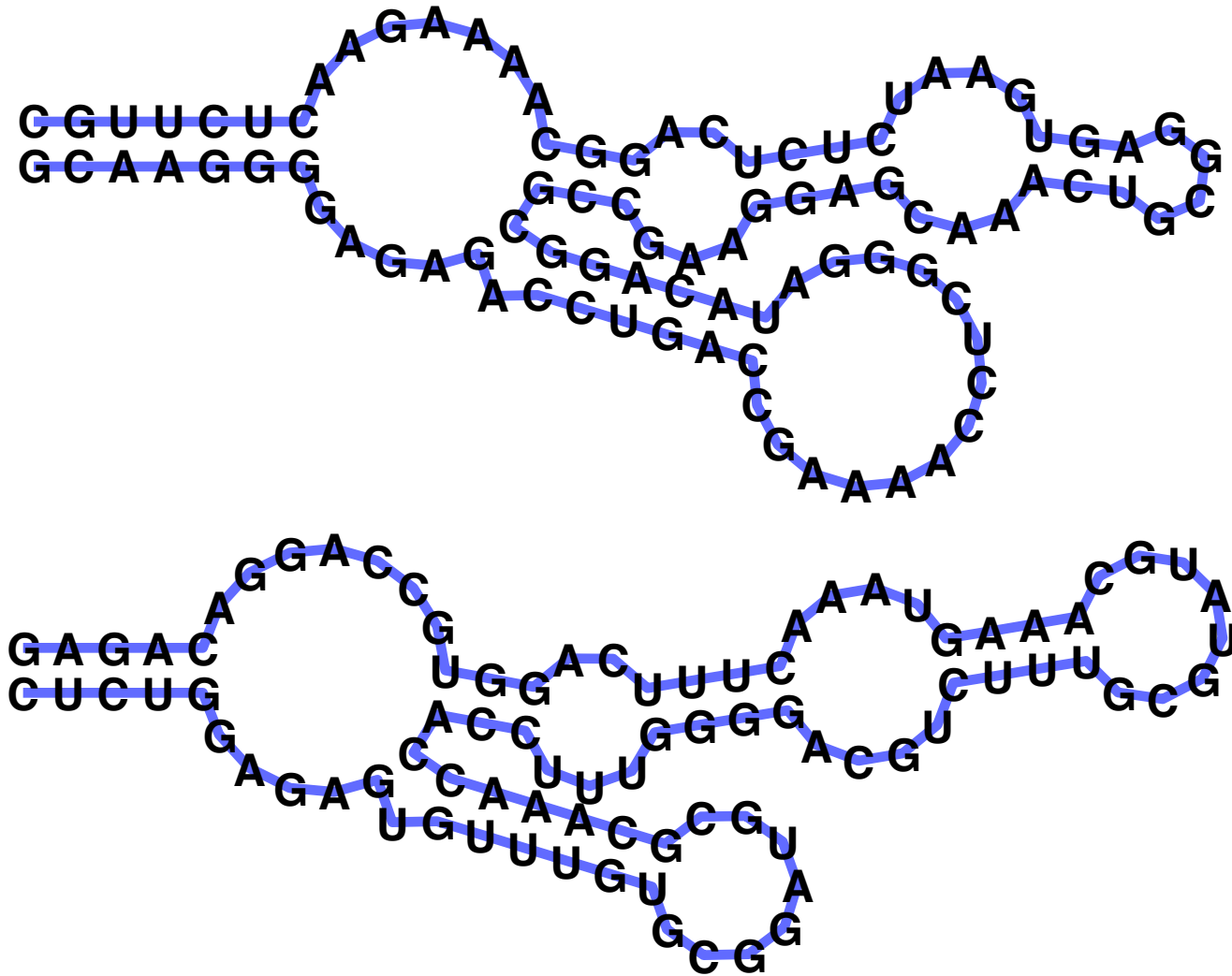
How do you recognize an interesting one?

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*₁₀₃

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

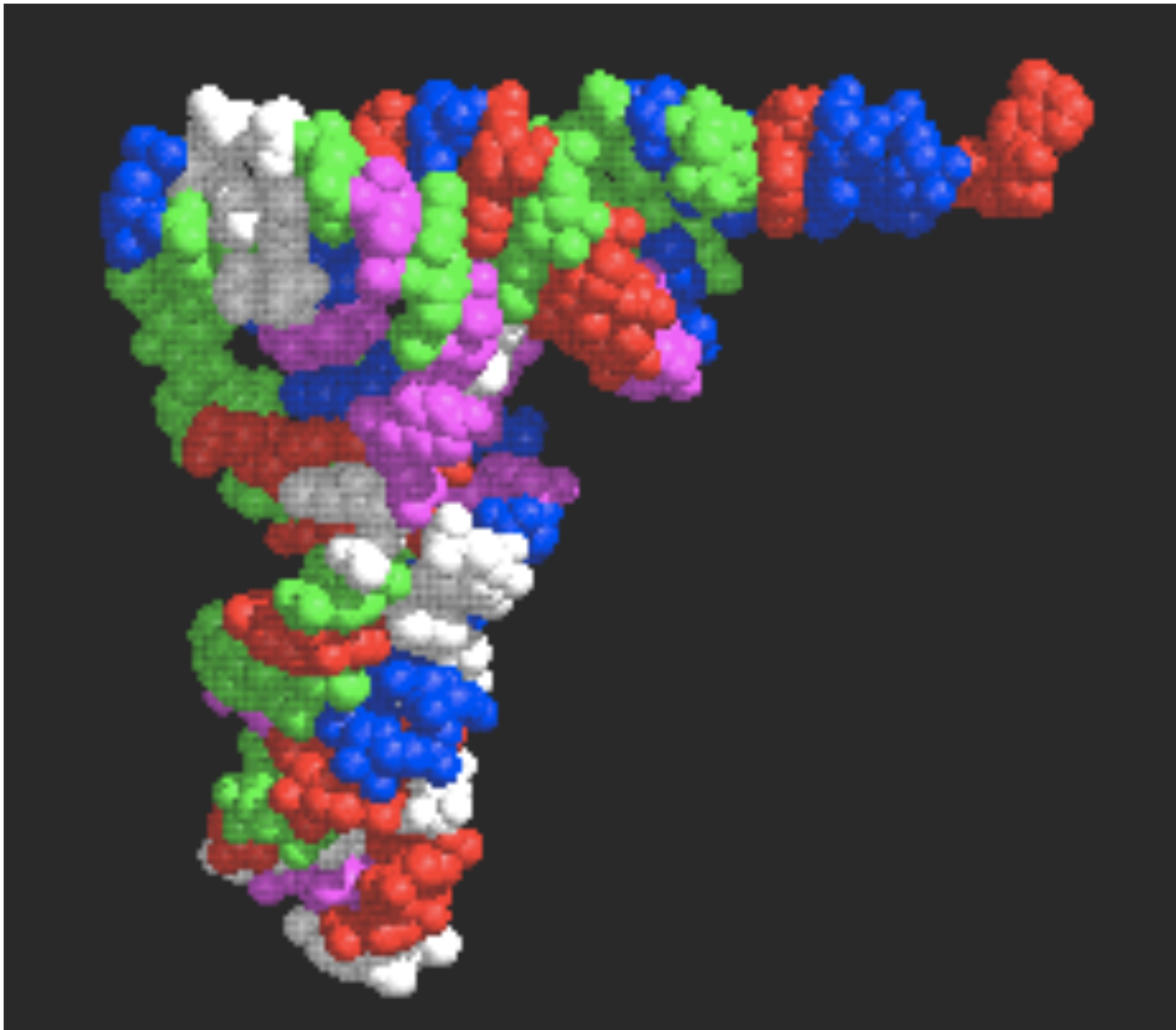
~ 2 kcal/mole

“Wobble Pair” G - U

~1 kcal/mole

Non-canonical Pairs (esp. if modified)

tRNA 3d Structure



tRNA - Alt. Representations

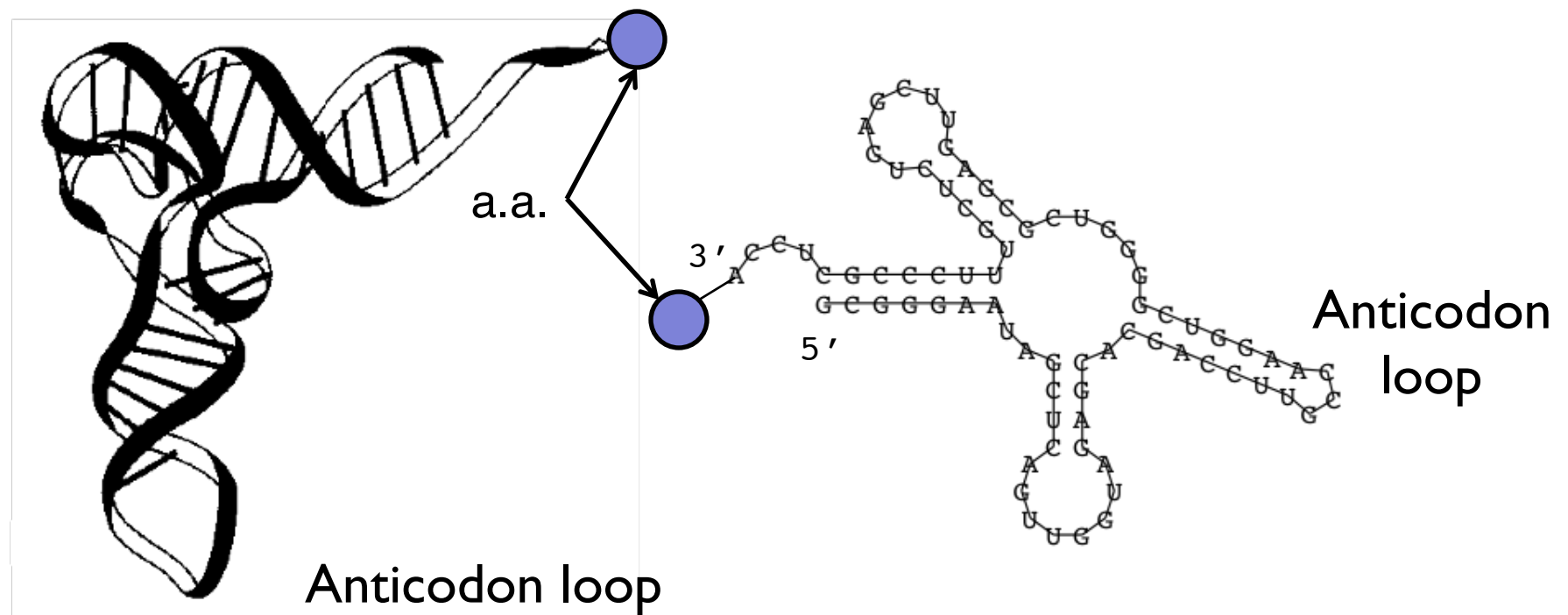
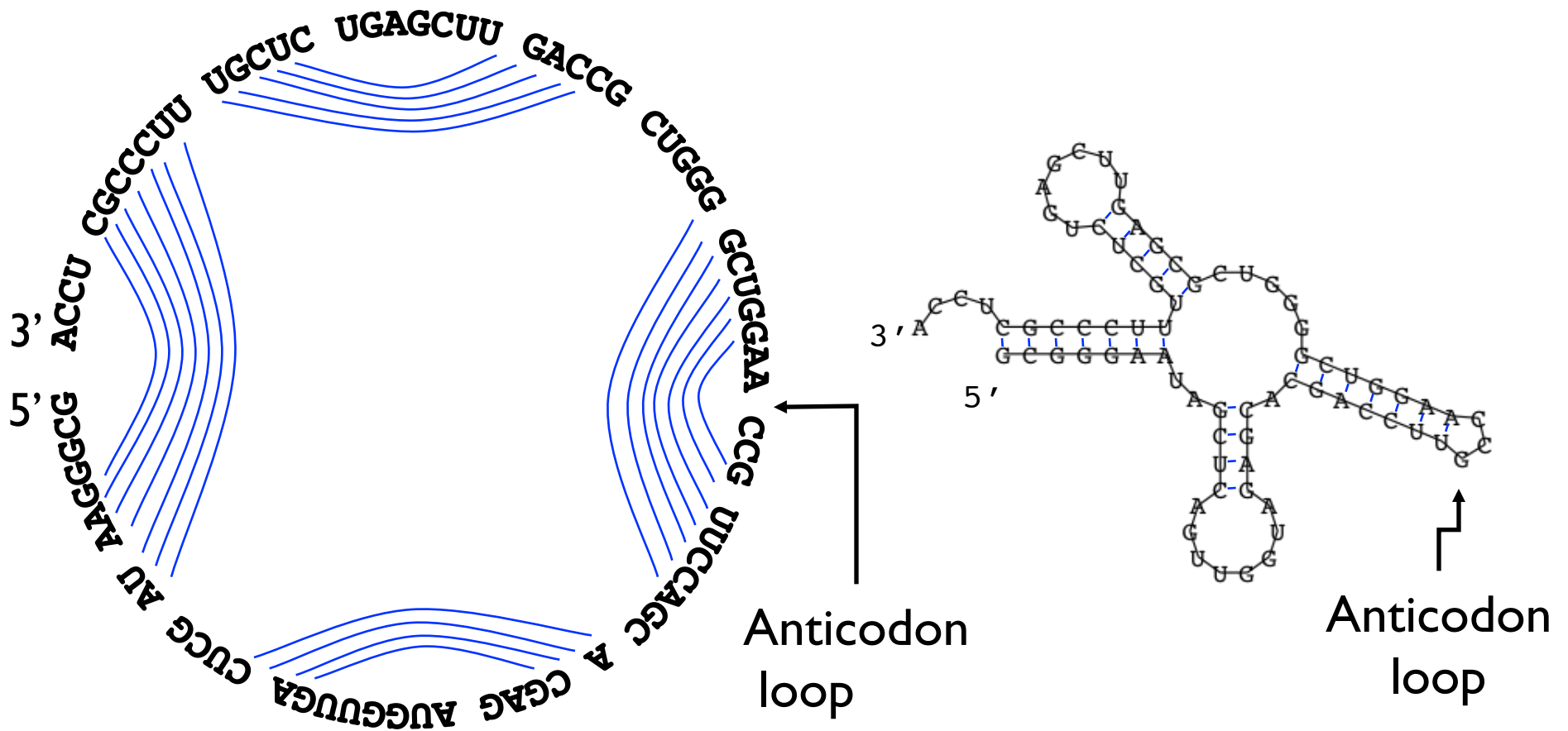


Figure 1: a) The spatial structure of the phenylalanine tRNA from yeast

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 $\{A, C, G, T/U\}$

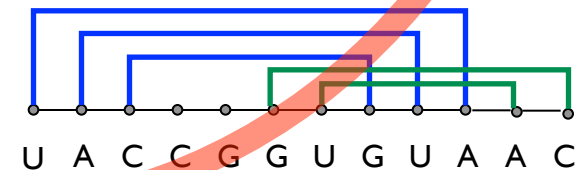
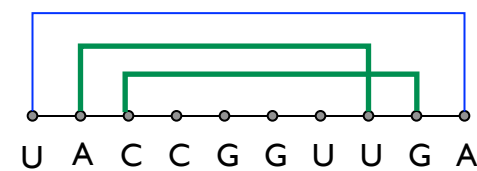
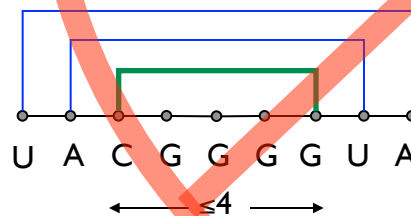
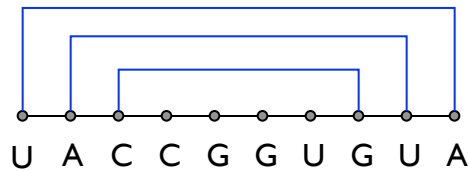
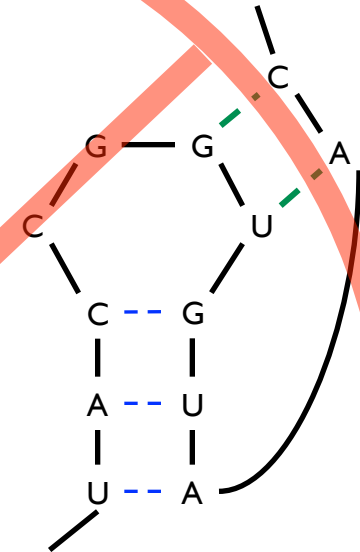
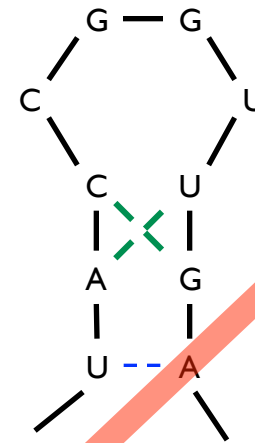
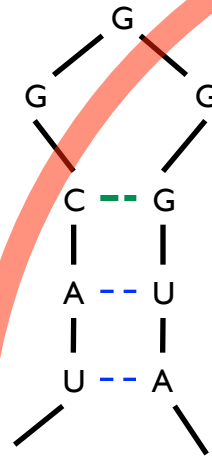
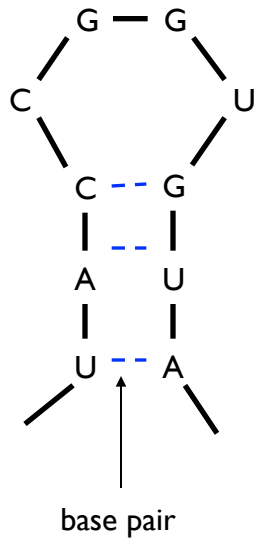
A **Secondary Structure** is a set of pairs $i \bullet j$ s.t.

$i < j-4$, 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.”

RNA Secondary Structure: Examples



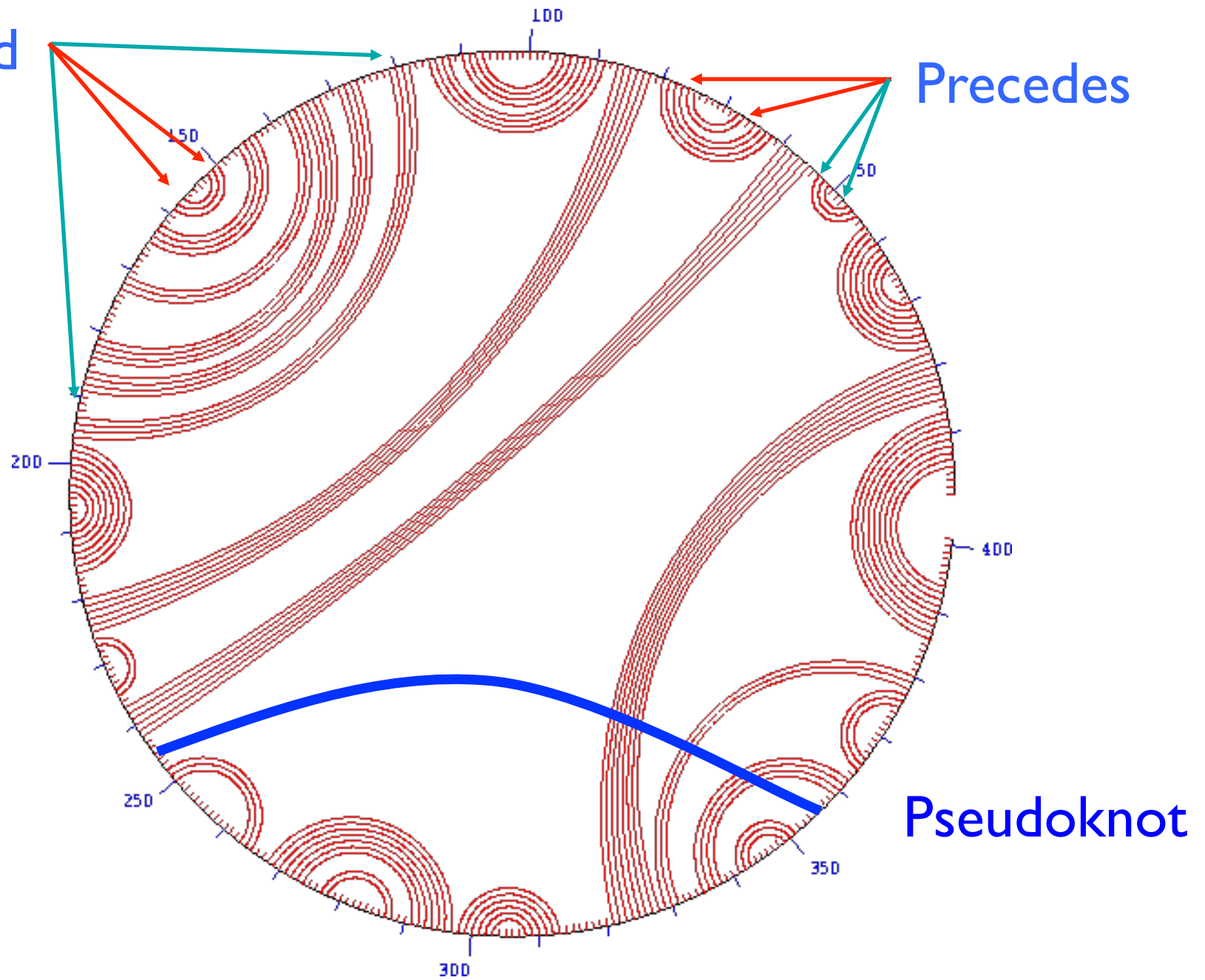
ok

sharp turn

crossing

Nested

Precedes



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

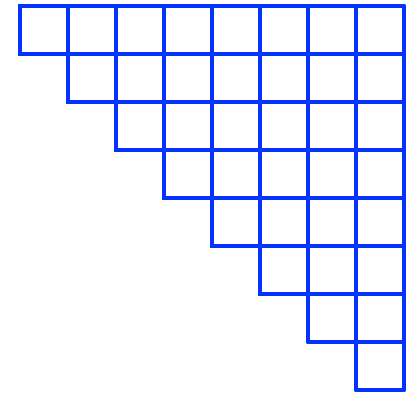
Nussinov: Max Pairing

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

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

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

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

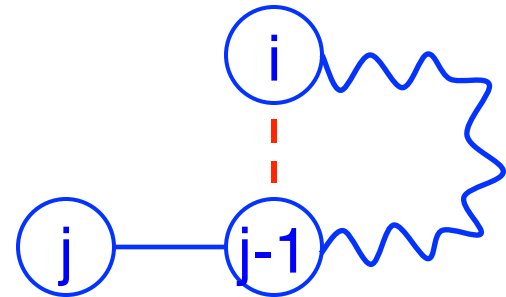


“Optimal pairing of $r_i \dots r_j$ ”

Two possibilities

j Unpaired:

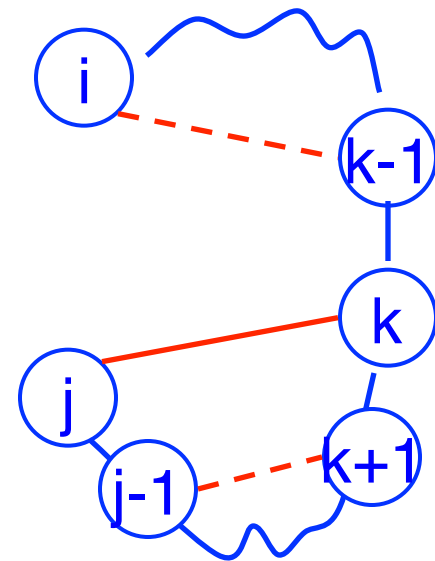
Find best pairing of $r_i \dots r_{j-1}$



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:

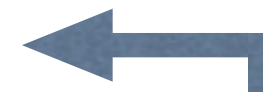
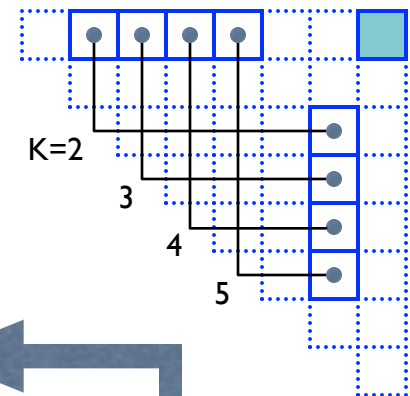
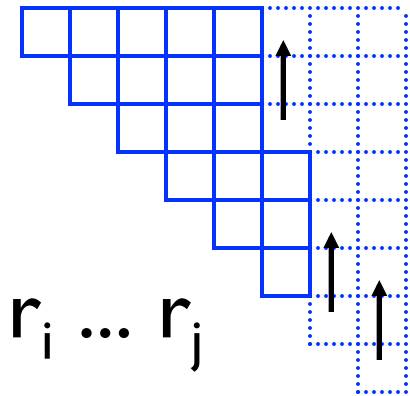
A Computation Order

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

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

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

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



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”