

**CSE 527**  
**Autumn 2007**  
**Lectures 17-18**

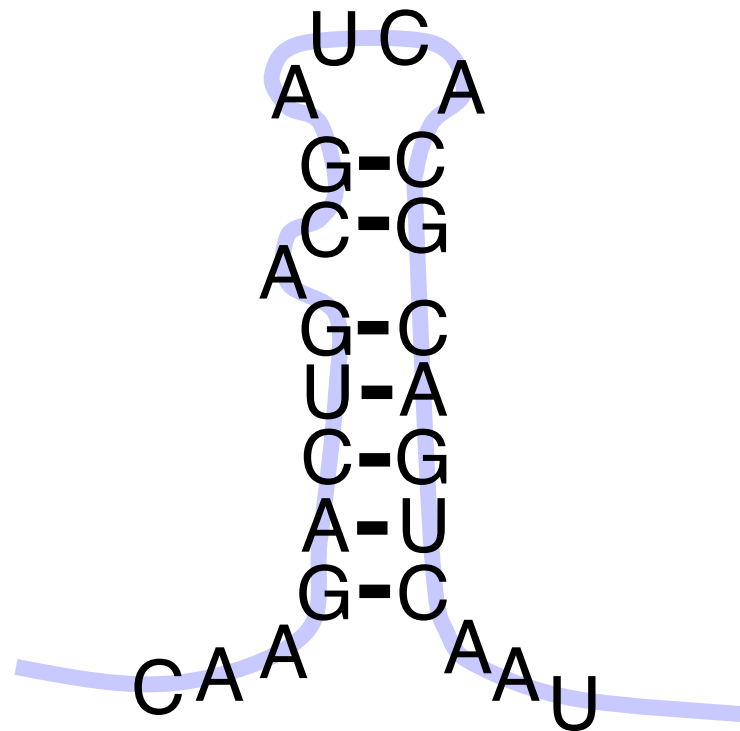
**RNA**  
**Secondary Structure Prediction**

# RNA Secondary Structure: RNA makes helices too

Base pairs

A-U

C-G







# 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 can form complex structures

RNA enzymes exist (ribozymes)

The “RNA world” hypothesis:  
1st life was RNA-based

# Outline

Biological roles for RNA

What is “secondary structure?”

How is it represented?

Why is it important?

Examples

Approaches

# RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

# RNA Pairing

## Watson-Crick Pairing

C - G

~ 3 kcal/mole

A - U

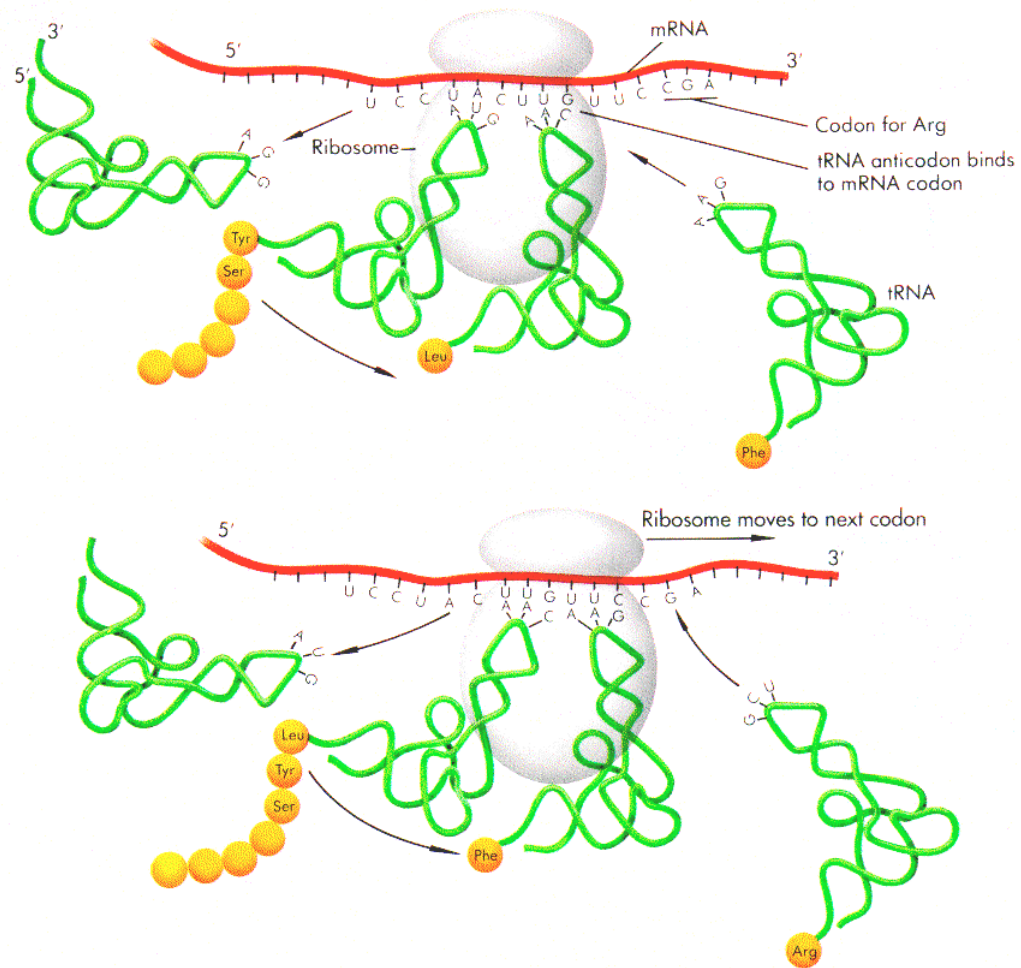
~ 2 kcal/mole

“Wobble Pair” G - U

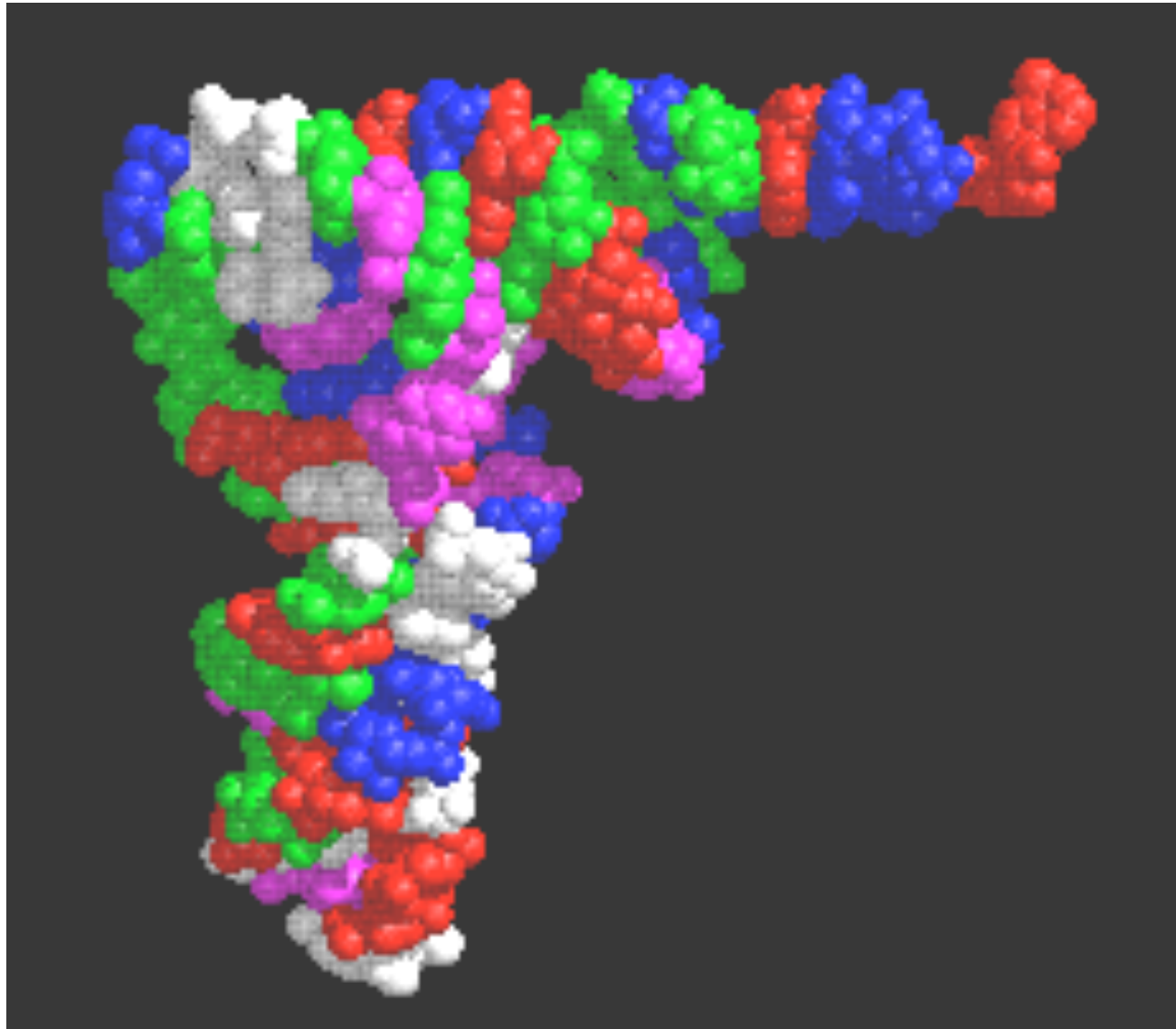
~1 kcal/mole

Non-canonical Pairs (esp. if modified)

# Ribosomes

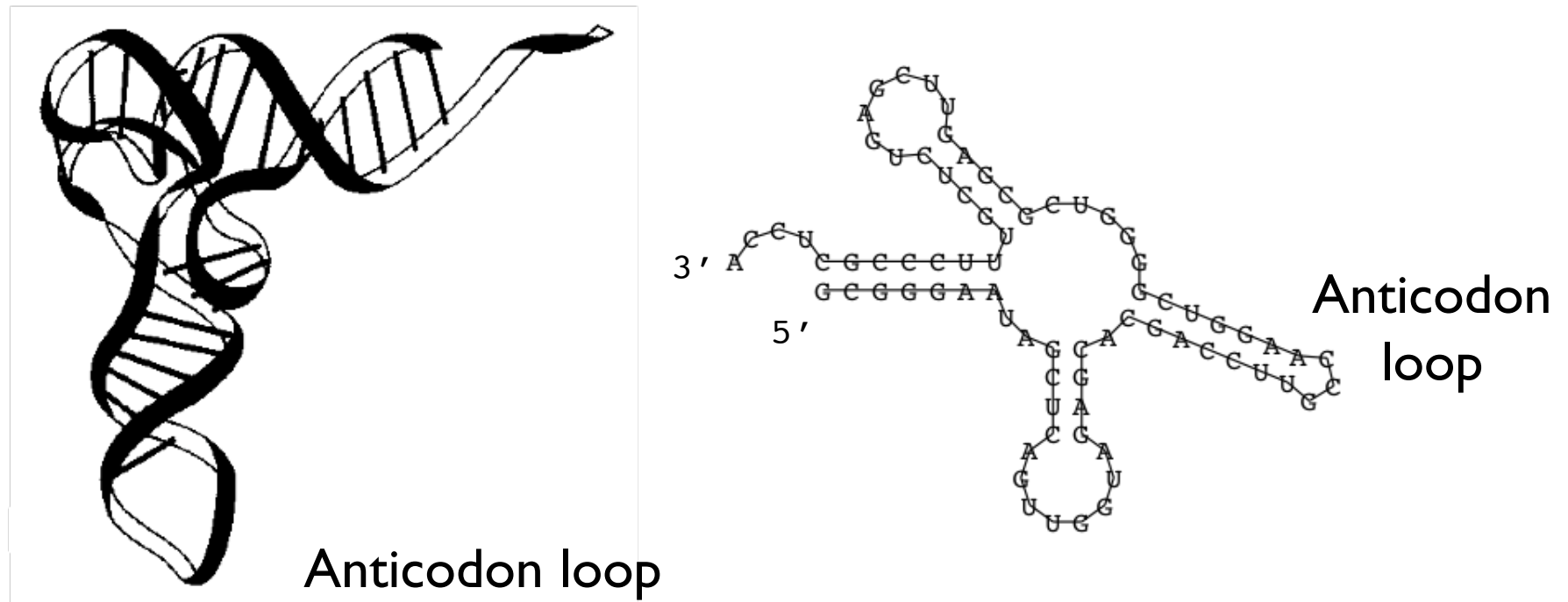


# tRNA 3d Structure





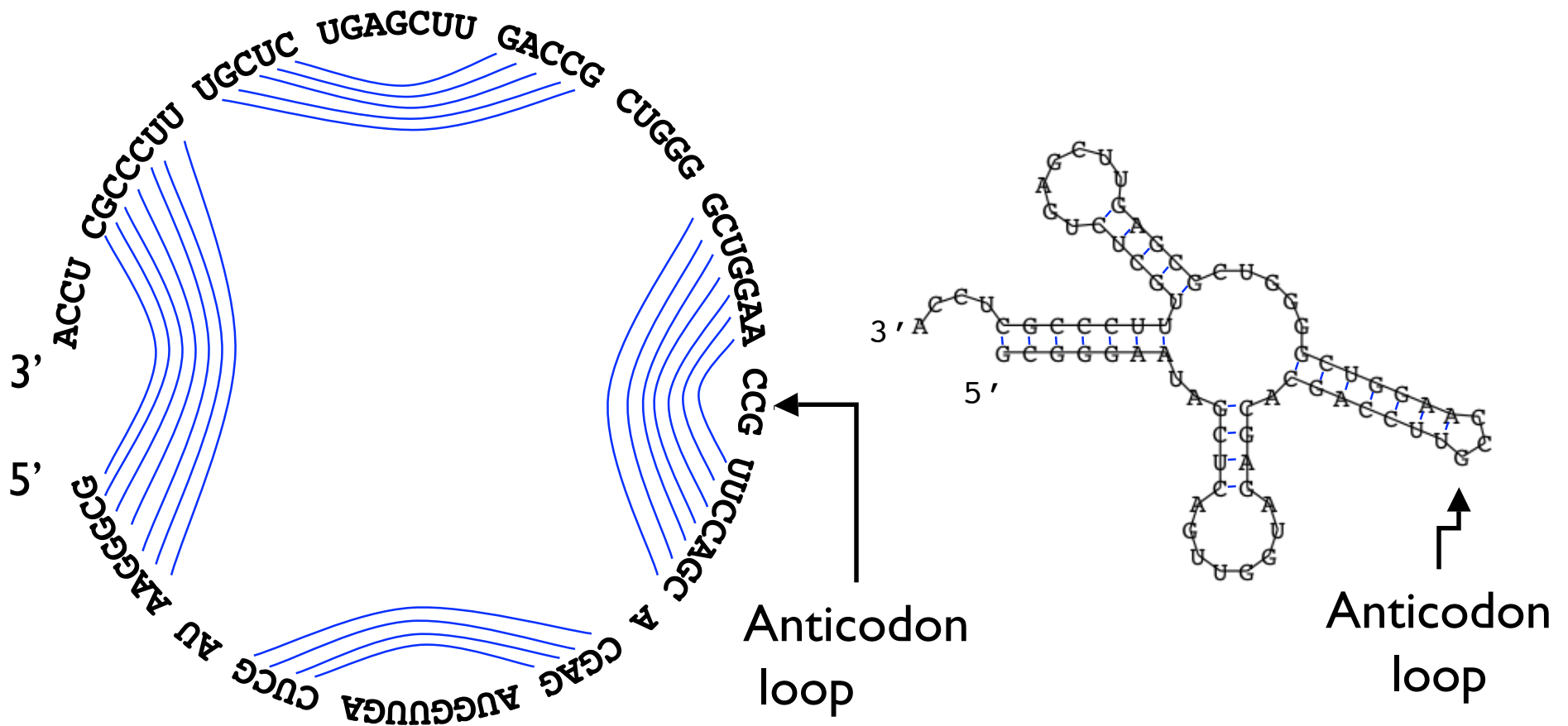
# tRNA - Alt. Representations



**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



# “Classical” RNAs

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

RNase MRP - rRNA processing; mito. rep. (~225 nt)

SRP - signal recognition particle; membrane targeting  
(~100-300 nt)

SECIS - selenocysteine insertion element (~65nt)

6S - ? (~175 nt)

# Semi-classical RNAs

(discovery in mid 90's)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)

# Recent discoveries

microRNAs (Nobel prize 2006, Fire & Mello)

riboswitches

many ribozymes

regulatory elements

...

*Hundreds* of families

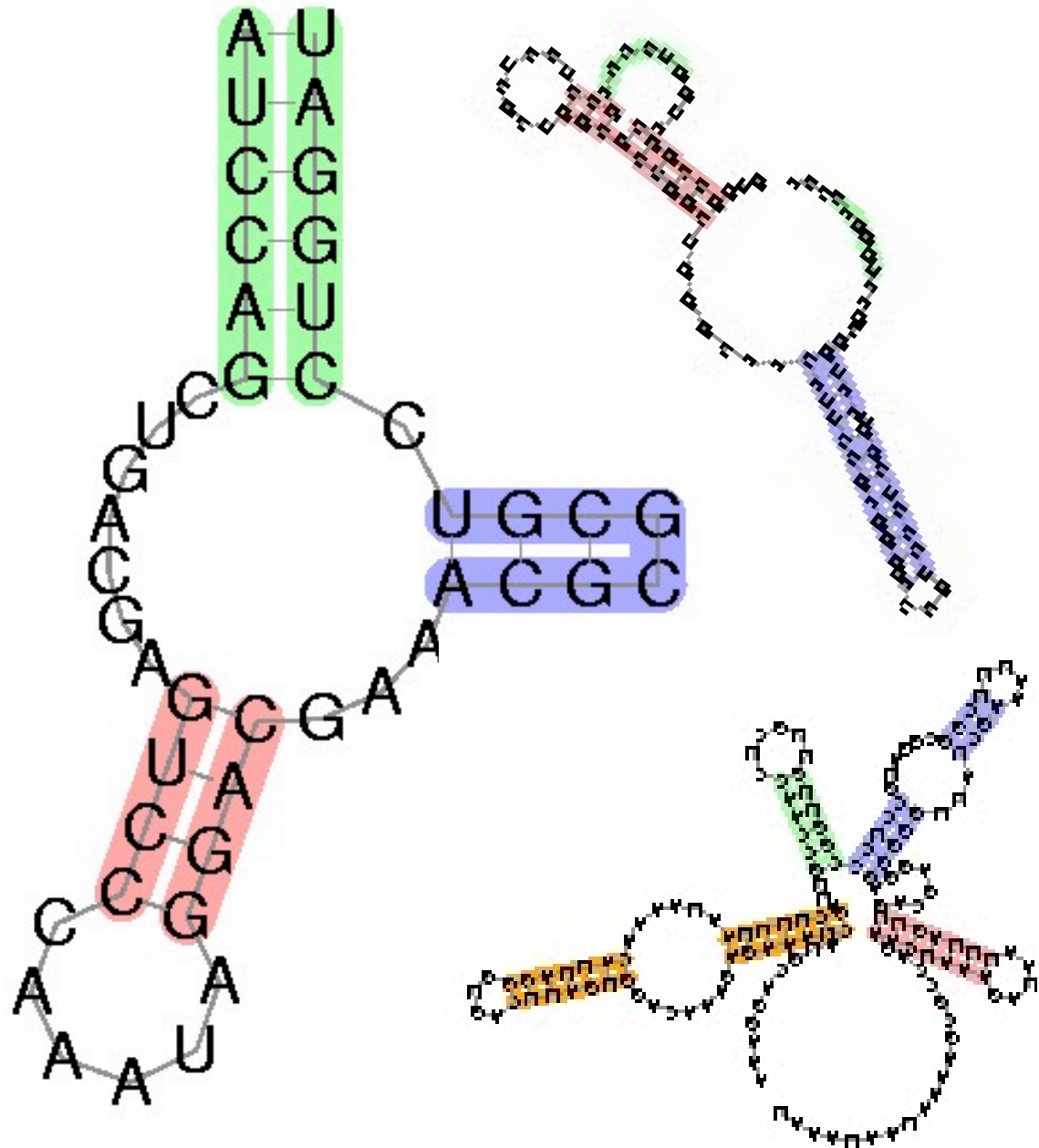
Rfam release 1, 1/2003: 25 families, 55k instances

Rfam release 7, 3/2005: 503 families, 300k instances

# Why?

RNA's fold,  
and function

Nature uses  
what works





# Noncoding RNAs

Dramatic discoveries in last 5 years

*100s of new families*

Many roles: Regulation, transport, stability, catalysis, ...

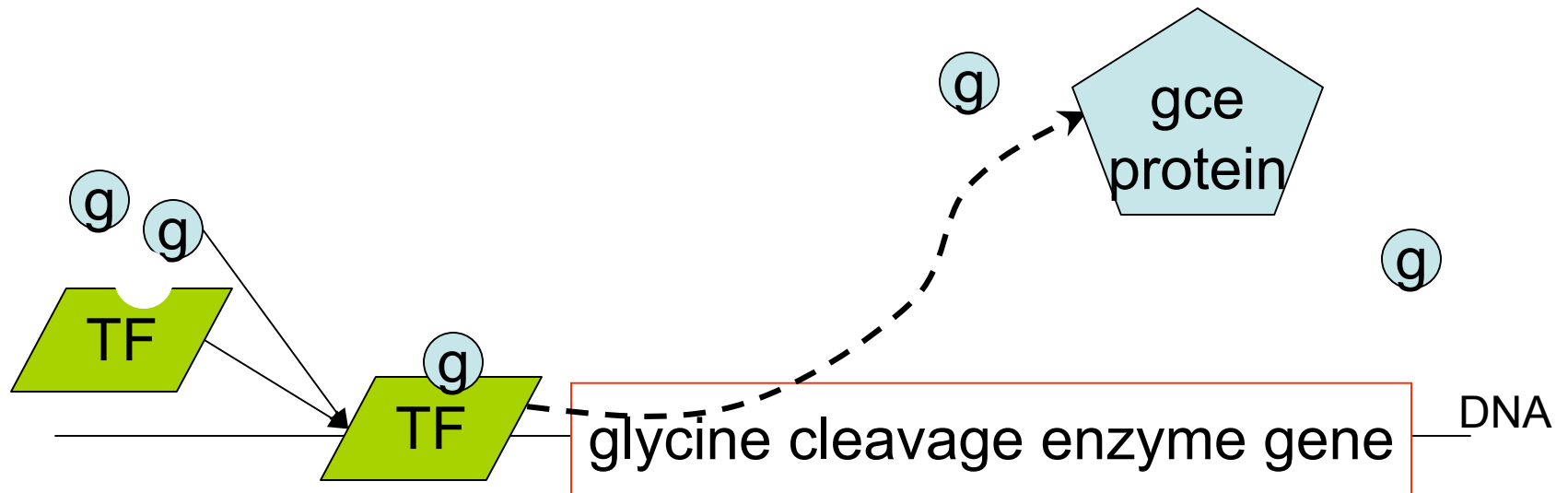
*1% of DNA codes for protein, but 90% of it is copied into RNA, i.e. ncRNA >> mRNA*

*Significance unclear, controversial*

# Example: Glycine Regulation

How is glycine level regulated?

Plausible answer:

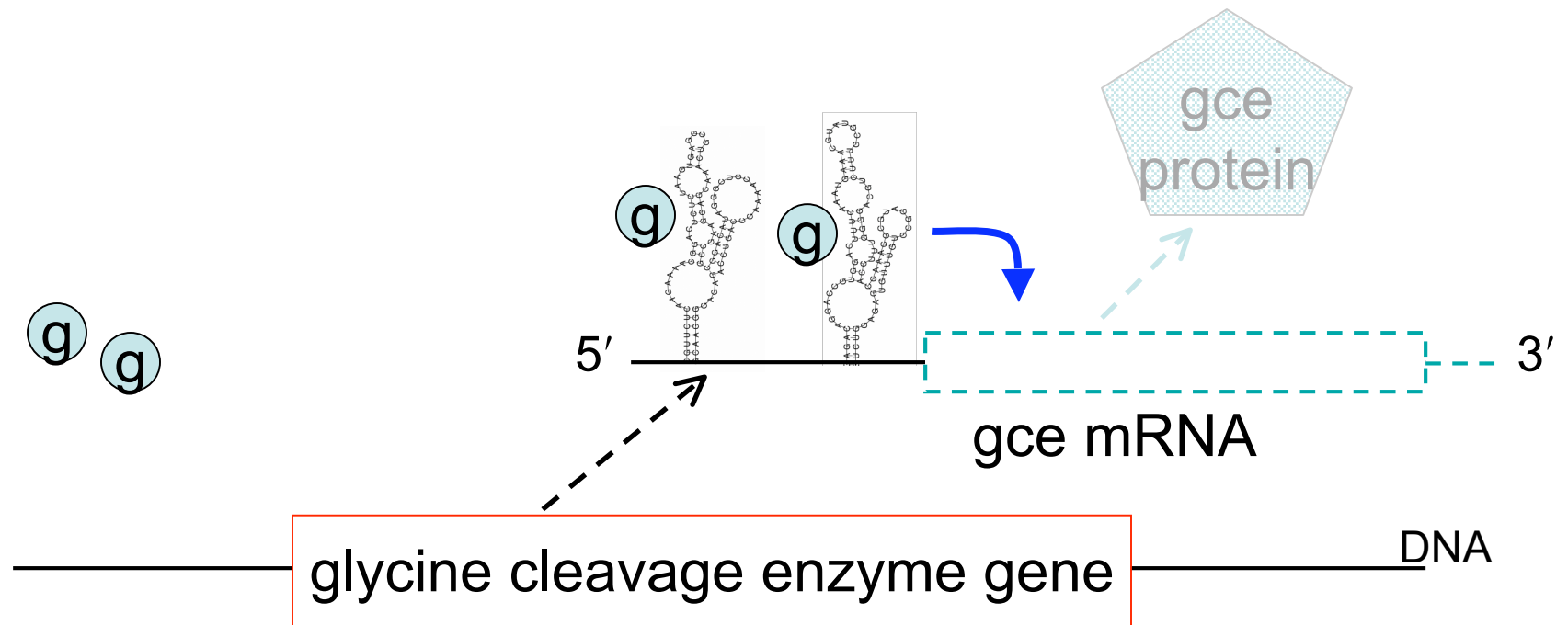


transcription factors (proteins) bind to DNA to turn nearby genes on or off

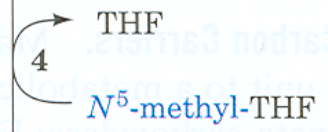
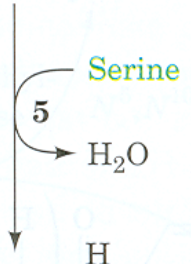
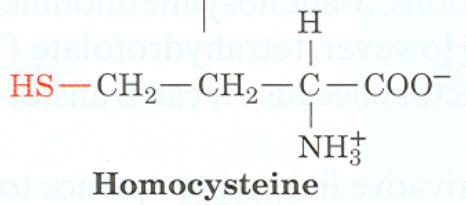
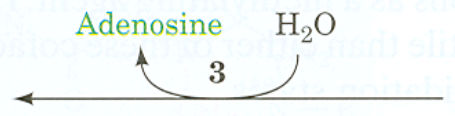
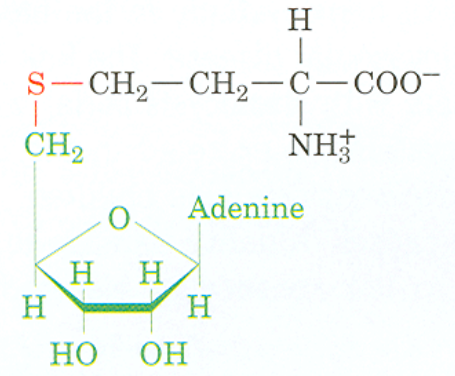
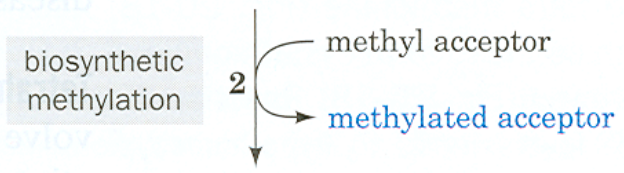
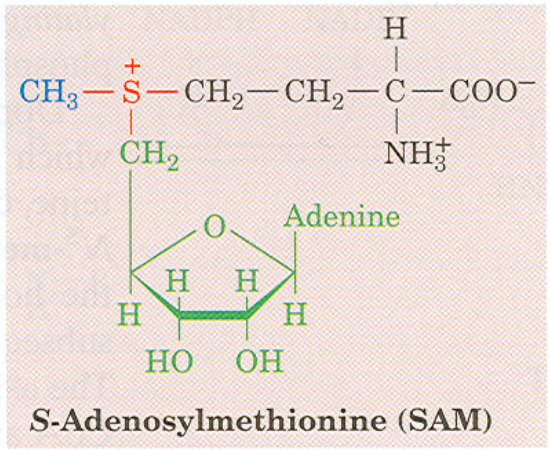
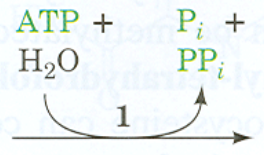
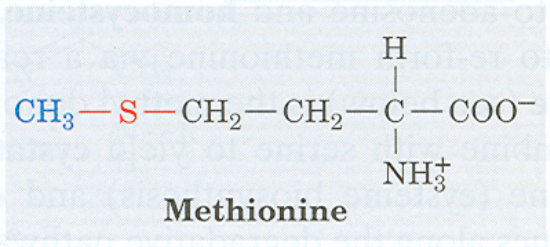


# The Glycine Riboswitch

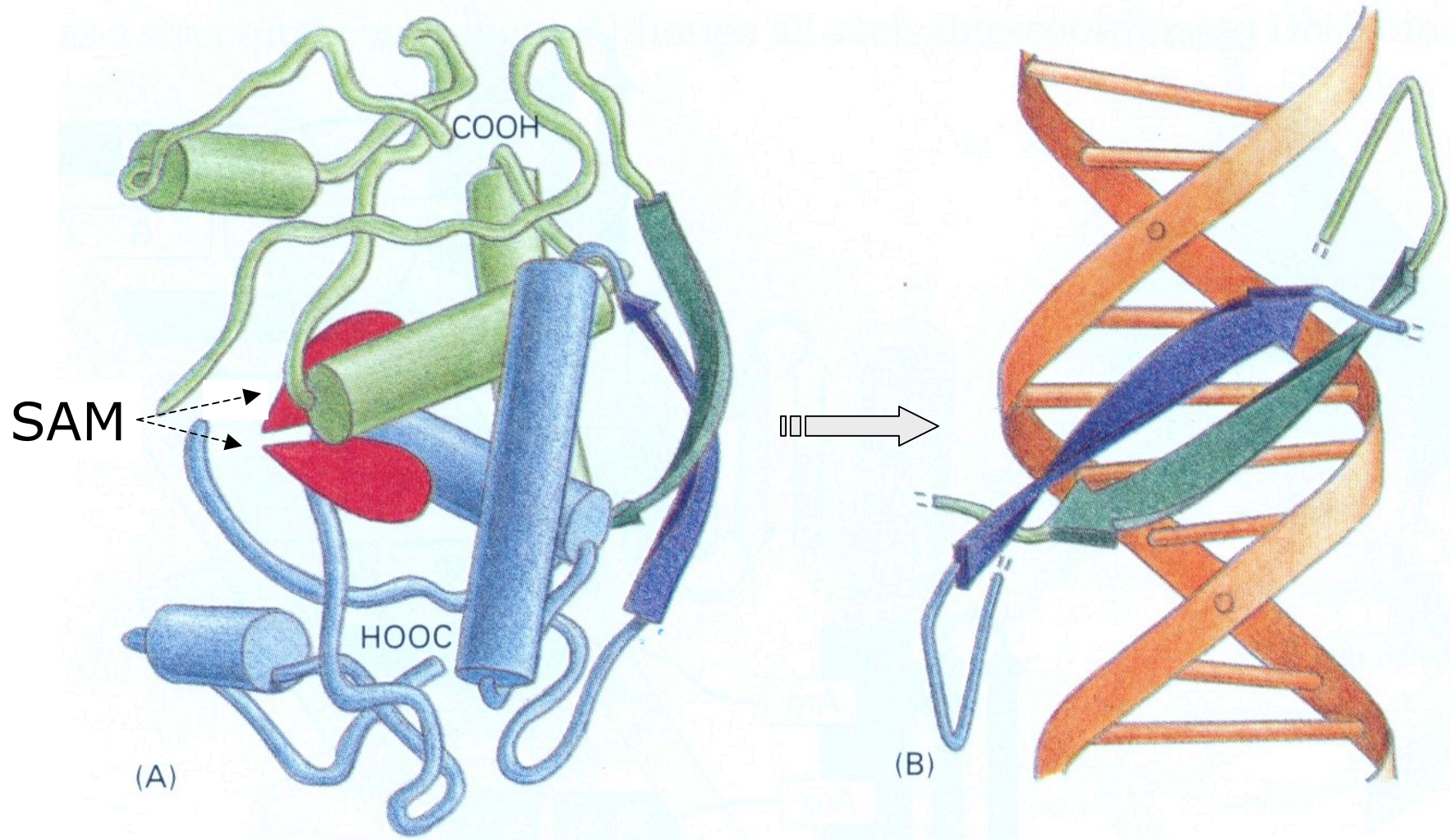
Actual answer (in many bacteria):



Mandal et al. Science 2004



# Gene Regulation: The MET Repressor



Protein

Alberts, et al, 3e.

DNA

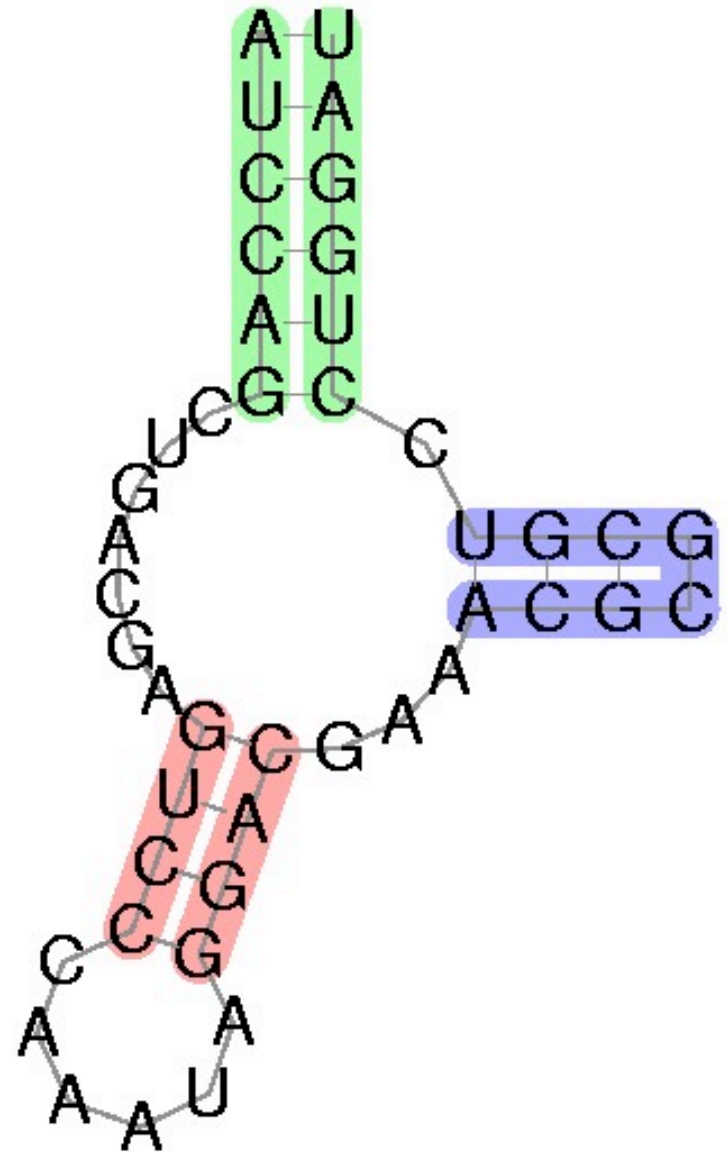






# The Hammerhead Ribozyme

Involved in “rolling circle replication” of viruses.



# Wanted

Good structure prediction tools

Good motif descriptions/models

Good, fast search tools

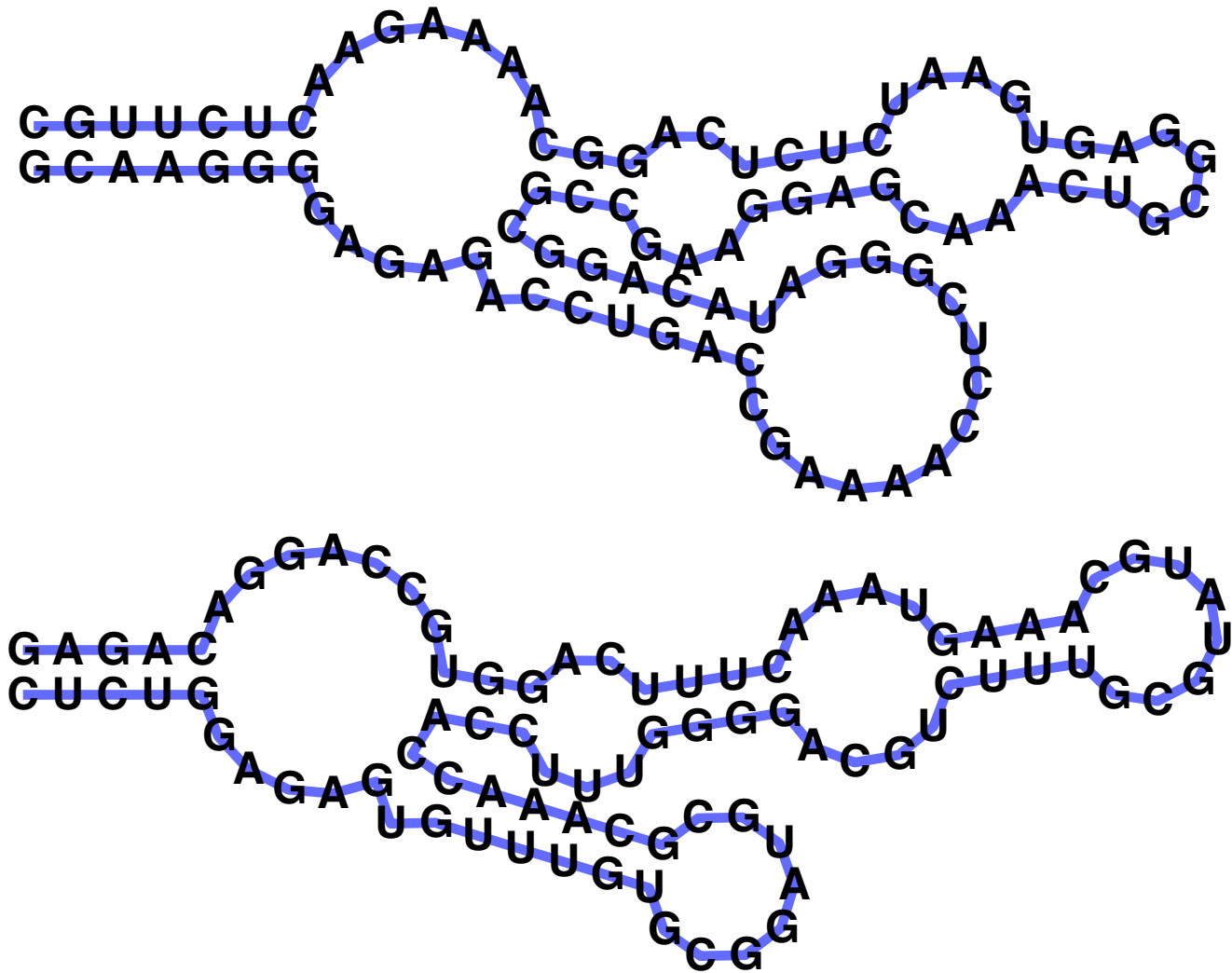
(“RNA BLAST”, etc.)

Good, fast motif discovery tools

(“RNA MEME”, etc.)

Importance of structure makes last 3 hard

# Why is RNA hard to deal with?



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



# Task I: Structure Prediction

# 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)

# Definitions

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

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

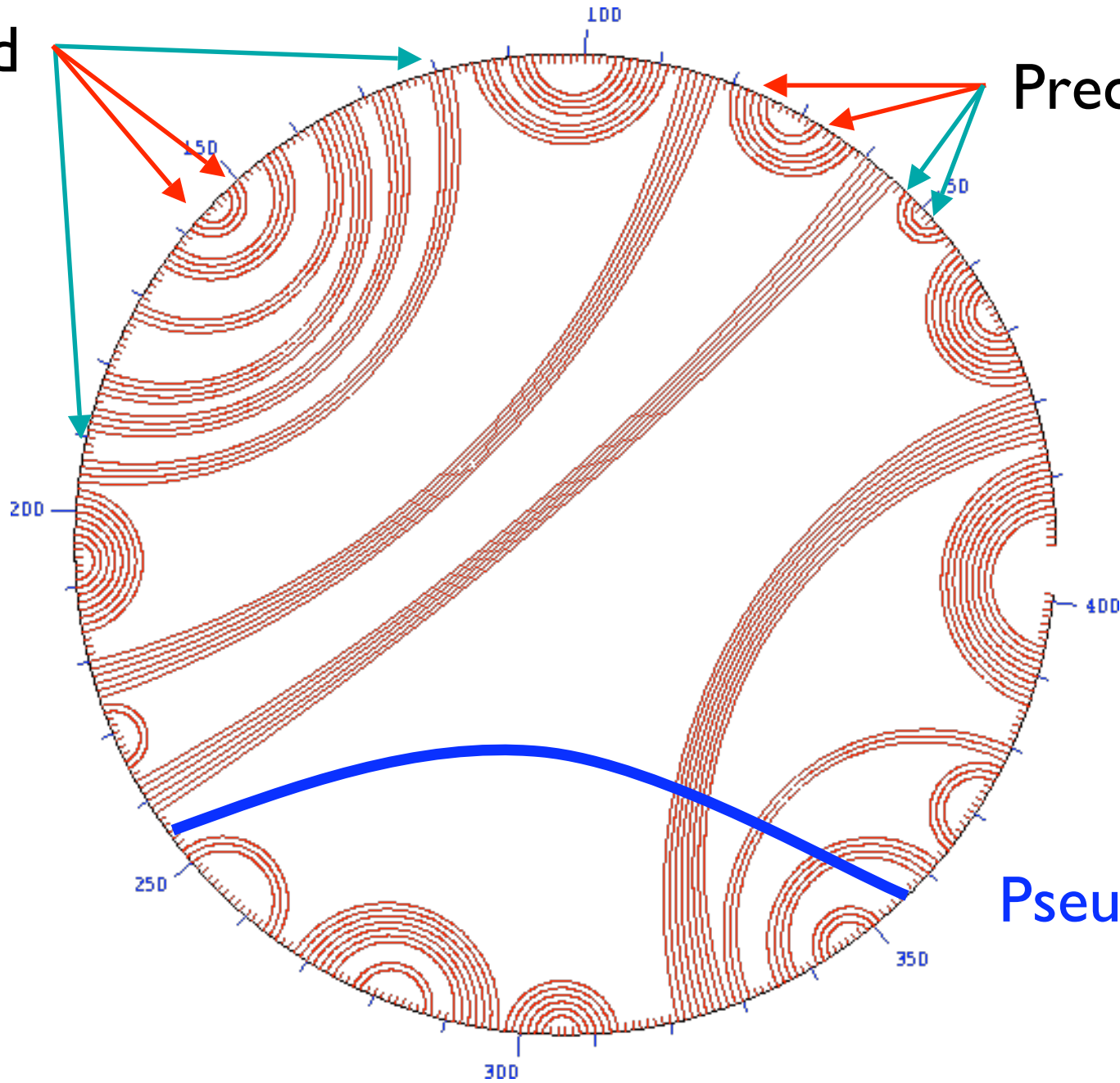
$i < j-4$ , and  $\left. \vphantom{i < j-4} \right\}$  no sharp turns

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

$j < i'$ , or  
 $i < i' < j' < j$   $\left. \vphantom{i < i' < j' < j} \right\}$  2nd pair follows 1st, or  
is nested within it;  
no “pseudoknots.”

Nested

Precedes



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

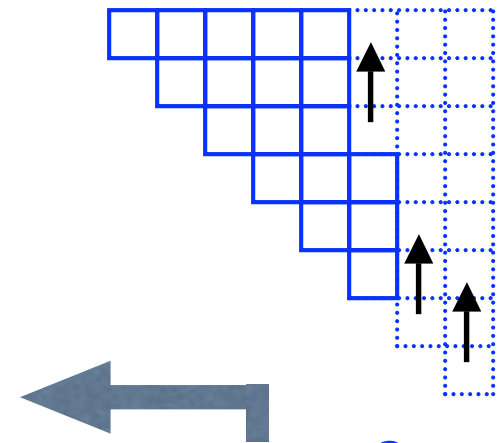
# Nussinov: Max Pairing

$B(i,j)$  = # 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:

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



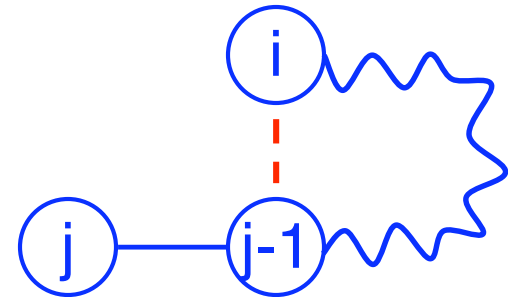
Time:  $O(n^3)$

# “Optimal pairing of $r_i \dots r_j$ ”

Two possibilities

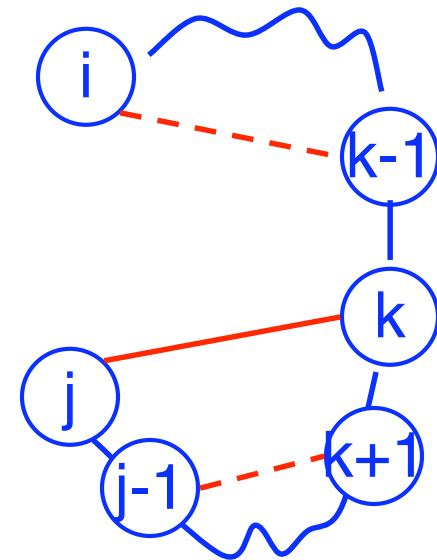
J Unpaired:

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



J Paired:

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?




# 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 j-k pair

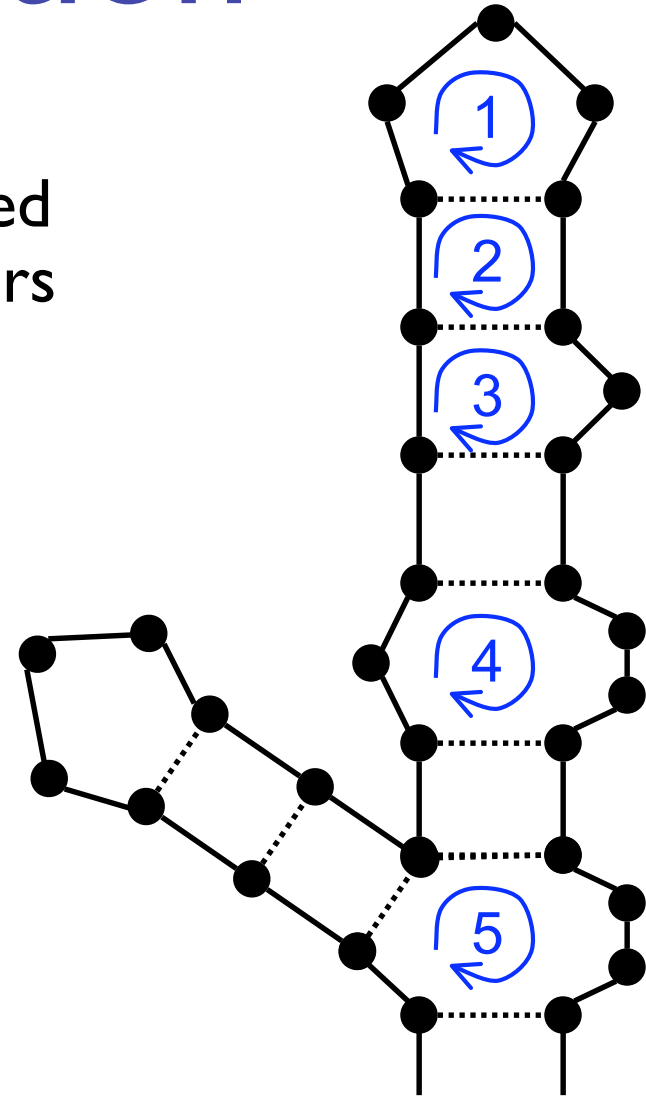
Time:  $O(n^3)$  

# Loop-based Energy Minimization

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

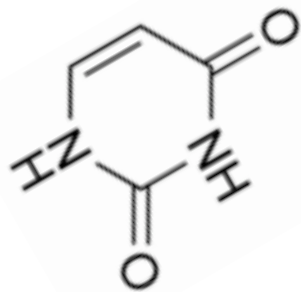
## Loop types

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

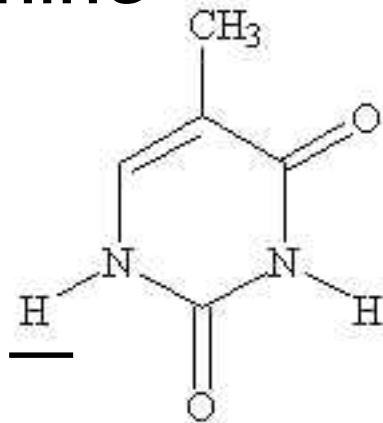


# Base Pairs and Stacking

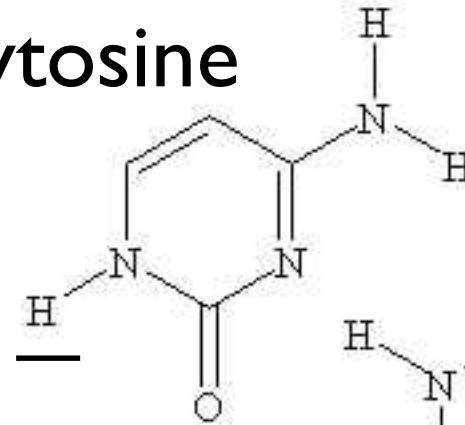
uracil



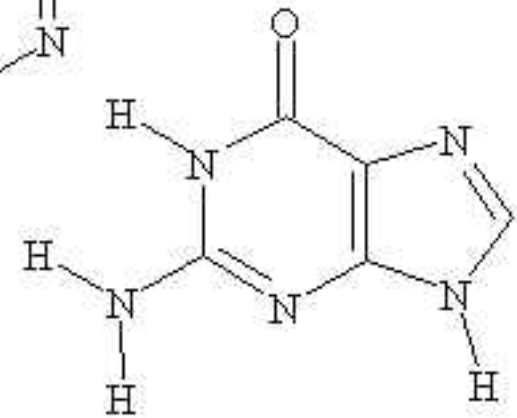
thymine



cytosine

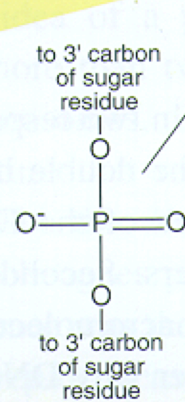
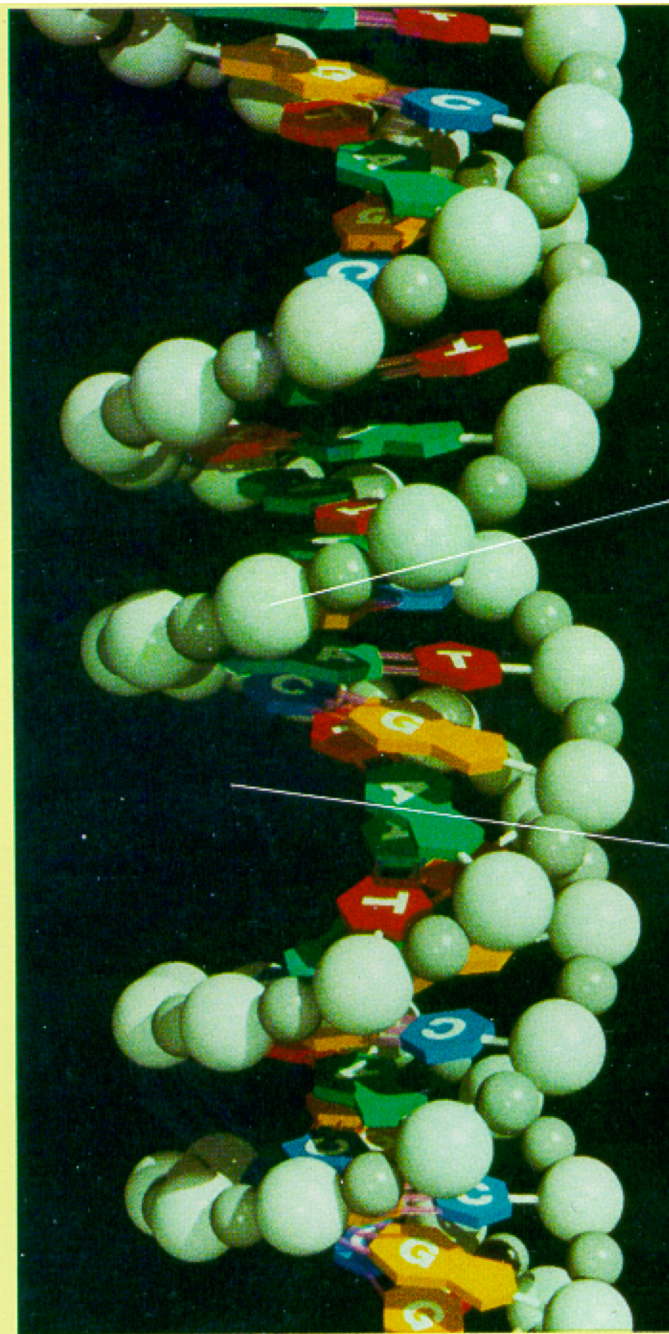


adenine

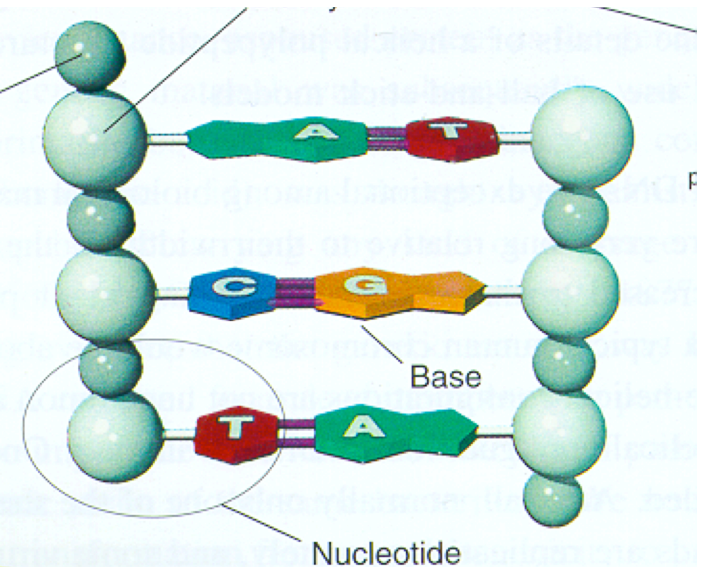


guanine

# The Double Helix



Phosphate group



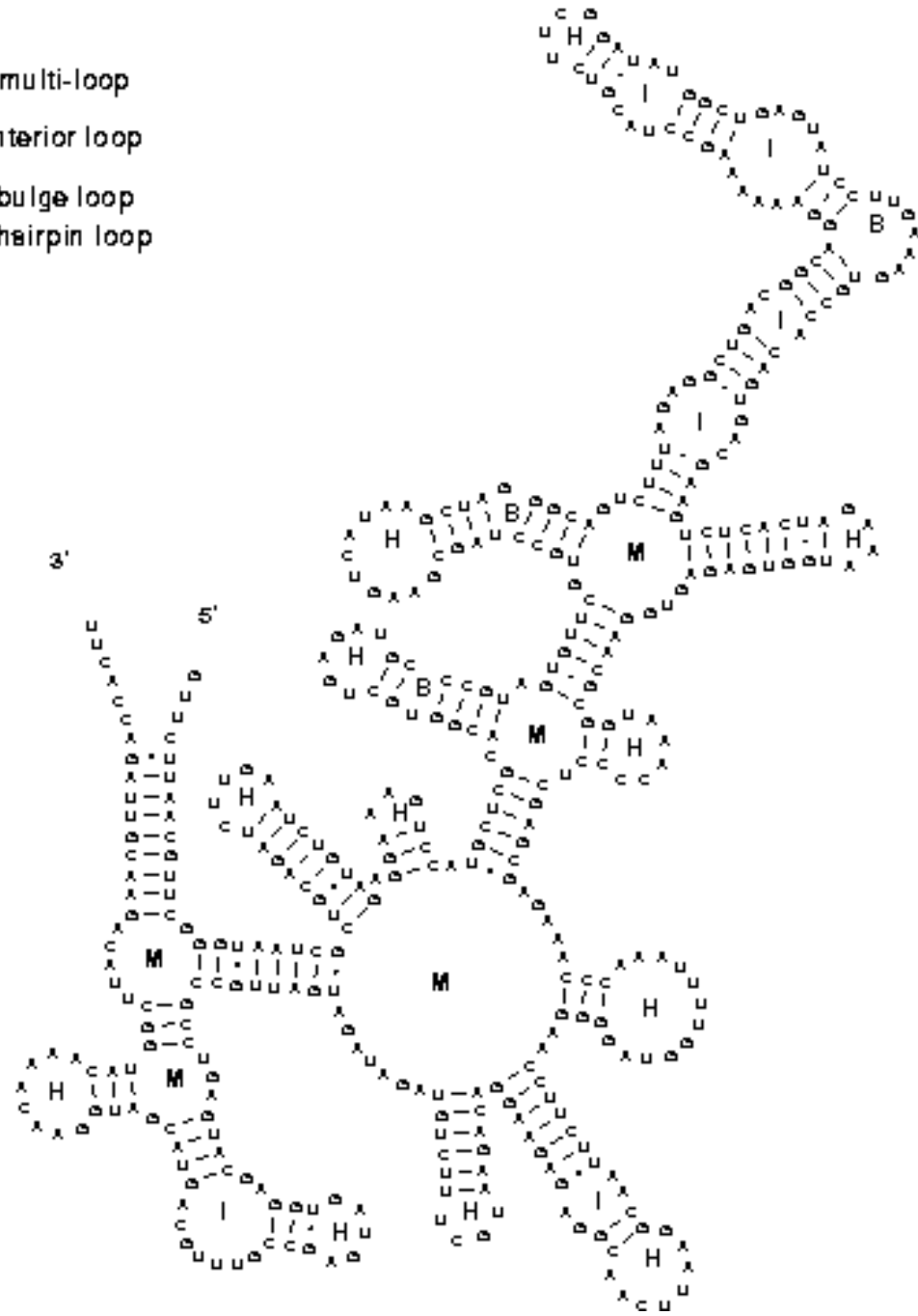
As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a) three complementary base pair chemist's viewpoint, each strand a polymer made up of four re-called deoxyribonucleotides

*Bacillus subtilis RNase P RNA*

- M** - multi-loop
- I** - interior loop
- B** - bulge loop
- H** - hairpin loop

# Loop Examples



# 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/  
interior    multi-  
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”



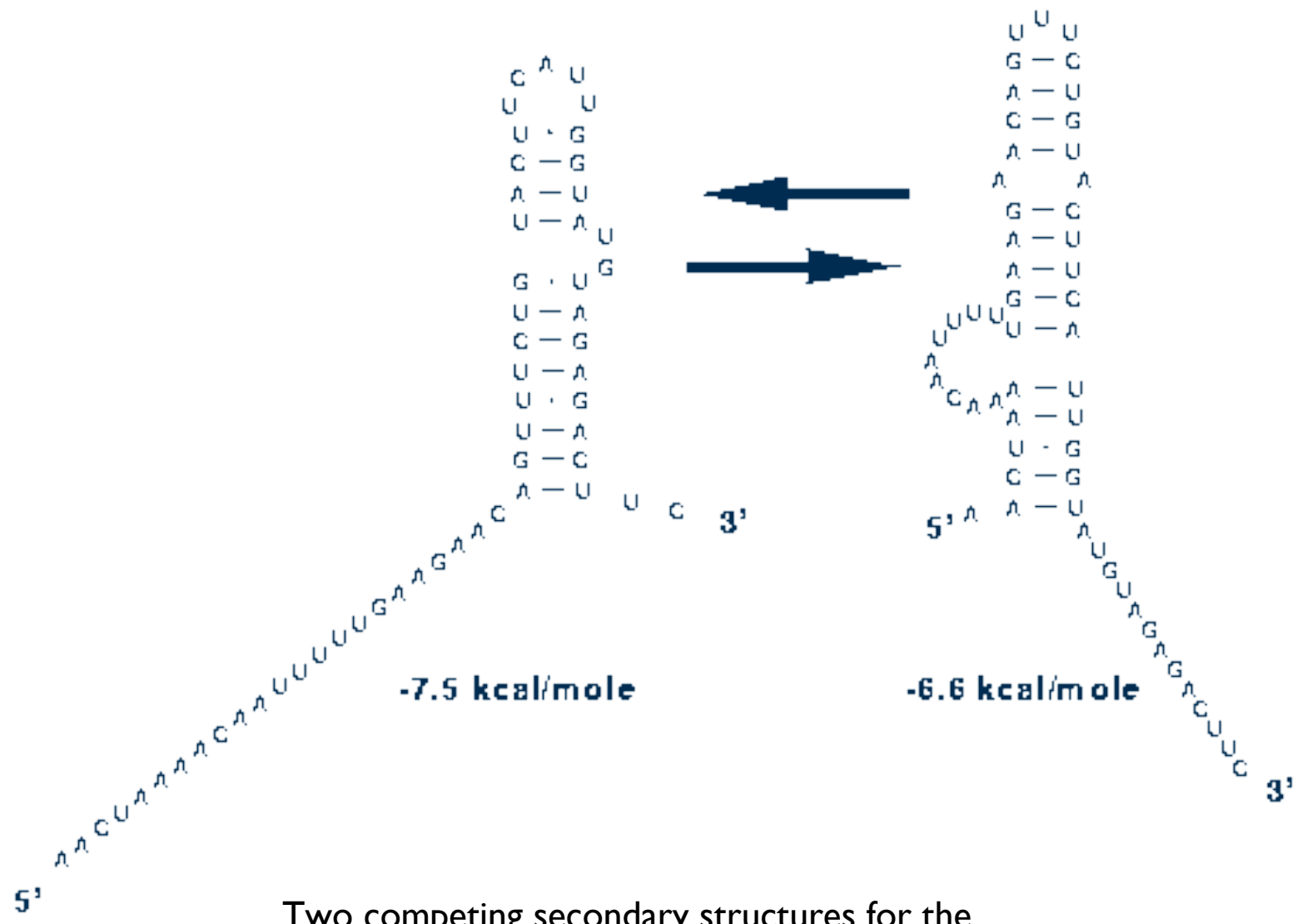
# Suboptimal Energy

There are always alternate folds with near-optimal energies. Thermodynamics: populations of identical molecules will exist in different folds; individual molecules even flicker among different folds

Mod to Zuker's algorithm finds subopt folds

McCaskill: more elaborate dyn. prog. algorithm calculates the "partition function," which defines the probability distribution over all these states.  
(Key addition: recurrence must count each possibility exactly once.)



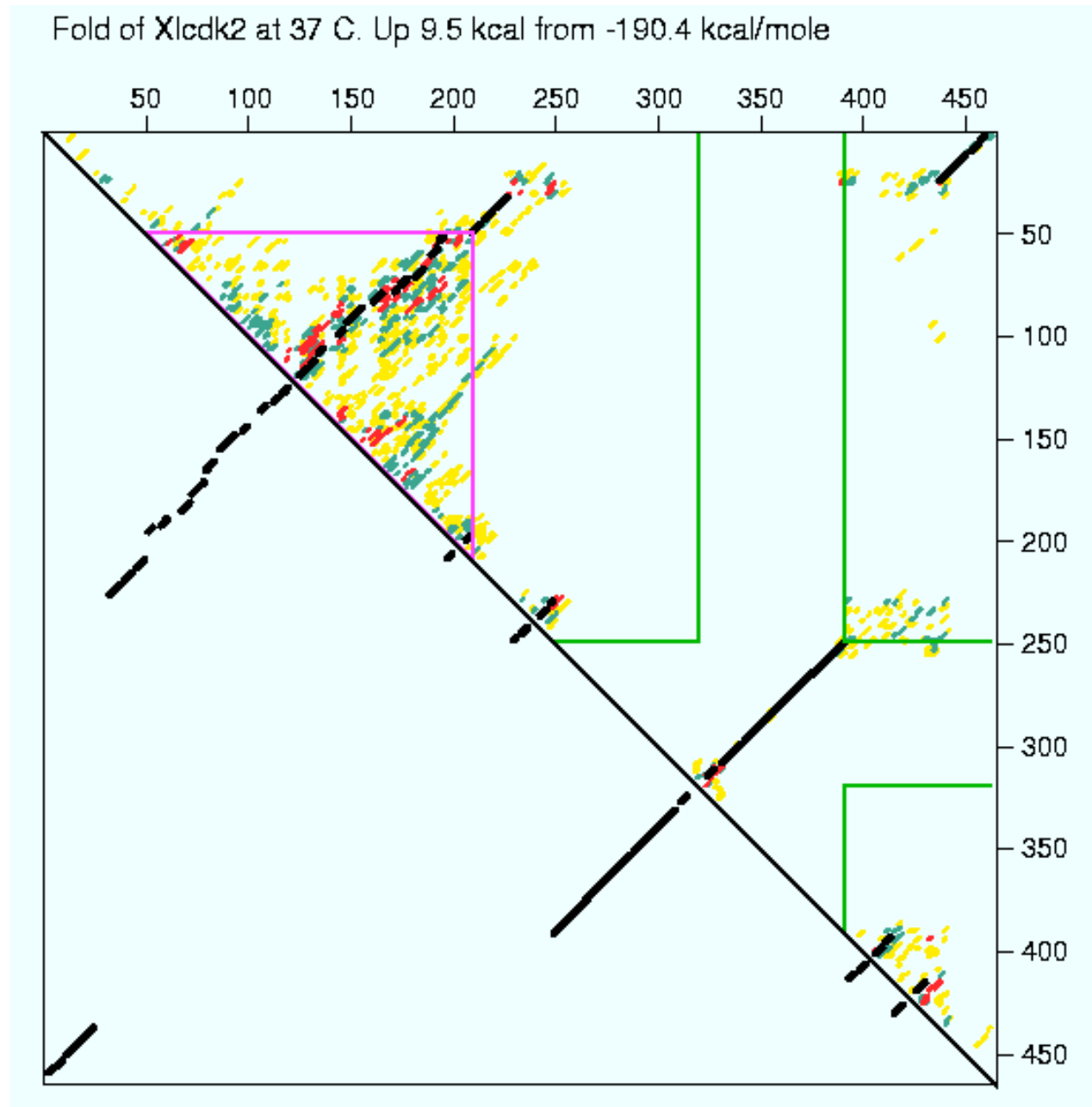


Two competing secondary structures for the *Leptomonas collosoma* spliced leader mRNA.

# Example of suboptimal folding

Black dots: pairs in opt fold

Colored dots: pairs in folds 2-5% worse than optimal fold



# Accuracy

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

Definitely useful, but obviously imperfect

# 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

# Approaches, II

Comparative sequence analysis

- + handles all pairings (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, due, e.g., to codon structure, which RNAs lack

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

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