# CSE 527
# Lecture 17, 11/24/04

RNA Secondary Structure Prediction

# Outline

- What is it
- How is it Represented
- Why is it important
- Examples
- Approaches

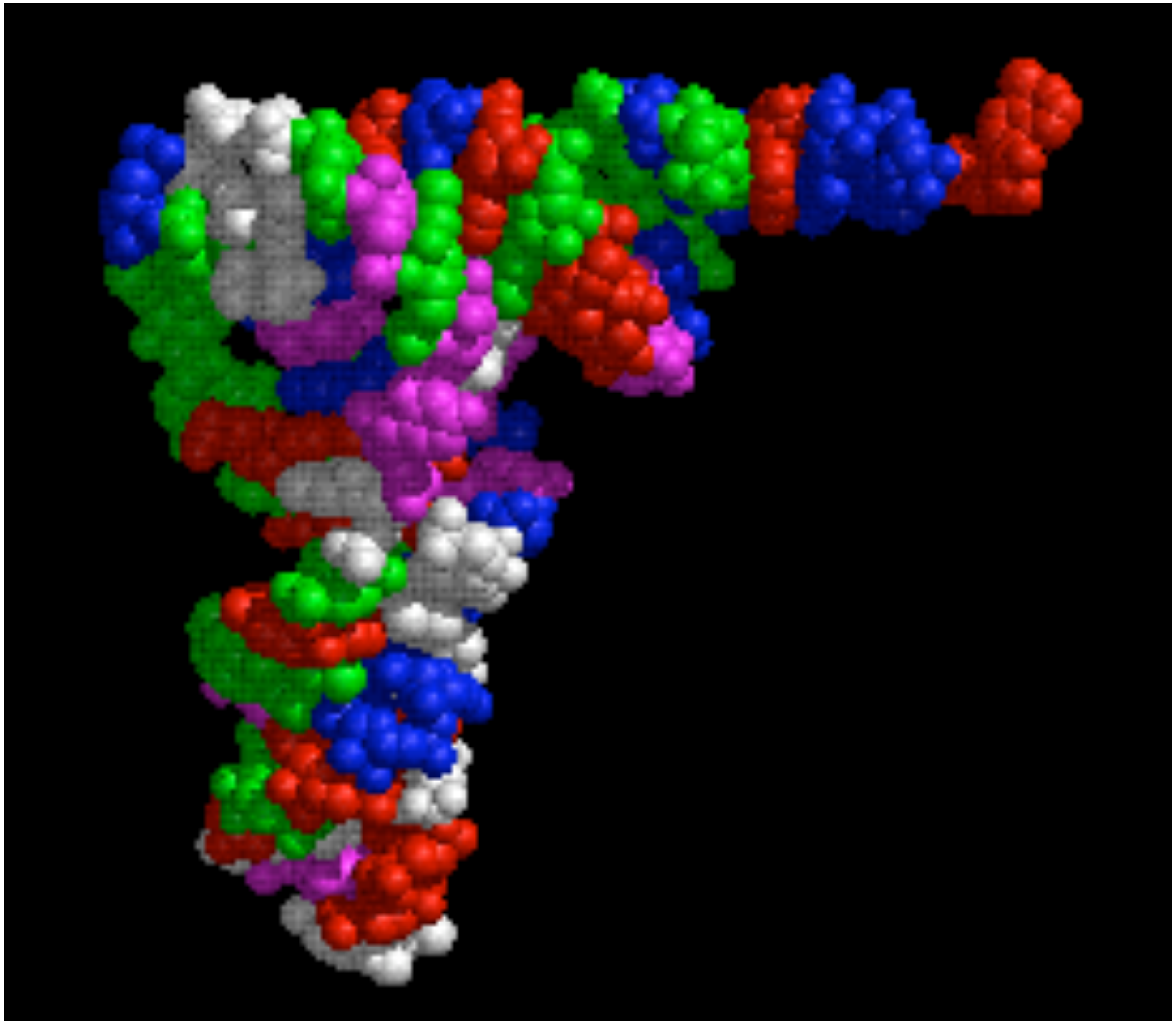
# RNA Structure

- 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)

# A tRNA 3d Structure



# tRNA - Alt. Representations

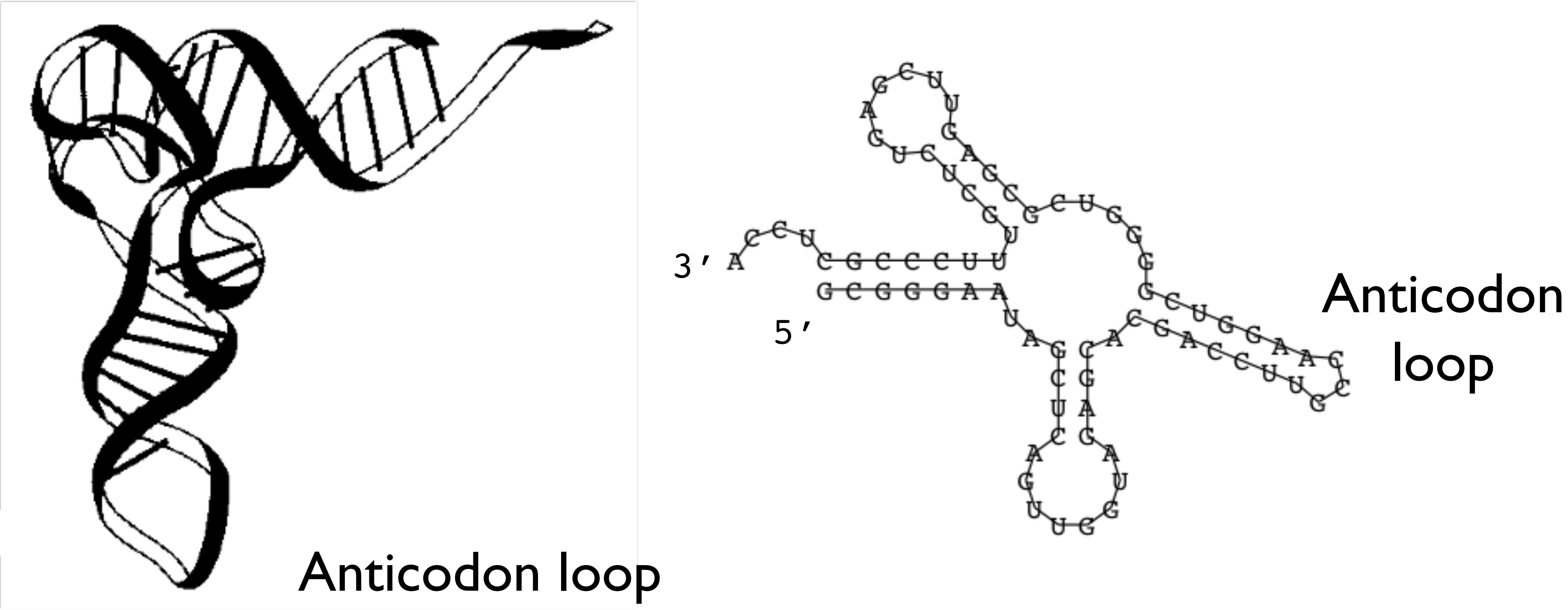
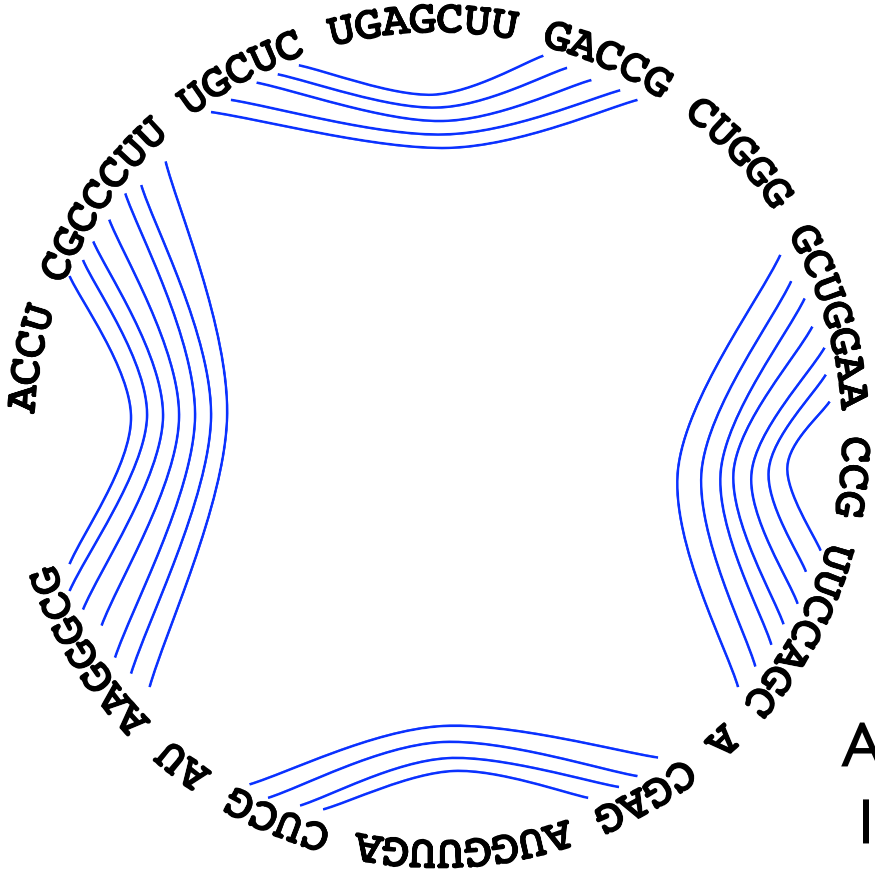


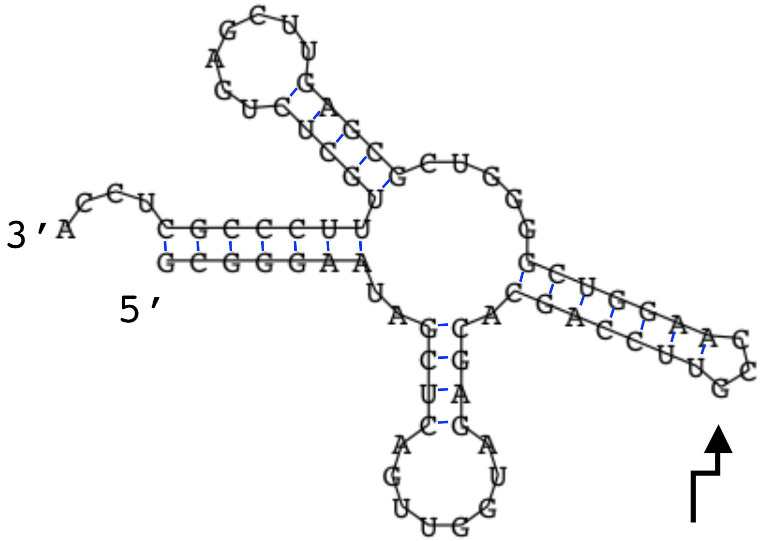
**Figure 1: a)** The spatial structure of the phenylalanine tRNA form yeast

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

# tRNA - Alt. Representations



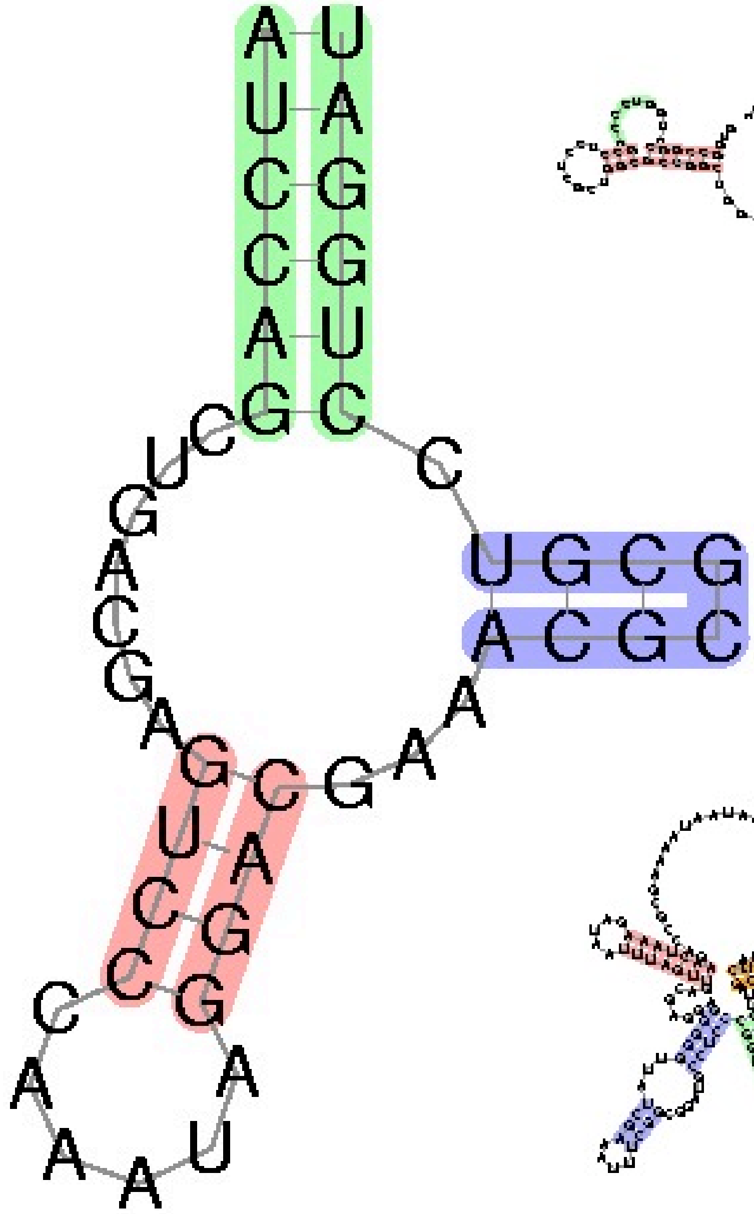
Anticodon loop

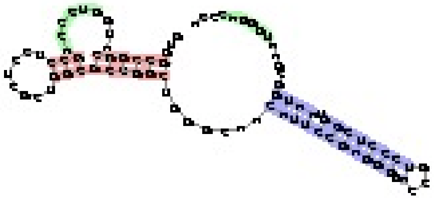


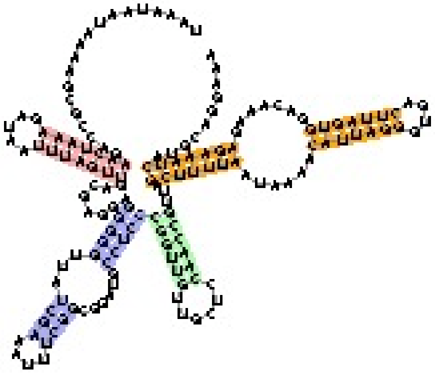
Anticodon loop

# Why?

- RNA’s fold, and function
- Nature uses what works

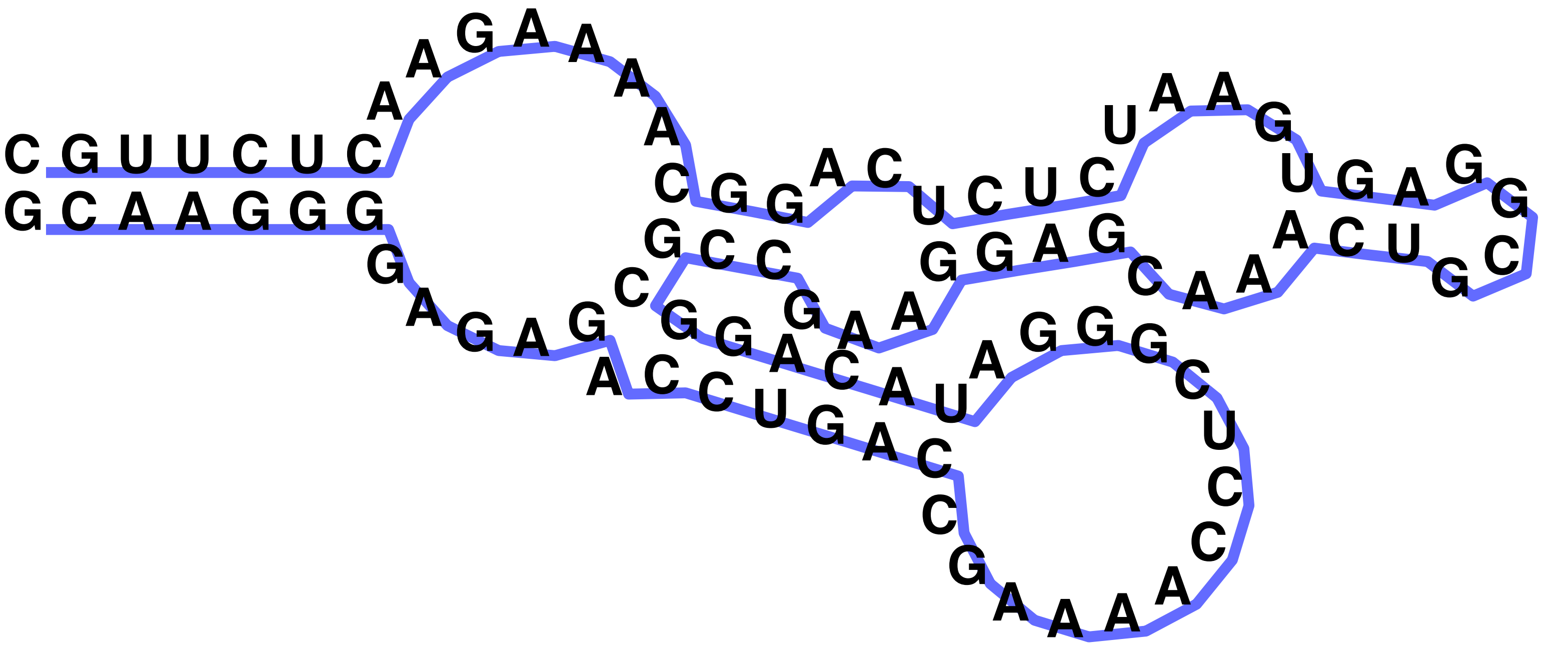


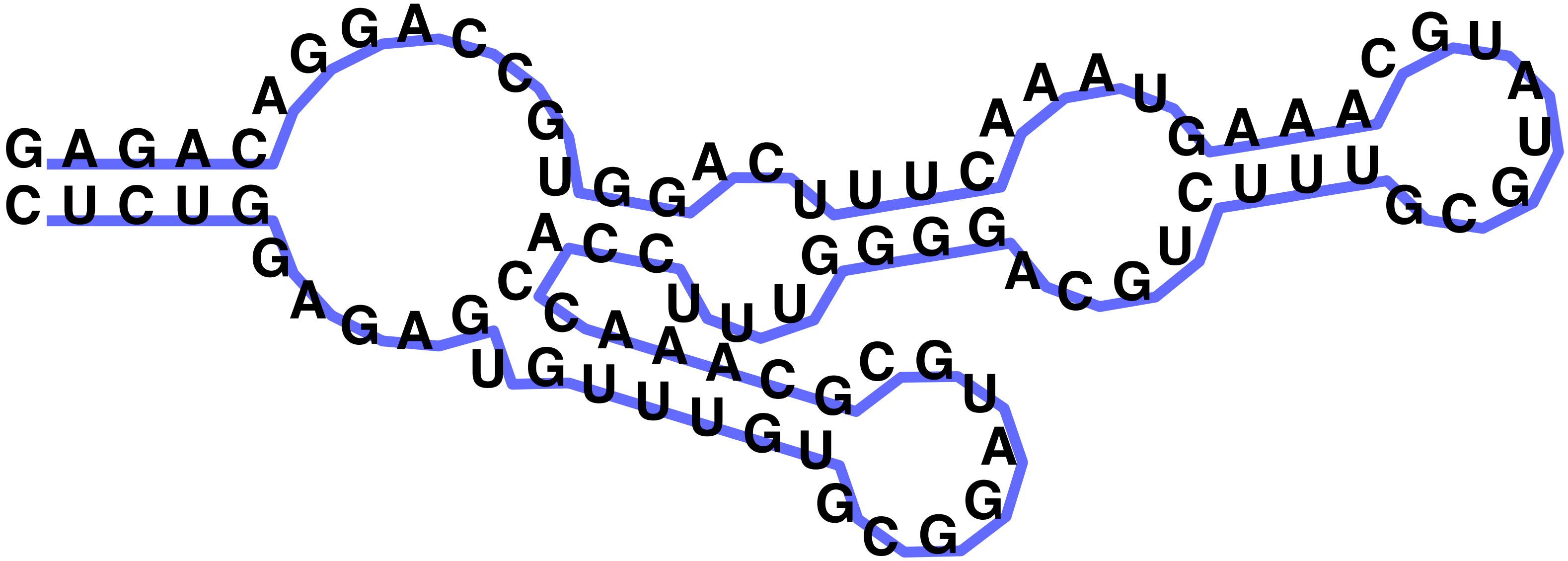




# Importance

- Ribozymes (RNA Enzymes)
- Retroviruses
- Effects on transcription, translation, splicing...
- Functional RNAs: rRNA, tRNA, snRNA, snoRNA, micro RNA, RNAi, riboswitches, regulatory elements in 3' & 5' UTRs, ...


CGUUCUCAAGAAAAACGGACUCUCUAAGUGAGGCGUCAAACGGAGGAAGCCGCGGACAUAGGGCUCCAAAAGCCAGUCCAGAGAGGGGAACG


GAGACAGGACCGUGGACUUUCAAAUGAAACGUAUGCGUUUCUGCAGGGGGUUUCCACCAAACGCGUAGGCGUGUUUGUGAGAGGUCUC

# RNA Pairing

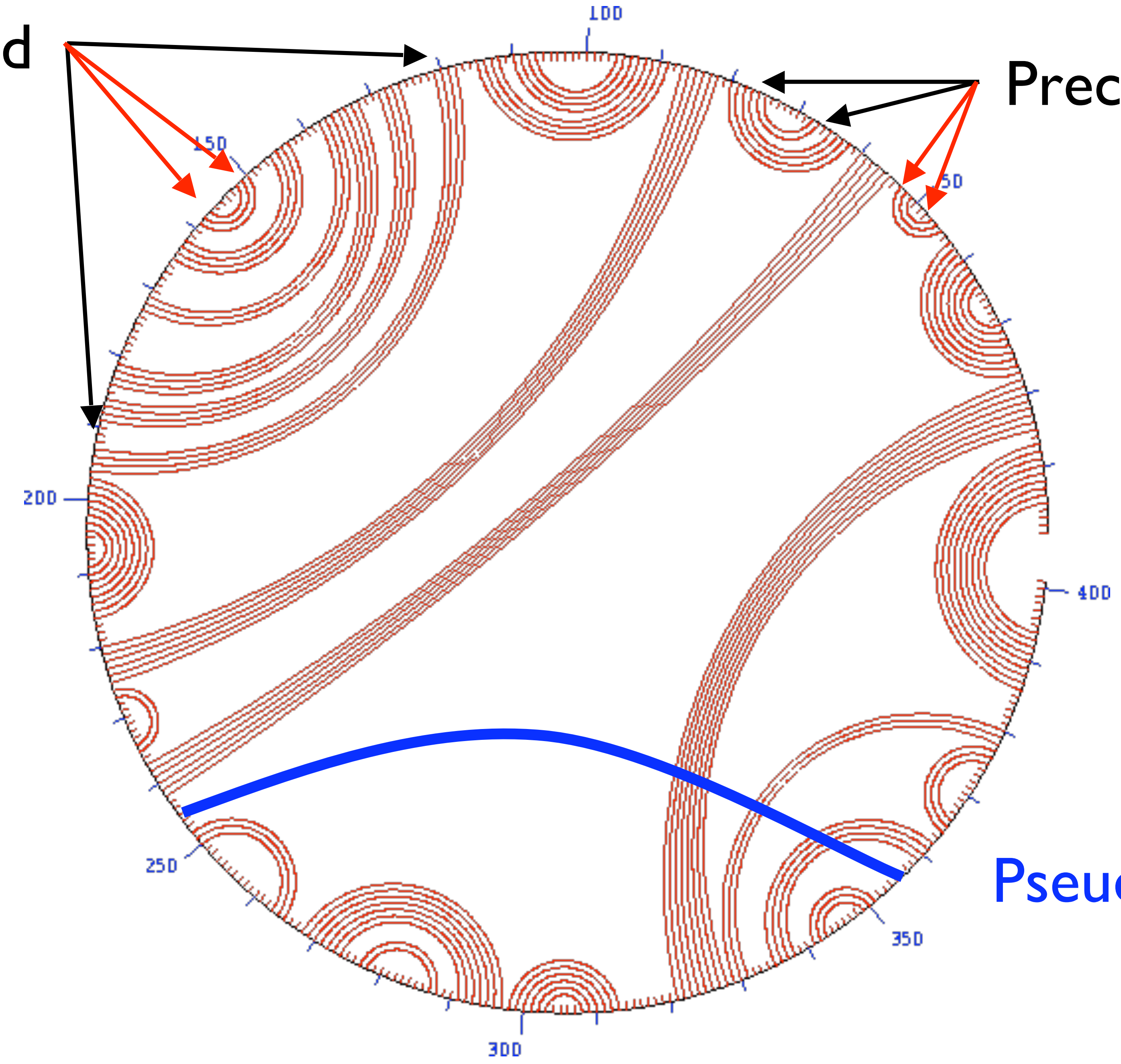
- Watson-Crick Pairing
  - C - G ~ 3 kcal/mole
  - A - U ~ 2 kcal/mole
- “Wobble Pair” G - U ~ 1 kcal/mole
- Non-canonical Pairs (esp. if modified)

# Definitions

- Sequence $^{5'}\ r_1\ r_2\ r_3 \ldots r_n\ ^{3'}$ in {A, C, G, T}
- A Secondary Structure is a set of pairs i•j s.t.

1. i < j-4
2. if i•j & i'•j' are two pairs with i ≤ i', then
   - A. i = i' & j = j', or
   - B. j < i', or
   - C. i < i' < j' < j

} First pair precedes 2nd, or is nested within it. No "pseudoknots."



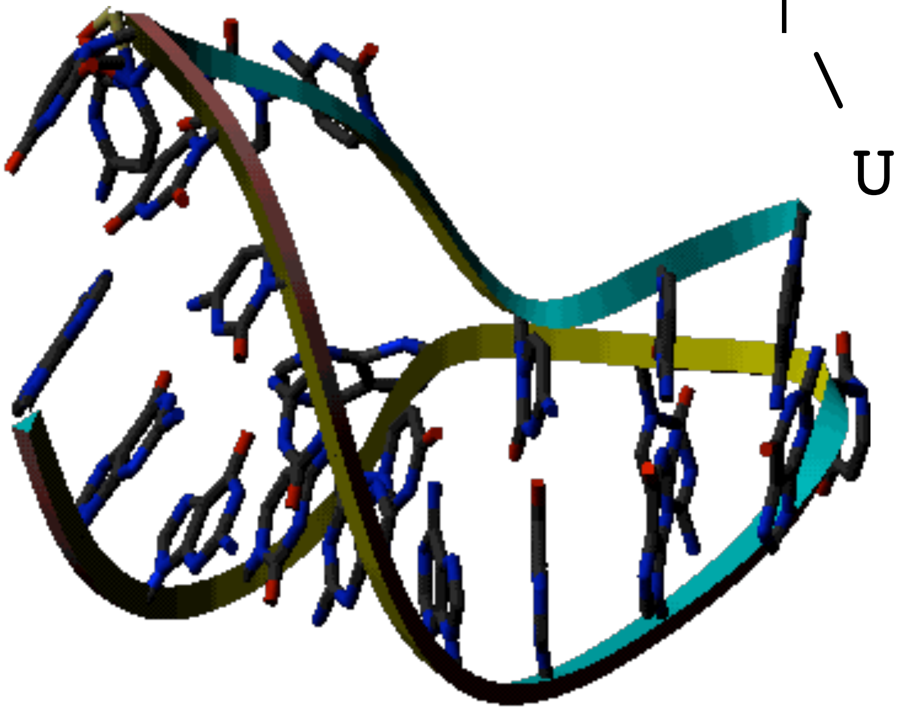
Nested

Precedes

Pseudoknot

# A Pseudoknot

```
               A-C
              /   \
3' - A-G-G-C-U     U
     U-C-C-G-A-G-G-G
     |         C-C-C - 5'
      \       /
       U-C-U-C
```



# 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

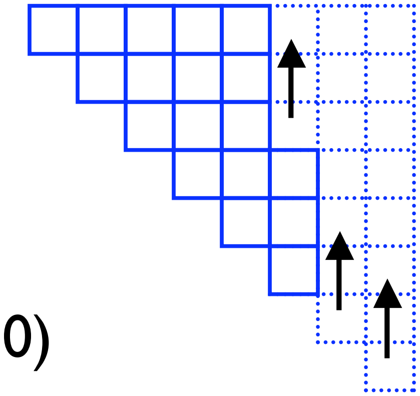
- Stochastic Context-free Grammars
  Roughly combines min energy & comparative, but
  no pseudoknots

- Physical experiments (x-ray crystalography, NMR)

# 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:

1. B(i+1,j)
2. B(i,j-1)
3. B(i+1,j-1) +(if $r_i$ pairs with $r_j$ then 1 else 0)
4. max { B(i,k)+B(k+1,j) | i < k < j }

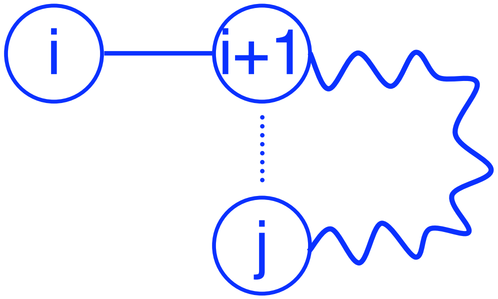


Time: $O(n^3)$

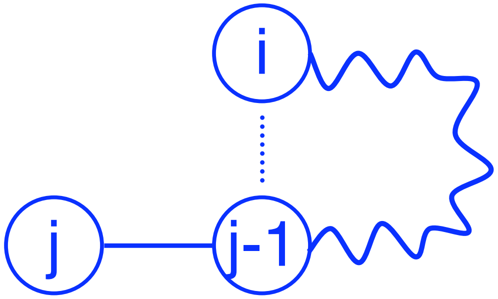
# “optimal pairing of $r_i$ … $r_j$”

Several (overlapping, but exhaustive) possibilities

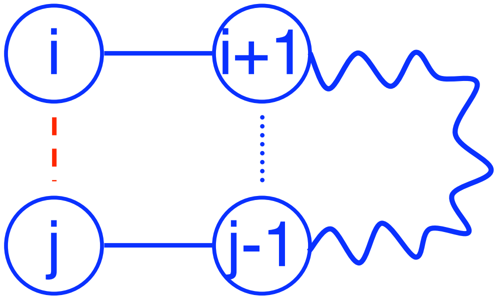
1. $r_i$ is unpaired; look at best way to pair $r_{i+1}$ … $r_j$



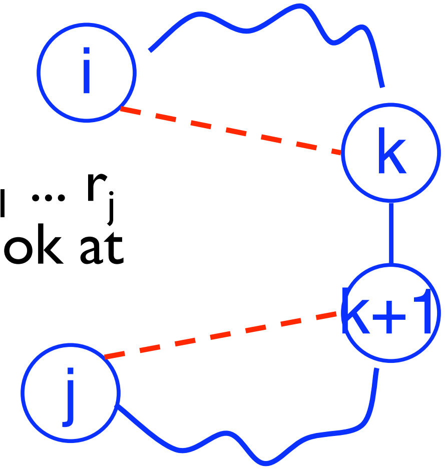
2. $r_j$ is unpaired; look at best way to pair $r_i$ … $r_{j-1}$



3. they pair with each other, so 1 + best $r_{i+1}$ … $r_{j-1}$



4. They pair, but *not* to each other; i pairs with k for some $i < k < j$; so look at best $r_i$ … $r_k$ + best $r_{k+1}$ … $r_j$ (don’t need to look at other k; why?)



# Pair-based Energy Minimization

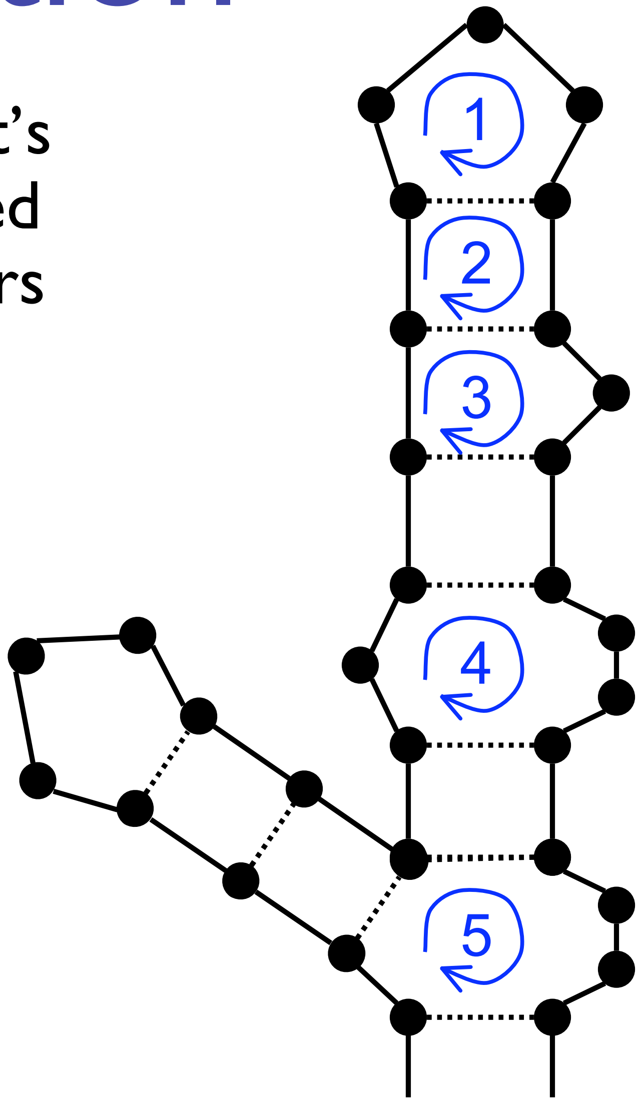
- E(i,j) = energy of pairs in optimal pairing of $r_i$ ... $r_j$
- E(i,j) = ∞ for all i, j with $i \geq j-4$; otherwise
- E(i,j) = min of:
  - E(i+1,j)
  - E(i,j-1)
  - E(i+1,j-1) + e($r_i$, $r_j$)
  - min { E(i,k)+E(k+1,j) | i < k < j }

energy of one pair

Time: O($n^3$)

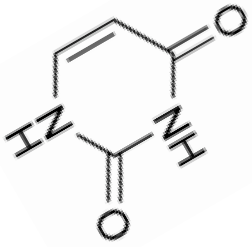
# Loop-based Energy Minimization

- 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

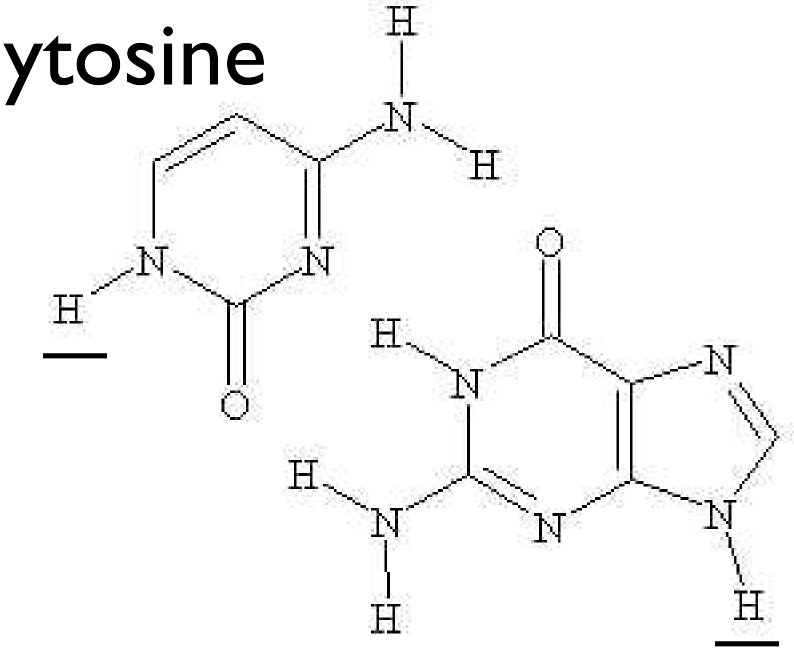


# Base Pairs and Stacking

uracil

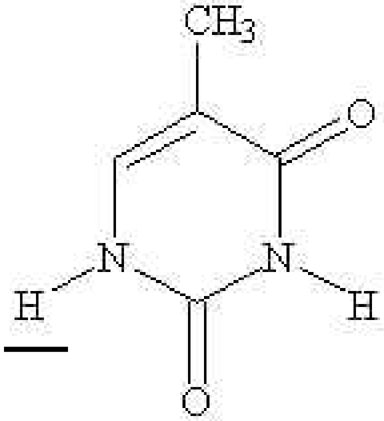


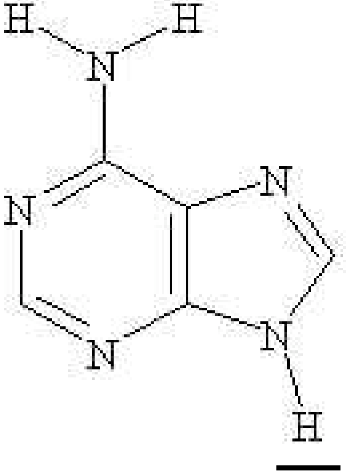
cytosine



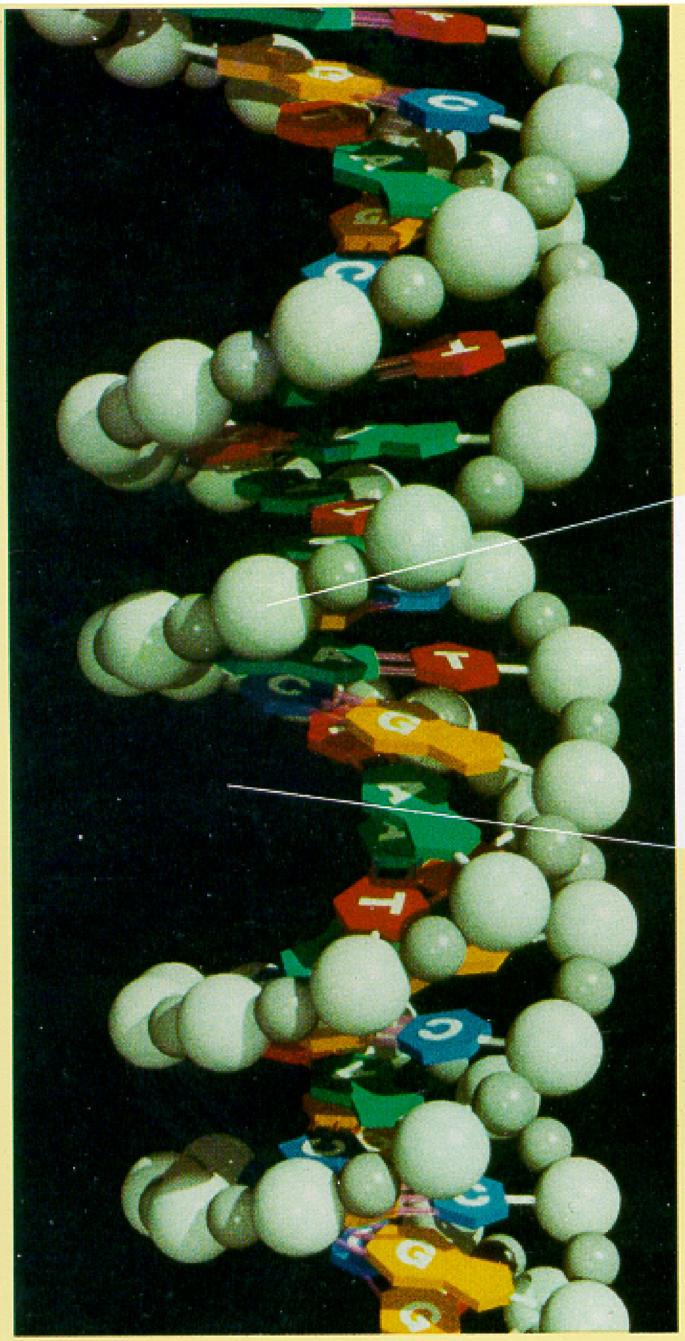
guanine

thymine



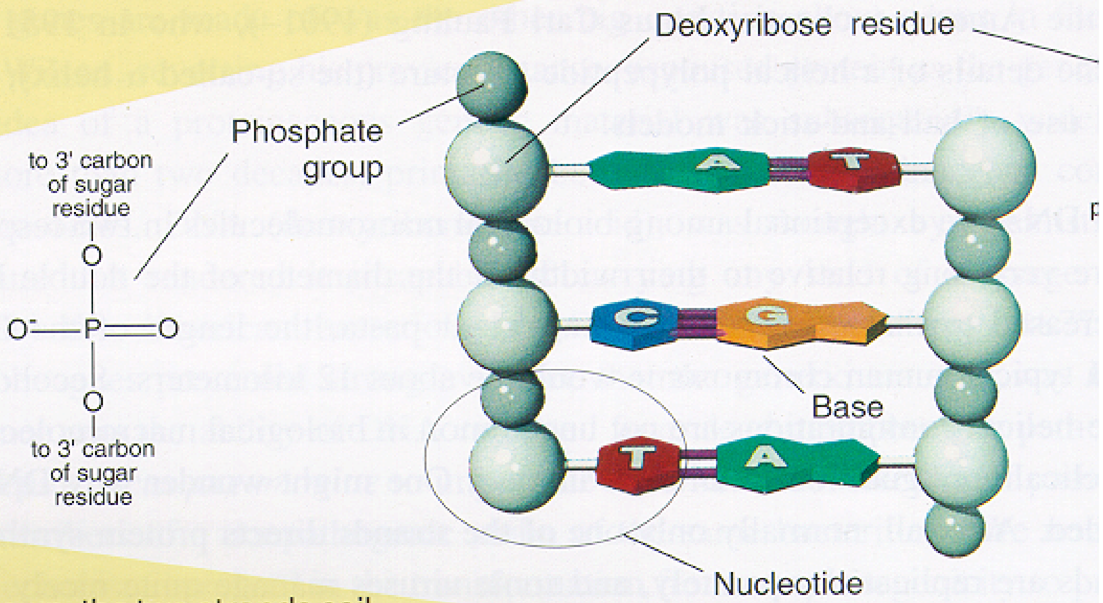


adenine



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

(b) Uncoiled DNA Fragment

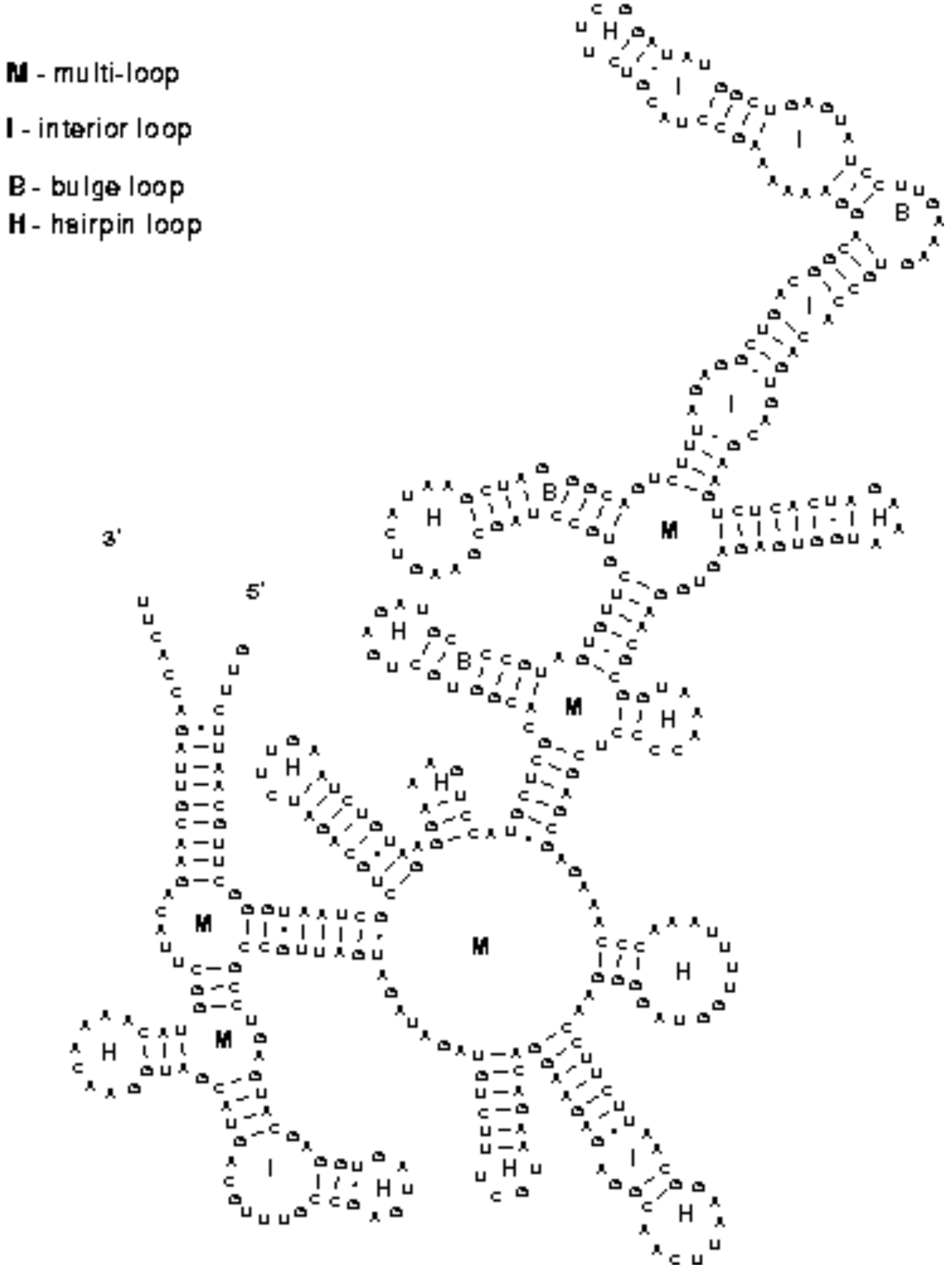


As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. 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 stra a polymer made up of four re called deoxyribonucleotides

# Loop Examples

*Bacillus subtilis RNase P RNA*



# Zuker: Loop-based Energy, I

- W(i,j) = energy of optimal pairing of $r_i \ldots r_j$
- V(i,j) = as above, but forcing pair i•j
- $W(i,j) = V(i,j) = \infty$ for all i, j with $i \geq j-4$
- $W(i,j) = \min(W(i+1,j),\ W(i,j-1),\ V(i+1,j-1),$
  $\min \{ E(i,k)+E(k+1,j) \mid i < k < j \} )$

# Zuker: Loop-based Energy, II

hairpin stack bulge/ interior multi- loop

- $V(i,j) = \min(eh(i,j),\ es(i,j)+V(i+1,j-1),\ VBI(i,j),\ VM(i,j))$
- $VM(i,j) = \min \{ W(i,k)+W(k+1,j) \mid i < k < j \} )$
- $VBI(i,j) = \min \{ ebi(i,j,i',j') + V(i', j') \mid$
  $i < i' < j' < j \ \&\ i'-i+j-j' > 2 \}$

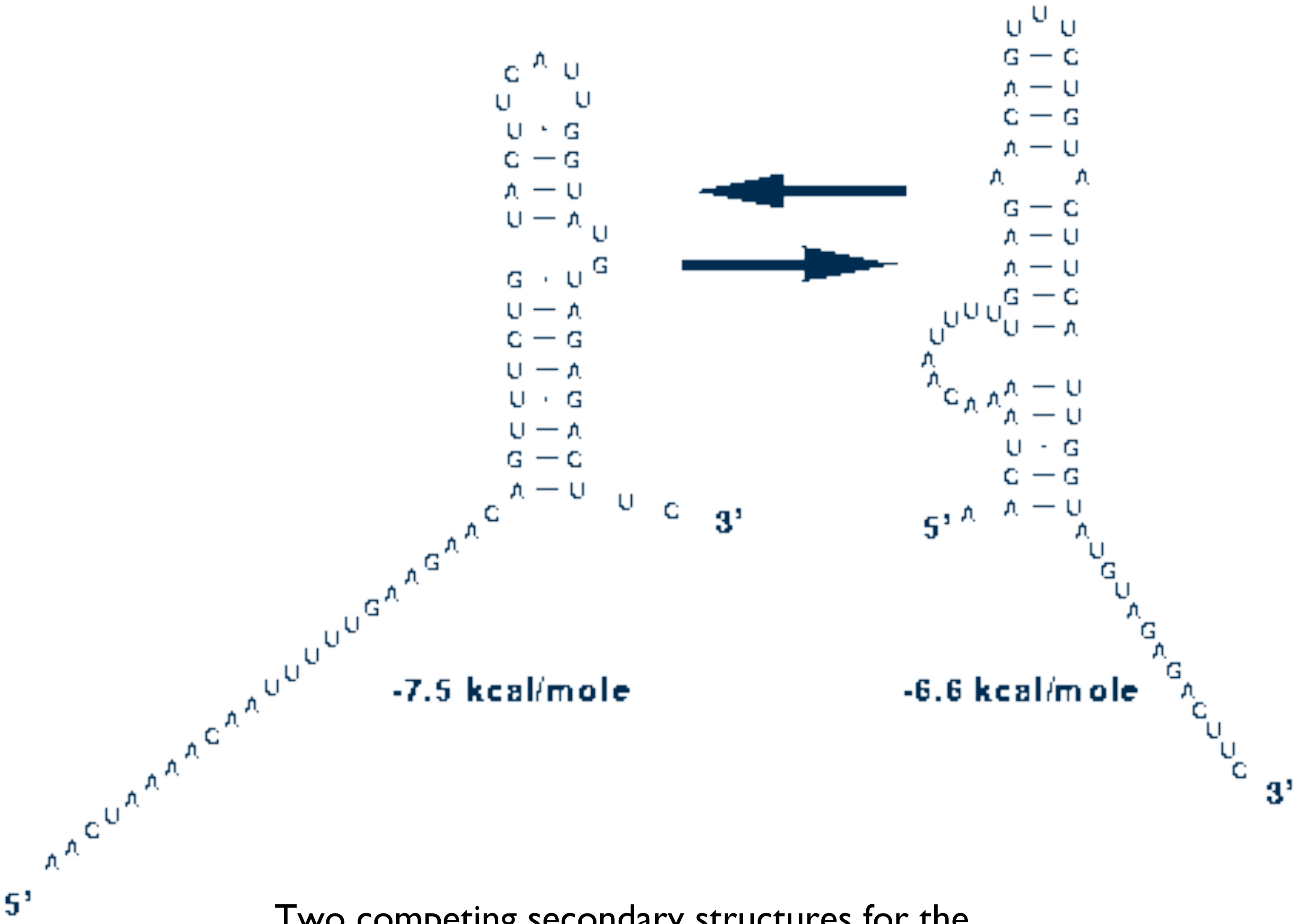
bulge/ interior

Time: $O(n^4)$

$O(n^3)$ possible if ebi(.) is “nice”

# Suboptimal Energy

- There are always alternate folds with near-optimal energies. Thermodynamics predicts that populations of identical molecules will exist in different folds; individual molecules even flicker among different folds
- Zuker's algorithm can be modified to find suboptimal folds
- McCaskill gives a more elaborate dynamic programming algorithm calculating the "partition function," which defines the probability distribution over all these states.

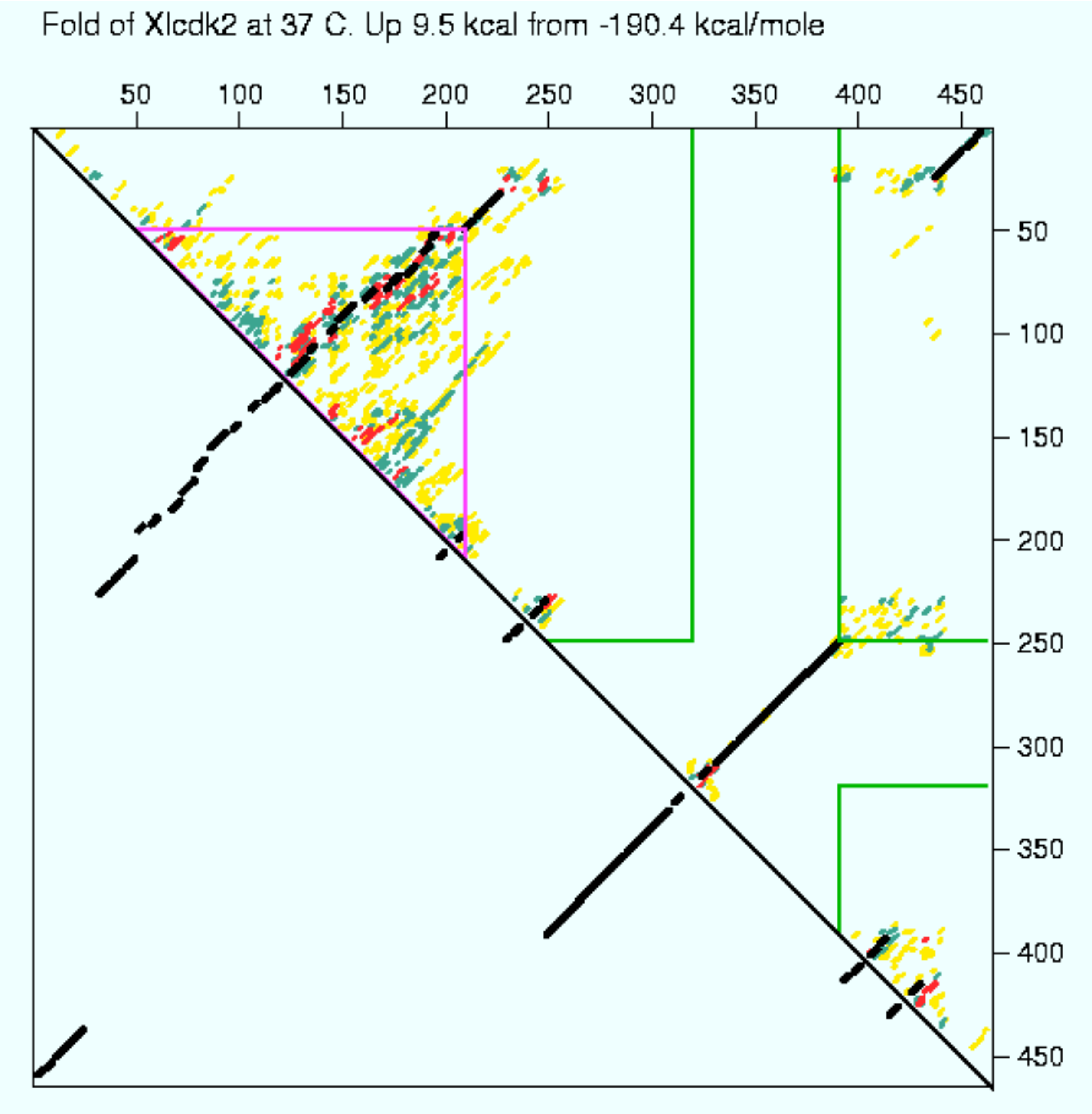


Two competing secondary structures for the Leptomonas collosoma spliced leader mRNA.

# Example of suboptimal folding

Black dots: pairs in opt fold

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



# A “Mountain” diagram

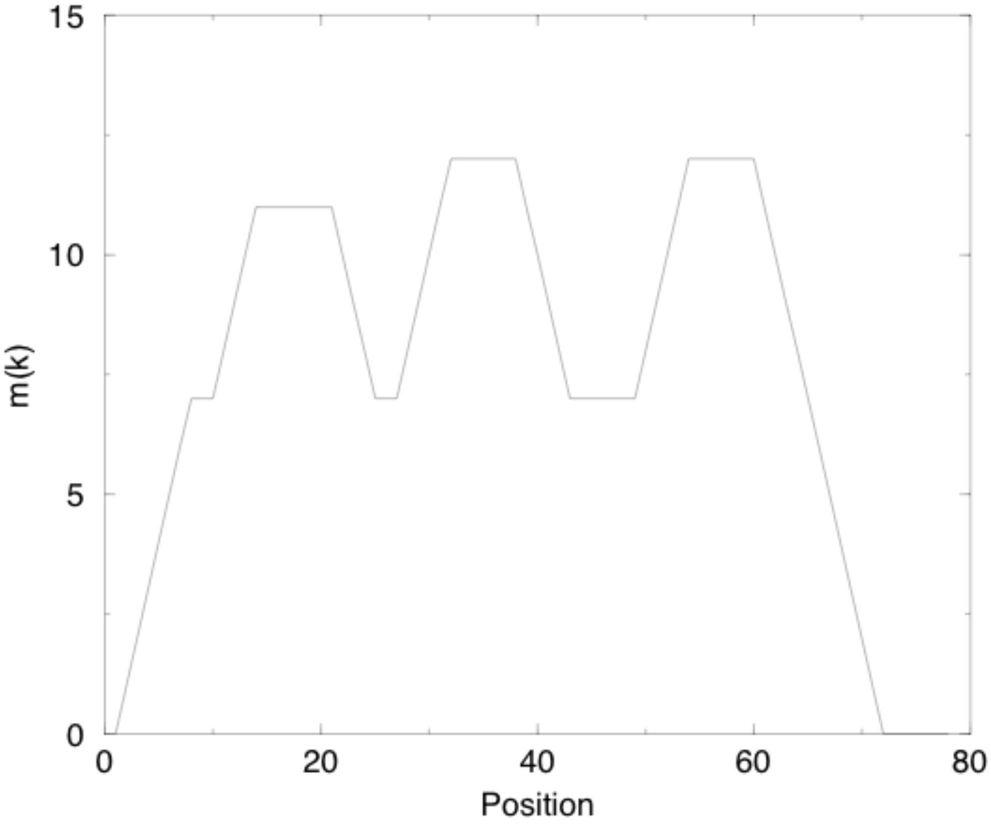


**Figure 3:** Mountain representation of the tRNA secondary structure shown in Figure 1. The three plateaus correspond to the three hairpin loops of the clover leave structure.