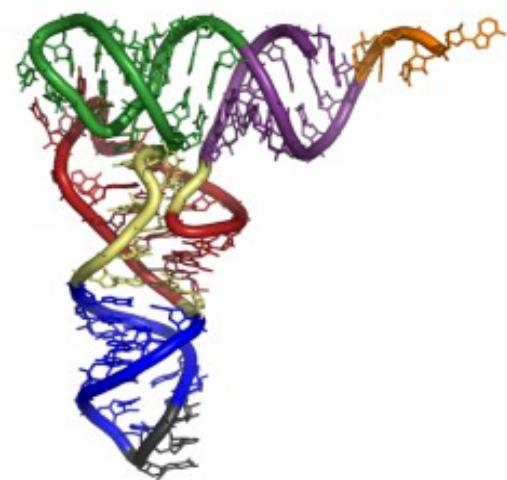
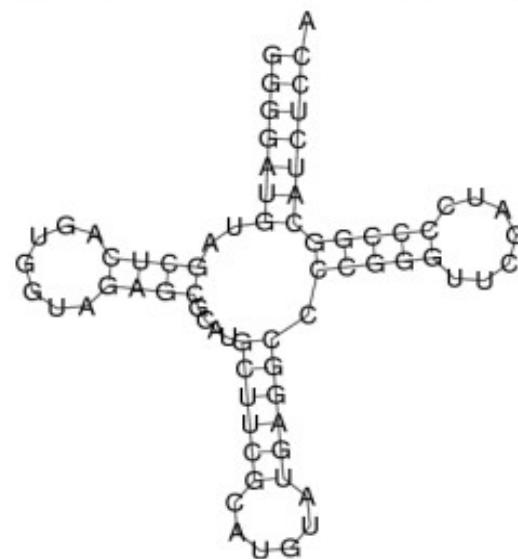
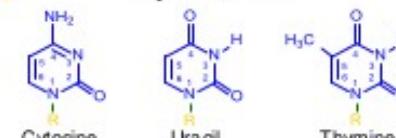
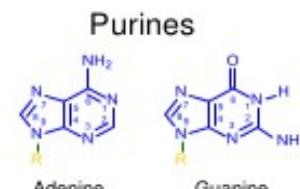
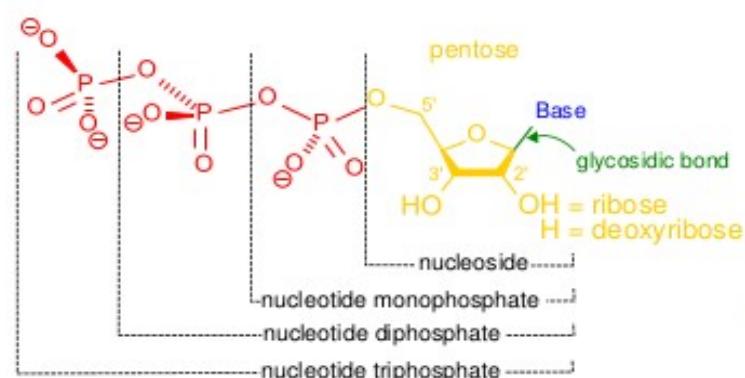


RNA Structure and RNA Structure Prediction



# Introducción a la estructura de ARN

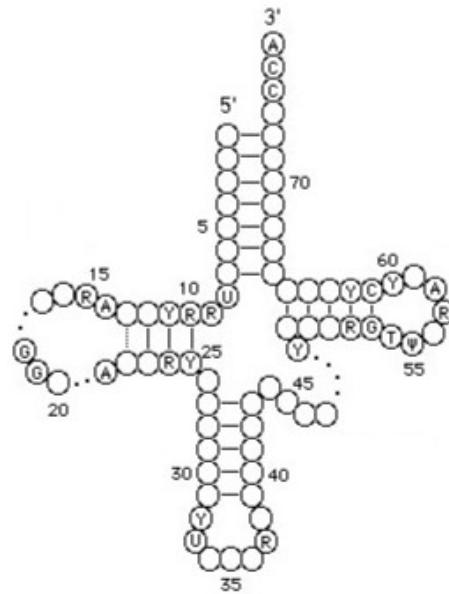
A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein.

The DNA sequence from which a functional non-coding RNA is transcribed is often called an RNA gene. Abundant and functionally important types of non-coding RNAs include transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), as well as small RNAs such as microRNAs, siRNAs, piRNAs, snoRNAs, snRNAs, exRNAs, scaRNAs and the long ncRNAs such as Xist and HOTAIR.

# Introducción a la estructura de ARN

## Various types of RNA

- messenger RNA (mRNA)
- transfer RNA (tRNA)
- Ribosomal RNA (rRNA)
- small interfering RNA (siRNA)
- micro RNA (miRNA)
- small nuclear RNA (snRNA)
- small nucleolar RNA (snoRNA)
- guide RNA (gRNA)
- efference RNA(eRNA)



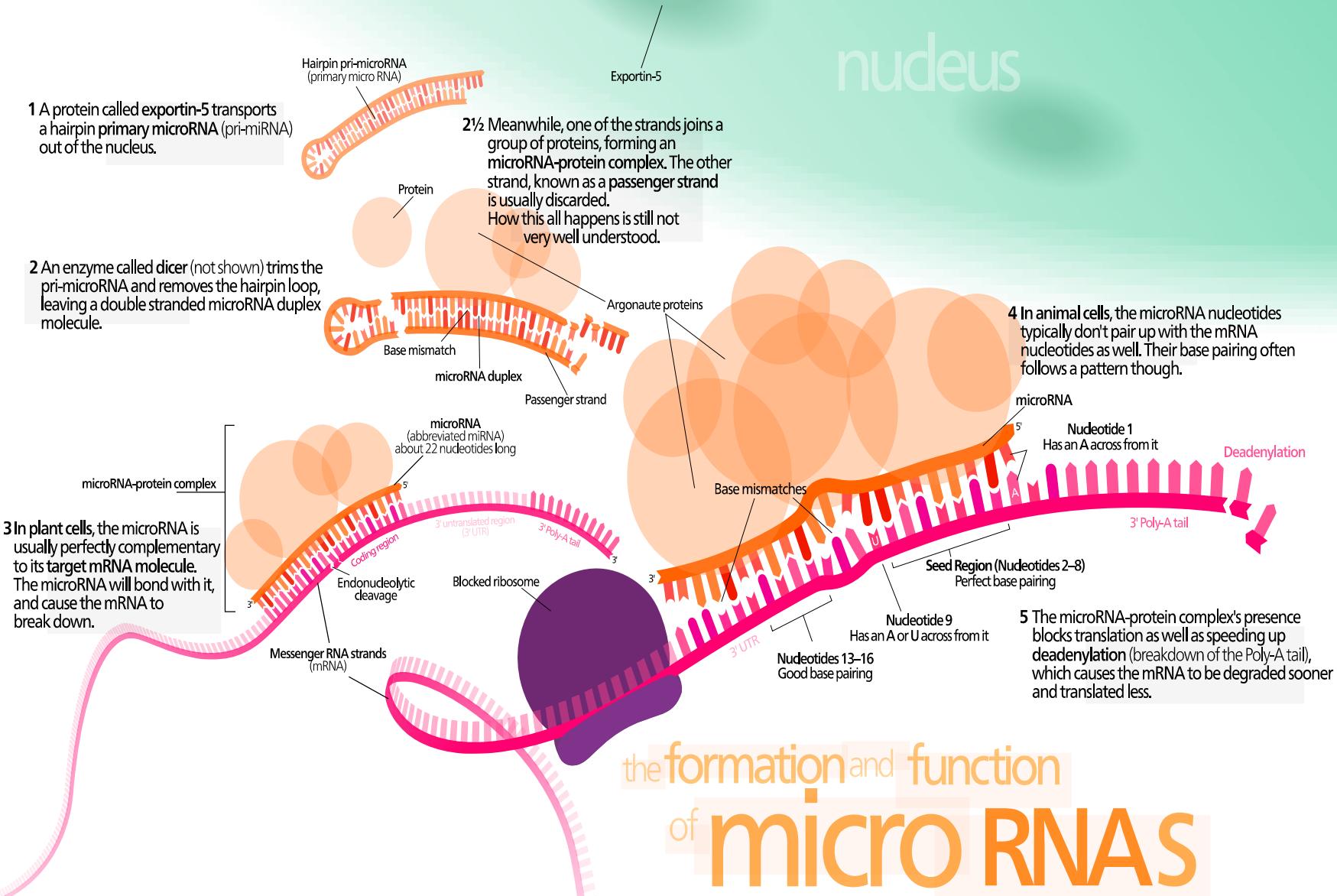
# Introducción a la estructura de ARN

## Non coding RNA (ncRNA)

- RNA that isn't translated into protein
- Includes:
  - tRNA, rRNA, snRNA, snoRNA, miRNA, gRNA, eRNA, pRNA, tmRNA
- What about mRNA?
  - 5' methylated cap
  - 5'-UTR (un-translated regions)
  - CDS (coding sequence)
  - 3'-UTR
  - Poly-A tail
- mRNA contains untranslated regions (5'UTR, 3'UTR), but UTRs are not considered ncRNA

# Introducción a la estructura de ARN

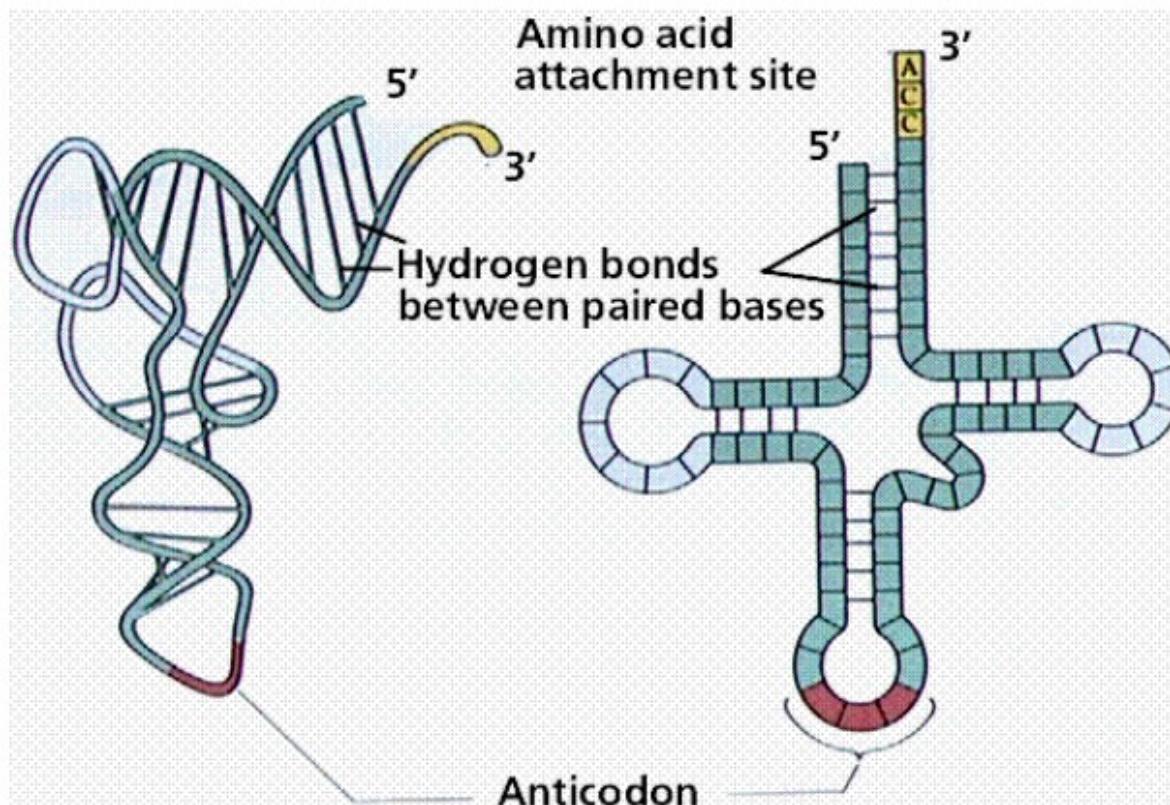
nudeus



# Introducción a la estructura de ARN

## RNA Secondary and Tertiary Structure

Example: tRNA



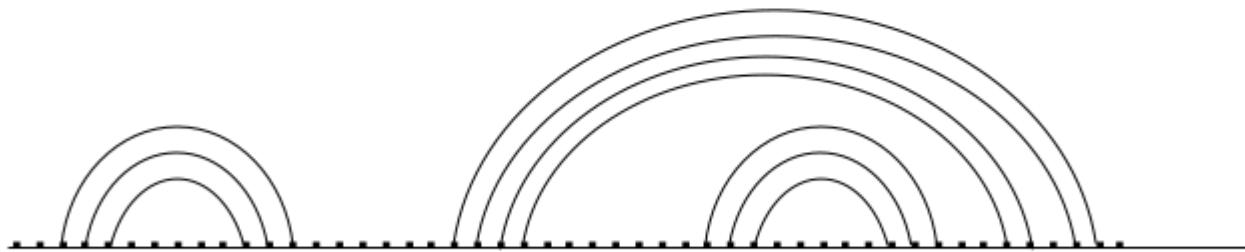
© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

# RNA Secondary Structure Notation

Parentheses notation

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

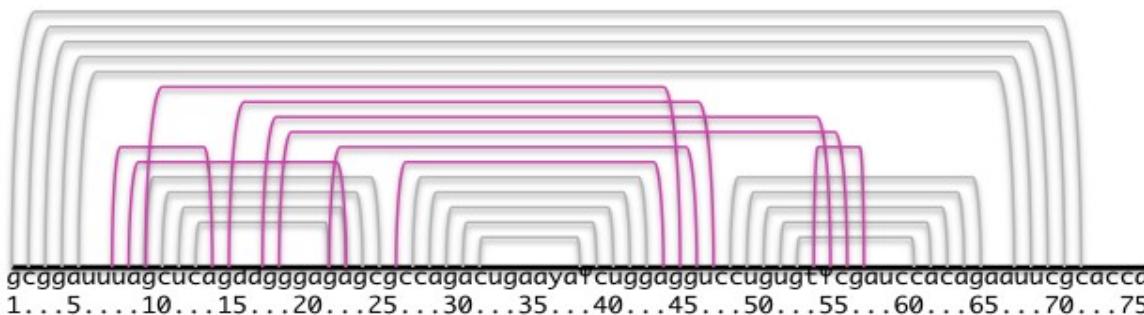
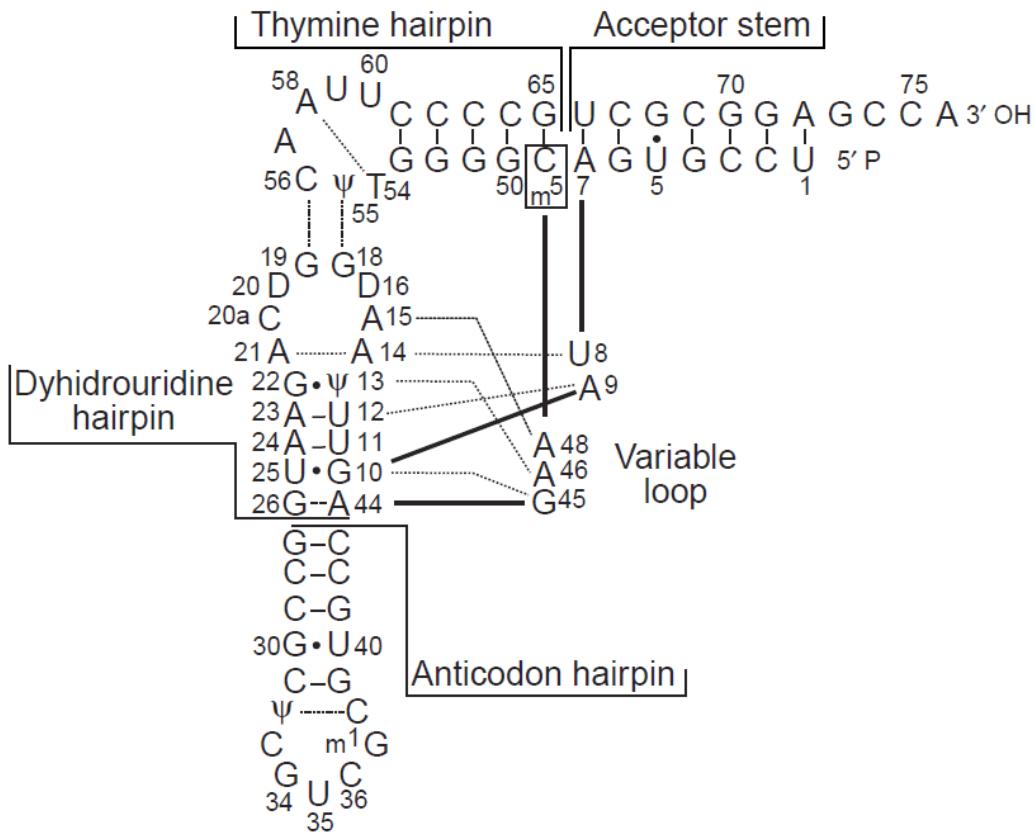
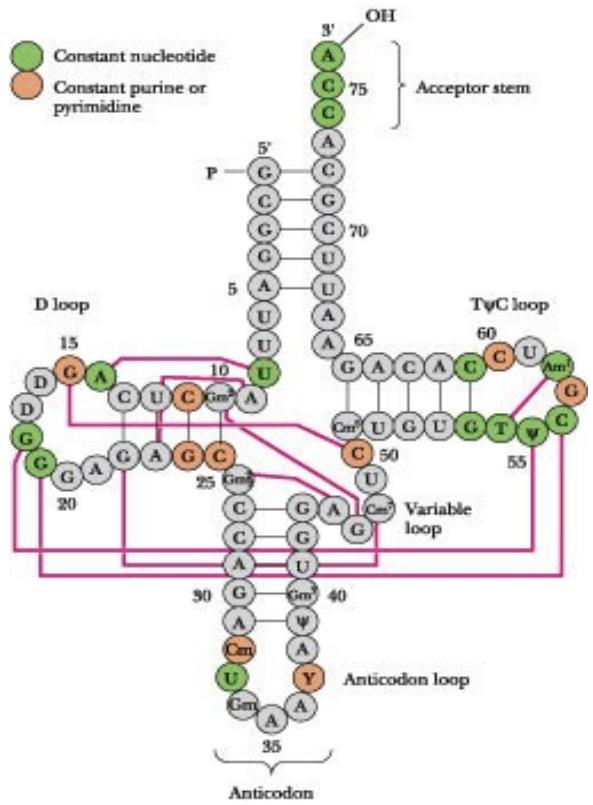
Arc ('rainbow') notation



What do these structures look like?

What is the difference between these two structures?

# Introducción a la estructura de Proteínas

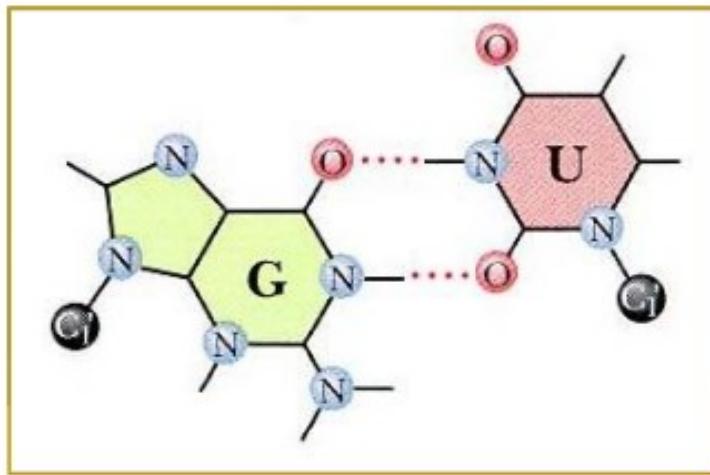


# Introducción a la estructura de Proteínas

## RNA Basics

- RNA bases: A, C, G, U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U
- Bases can only pair with one other base

'wobble pairing'

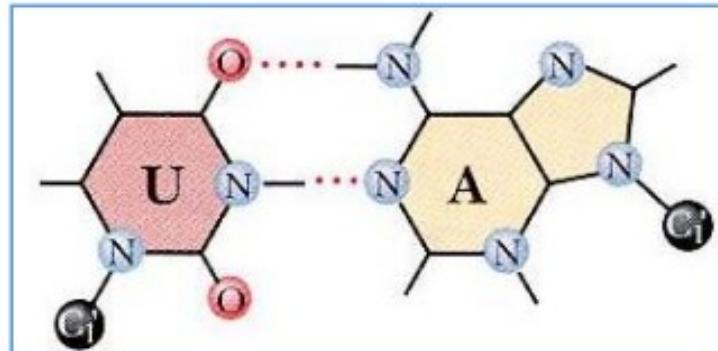


# Introducción a la estructura de Proteínas

## RNA Basics

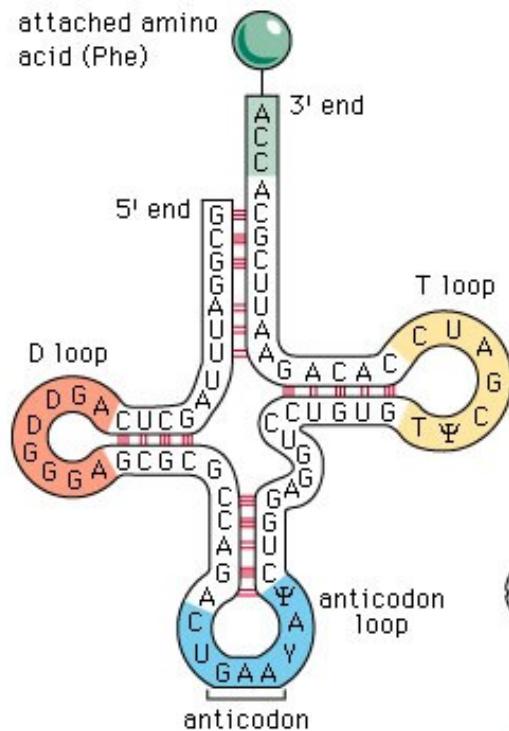
- RNA bases: A, C, G, U
- Watson-Crick Pair
  - A-U ( $\sim 2$  kcal/mol)
  - G-C ( $\sim 3$  kcal/mol)
- Wobble pair
  - G-U ( $\sim 1$  kcal/mol)
- Non-Canonical pairs (modified suitably)
- Bases can only pair with one other base

Two hydrogen bonds

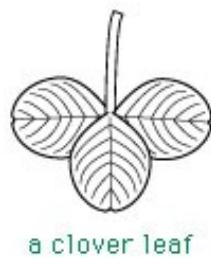


# Introducción a la estructura de ARN

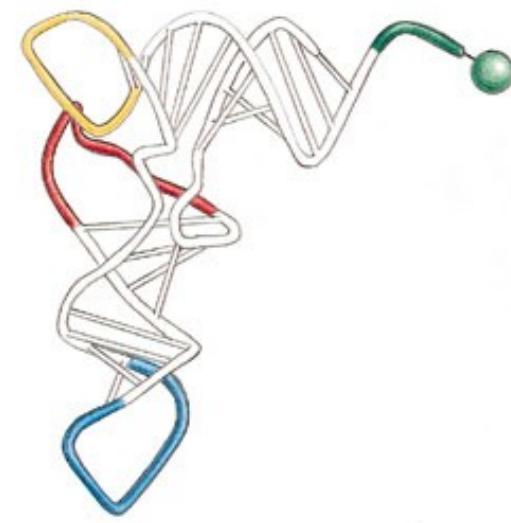
La Estructura Secundaria es importante para la estructura terciaria



(A)



(B)



(C)

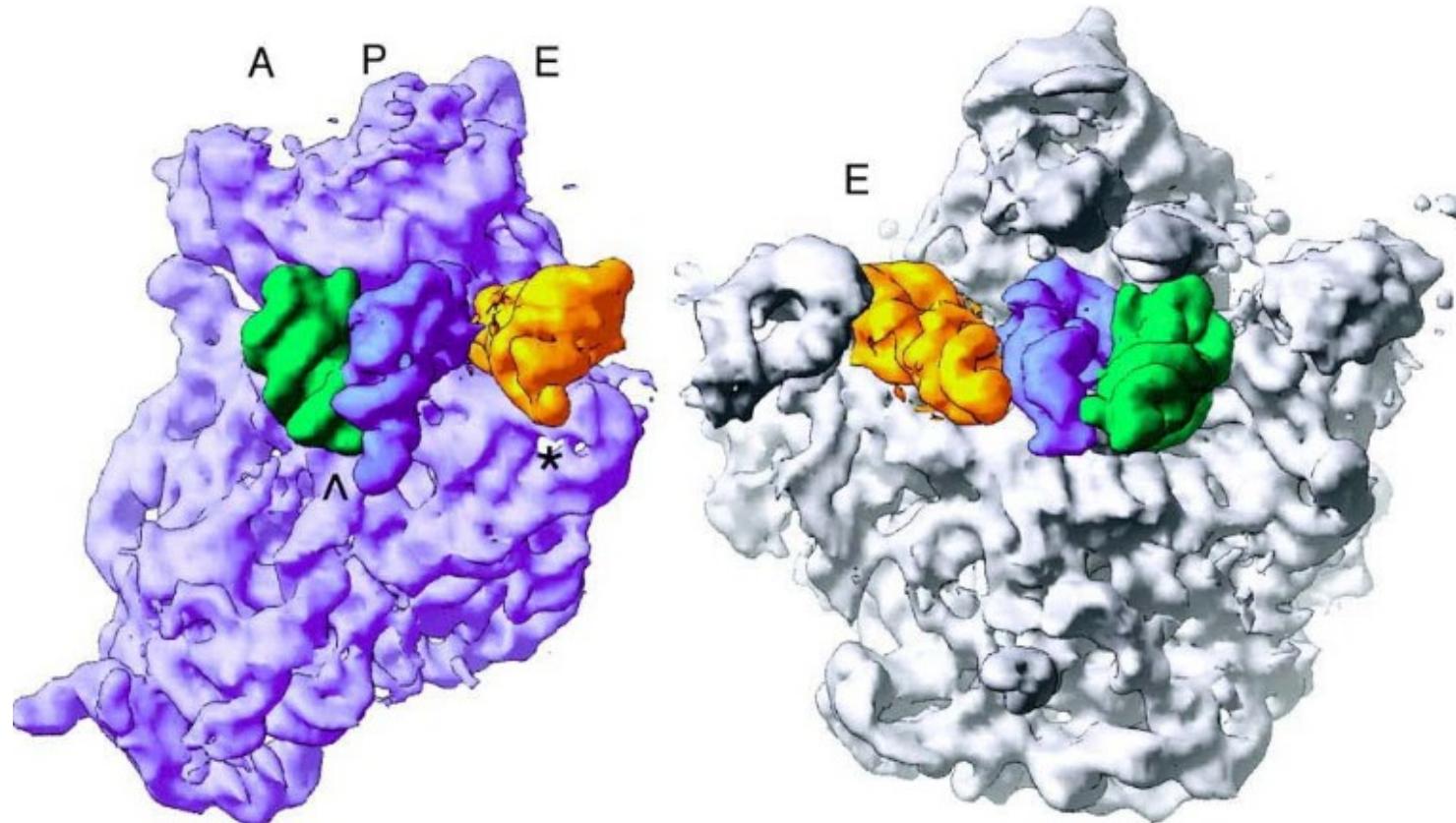
5' GCGGAUUUAGCU**CAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA** 3'  
(D) anticodon

# Introducción a la estructura de Proteínas



# Introducción a la estructura de ARN

## Ribosome at 7 Å with tRNAs

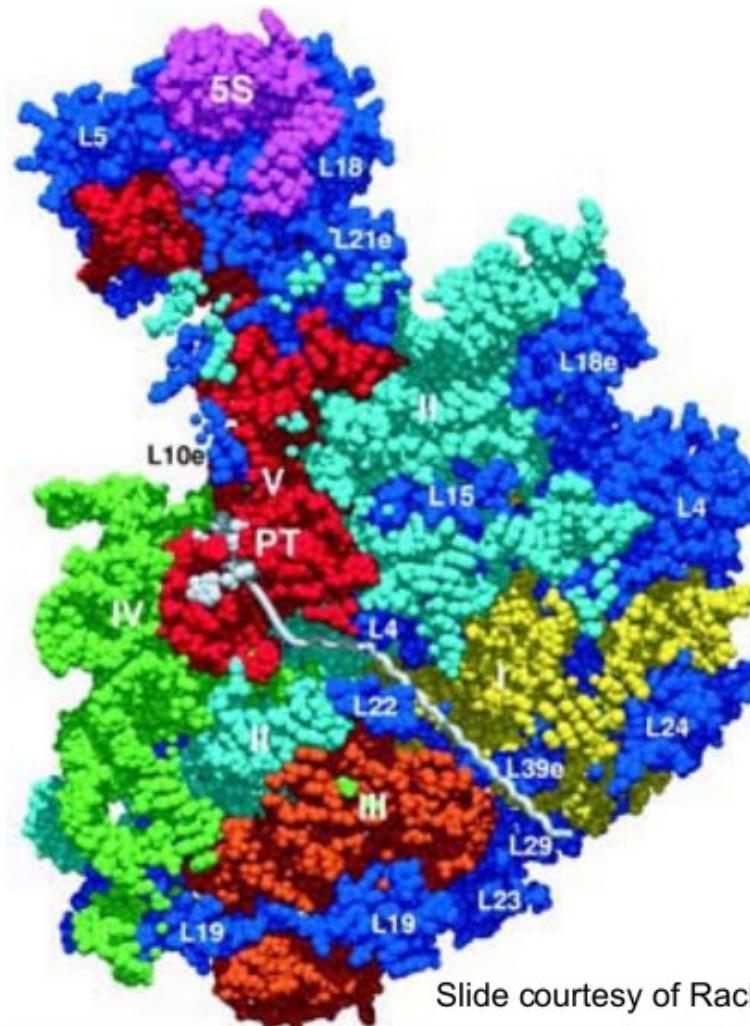


© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.  
Source: Cate, Jamie H., Marat M. Yusupov, et al. "[X-ray Crystal Structures of 70S Ribosome Functional Complexes](#)." *Science* 285, no. 5436 (1999): 2095-104.

# Introducción a la estructura de Proteínas

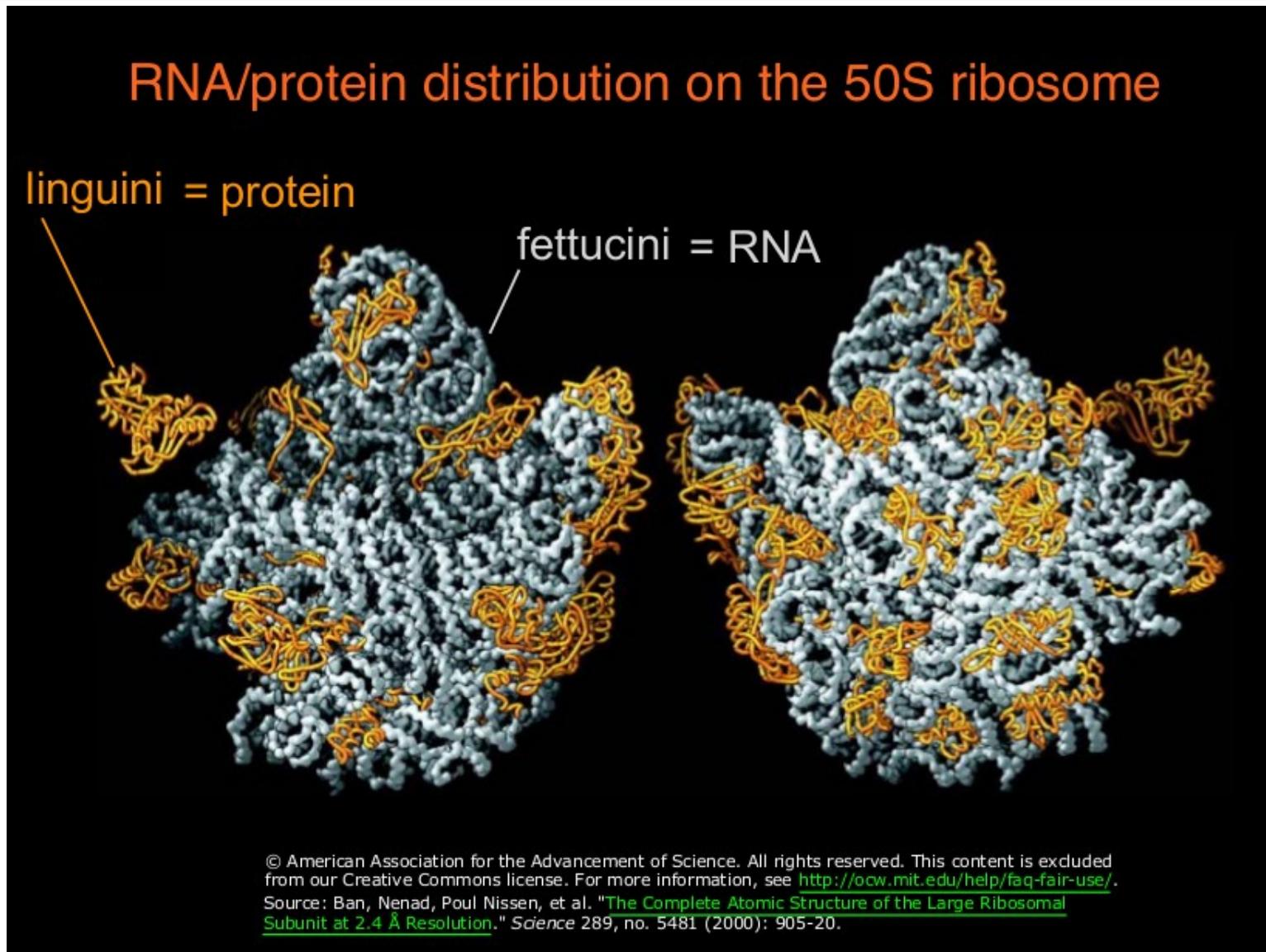
**Can build  
useful  
structures  
out of RNA**

**The exit channel  
for the growing  
polypeptide**



Slide courtesy of Rachel Green

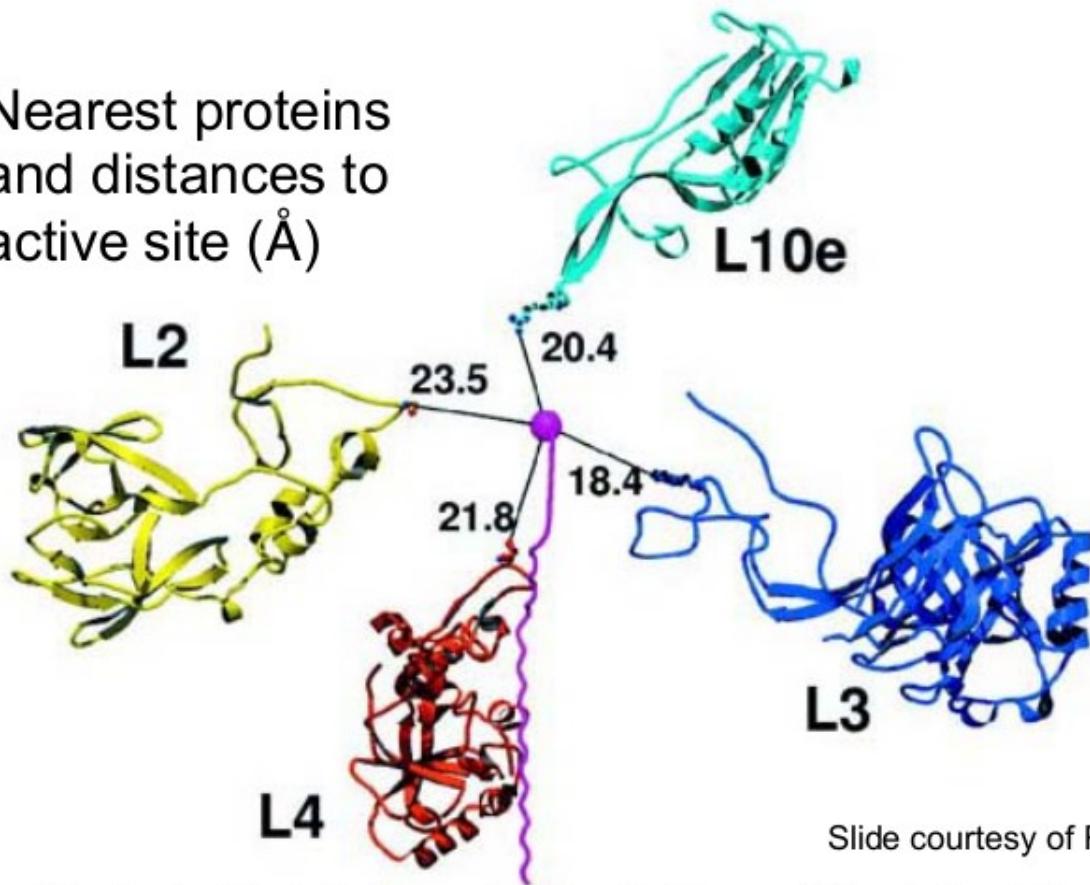
# Introducción a la estructura de ARN



# Introducción a la estructura de ARN

## The ribosome is a ribozyme

Nearest proteins  
and distances to  
active site (Å)



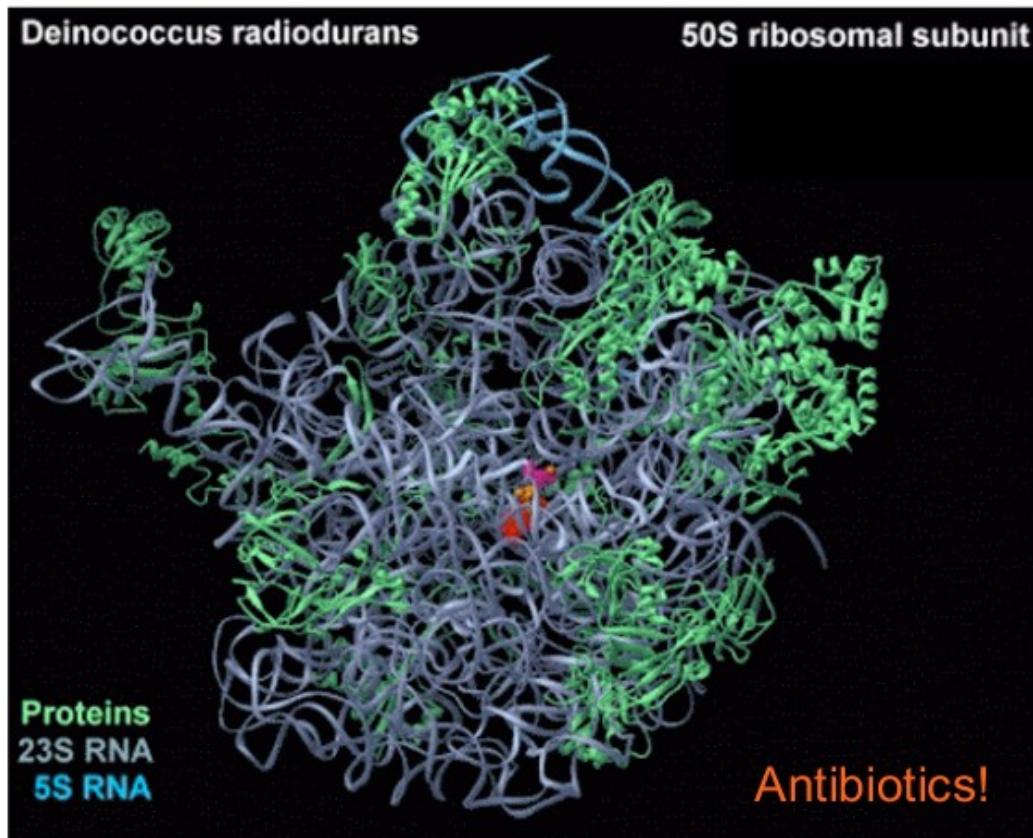
Slide courtesy of Rachel Green

Ribozymes (ribonucleic acid enzymes) are **RNA** molecules that are capable of catalyzing specific biochemical reactions, similar to the action of protein enzymes.

The 1982 discovery of ribozymes demonstrated that RNA can be both genetic material (like **DNA**) and a biological **catalyst** (like protein enzymes)

# Introducción a la estructura de Proteínas

What are the practical applications of knowing the ribosome structure?



© sources unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

## ncRNAs: Challenges for Computational Biology

- Prediction of ncRNA structure
- Identification of ncRNA genes
- Prediction of ncRNA functions

# Definitions

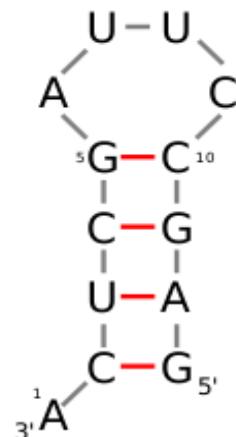
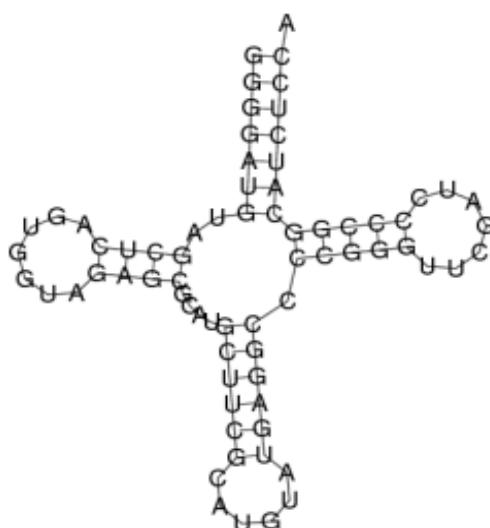
## Definition (RNA Structure)

Let  $S \in \{A, C, G, U\}^*$  be an *RNA sequence* of length  $n = |S|$ . An *RNA structure of S* is a set of *base pairs*

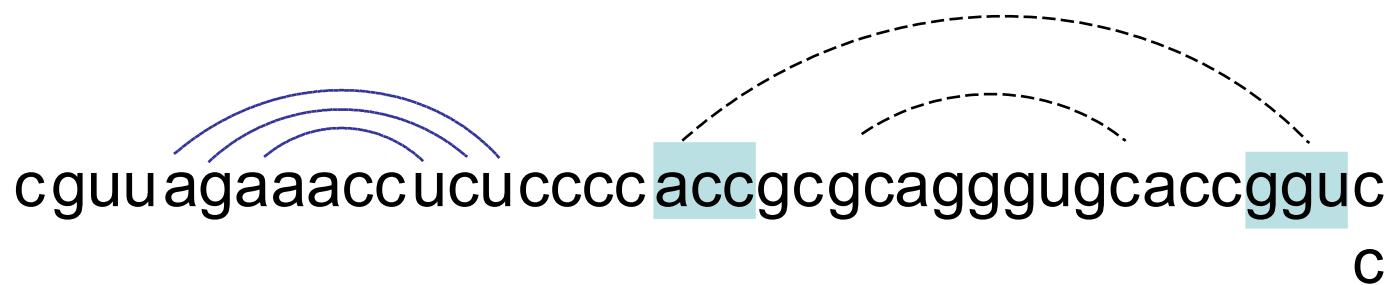
$$P \subseteq \{(i, j) \mid 1 \leq i < j \leq n, S_i \text{ and } S_j \text{ complementary}\}$$

such that the degree of  $P$  is at most one, i.e.

for all  $(i, j), (i', j') \in P : (i = i' \Leftrightarrow j = j') \text{ and } i \neq j'$ .

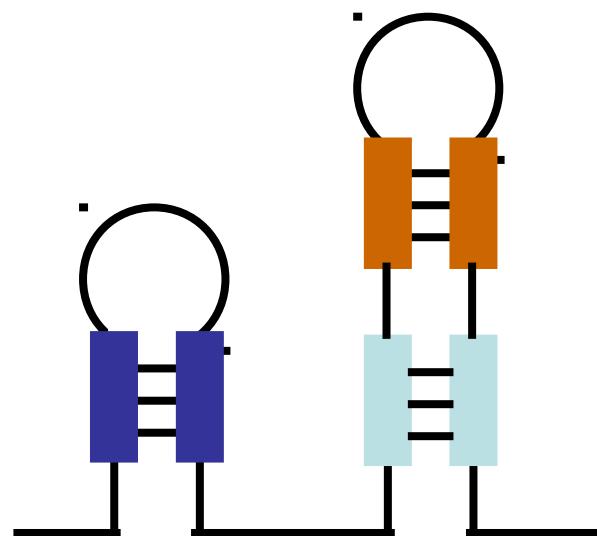


# Introducción a la estructura de ARN



stem (double helix):  
stacked base pairs

loop: strand of  
unpaired bases



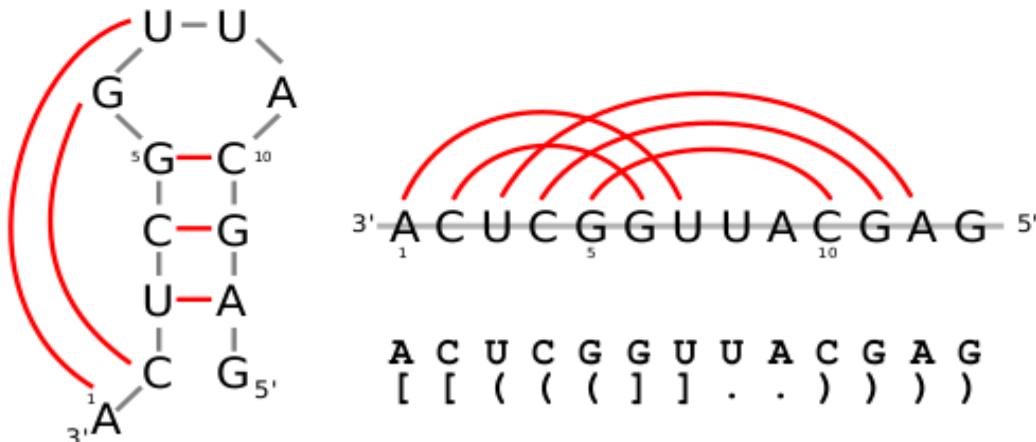
## Definitions II

### Definition (Crossing)

Two base pairs  $(i, j)$  and  $(i', j')$  are *crossing* iff

$$i < i' < j < j' \quad \text{or} \quad i' < i < j' < j.$$

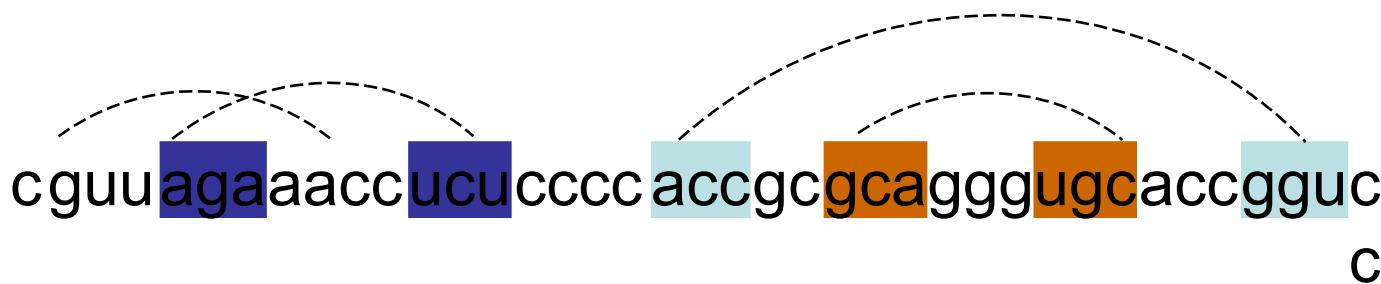
An RNA structure  $P$  (of an arbitrary RNA sequence  $S$ ) is *crossing* iff  $P$  contains (at least) two crossing base pairs. Otherwise,  $P$  is called *non-crossing* or *nested*.



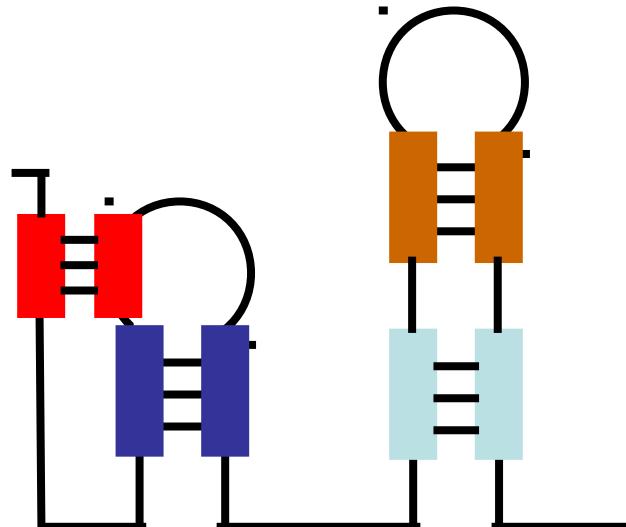
$$P = \{(1, 7), (2, 6), (3, 12), (4, 11)\}$$

# Introducción a la estructura de ARN

## Stems in crossing patterns

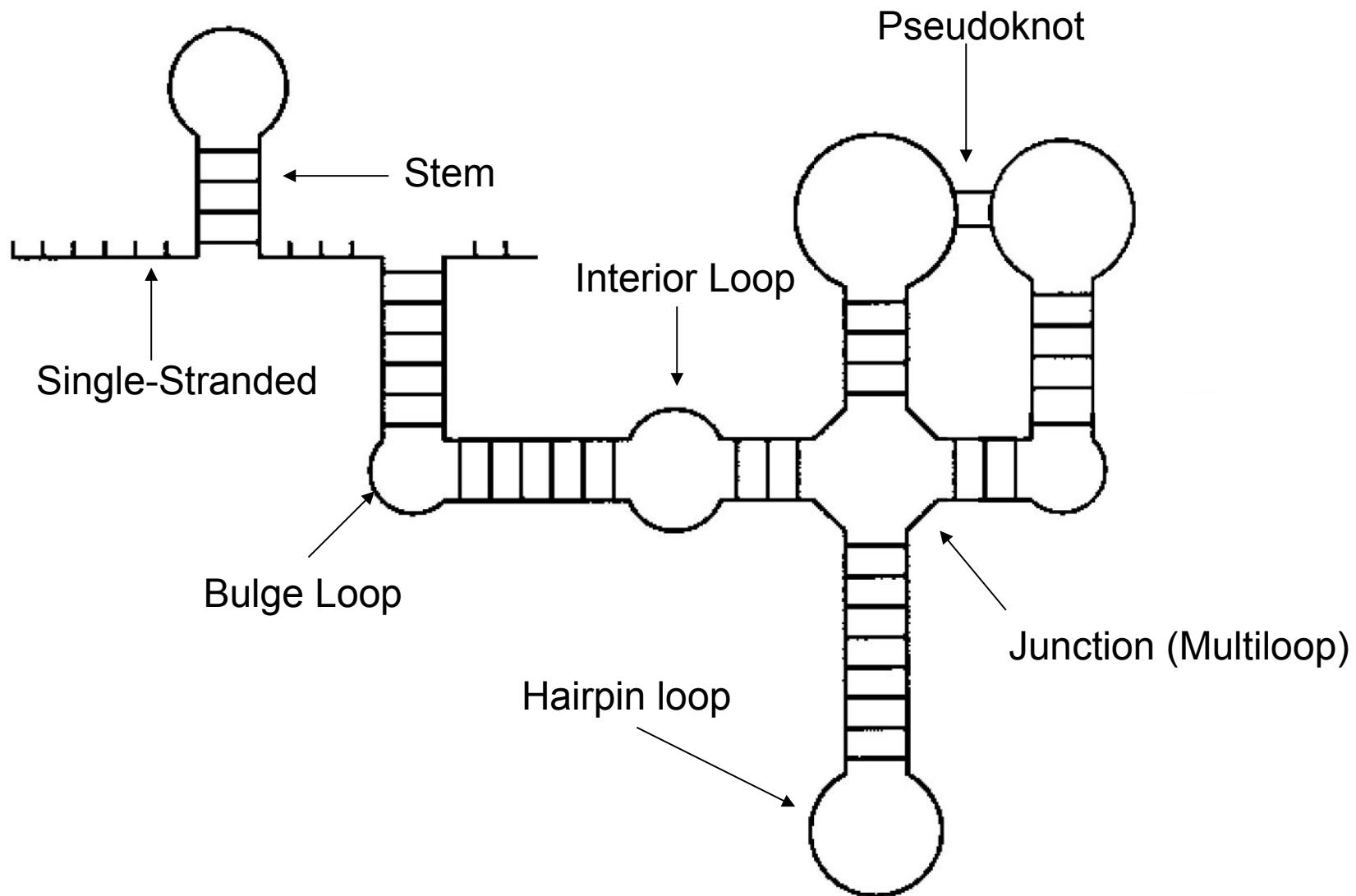


Pseudoknots: **crossing**  
patterns of stems



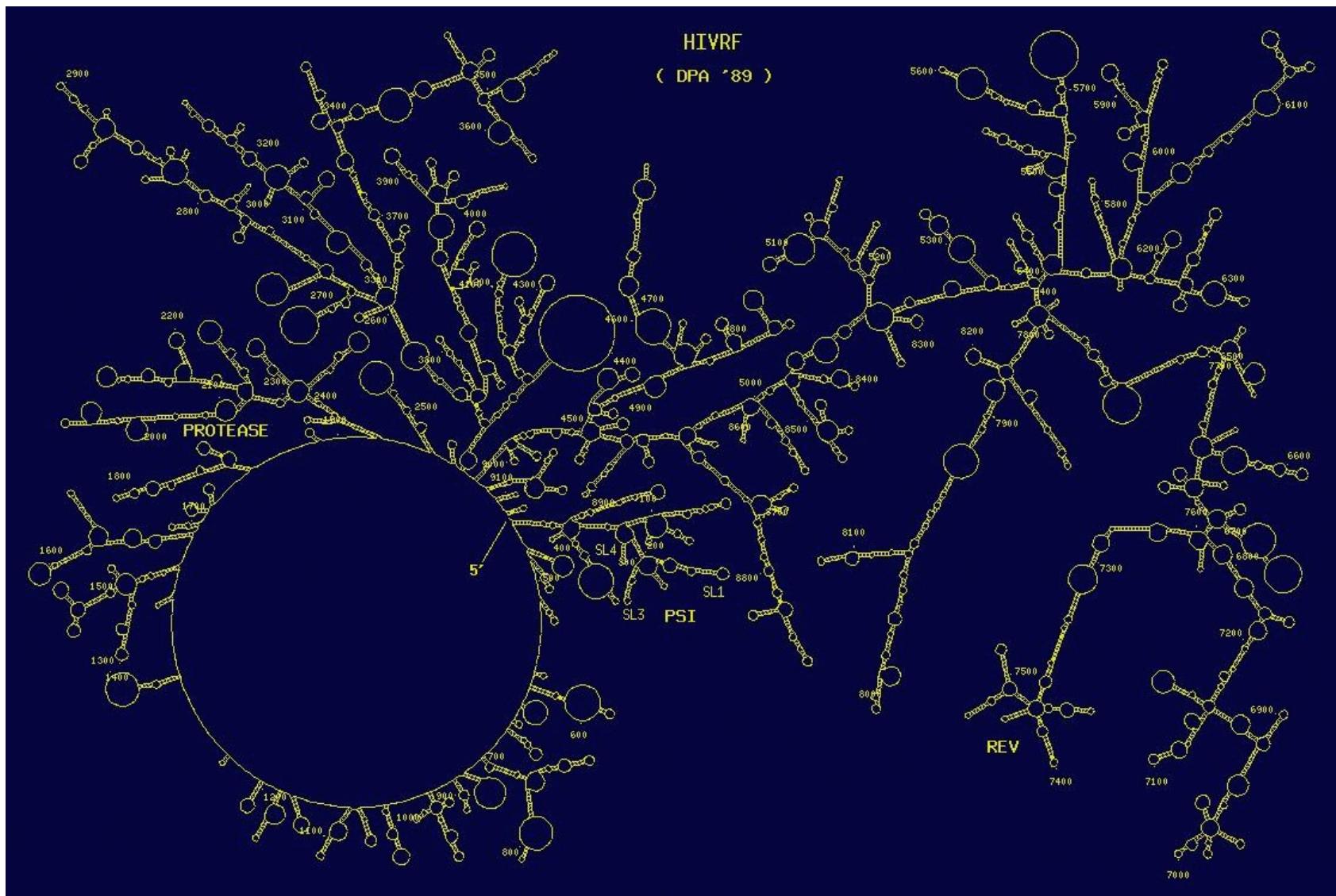
# Introducción a la estructura de ARN

## RNA secondary structure elements



# Introducción a la estructura de Proteínas

## RNA secondary structure Example



# Introducción a la estructura de ARN

## RNA secondary structure prediction

1. ab initio structure prediction  
to predict the structure of a single sequence
2. Consensus structure prediction  
to predict the structure shared by more than one sequences
3. Statistical model-based prediction and alignment  
to search for desirable structures on genomes or data bases

## What is needed for accurate inference of RNA secondary structure by covariation?

- Secondary structure more highly conserved than primary sequence
- Sufficient divergence between homologs for many variations to have occurred, but not so much that can't be aligned
- Sufficient number of homologs sequenced

## Energy Minimization Approach

$$\Delta G_{\text{folding}} = G_{\text{unfolded}} - G_{\text{folded}}$$

There are typically many possible folded states

# Introducción a la estructura de ARN – Alg. Nussinov

## How Do Energy Minimization Algorithms Work?

Consider Simple Model: Base Pair Maximization

### Scoring System:

+1 for base pair (C:G, A:U)

0 for anything else

Maximizing score equivalent to minimizing folding free energy for a model which assigns same enthalpy to all allowed base pairs (and ignores details such as base stacking, loops, entropy)

Nussinov algorithm: recursive maximization of base pairing

# Introducción a la estructura de ARN

## 1. *ab initio* structure prediction

- Hydrogen bonds consume energy contained in the molecule.
- The smaller the free energy is, the more stable the structure folded.

# Introducción a la estructura de ARN

## *ab initio* structure prediction (cont')

- Consider only canonical base pairs A-U, C-G, and G-U.  
Base pairings reduce the amount of free energy contained in the molecule.
- Maximizing the number of base pairs would minimize the free energy in the molecule.  
(Only an approximate model)

# Introducción a la estructura de ARN

## *ab initio* structure prediction (cont')

- But how to count?

An RNA could be very long; there may be many possible ways that base pairs can be formed:

e.g., .....**A**CGGUAC**G**U**C**.....

conflicting pairs **A-U**, **A-U**

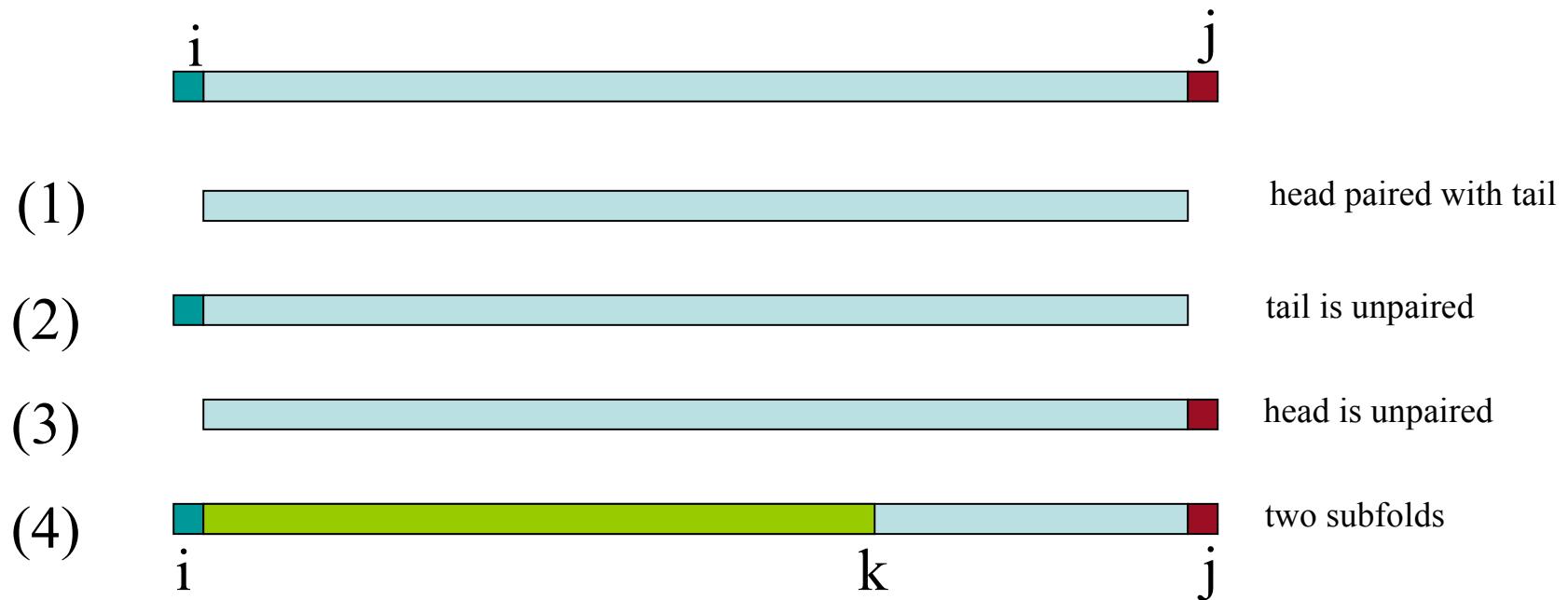
**G-C**, **G-C**

etc.

Even the number of non-conflicting combinations of base pairs is exponentially large.

# Introducción a la estructura de ARN

## *ab Initio* structure prediction (cont')



# Introducción a la estructura de ARN

looking at shorter (e.g., very short) subsequences  
in a long sequence ACGGU...ACGUC

For subsequences of length 1,

A, C, G, G, U, ..., A, C, G, U, C

#of base pairs 0, 0, 0, 0, 0, ..., 0, 0, 0, 0, 0

- For subsequences of length 2,

AC, CG, GG, GU, ..., AC, CG, GU, UC

# 0, 1, 0, 1, ..., 0, 1, 1, 0

- For subsequence of length 3,

ACG, CGG, GGU, ..., UAC, ACG, CGU, GUC, UUC

? e.g., GUC (1) G-C + U  $\rightarrow 1+0=1$  head-tail

(2) G + UC  $\rightarrow 0+0=0$  head unpaired

(3) GU + C  $\rightarrow 1+0=1$  tail unpaired

(4) GU + C  $\rightarrow 1+0=1$  split

(5) G + UC  $\rightarrow 0+0=0$  split

# Introducción a la estructura de Proteínas

Example: .....ACGGUACCGU.....

i            j    ==> max of {cases 1, 2, 3, 4}

1. Head-tail paired, count = 1 + max count in subsequence

CGGUACCG

i+1        j-1

2. Head unpaired, count = max count in subsequence

CGGUACCGU

i+1        j

- Tail unpaired, count = max count in subsequence

ACGGUACG

i        j-1

- Split (why needed and where to split ?)

ACGGUACCGU      when k=i+2

i        j    ==> ACG + GUACCGU

<---- k --->    count = max count in ACG  
                          + max count in GUACCGU

# Introducción a la estructura de ARN

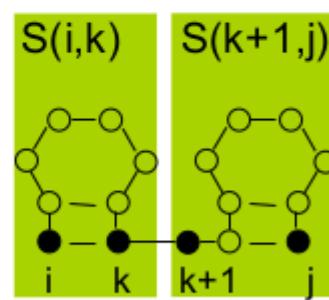
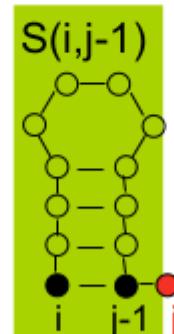
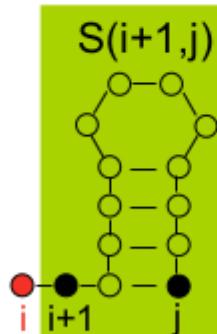
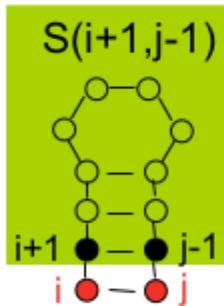
## Recursive Maximization of Base Pairing

Given an RNA sequence of length N

Define  $S(i,j)$  to be the score of the best structure for the subsequence  $(i, j)$

Notice that  $S(i,j)$  can be defined recursively in terms of optimal scores of smaller subsequences of the interval  $(i,j)$

There are four possible ways that the score of the optimal structure on  $(i,j)$  can relate to scores of optimal structures of nested subsequences:



1.  $i, j$  pair

2.  $i$  unpaired

3.  $j$  unpaired

4. bifurcation

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Eddy, Sean R. "[How do RNA Folding Algorithms Work?](#)" *Nature Biotechnology* 22, no. 11 (2004): 1457-8.

# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization Algorithm

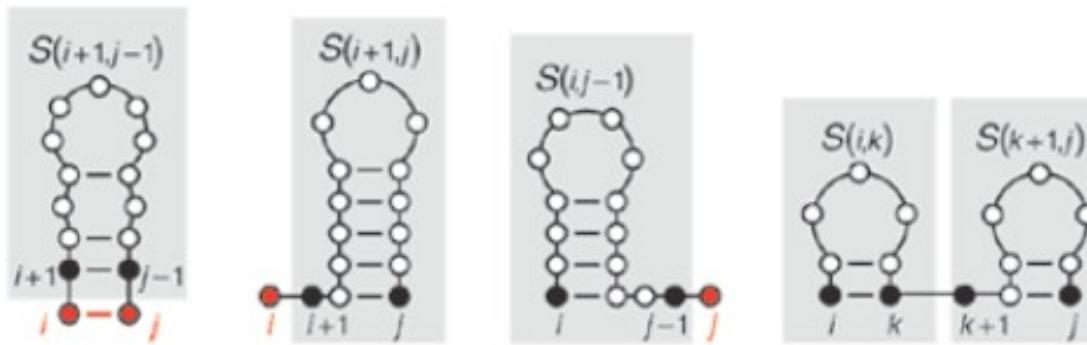
$S(i,j)$  = score of the optimal structure for the subsequence  $(i, j)$

$$S(i,j) = \max \begin{cases} S(i+1,j-1) + 1 & (\text{if } i,j \text{ base pair}) \\ S(i+1,j) & (i \text{ is unpaired}) \\ S(i,j-1) & (j \text{ is unpaired}) \\ \max_{i < k < j} S(i,k) + S(k+1,j) & (\text{bifurcation}) \end{cases}$$

- 1) Initialize an  $N \times N$  matrix  $S$  with  $S(i,i) = S(i,i-1) = 0$
- 2) Fill in  $S(i,j)$  matrix recursively from the diagonal up and to the right  
(keep track of which choice was made at each step)
- 3) Trace back from  $S(1,N)$  (upper right corner of matrix) to diagonal to determine optimal structure

# Dynamic Programming for Base Pair Maximization

Recursive definition of the best score for a sub-sequence  $i,j$  looks at four possibilities:



Dynamic programming algorithm for all sub-sequences  $i,j$ , from smallest to largest:

		j →								
		G	G	G	A	A	A	U	C	C
i ↓	G	0								
	G	0	0							
A		0	0							
A			0	0						
A				0	0					
U					0	0				
C						0	0			
C							0	0		

Initialization:

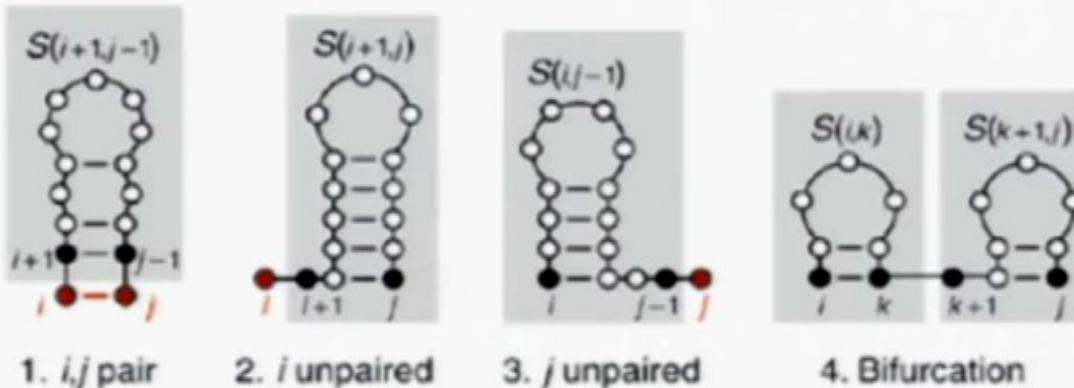
Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Eddy, Sean R. "[How do RNA Folding Algorithms Work?](#)" *Nature Biotechnology* 22, no. 11 (2004): 1457-8.

# Introducción a la estructura de ARN

## Dynamic Programming for Base Pair Maximization

Recursive definition of the best score for a sub-sequence  $i,j$  looks at four possibilities:



Dynamic programming algorithm for all sub-sequences  $i,j$ , from smallest to largest:

		$j \rightarrow$									
		G	G	G	A	A	A	U	C	C	
i \downarrow		G	0	0	0	0	0	0	0	0	
		G	0	0	0	0	0	0	0	0	
		G	0	0	0	0	0	0	0	0	
		A	0	0	0	0	0	0	0	0	
		A	0	0	0	0	0	0	0	0	
		A	0	0	0	0	0	0	0	0	
		U	0	0	0	0	0	0	0	0	
		C	0	0	0	0	0	0	0	0	
		C	0	0	0	0	0	0	0	0	

Initialization:

		$j \rightarrow$									
		G	G	G	A	A	A	U	C	C	
i \downarrow		G	0	0	0	0	0	0	1	2	3
		G	0	0	0	0	0	0	1	2	3
		G	0	0	0	0	0	0	1	2	3
		A	0	0	0	0	0	0	1	2	3
		A	0	0	0	0	0	0	1	2	3
		A	0	0	0	0	0	0	1	2	3
		U	0	0	0	0	0	0	0	0	0
		C	0	0	0	0	0	0	0	0	0
		C	0	0	0	0	0	0	0	0	0

recursive fill:

		$j \rightarrow$										
		G	G	G	A	A	A	U	C	C		
i \downarrow		G	0	0	0	0	0	0	0	1	2	3
		G	0	0	0	0	0	0	0	0	1	2
		G	0	0	0	0	0	0	0	0	1	2
		A	0	0	0	0	0	0	0	0	1	2
		A	0	0	0	0	0	0	0	0	1	2
		A	0	0	0	0	0	0	0	0	1	2
		U	0	0	0	0	0	0	0	0	0	0
		C	0	0	0	0	0	0	0	0	0	0
		C	0	0	0	0	0	0	0	0	0	0

traceback:

# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:  
Maximizing Base Pairing

$$S(i,j) = \max \begin{cases} S(i + 1, j - 1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i + 1, j) \\ S(i, j - 1) \\ \max_{i < k < j} S(i, k) + S(k + 1, j) \end{cases}$$

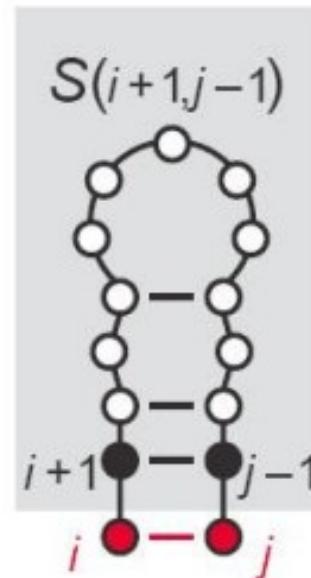
$S(i,j)$  is the folding of the subsequence of the RNA strand from index  $i$  to index  $j$  which results in the highest number of base pairs

## Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:  
Maximizing Base Pairing

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$

Base pair at  $i$  and  $j$



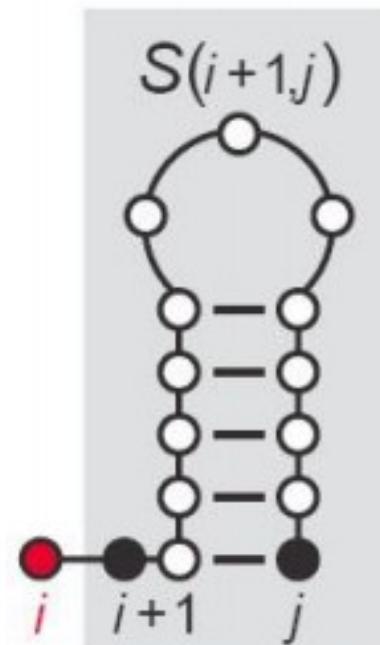
# Introducción a la estructura de Proteínas

## Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:  
Maximizing Base Pairing

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$

Unmatched at  $i$

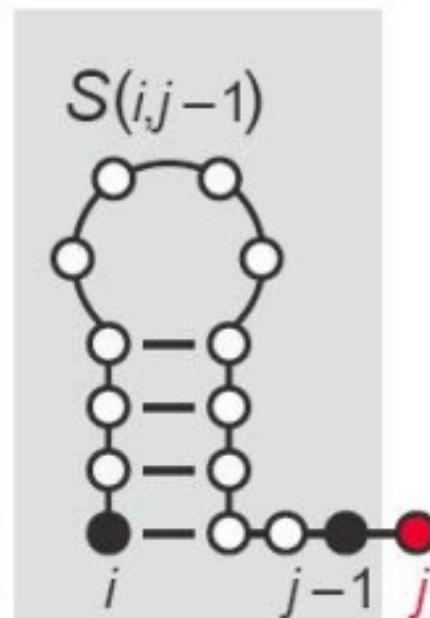


## Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:  
Maximizing Base Pairing

$$S(i,j) = \max \begin{cases} S(i + 1, j - 1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i + 1, j) \\ S(i, j - 1) \\ \max_{i < k < j} S(i, k) + S(k + 1, j) \end{cases}$$

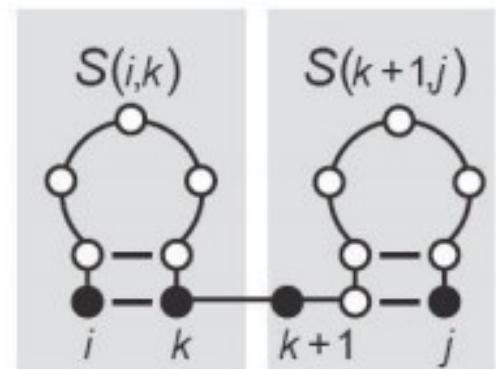
Unmatched at j



## Base Pair Maximization – Dynamic Programming Algorithm

Simple Example:  
Maximizing Base Pairing

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$

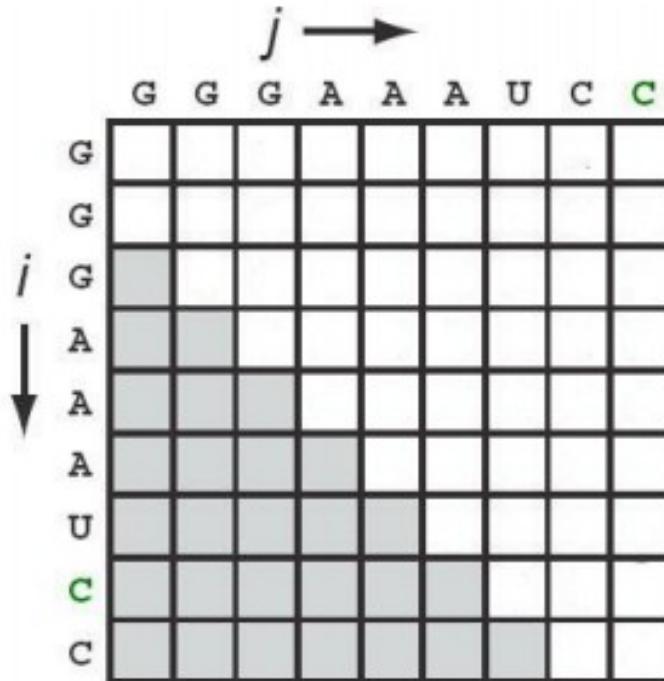


Bifurcation

# Introducción a la estructura de ARN

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension



# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
    - Align RNA strand to itself
    - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Initialize first two diagonal arrays to 0

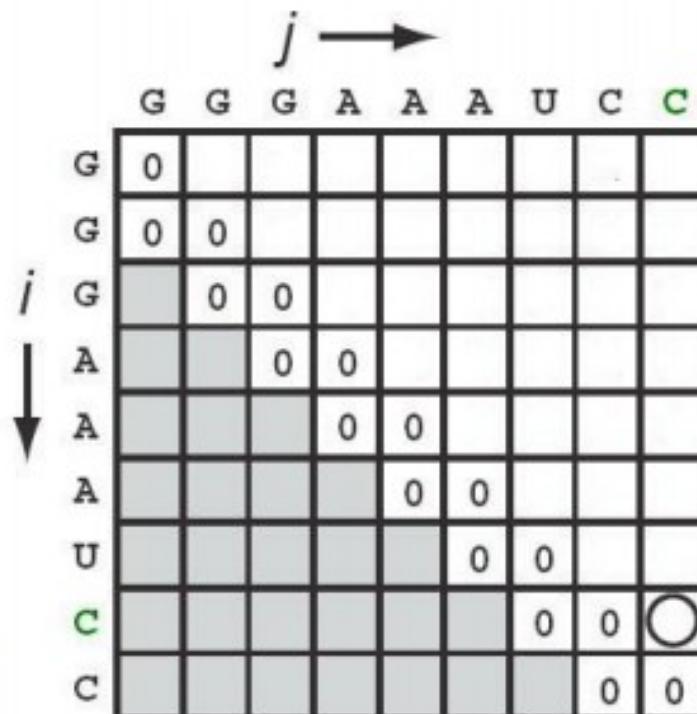
		j →									
		G	G	G	A	A	A	U	C	C	C
i ↓	G	0									
	G	0	0								
G		0	0								
A			0	0							
A				0	0						
A					0	0					
U						0	0				
C							0	0			
C								0	0		

# Introducción a la estructura de ARN

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
    - Align RNA strand to itself
    - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Fill in squares sweeping diagonally



# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
    - Align RNA strand to itself
    - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Bases cannot pair, similar to unmatched alignment

# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bases can pair, similar to matched alignment

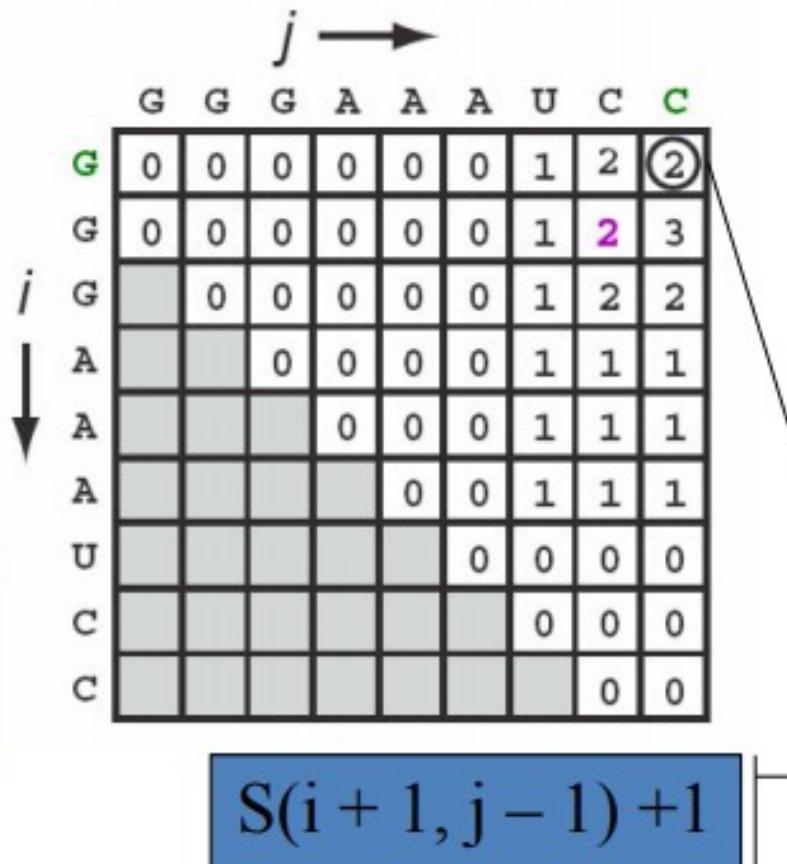
		j →								
		G	G	G	A	A	A	U	C	C
i ↓	G	0	0	0	0	0	0	1	2	0
	G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	2	2
A			0	0	0	0	1	1	1	1
A				0	0	0	1	1	1	1
A					0	0	1	1	1	1
U						0	0	0	0	0
C							0	0	0	0
C								0	0	0

# Introducción a la estructura de ARN

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Dynamic Programming – possible paths

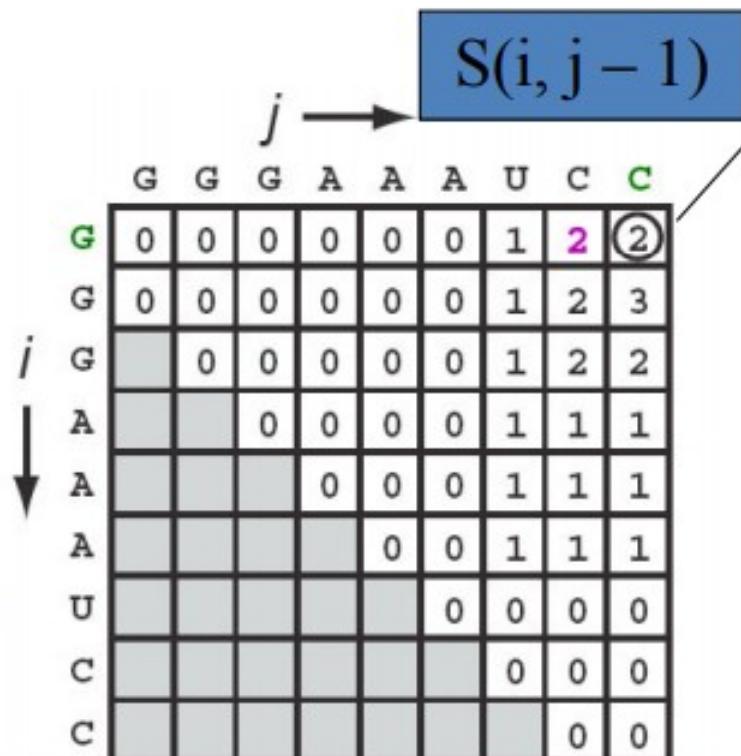


# Introducción a la estructura de ARN – Alg. NUSSINOV

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Dynamic Programming – possible paths

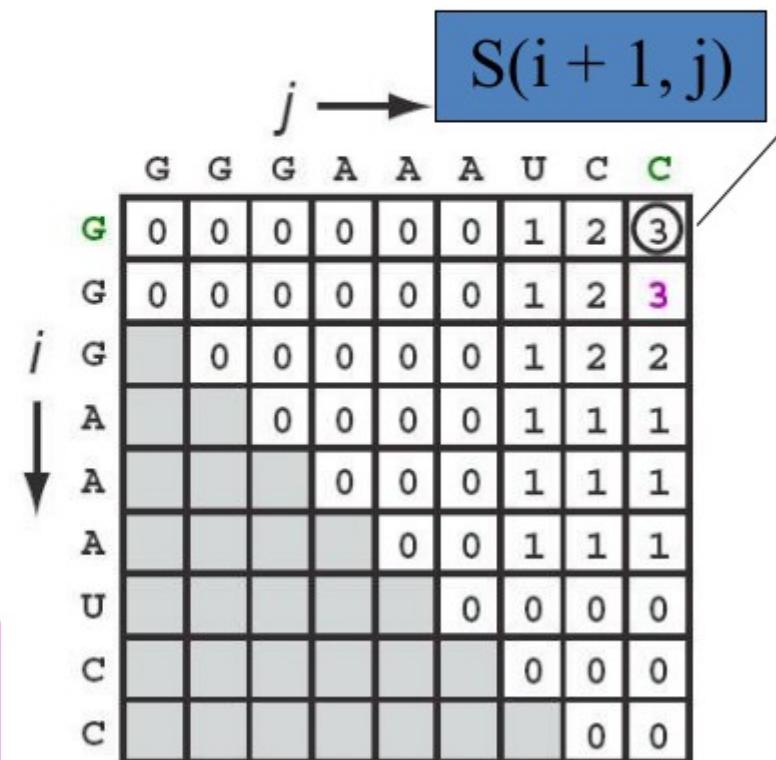


# Introducción a la estructura de ARN – Alg. NUSSINOV

# Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
    - Align RNA strand to itself
    - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

## Dynamic Programming – possible paths

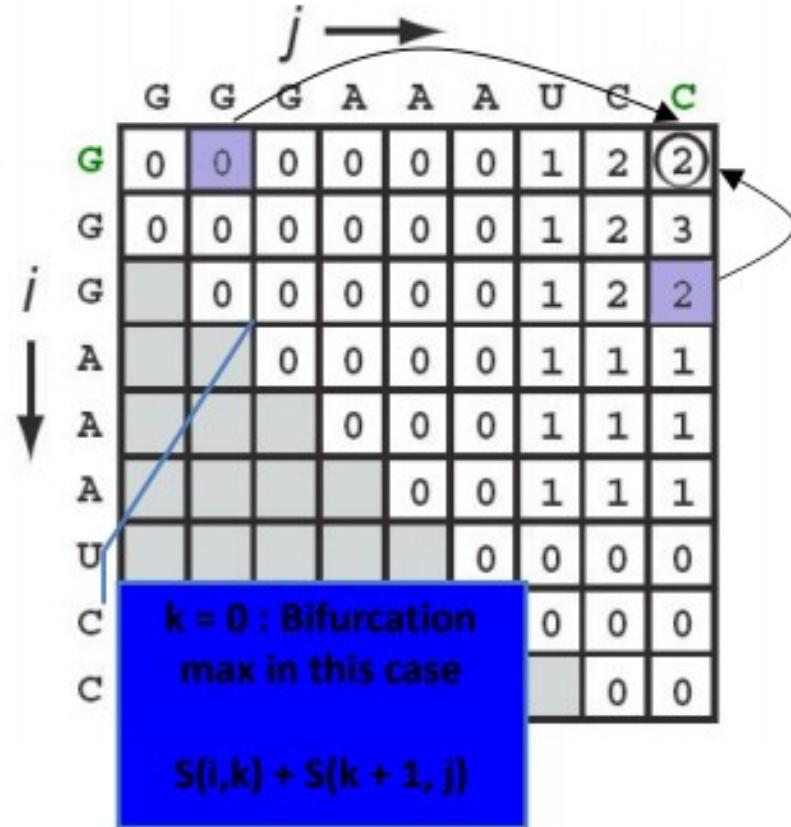


# Introducción a la estructura de ARN

## Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
  - Align RNA strand to itself
  - Score increases for feasible base pairs
- Each score independent of overall structure
- Bifurcation adds extra dimension

Bifurcation – add values for all k

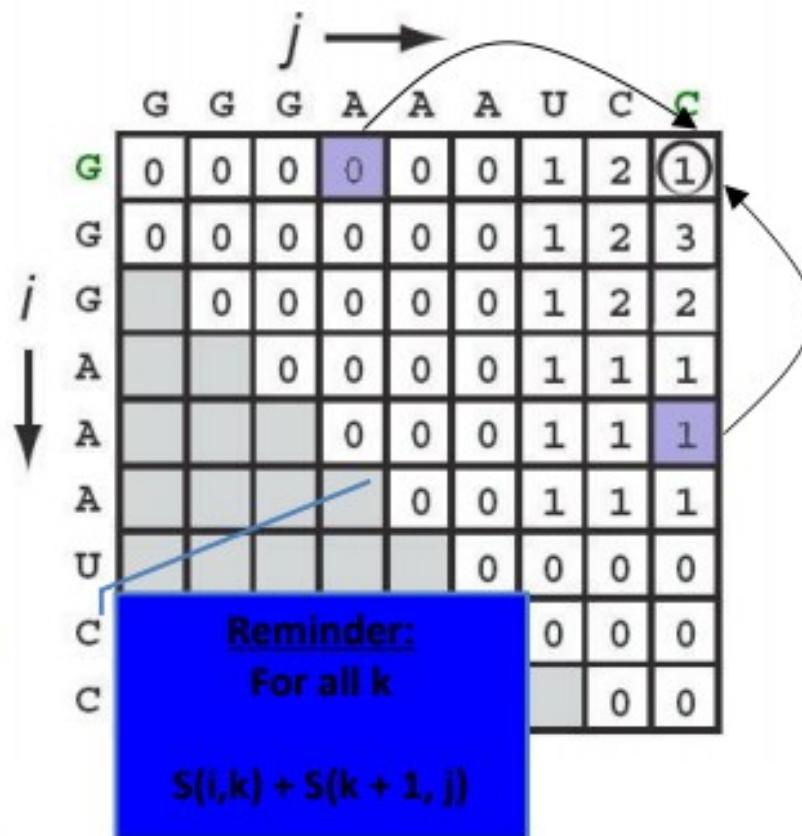


# Introducción a la estructura de ARN – Alg. NUSSINOV

# Base Pair Maximization – Dynamic Programming Algorithm

- Alignment Method
    - Align RNA strand to itself
    - Score increases for feasible base pairs
  - Each score independent of overall structure
  - Bifurcation adds extra dimension

Bifurcation – add values  
for all k



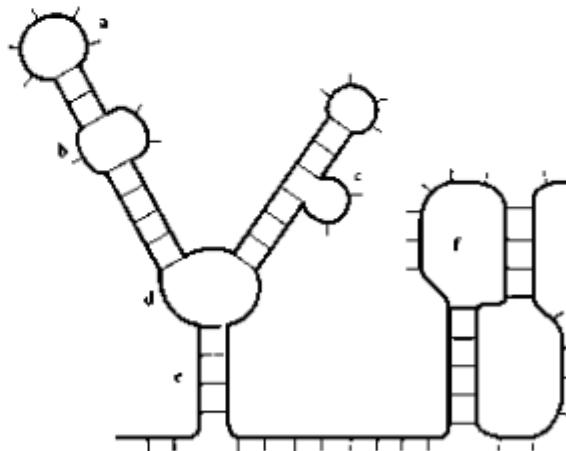
# Introducción a la estructura de ARN

## Base Pair Maximization Algorithm Issues

- What is computational complexity of algorithm?  
(for sequence of length  $N$ )

Answer: Memory -  $O(N^2)$  Time -  $O(N^3)$

- Can it handle pseudoknots?



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Answer: No. Pseudoknots invalidate recursion for  $S(i,j)$

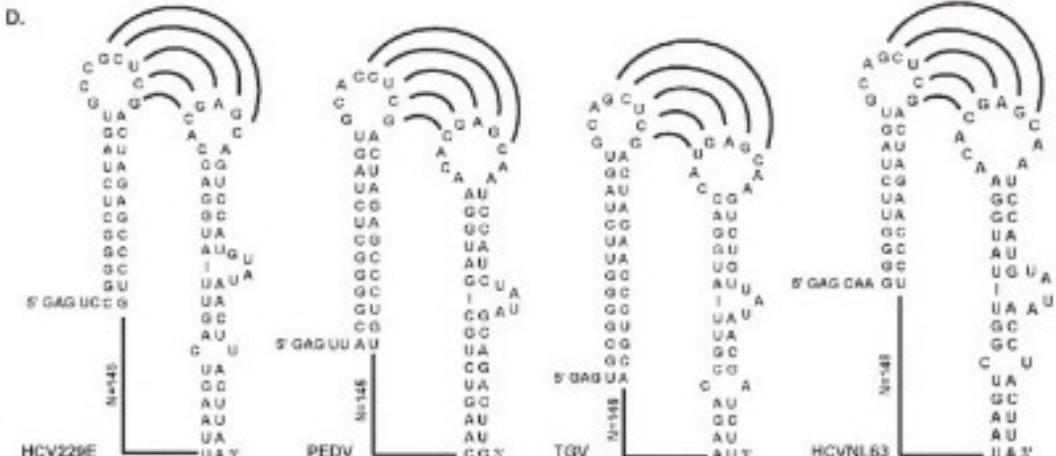
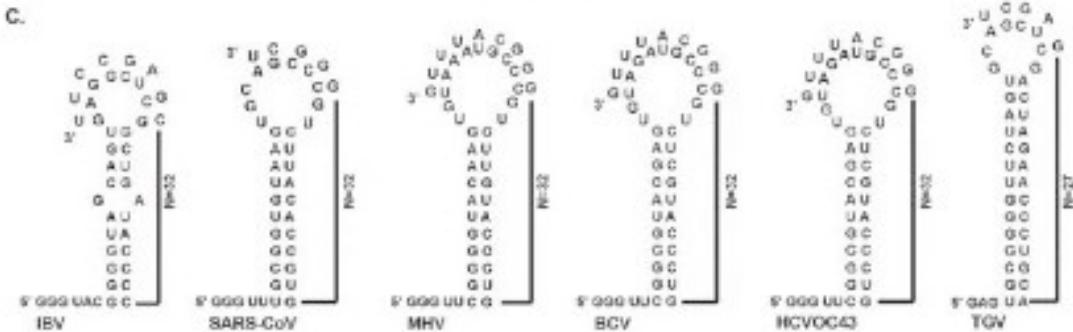
# Viral Pseudoknots and “Kissing loops”

A.

	STEMI	STEMII	STEMI	STEMII
BCV	GGTTTAAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
HSV	CCTTTAAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
BCVOC43	GGTTTAAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
SARS	GTTTTAAAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
IBV	TTAATTAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
BCV229E	TTAATTAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
PEDV	TTAATTAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
TGV	TTAATTAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			
BCVNL63	TTAATTAACGGGTTGGGGTAAGGGTGATGAGGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG			

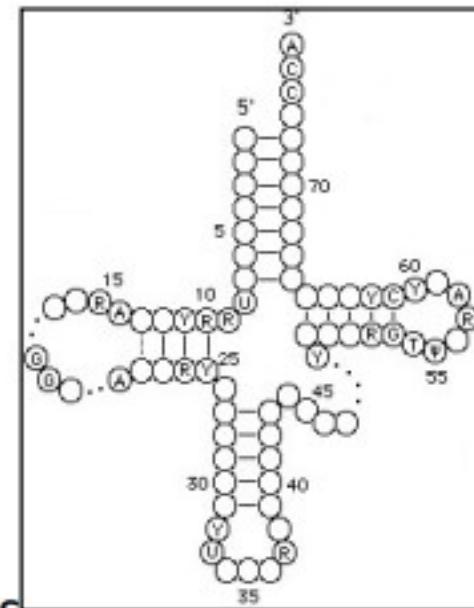
B.

	STEMIII	STEMII	STEMIII
BCV	AAAGGAGACAGACTGAACTATAT-AATGAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
HSV	AAAGGAGACATTTAGGAGTGT-TATAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
BCVOC43	AAAGGAGACAGACTGAACTATAT-AATGAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
SARS	AAAGGAGCACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
IBV	AAAGGAGCACTGAACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
BCV229E	AAAGGAGCACTGAACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
PEDV	AAAGGAGCACTGAACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
TGV	AAAGGAGCACTGAACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		
BCVNL63	AAAGGAGCACTGAACTATATGTTTAACTAACACATGAAAGGAGTGGGGTGCGCTGGGGTATCCTCTGATGTCACATTAAAGGGGATTG		



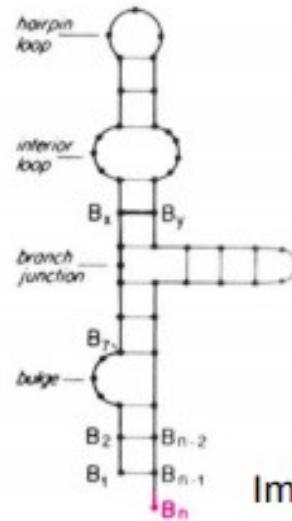
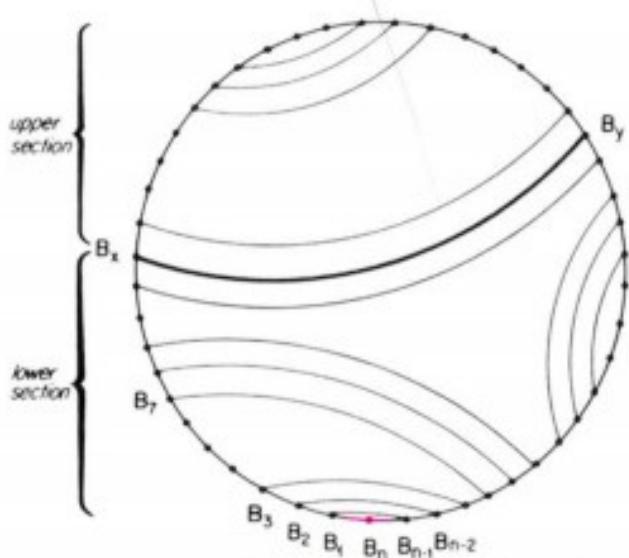
## Energy Minimization

- Thermodynamic Stability
  - Estimated using experimental techniques
  - Theory : Most Stable is the Most likely
- No Pseudoknots due to algorithm limitations
- Uses Dynamic Programming alignment technique
- Attempts to maximize the score taking into account thermodynamics
- MFOLD and ViennaRNA



# Introducción a la estructura de Proteínas

## Energy Minimization Results

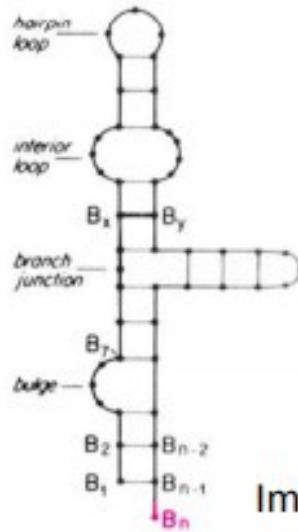
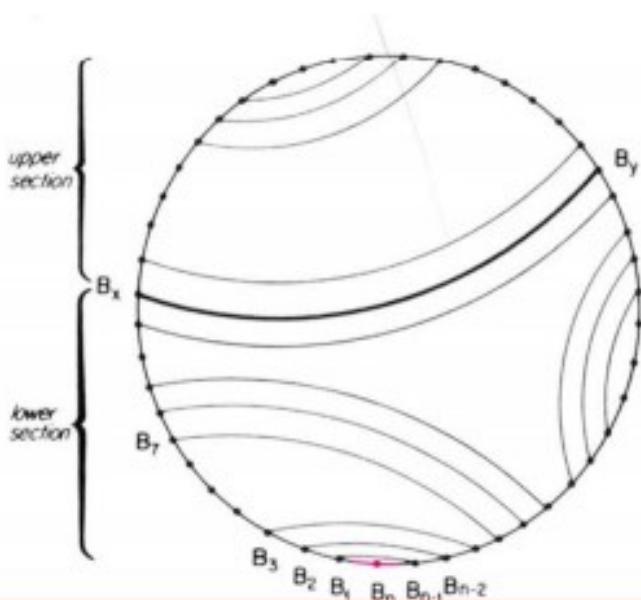


Images – David Mount

- Linear RNA strand folded back on itself to create secondary structure
- Circularized representation uses this requirement
  - Arcs represent base pairing

# Introducción a la estructura de ARN

## Energy Minimization Results



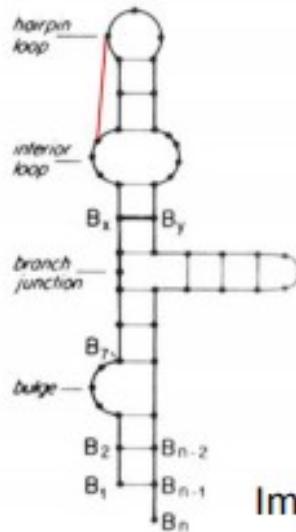
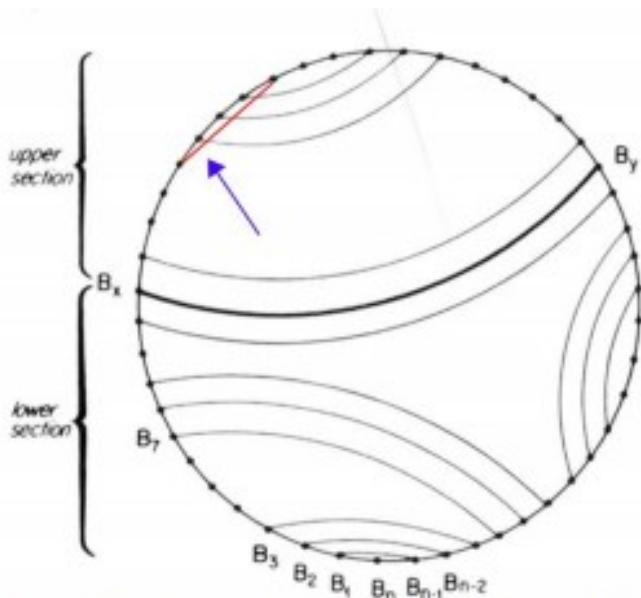
Images – David Mount

- All loops must have at least 3 bases in them  
Equivalent to having 3 base pairs between all arcs

**Exception:** Location where the beginning and end of RNA come together in circularized representation

# Introducción a la estructura de Proteínas

## Trouble with Pseudoknots



Images – David Mount

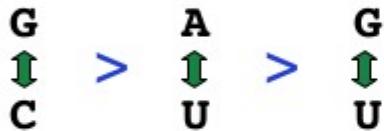
- Pseudoknots cause a breakdown in the Dynamic Programming Algorithm.
- In order to form a pseudoknot, checks must be made to ensure base is not already paired – this breaks down the recurrence relations

# Introducción a la estructura de ARN

## RNA Energetics I

Free energy contributions to helix formation come from:

- base pairing:

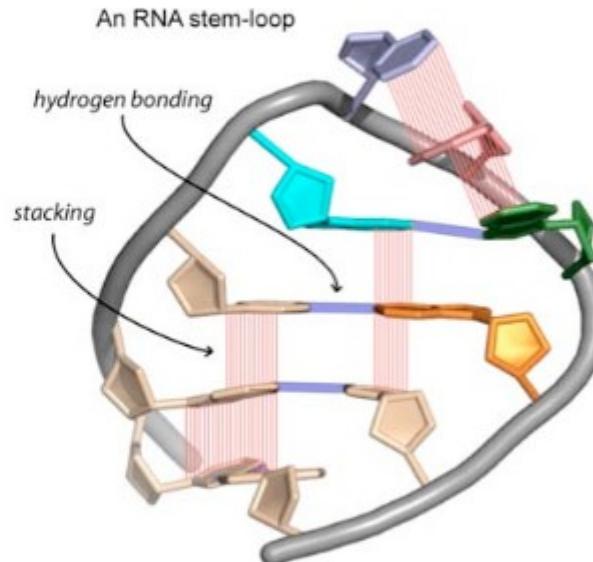


- base stacking:



Base stacking contributes more to free energy than base pairing

3' ...CCAUUCAUAG...5'  
|||  
5' ...CGUGAGU...3'



© American Chemical Society. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.  
Source: Mohan, Srividya, Chiaolong Hsiao, et al. "RNA Tetraloop Folding Reveals Tension Between Backbone Restraints and Molecular Interactions." *Journal of the American Chemical Society* 132, no. 36 (2010): 12679-89.

## RNA Energetics III

### Folding by Energy Minimization

A more complex dynamic programming algorithm is used - similar in spirit to the Nussinov base pair maximization algorithm

Gives:

- minimum energy fold
- suboptimal folds (e.g., five lowest  $\Delta G$  folds)
- probabilities of particular base pairs
- full partition function

Accuracy: ~70% of base pairs correct

# Introducción a la estructura de Proteínas

## Links & References

The Mfold web server:

<http://mfold.rna.albany.edu/?q=mfold/rna-folding-form>

The Vienna RNAfold package (free for download)

<http://www.tbi.univie.ac.at/~ivo/RNA/>

**RNA folding references:**

M. Zuker, et al. In *RNA Biochemistry and Biotechnology* (1999)

D.H. Mathews et al. *J. Mol. Biol.* **288**, 911-940 (1999)

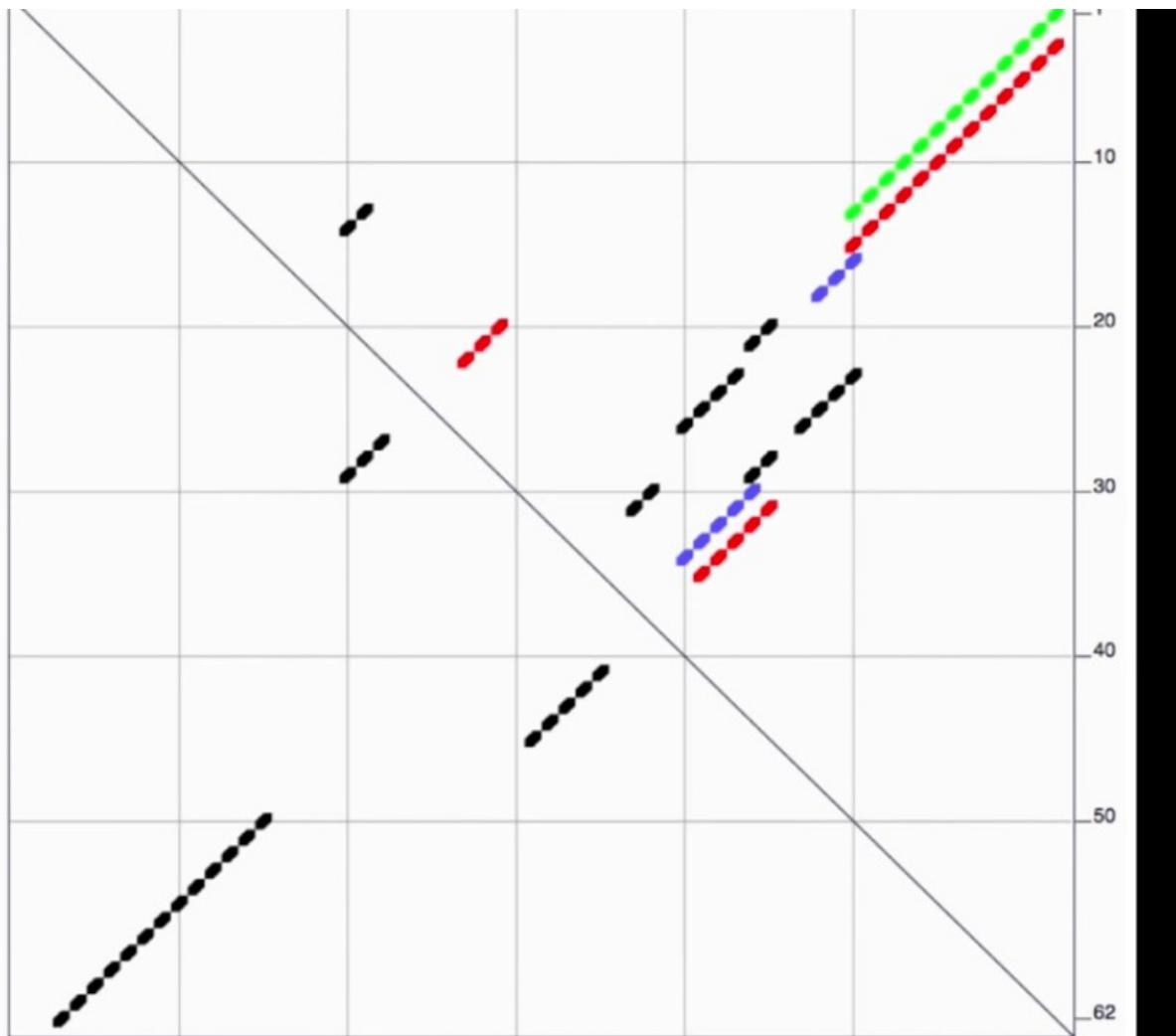
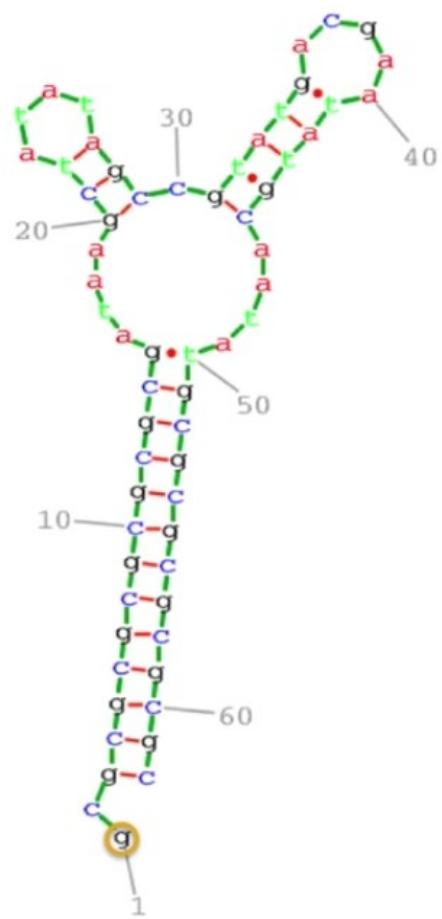
Vienna package by Ivo Hofacker

# Introducción a la estructura de Proteínas

## RNA Secondary Structure Prediction by Energy Minimization Summary

- Assumes folding energy decomposable into independent contributions of small units of structure
- Algorithms are guaranteed to find minimal free energy structure defined by the model
- In practice, algorithms predict ~70% of bp correct
- Errors result from
  - imprecision of the model/parameters
  - differences between *in vitro* and *in vivo* conditions
  - *in vivo* structure may not always have minimum free energy

# Introducción a la estructura de Proteínas



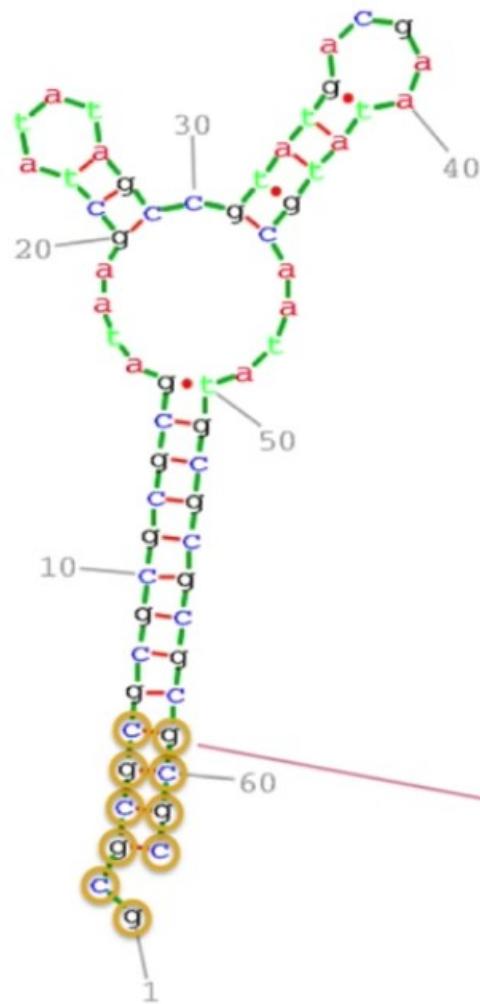
RNA – Dot Plot

# Introducción a la estructura de Proteínas

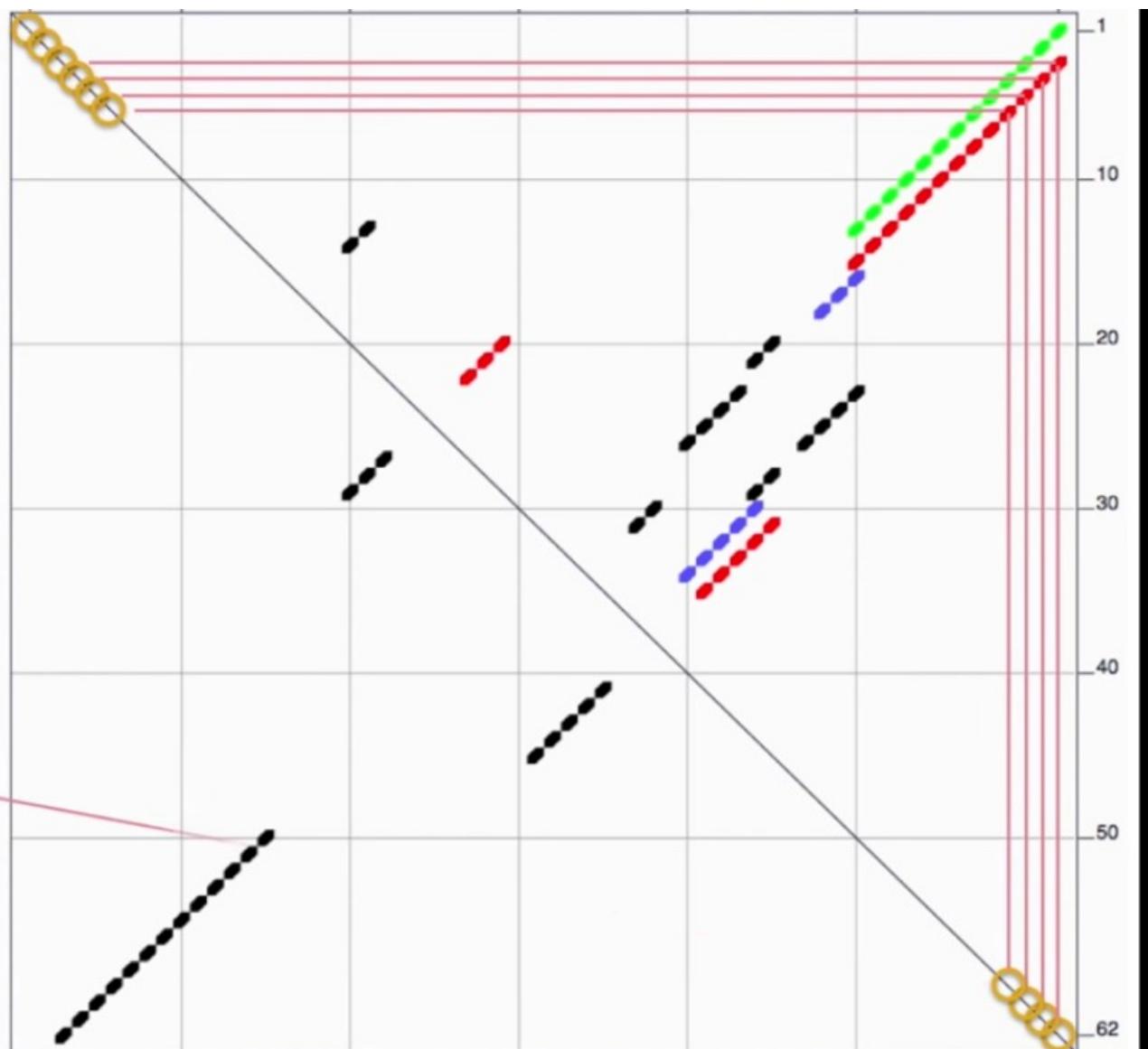


RNA – Dot Plot

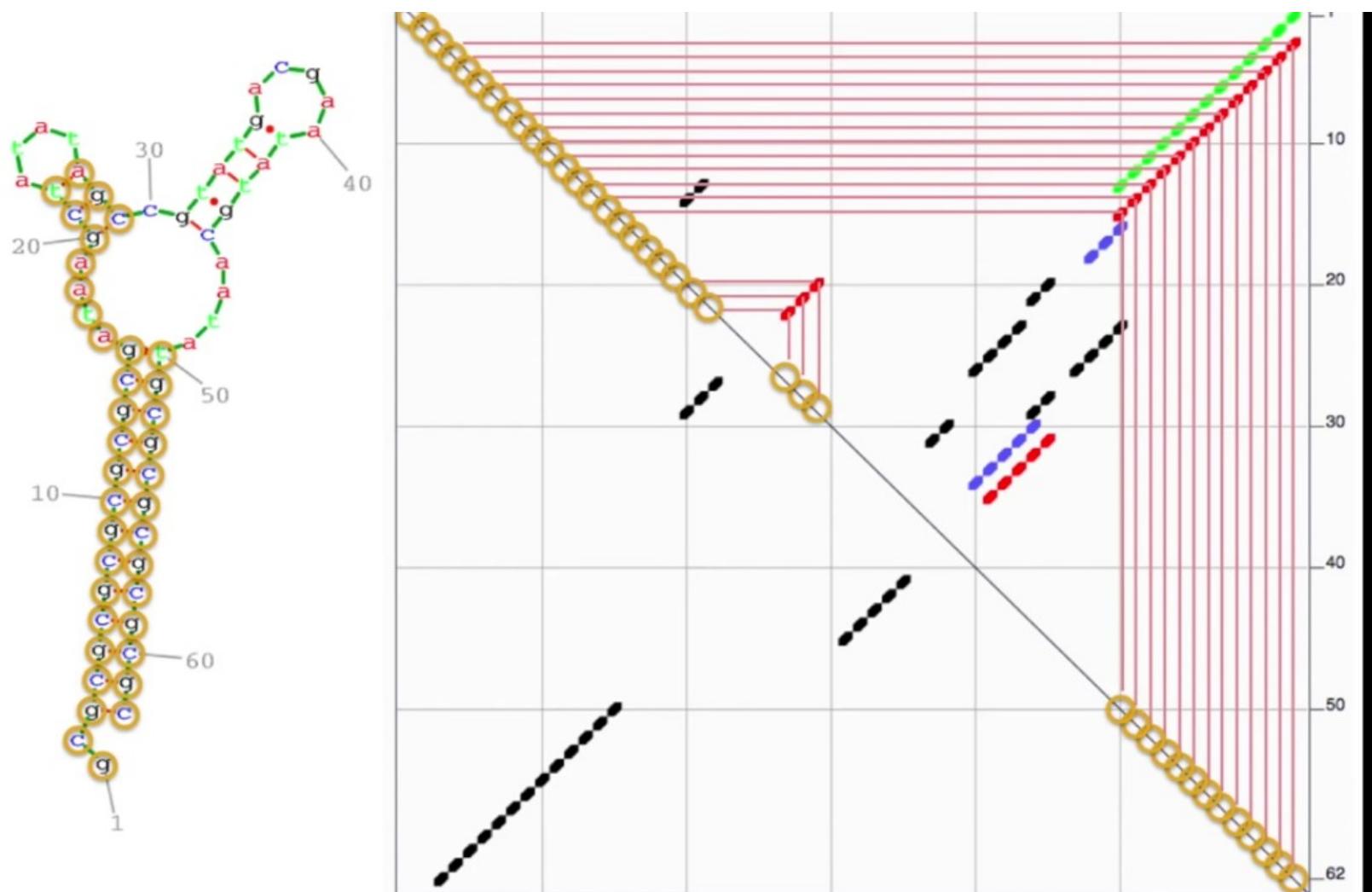
# Introducción a la estructura de Proteínas



RNA – Dot Plot



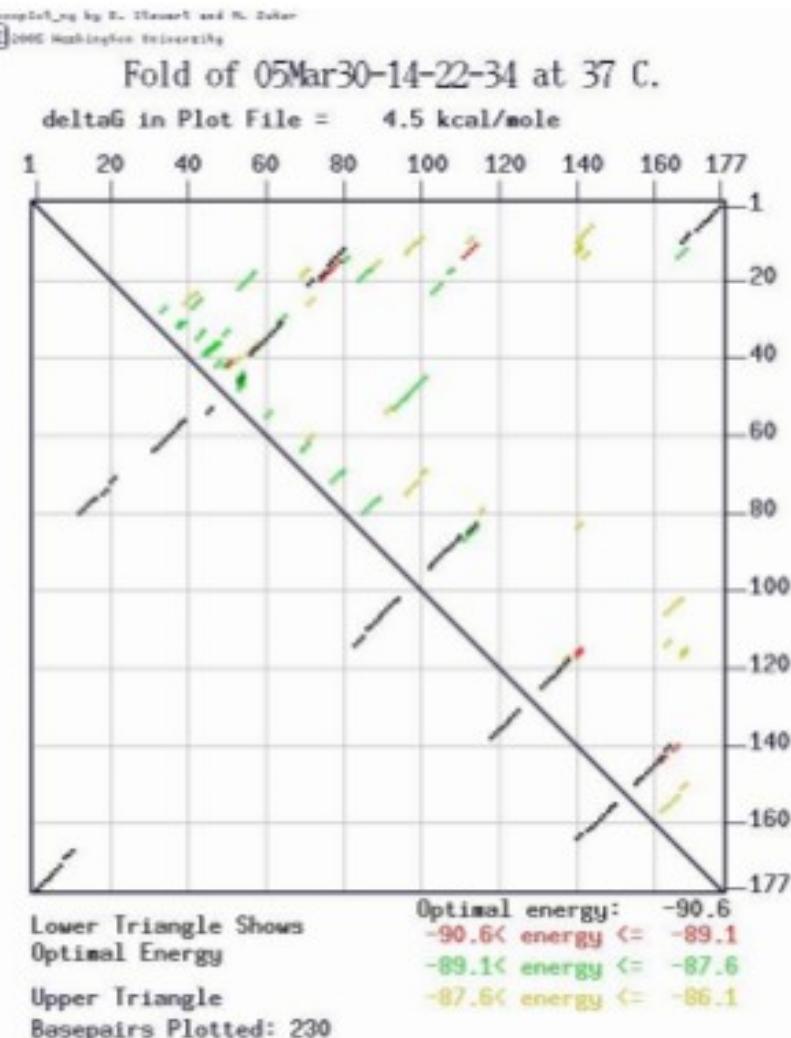
# Introducción a la estructura de Proteínas



RNA – Dot Plot

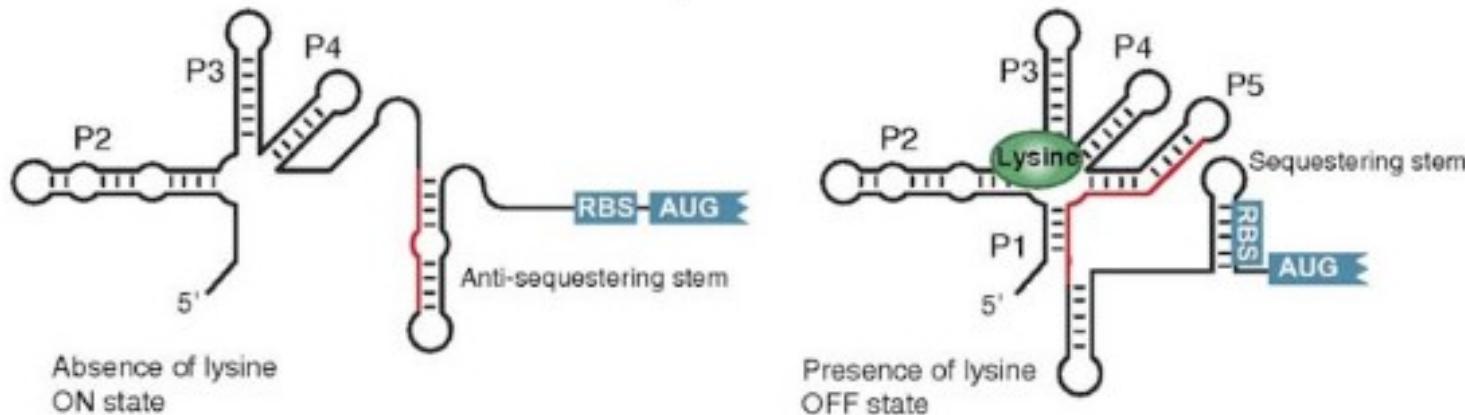
# Introducción a la estructura de Proteínas

Energy dot plot for  
a lysine riboswitch



# Introducción a la estructura de Proteínas

## Function of the lysine riboswitch



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Lysine interacts with the junctional core of the riboswitch and is specifically recognized through shape-complementarity within the elongated binding pocket and through several direct and K<sup>+</sup>-mediated hydrogen bonds to its charged ends.

Controls expression of enzymes involved in biosynthesis and transport of lysine

# Introducción a la estructura de Proteínas

# Introducción a la estructura de Proteínas

## Referencias:

<http://www.uniprot.org/>

- UNIPROT

<http://www.rcsb.org/pdb/home/home.do>

PDB – Protein Data Bank

<https://www.ncbi.nlm.nih.gov/protein>

Protein Database at NCBI

<http://www.ebi.ac.uk/pdbe/pisa/>

Protein Data Bank