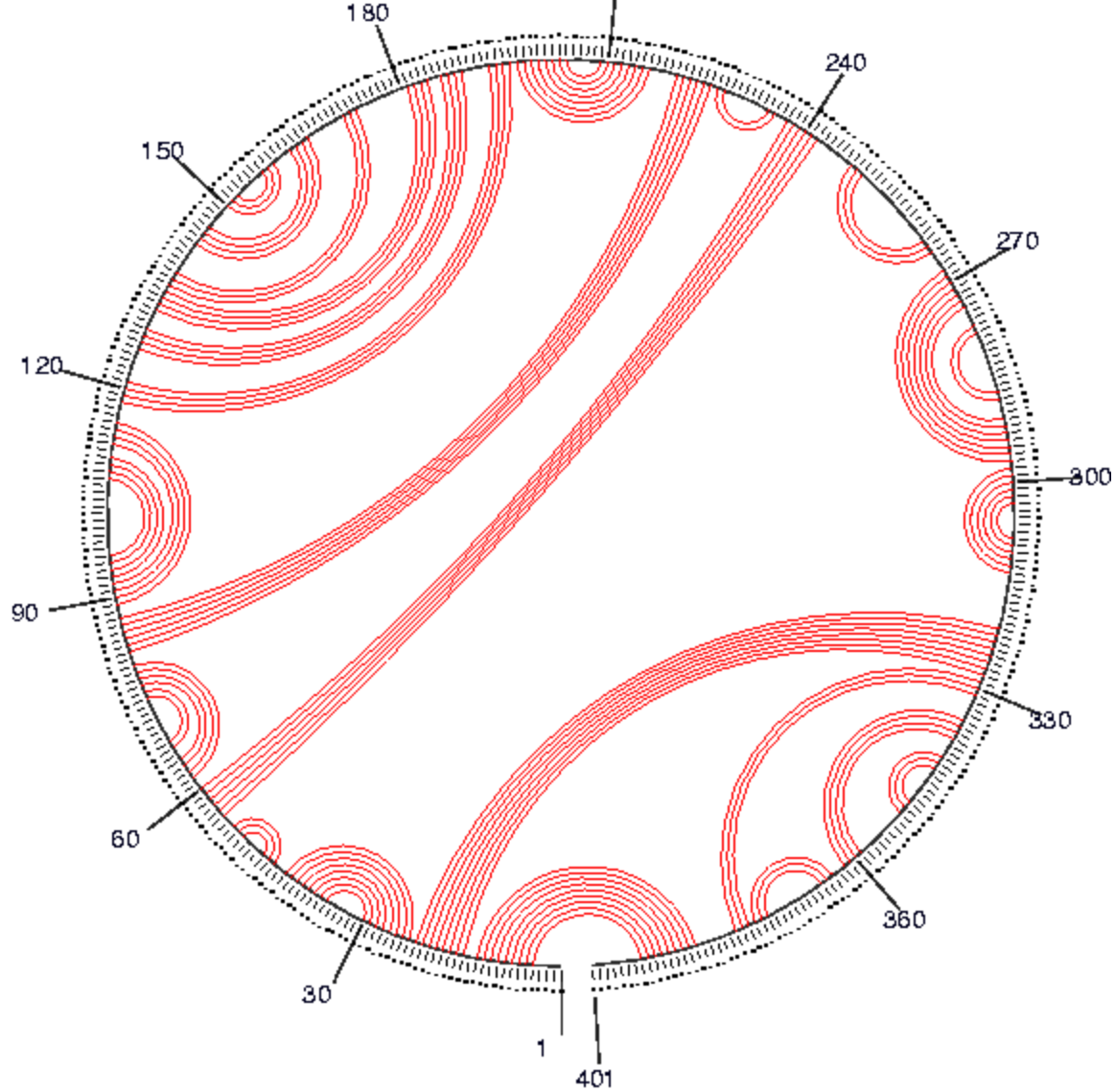# GLOBEX Bioinformatics (Summer 2015)
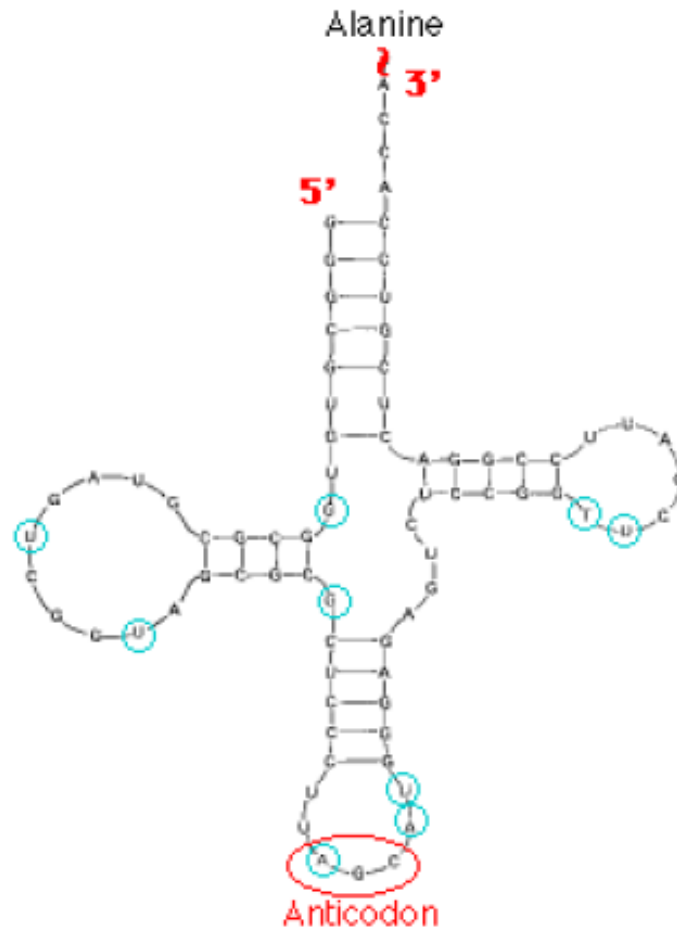
RNA secondary structure

Facts about RNAs

- Mainly as "information carrier" in protein synthesis
  - mRNA
  - tRNA
- Also as catalysts
  - Ribozymes
- Also as regulator
  - RNAi
- Single strand
  - Intra-molecular pairing
  - Secondary structure
  - Essential for sequence stability and function

- Similarity in secondary structure plays a more important role than primary sequence in determining homology for RNAs
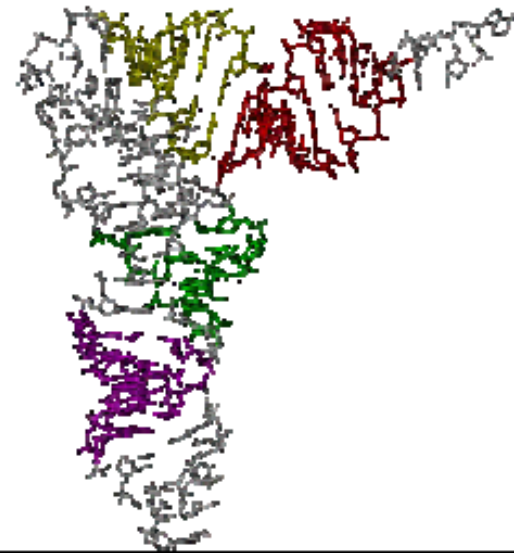
Bacillus subtilis RNase P RNA

M - multi-loop

I - interior loop

B - bulge loop
H - hairpin loop

3

ENERGY =   -85.7   Bacillus subtilis RNase P RNA

4

tRNA Secondary Structure

Alanine

tertiary structure

Anticodon

RNA stem loop

hybridization pairing: A-U, and C-G

```
     A   A                      C   A                    C   A

  G        A                 G        A               G        A
    G  —  C                   U  —  A                 U   x   C
    A  —  U                   C  —  G                 C   x   U
    C  —  G                   G  —  C                 G   x   G

       seq1                      seq2                    seq3
```

where "-" stands for a pairing, and "x" for no pairing.

In the above example, seq1 and seq2 fold into a similar structure, whereas seq3 does not.

Pairwise alignments disregarding such structural restrictions may be misleading; seq2 and seq3 have 70% sequence identity, seq1 and seq3 have 60%, whereas seq1 and seq2 have only 30%.

```
seq1   C  A  G  G  A  A  A  C  U  G
seq2   G  C  U  G  C  A  A  A  G  C
seq3   G  C  U  G  C  A  A  C  U  G
```

Representations for palindromic structures

1)

```
        A   A
     G         A
      G  —  C
      A  —  U
      C  —  G
```

2)

```
C  A  G  G  A  A  A  C  U  G
```

3)

```
C  A  G  G  A  A  A  C  U  G
(  (  (  .  .  .  .  )  )  )
```

There is a 1-1 correspondence between RNA secondary structures and well-balanced parenthesis expressions, where the balancing parentheses correspond to base pairings via hydrogen bonds.

Secondary structure prediction

Base pair maximization algorithm [Nussinov]

$m_{i,j}$: maximal # of base pairs that can be formed for sequence $s_i \ldots s_j$.

$d_{i,j}$ = 1, if s_i and s_j are paired
    = 0, otherwise

Initialization

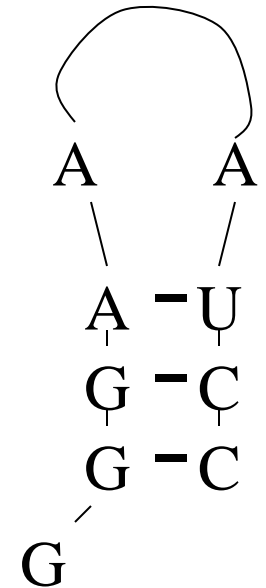$m_{i,i-1} = 0$ for i= [2, L]
$m_{i,i} = 0$ for i = [1,L]

Recursion

$m_{i,j}$= max {$m_{i+1,j}$,
          $m_{i,j-1}$,
          $m_{i+1,j-1}+ d_{i,j}$,
          $\max_{i<k<j} \{m_{i,k}+ m_{k+1,j}\}$ // bifurcation
          }

Dynamic programming table

		1 G	2 G	3 G	4 A	5 A	6 A	7 U	8 C	9 C
1	G	0→						1	2	③
2	G		0					1	2	③
3	G			0				1	②	2
4	A				0			①	1	1
5	A					0	⓪	1	1	1
6	A						⓪	1	1	1
7	U							0		
8	C								0	
9	C									0

A   A

A ─ U

G ─ C

G ─ C

G

Time Complexity: $O(L^3)$, where L is sequence length

Space complexity: $O(L^2)$

9

Secondary structure prediction

Minimum energry algorithm [Zuker]

$E_{i,j}$: minimum energy for an optimal fold formed by sequence $s_i \ldots s_j$.
$e_{i,j} = -5$, if s_i , s_j is CG or GC
$\quad = -4$, if s_i , s_j is AU or UA
$\quad = -1$, if s_i , s_j is GU or UG

Initialization
$\quad E_{i,i-1} = 0 \qquad$ for i= [2, L]
$\quad E_{i,i} = 0 \qquad$ for i = [1,L]
Recursion
$E_{i,j} = \min \{ E_{i+1,j},$
$\qquad\qquad E_{i,j-1},$
$\qquad\qquad E_{i+1,j-1} + e_{i,j},$
$\qquad\qquad \min_{i<k<j} \{ E_{i,k} + E_{k+1,j} \} \qquad$ // bifurcation
$\qquad\qquad \}$

Resources

Lecture notes:

http://www.bioinfo.rpi.edu/~zukerm/lectures/RNAfold-html/

RNA informatics:

http://www-lbit.iro.umontreal.ca/RNA_Links/RNA.shtml

Software:

-Vienna RNA fold

- Mfold (http://www.bioinfo.rpi.edu/zukerm/export/)