RNA secondary structure

S.R. Eddy. *Nature Biotechnology*, 22:1457-1458, 2004

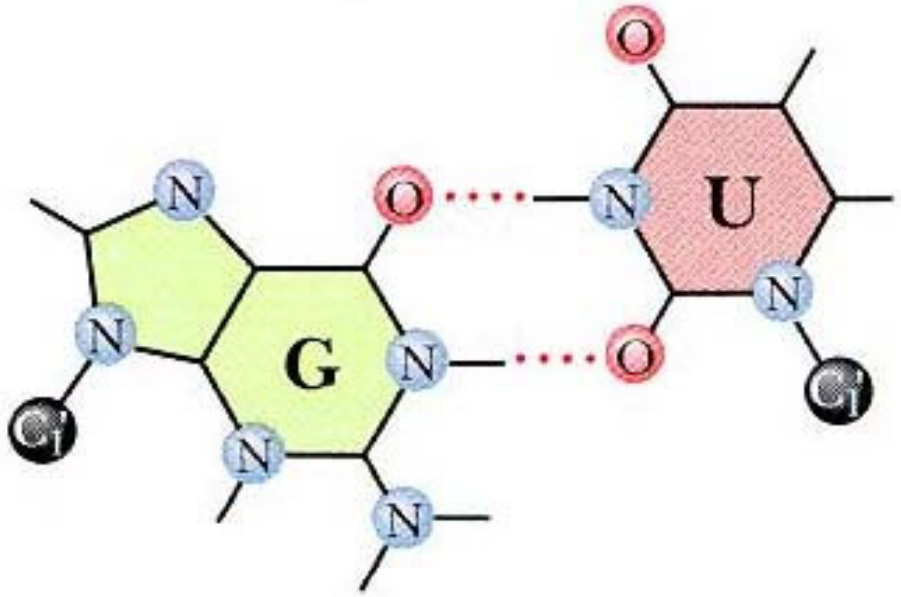
Modified from Jonathan D. Wren's *slides*

RNA Basics

- RNA bases A,C,G,U
- Canonical Base Pairs
  - A-U
  - G-C
  - G-U

    "wobble" pairing
  - *Bases can only pair with **one** other base*

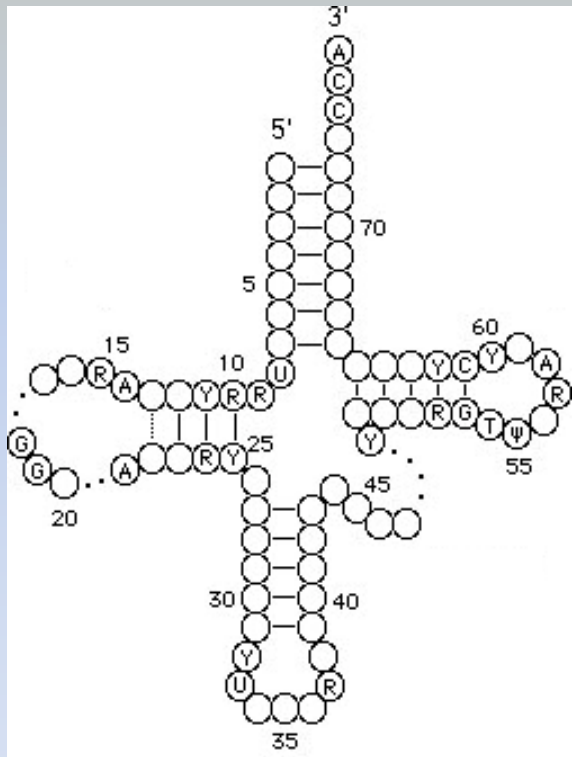
"wobble" pairing – less stable



http://www.bioalgorithms.info/

RNA types

- transfer RNA (tRNA)
- messenger RNA (mRNA)
- ribosomal RNA (rRNA)
- small interfering RNA (siRNA)
- micro RNA (miRNA)
- small nucleolar RNA (snoRNA)
- Long non-coding RNA (lncRNA)
- …….

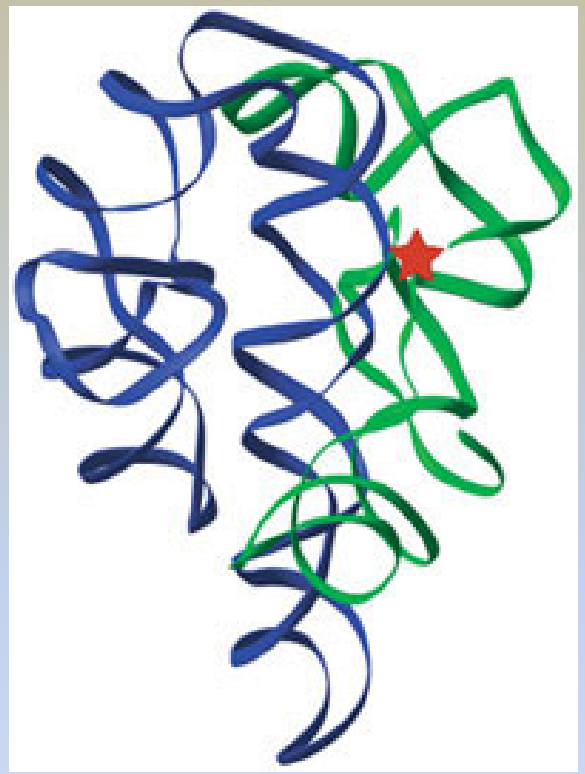


RNA has many biological functions

- Enzymatic reaction (protein synthesis)
- Control of mRNA stability (UTR)
- Control of splicing (snRNP)
- Control of translation (microRNA)

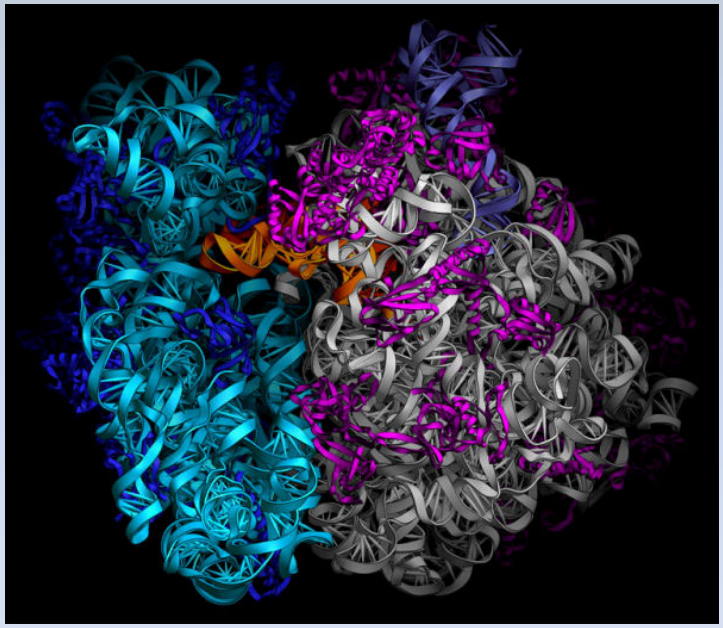
The function of the RNA molecule depends on its folded structure

Ribozyme



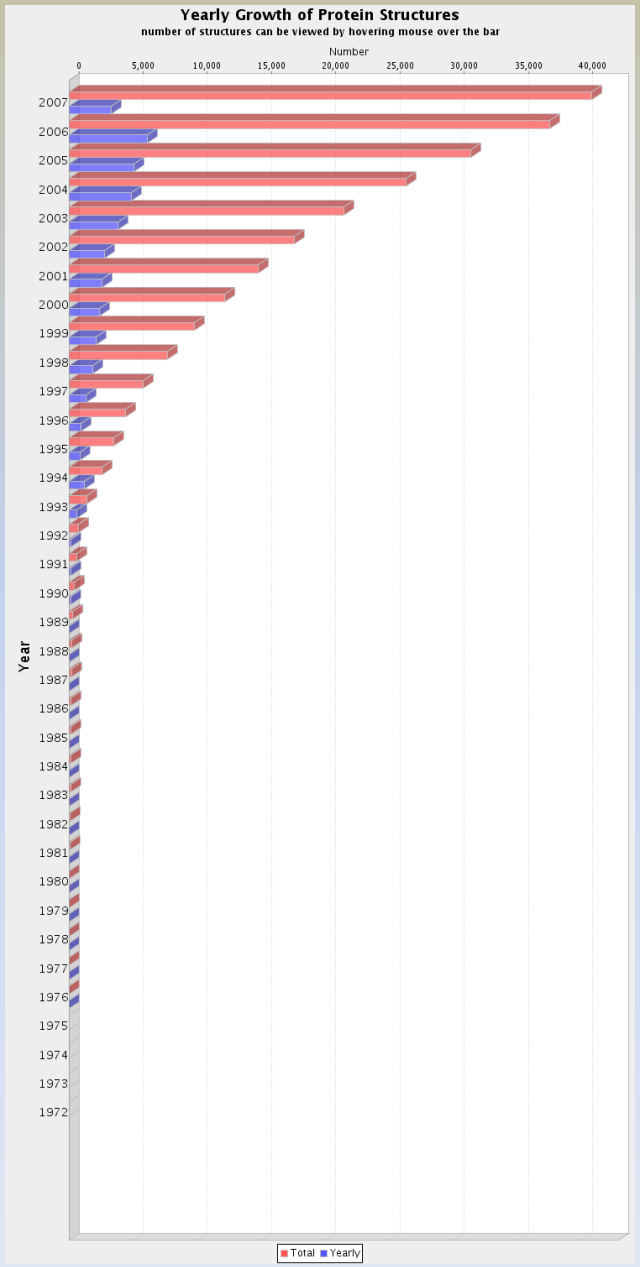
Nobel prize
1989

Ribosome

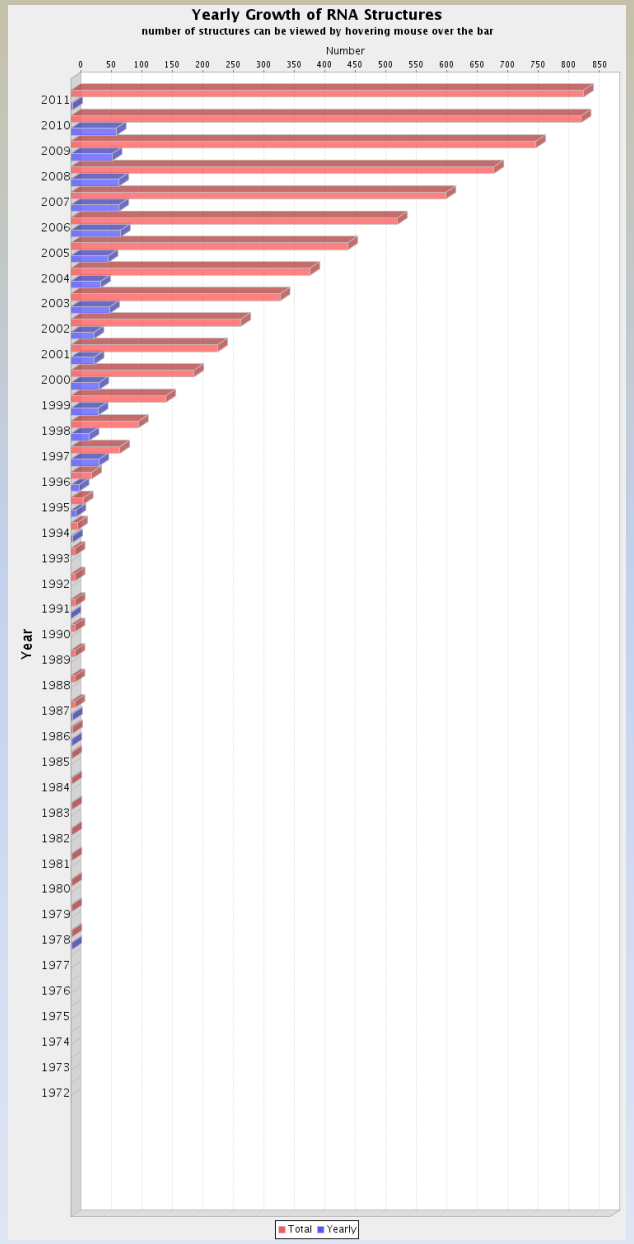


Nobel prize
2009

Protein structures

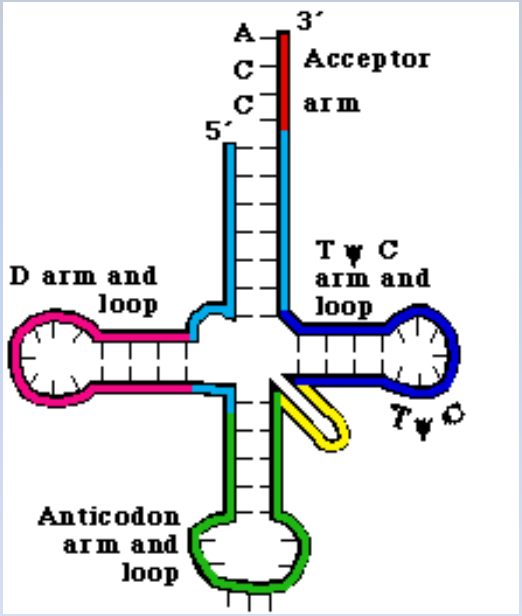


RNA structures

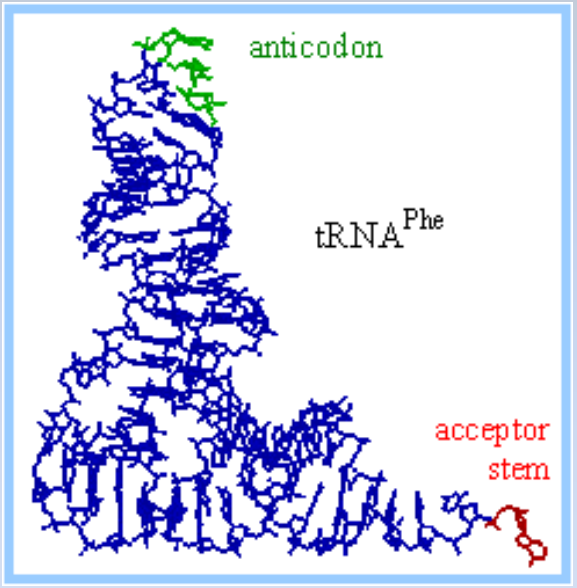


RNA Structural levels

Secondary Structure



Tertiary Structure



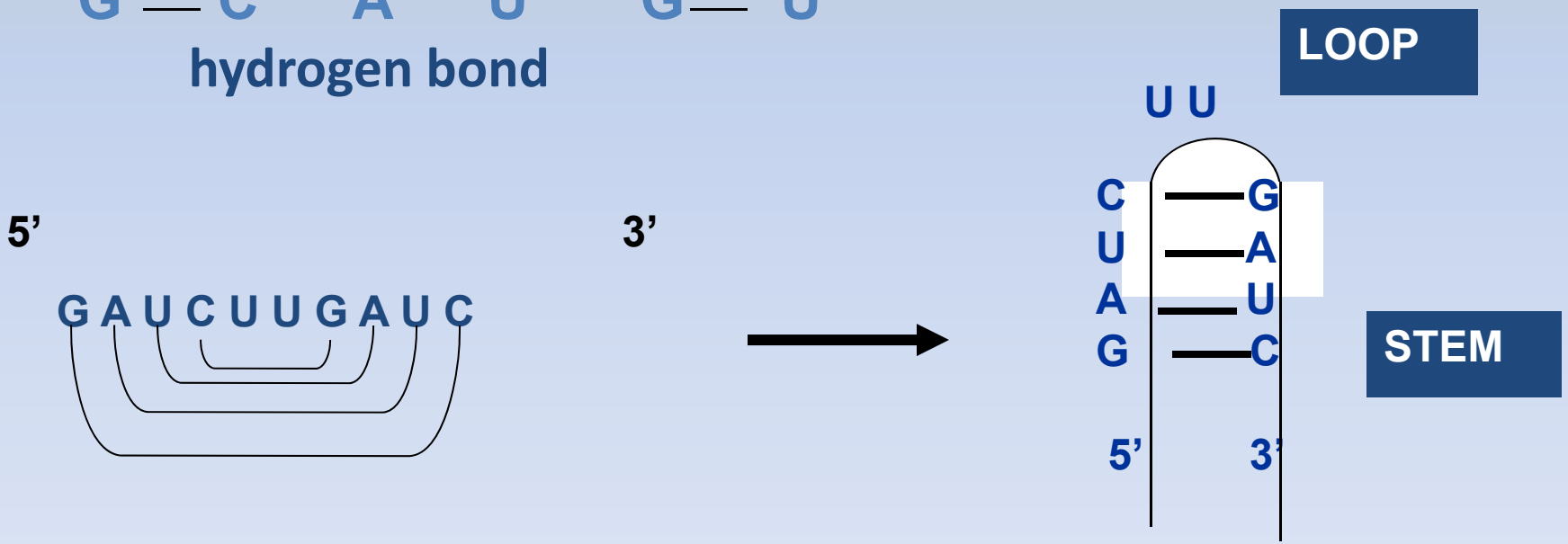
tRNA

RNA Secondary Structure

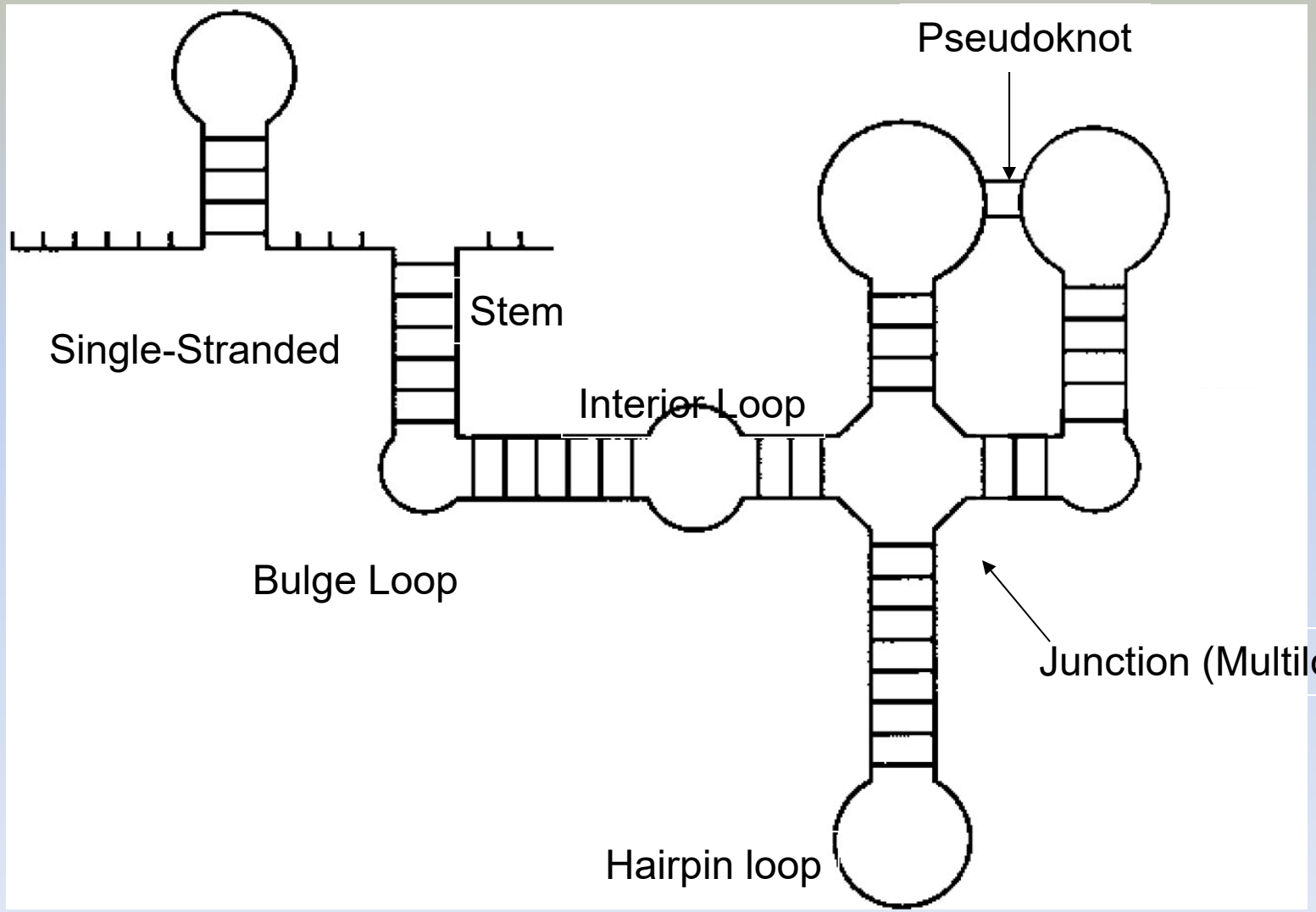
- The RNA molecule folds on itself.
- The base pairing is as follows:

G ≡ C A = U G = U

hydrogen bond

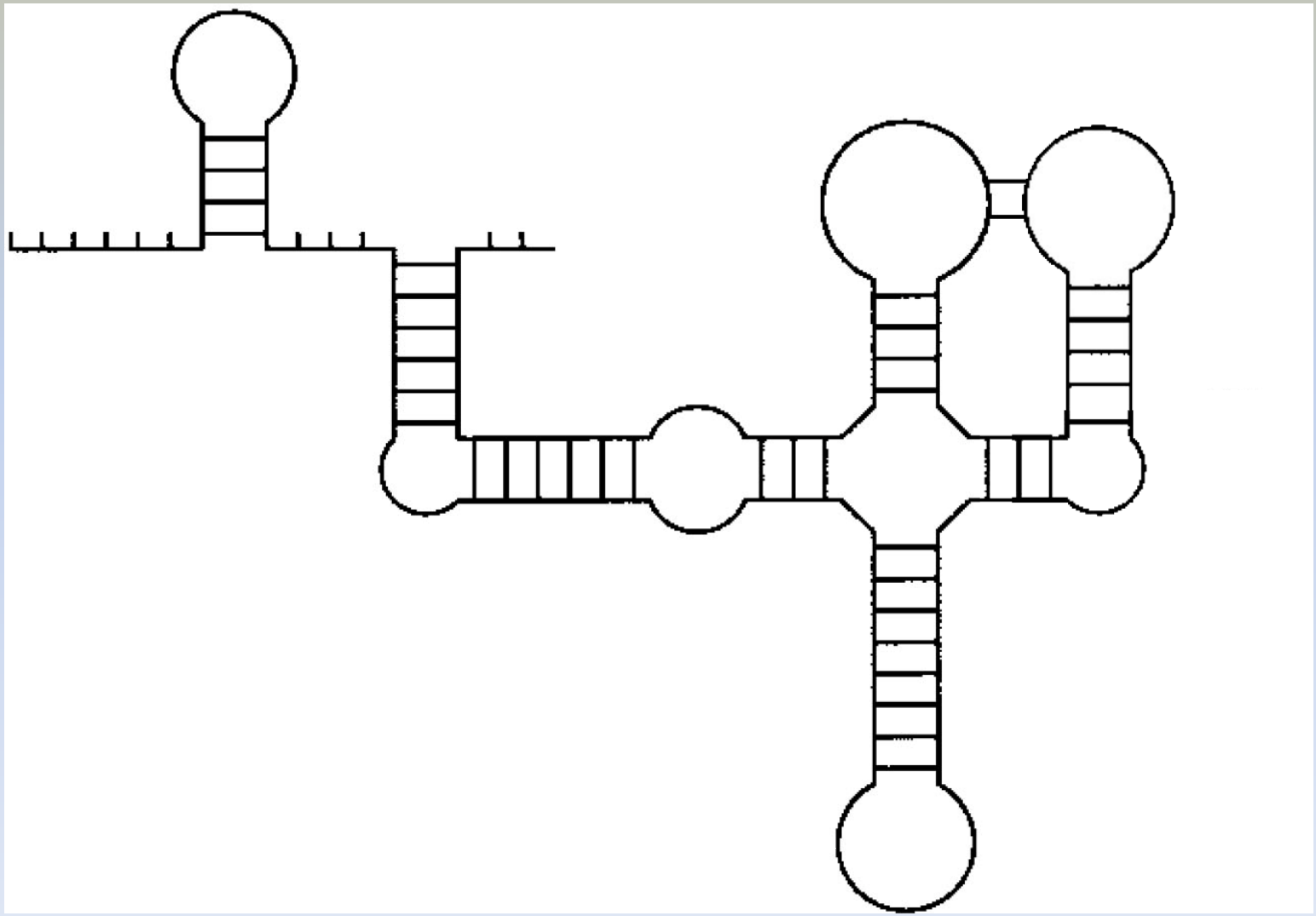


RNA Secondary Structure



Image– Wuchty

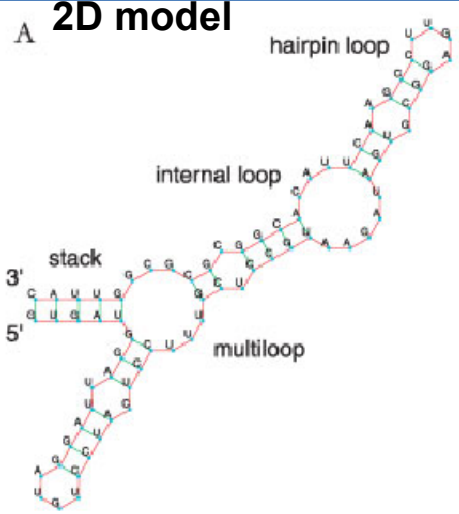
RNA Secondary Structure (test)



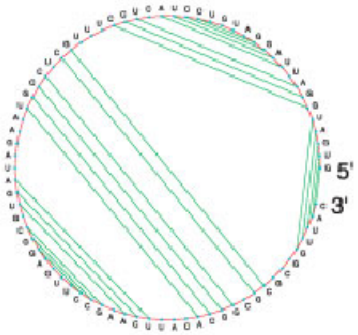
Image– Wuchty

RNA Structure Representations

A **2D model**



B **Circle with lines**



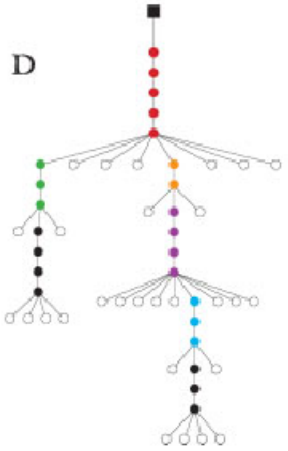
Ordered tree

C

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

Balanced nested parenthesis

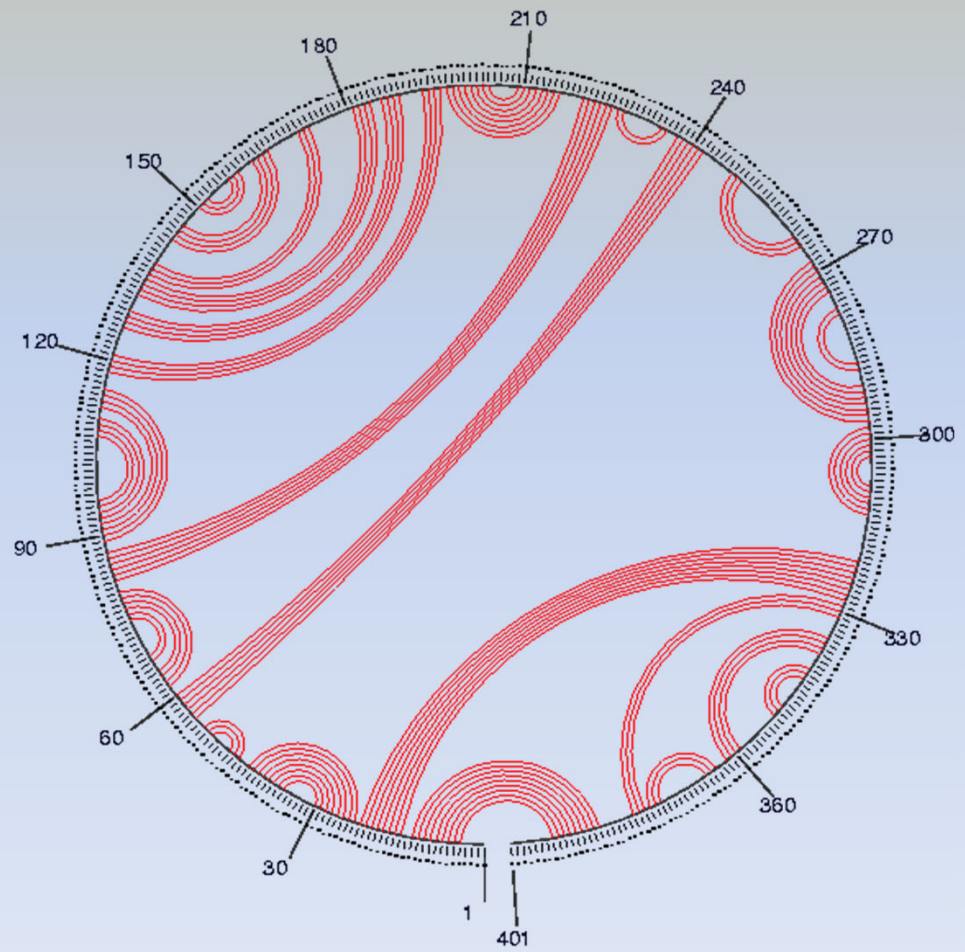
D



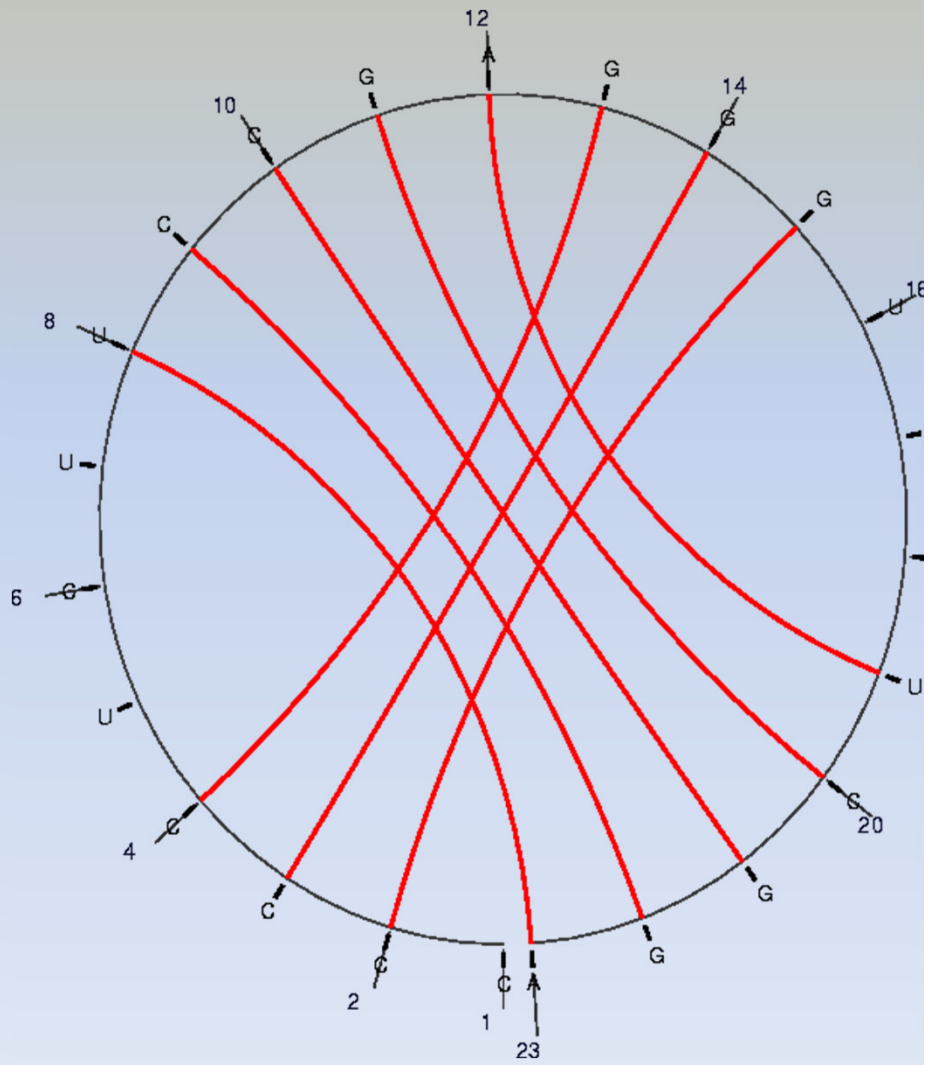
E **Mountains**

RNA secondary structure representation

No pseudoknots

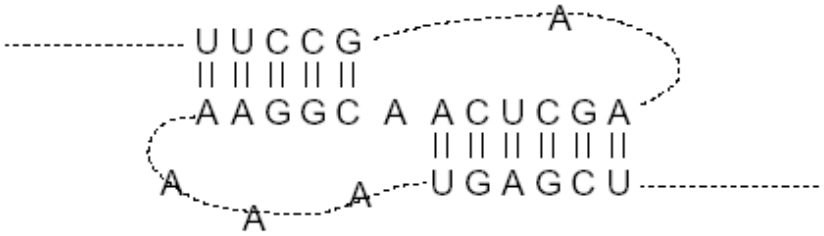


Pseudoknots

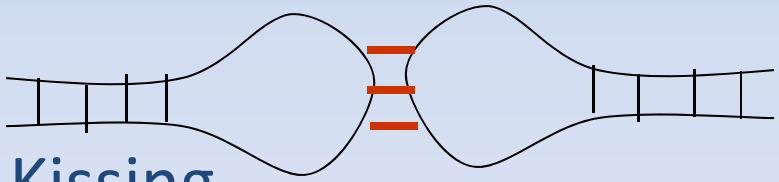


Examples of known interactions of RNA secondary structural elements

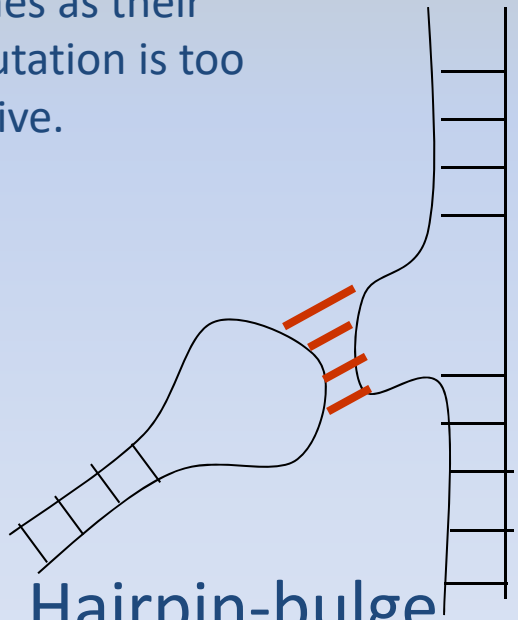
Pseudo-knot



These patterns are excluded from the prediction schemes as their computation is too intensive.



Kissing hairpins

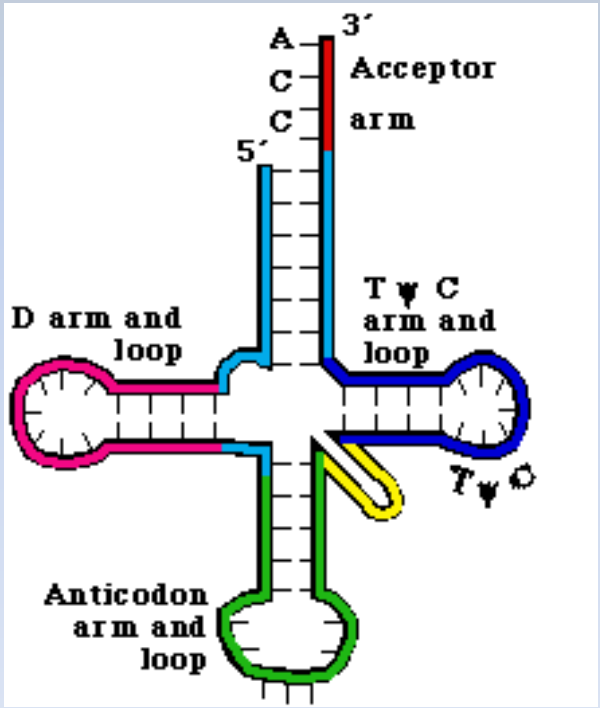


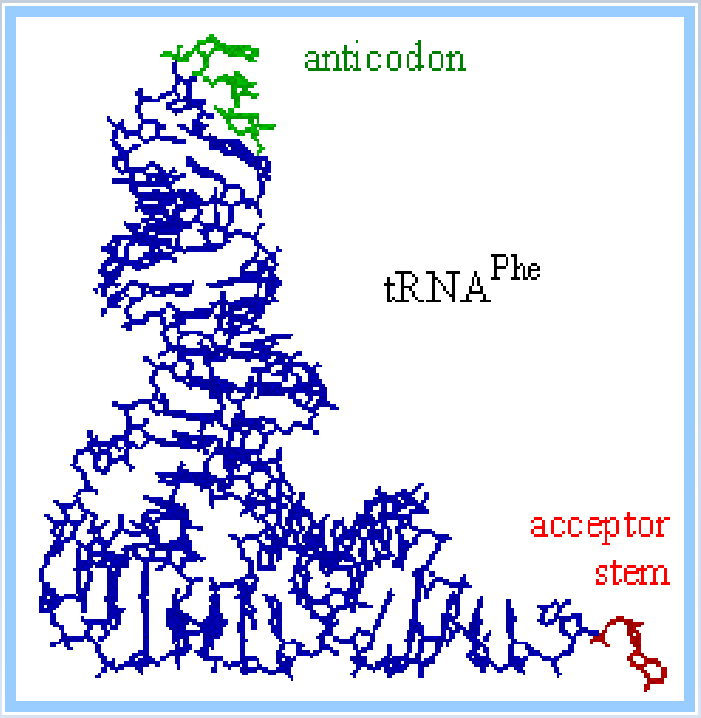
Hairpin-bulge contact

RNA secondary Structure is only an approximation

We have to keep in mind that these 2D representations are an approximation – the real structure has 3D

tRNA



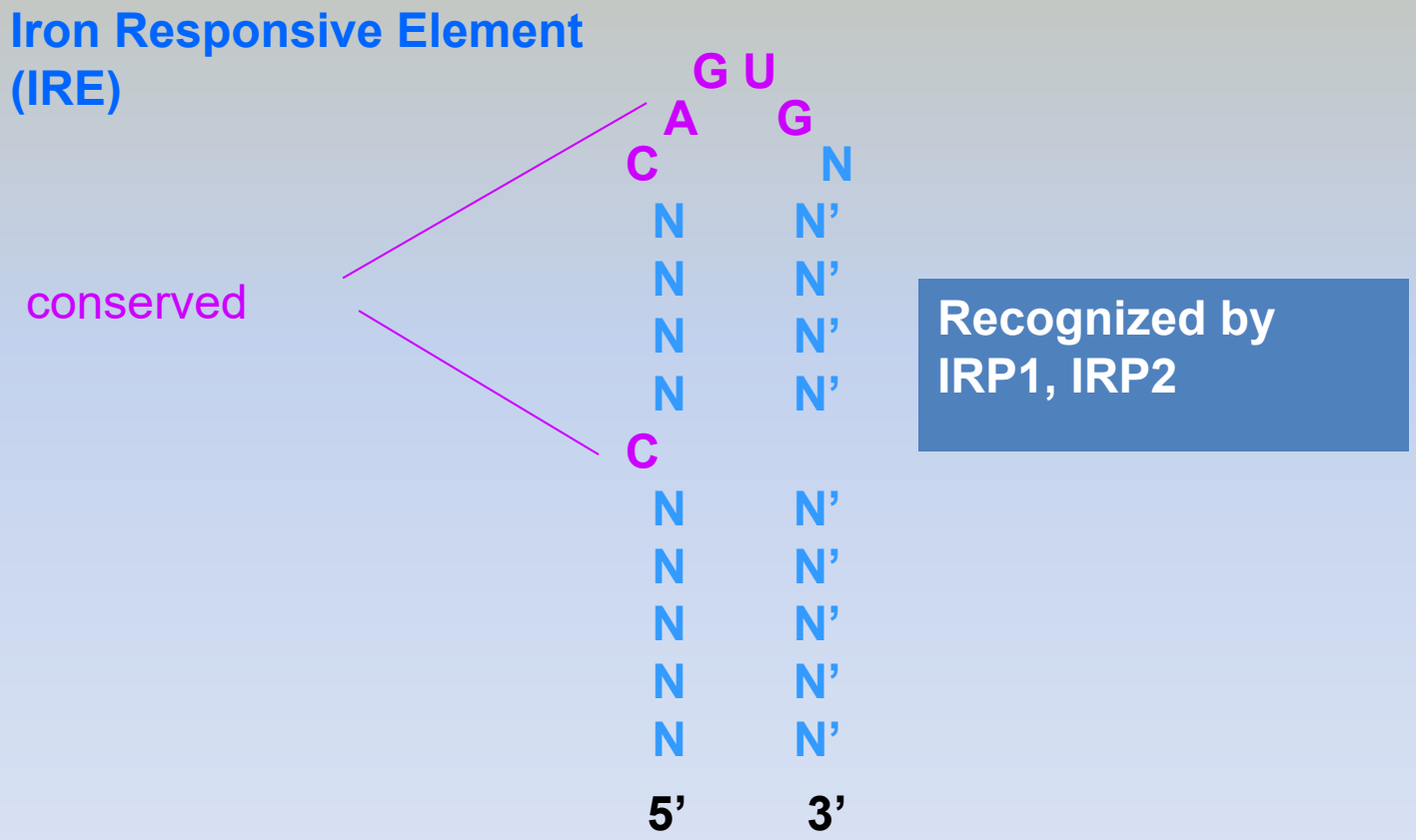


Some biological functions of non-coding RNA

- RNA splicing (snRNAs)
- Guide RNAs (RNA editing)
- Catalysis
- Telomere maintenance
- Control of translation (miRNAs)

The function of the RNA molecule depends on its folded structure

Control of iron levels by mRNA secondary structure

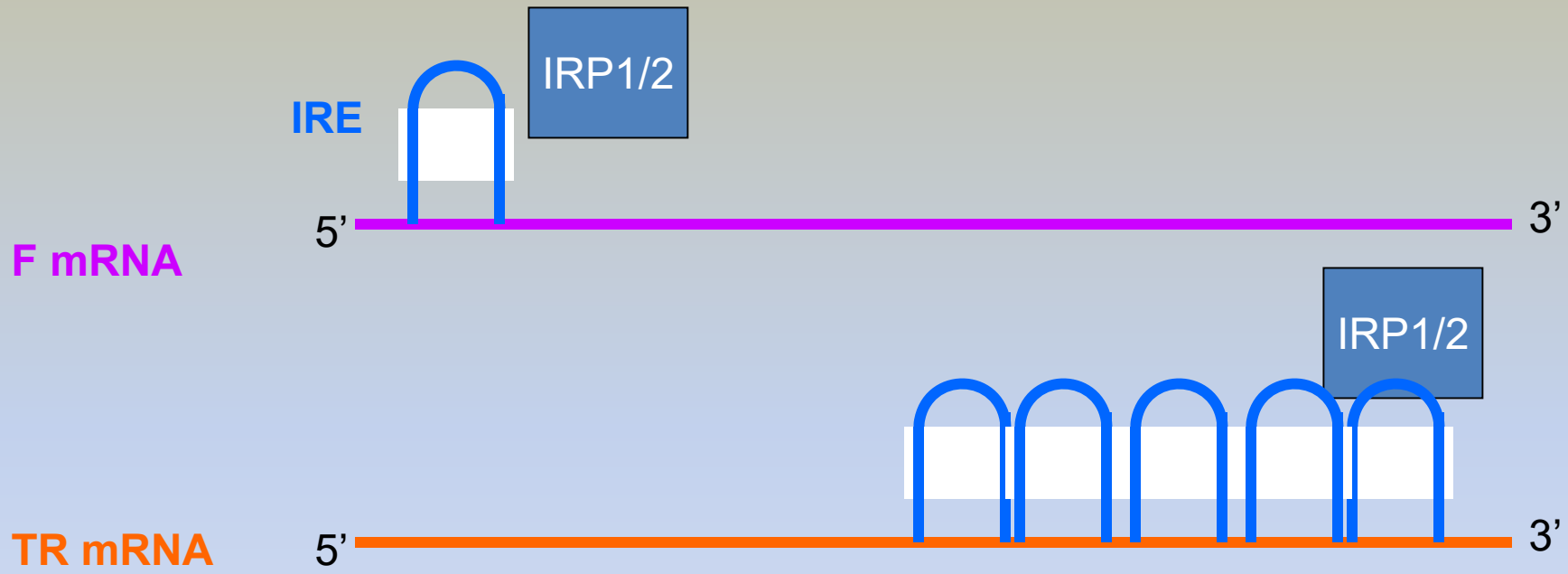


When cells are deprived of iron, IRP binds to the IRE. If IRE is located at 5'UTR, IRP binding will inhibit translation initiation, else if IRE is at 3'UTR, IRP binding will stabilize mRNA and prevent it from degradation

F: Ferritin = iron storage

TR: Transferrin receptor = iron uptake

Ferritin can store 2,250 iron ions in its globular shell



Low Iron

IRE-IRP inhibits translation of Ferritin

IRE-IRP Inhibition of degradation of TR

High Iron

IRE-IRP off -> Ferritin translated

Transferrin receptor degraded

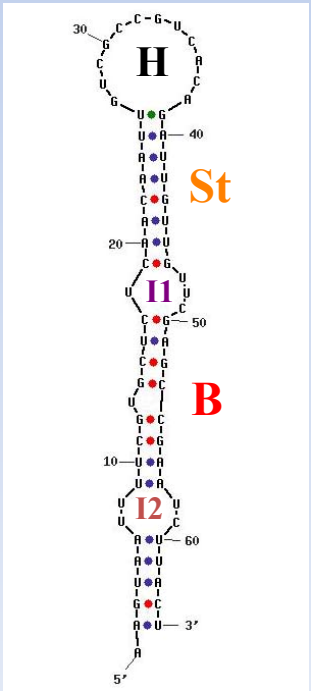
Structure-based similarity

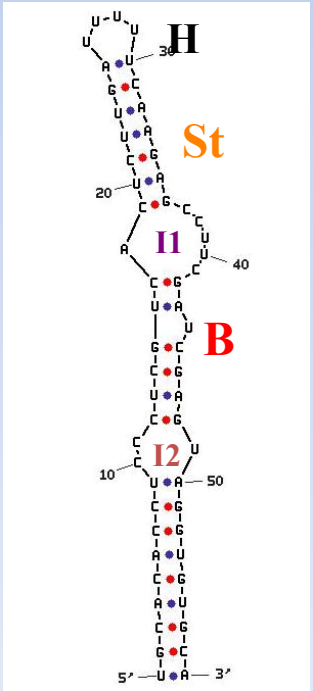
Sequence Similarity **%ID = 34%**

```
gurken      AAGTAATTTTCGTGCTCTCAACAATTGTCGCCGTCACAGATTGTTGTTCGAGCCGAATCTTACT 64
Ifactor     ---TGCACACCTCCCTCGTCACTCTTGATTTT-TCAAGAGCCTTCGATCGAGTAGGTGTGCA-- 58
               *      *   ***   **  ***      ***       * * *****  *      *
```

Structural Similarity

gurken
64nt stem loop





I Factor
58nt stem loop

RNA secondary structure prediction

Dynamic programming & free energy minimization

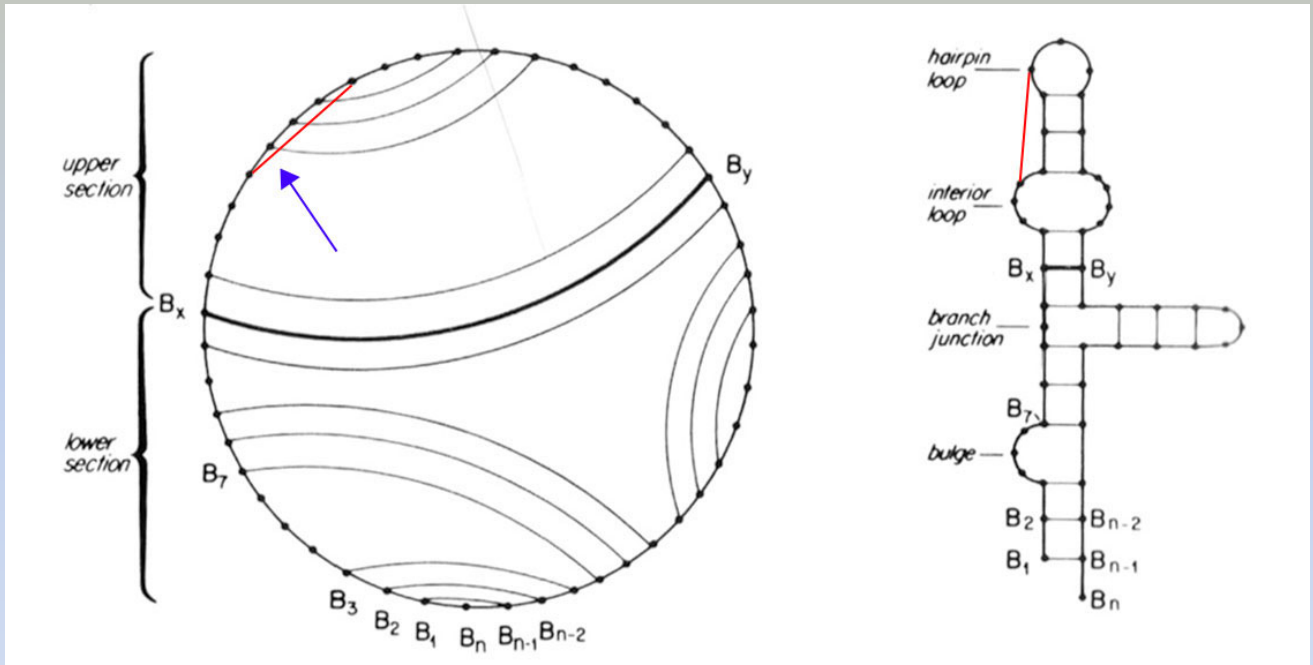
Predicting RNA Secondary Structure

- According to base pairing rules only, (A-U, G-C and wobble pairs G-U) sequences can potentially form many different structures
- An energy value is associated with each possible structure
- *Predict the structure with the minimal free energy (MFE)*

Simplifying Assumptions for Structure Prediction

- RNA folds into one minimum free-energy structure
- There are no knots (base pairs never cross)
- The energy of a particular base pair in a double stranded regions sequence independent
  - Neighbors do not influence the energy

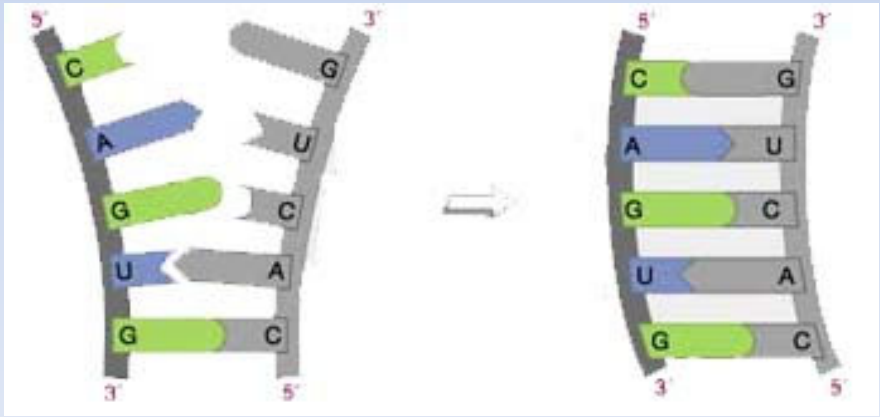
Trouble with Pseudoknots



- 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

Sequence alignment as a method to determine structure

- Bases pair in order to form backbones and determine the secondary structure
- Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure



Base Pair Maximization – Dynamic Programming Algorithm

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

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

Base pair at i and j

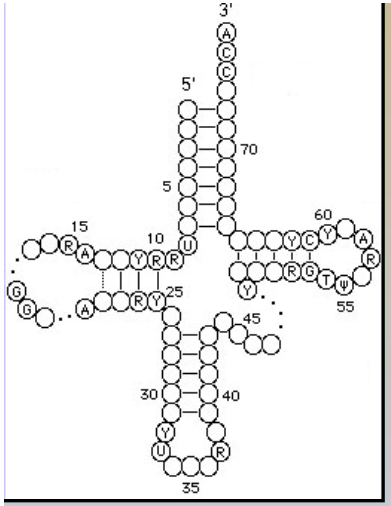
Unmatched at i

Umatched at j

Bifurcation

Base Pair Maximization - Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure
  - May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches – not biologically reasonable



Energy Minimization

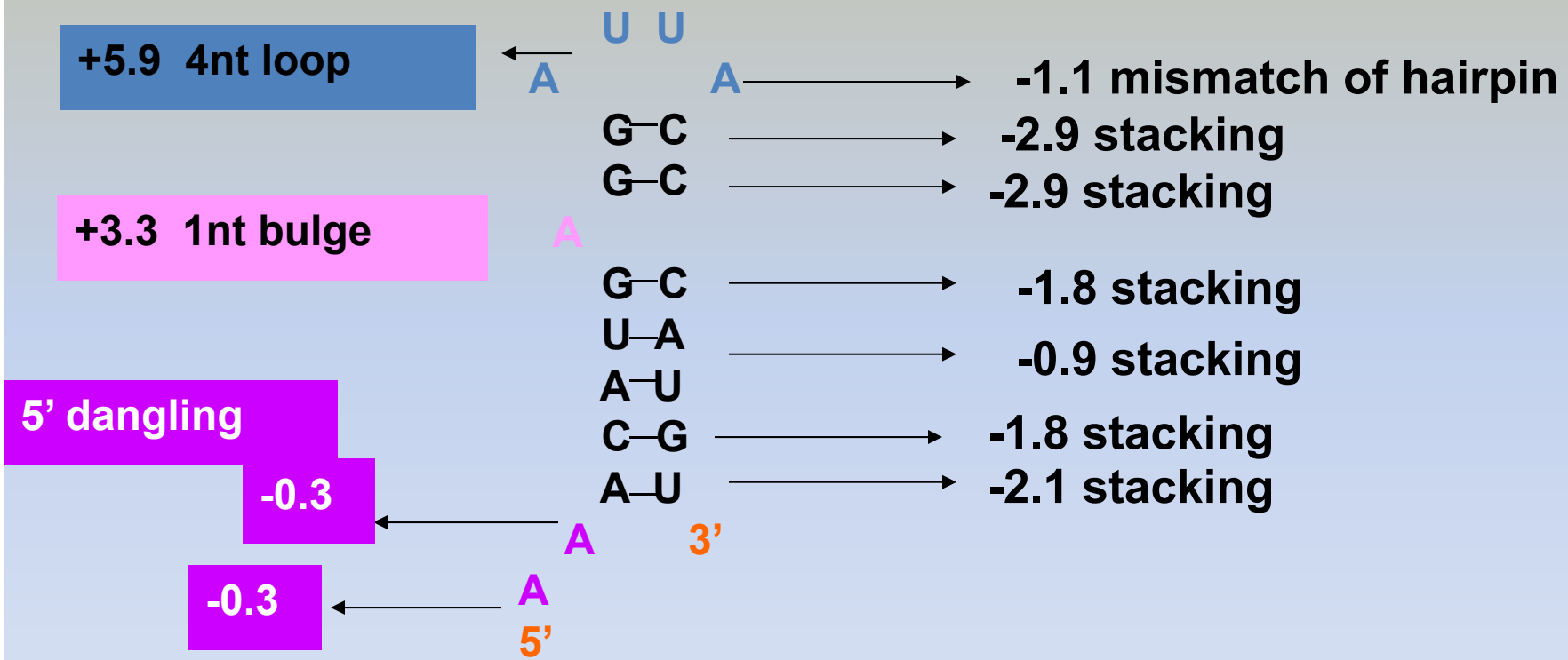
- Thermodynamic Stability
  - Estimated using experimental techniques
  - Theory : Most stable is the most likely
- No pseudoknots due to algorithm limitations
- Uses Dynamic Programming alignment technique
- Attempts to maximize the score taking into account thermodynamics
- MFOLD and ViennaRNA

Thermodynamics

- Gibbs Free Energy, *G*
- Describes the energetics of molecules in aqueous solution. The change in free energy, ΔG, for a chemical process, such as nucleic acid folding, can be used to determine the direction of the process:
- $\Delta G=0$: equilibrium
- $\Delta G>0$: unfavorable process
- $\Delta G<0$: favorable process
- Thus the natural tendency for biomolecules in solution is to minimize free energy of the entire system (biomolecules + solvent).

- $\Delta G = \Delta H - T\Delta S$
- ΔH is enthalpy, ΔS is entropy, and T is the temperature in Kelvin.
- Molecular interactions, such as hydrogen bonds, van der Waals and electrostatic interactions contribute to the ΔH term. ΔS describes the change of order of the system.
- Thus, both molecular interactions as well as the order of the system determine the direction of a chemical process.
- **For any nucleic acid solution, it is extremely difficult to calculate the free energy from first principle**

Free energy computation



ΔG= -4.6 KCAL/MOL

Adding Complexity to Energy Calculations

- Stacking energy - We assign negative energies to these *between base pair* regions.
  - Energy is influenced by the previous base pair (not by the base pairs further down).
  - These energies are estimated experimentally from small synthetic RNAs.
- Positive energy - added for destabilizing regions such as bulges, loops, etc.
- More than one structure can be predicted

Energy Minimization Drawbacks

- Compute only one optimal structure
- Usual drawbacks of purely mathematical approaches
  - Similar difficulties in other algorithms
    - Protein structure
    - Exon finding

Prediction Tools based on Energy Calculation

Fold, Mfold

Zucker & Stiegler (1981) *Nuc. Acids Res*. 9: 133-48

Zucker (1989) *Science* 244:48-52

RNAfold

Vienna RNA secondary structure server

Hofacker (2003) *Nuc. Acids Res*. 31: 3429-31

Submitting RNA to MFOLD

- Enter the sequence to be folded in the box.

All non-alphabet characters will be removed.
FASTA format may be used.

```
catttatcagcgacagagctcactataaatggtgttttttaatagaata
taatttatcggaagcagtgccttccataattatgacagttatactgtcggt
ttttttaaataaaagcagcatctgctaataaaacccaacagatactgga
agttttgcatttatggtcaacacttaagggttttagaaaacagccgtcag
ccaaatgtaattgaataaagttgaagctaagatttagagatgaatttaaat
ttaattaggggttgctaagaagcgagcactgaccagataagaatgctggt
tttcctaaatgcagtgaattgtgaccaagttataaatcaatgtcacttaa
aggctgtggtagtactcctgcaaaattttatagctcagtttatccaaggt
gtaactctaattcccatttgcaaaatttccagtacctttgtcacaatcct
aacacattatcgggagcagtgtcttccataatgtataaagaacaaggtag
tttttacctaccacagtgtctgtatcggagacagtgatctccatatgtta
cactaagggtgtaagtaattatcgggaacagtgtttcccataattt
```

Paste your sequence

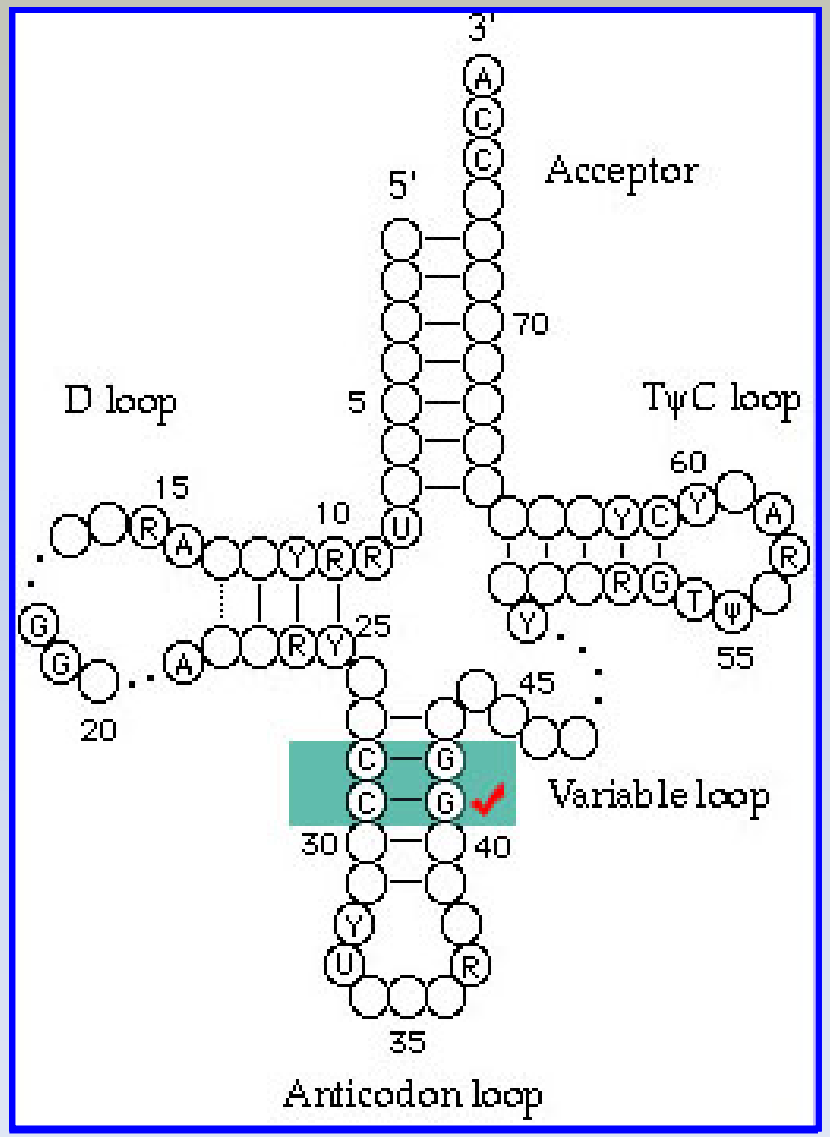
- Enter **constraint information** in the box at the right. (optional) You may:
  1. force bases $i,i+1,\ldots,i+k-1$ to be double stranded by entering:
     F i 0 k on 1 line in the constraint box.
  2. force consecutive base pairs $i,j,i+1,j-1,\ldots,i+k-1,j-k+1$ by entering:
     F i j k on 1 line in the constraint box.
  3. force bases $i,i+1,\ldots,i+k-1$ to be single stranded by entering:
     P i 0 k on 1 line in the constraint box.
  4. prohibit the consecutive base pairs
     $i,j,i+1,j-1,\ldots,i+k-1,j-k+1$ by entering:
     P i j k on 1 line in the constraint box.
  5. prohibit bases i to j from pairing with bases k to l by entering:
     P i-j k-l on 1 line in the constraint box.

Use default parameters

Scroll wayyyy down and hit "Fold RNA"

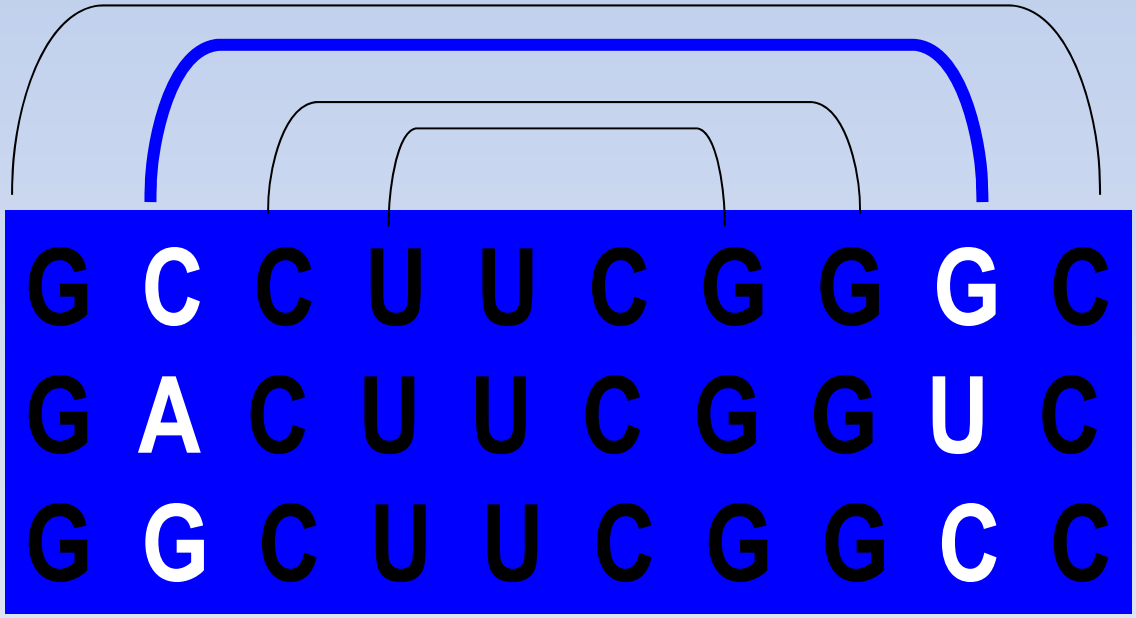
- The RNA sequence is linear .
- Folding temperature is fixed at 37°.
- Ionic conditions: 1M NaCl, no divalent ions.
- Enter the *percent suboptimality* number. 5
- Enter an upper bound on the number of computed foldings. 50
- Enter the window parameter if you wish. default
- Enter the maximum interior/bulge loop size 30

Compensatory substitutions



Covariation ensures ability to base pair is maintained and RNA structure is conserved

Evolutionary conservation of RNA molecules can be revealed by identification of compensatory mutations



U C
U G
C G
N N'
G C

Insight from Multiple Alignment

Information from multiple alignment about the probability of positions i,j to be base-paired.

- Conservation – no additional information
- Consistent mutations (GC→ GU) – support stem
- Inconsistent mutations – does not support stem.
- Compensatory mutations – support stem.

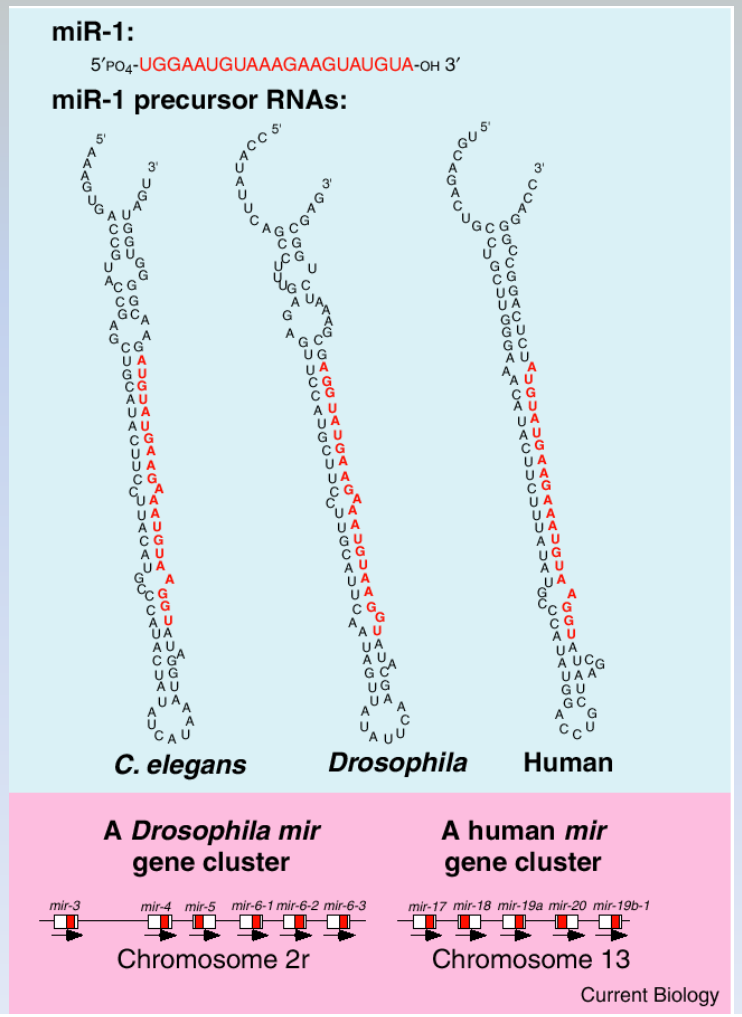
RNA families

- Rfam : General non-coding RNA database
- 379 families annotating 280,000 regions

http://www.sanger.ac.uk/Software/Rfam/

Includes many families of non-coding RNAs and functional motifs, as well as their alignment and secondary structures

An example of an RNA family miR-1 MicroRNAs



Summary

- MFOLD and other RNA secondary structure prediction tools rarely give the right answer first (or at all)
  - Too many possible structures in the low energy neighbourhood
- Can be used as a “first-pass” tool
  - Eyeball key conserved motifs
  - Collect sequences to build a consensus
- Often need to adjust parameters
  - Use prior knowledge to force base pairing
- Motif-searching tools can be used to identify conserved secondary structure motifs in a sequence database
  - Retrieves more results than sequence-based searches