

RNA Secondary Structure

Basis of RNA

- ⦿ RNA bases A,C,G,U
- ⦿ Primary Structure of RNA : A sequence of the bases A,G,C and U
- ⦿ Base Pairs
 - A-U
 - G-C
 - non-canonical pairs can occur in RNA
 - G-U is the most common
- ⦿ Stability
 - $G-C > A-U > G-U$

Basis of RNA

- ⦿ single stranded; strand folds upon itself to form base pairs; can have a diverse form of secondary structure
- ⦿ compared to base sequences, structure conservation is most important with RNA

Structure Rules

- ⦿ Base pairing stabilize the structure
- ⦿ Unpaired sections (loops) destabilize the structure
- ⦿ when a base in one position changes, the base it pairs to must also change to maintain the same structure (*covariation*)

Representations of Secondary Structure

Most basepairs are

non-crossing

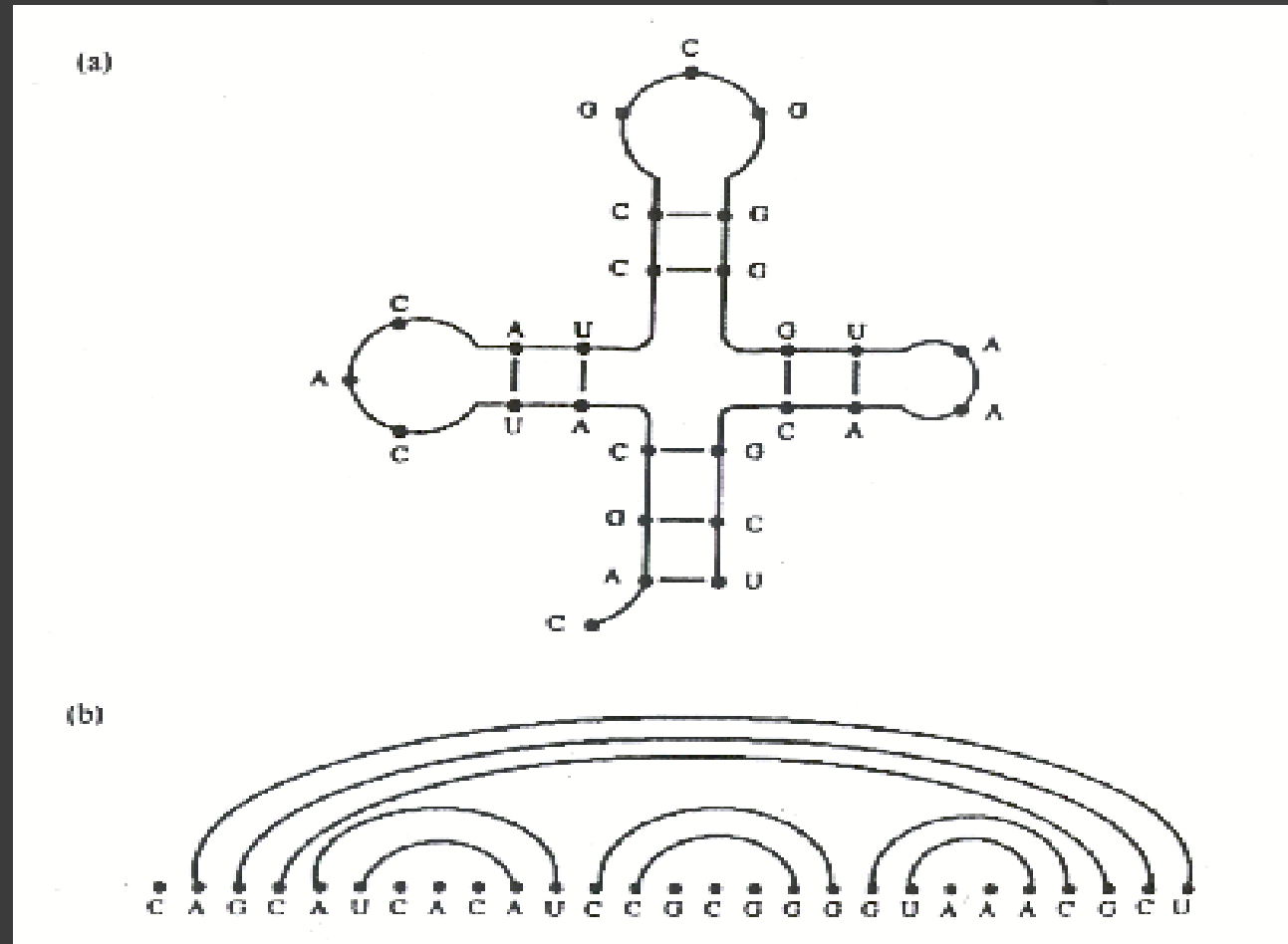
basepairs:

Any two pairs (i, j)

and (i', j') \rightarrow

$i < i' < j' < j$

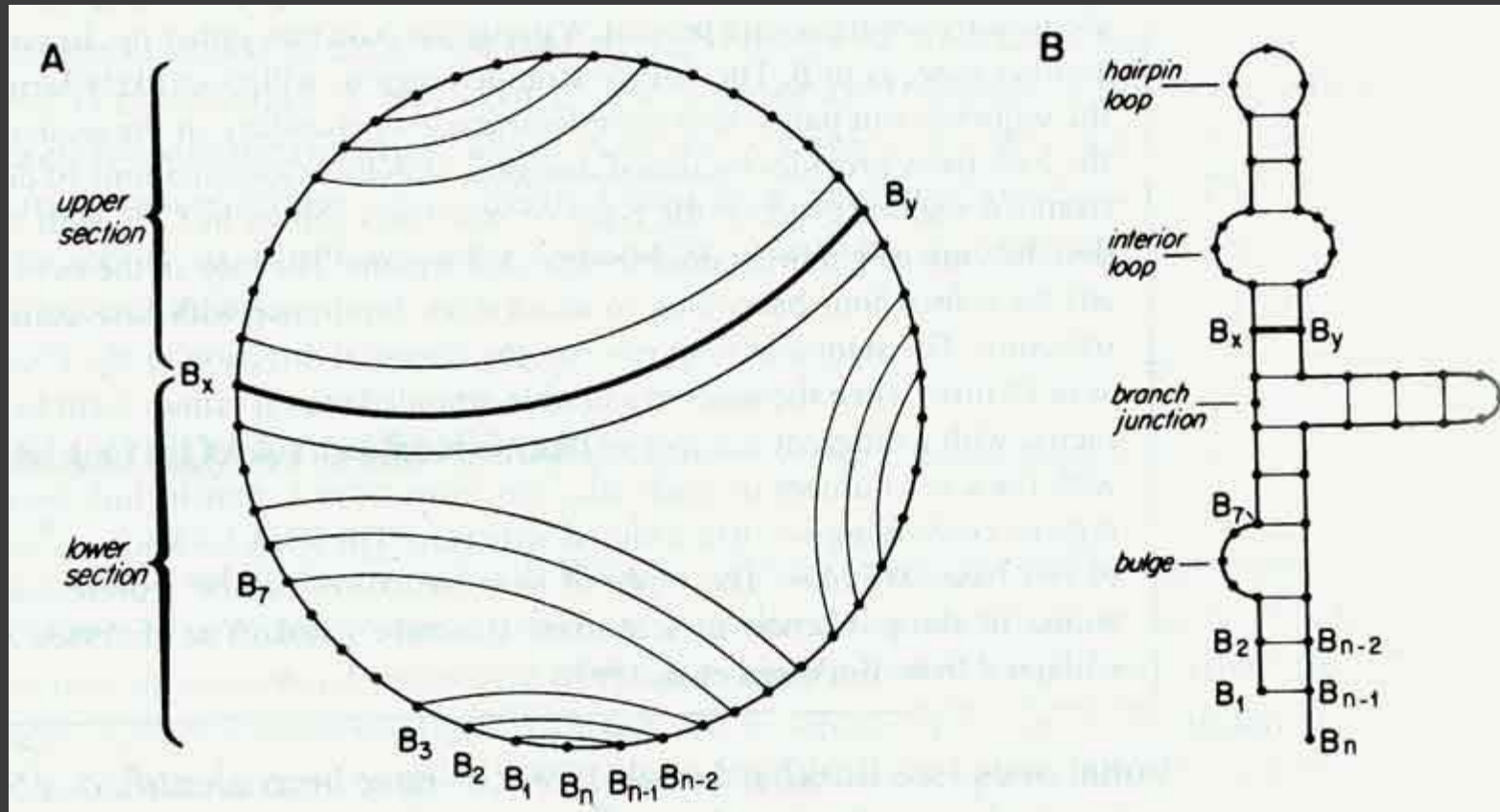
or $i' < i < j < j'$



Circular Representation of Base Pairs

- ⦿ base pairs of a secondary structure represented by a circle
- ⦿ arc drawn for each base pairing in the structure

Circular Representation of Base Pairs



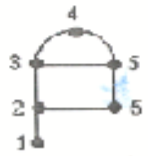
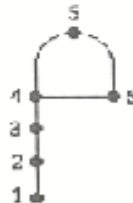
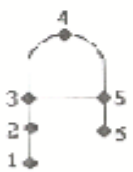
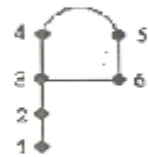
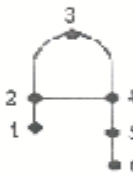
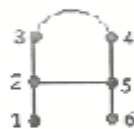
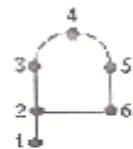
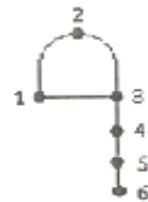
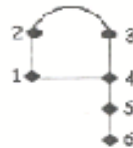
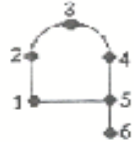
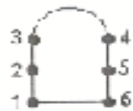
RNA Combinatorics

The number of RNA secondary structures for the sequence [1,n]

$$s(0) = s(1) = s(2) = 1$$

$$s(n+1) = s(n) + \sum_{j=1}^{n-1} s(j-1)s(n-j), (n \geq 2)$$

Recurrence Relation

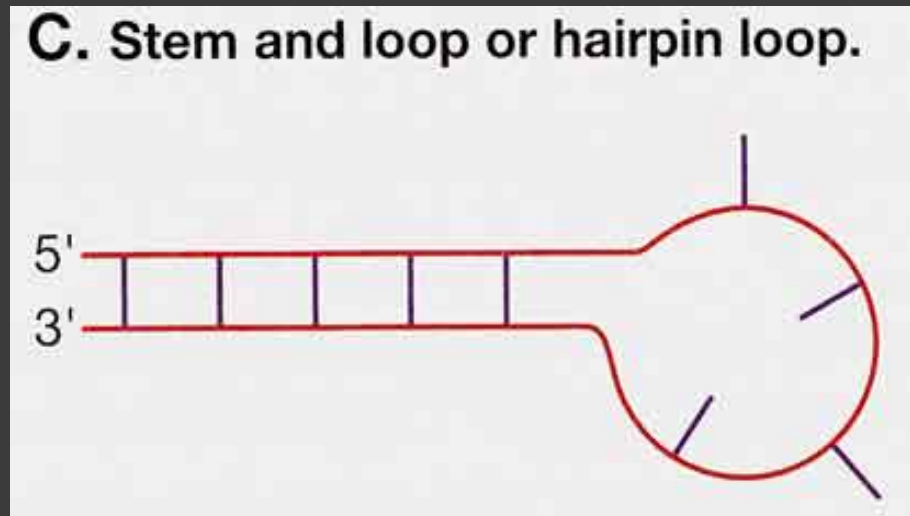


- ⦿ based on the combinatorics, there are approximately 1.3 billion possible RNA structures of length $n = 27$.

- ◎ Types of different regions in RNA secondary structure

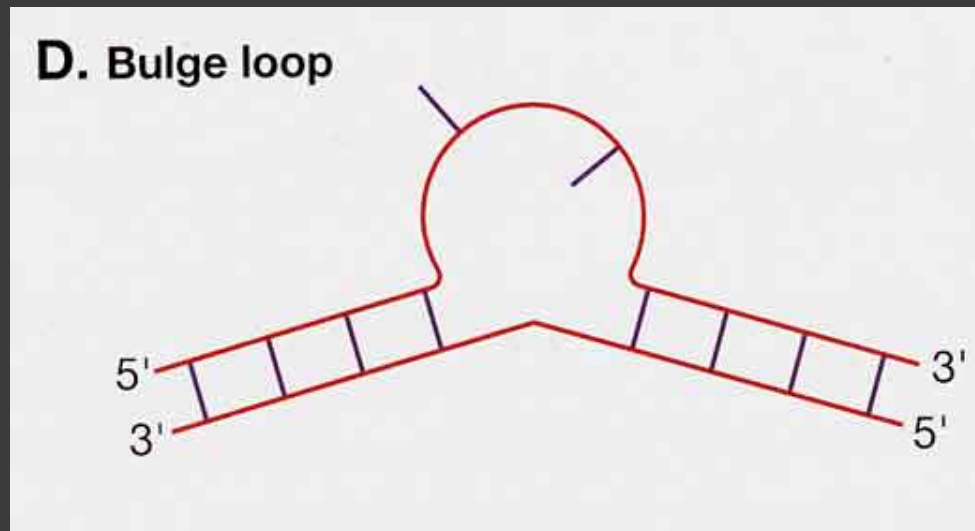
Hairpin Loop

- generally at least 4 bases long for each loop



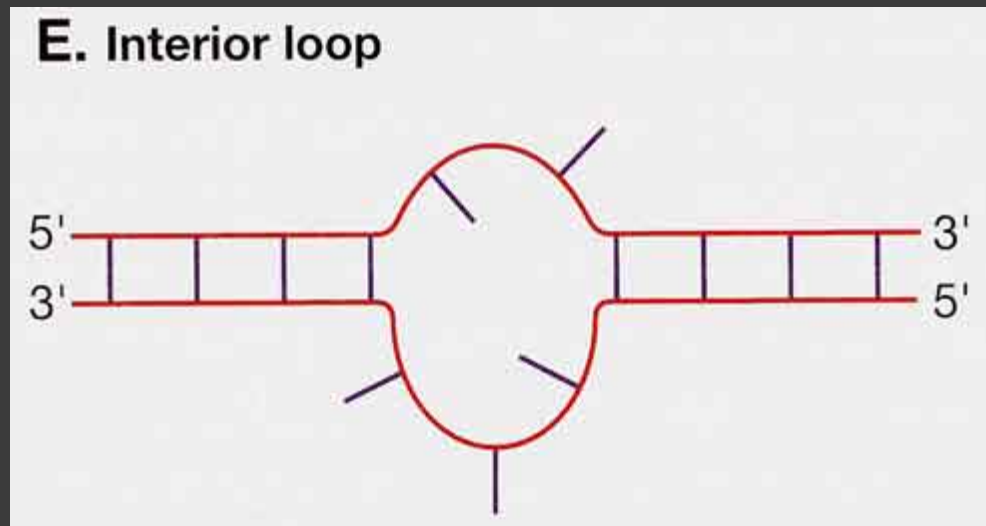
Bulge Loops

- occur when bases on one side of the structure cannot form base pairs



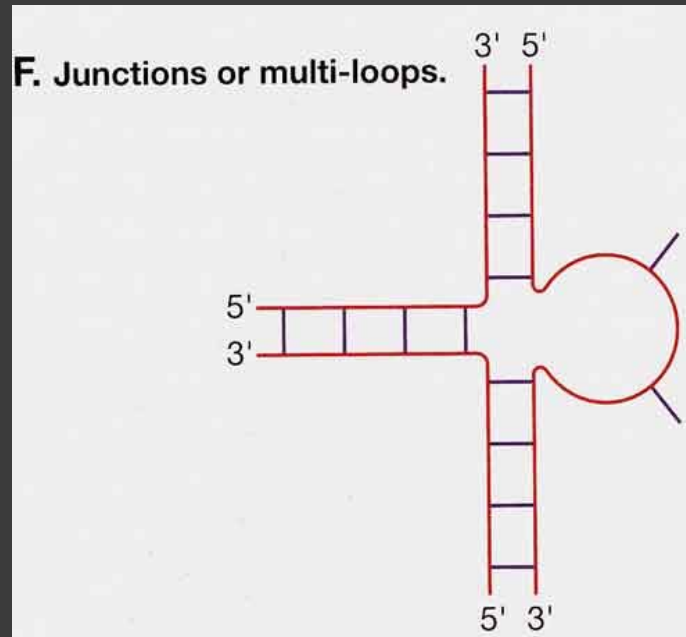
Interior Loops

- occur when bases on both sides of the structure cannot form base pairs

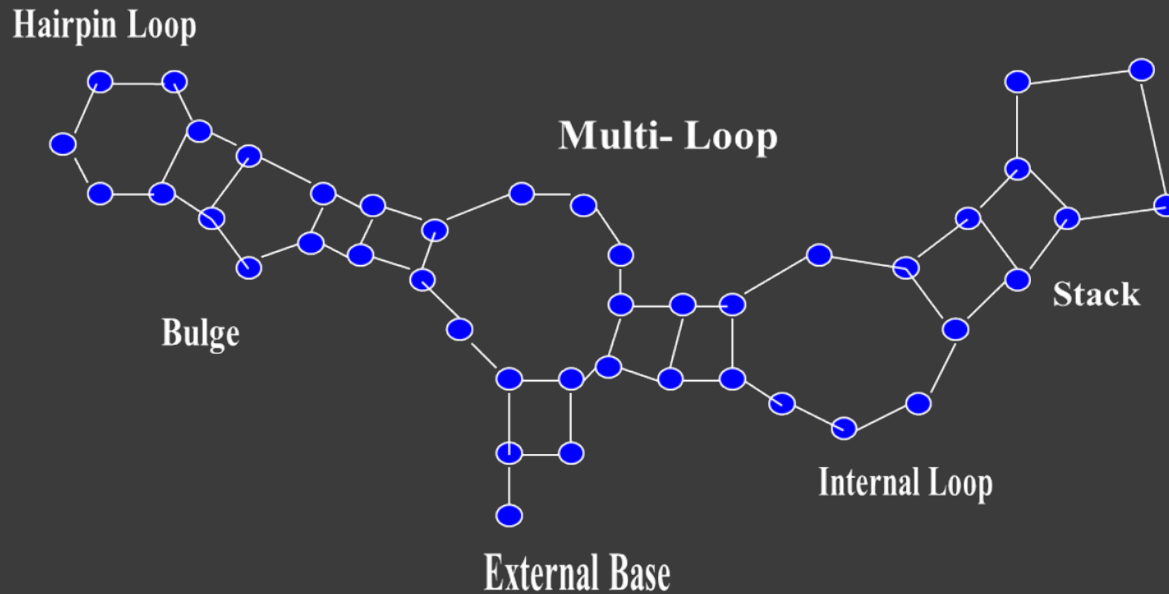


Junctions (Multi-loops)

- two or more double-stranded regions converge to form a closed structure



RNA Secondary Structure



- **Stacks:** continuous nested basepairs. (energetically favorable)
- **Non-basepaired loops:**
 - Hairpin loop.
 - Bulge.
 - Internal loop.
 - Multiloop.

RNA structure prediction methods

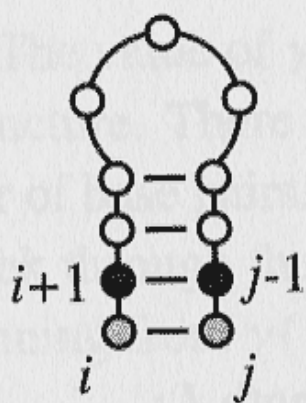
- ⦿ Maximize Base Pairs
- ⦿ Minimize Energy

Maximize Base Pairs

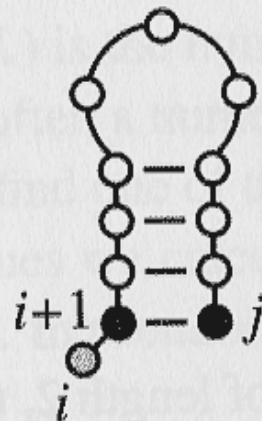
- ⦿ Given a RNA sequence, determine the set of maximal basepairs(no basepair across each other)
- ⦿ Align bases according to their ability to pair with each other gives an approach to determining the optimal structure
- ⦿ dynamic programming approach
- ⦿ Nussinov Algorithm

Nussinov Algorithm

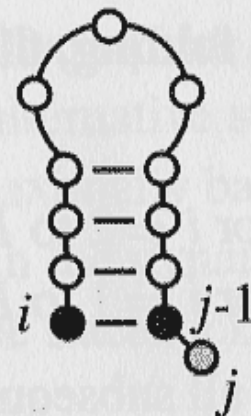
- Four ways to get the optimal structure between position i and j from the optimal substructure



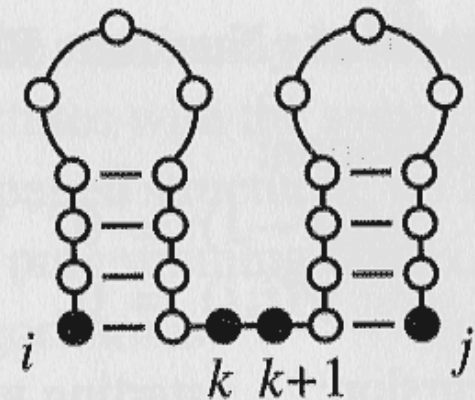
i, j pair



i unpaired



j unpaired



bifurcation

Nussinov Algorithm

- ⦿ compares a sequence against itself in a $n \times n$ matrix
- ⦿ Find the maximum of the scores for the four possible structures at a particular position

Nussinov Algorithm

- ⦿ this method will not necessarily generate the most stable structure
- ⦿ may have scattered matches which are not biologically reasonable
- ⦿ Does not give accurate structure predictions

Minimize Energy

- ⦿ All possible choices of complementary sequences are considered
- ⦿ consider all possible choices of complementary sequences to find the most stable structure
- ⦿ Stacks (contiguous nested base pairs) are the dominant stabilizing force – contribute to the negative free energy
- ⦿ Unpaired bases form destabilizing loops, contributing the positive free energy.
 - Hairpin loops, bulge/internal loops, and multiloops.
- ⦿ Uses Dynamic Programming alignment technique

Minimize Energy

- ⦿ Energy minimization algorithm predicts the secondary structure by minimizing the free energy (ΔG)
- ⦿ ΔG calculated as sum of individual contributions of:
 - loops
 - base pairs
 - secondary structure elements

● Free-energy values (kcal/mole at 37°C)

Stacking Energies for base pairs						
	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

- Free-energy values (kcal/mole at 37°C)



	Destabilizing Energies for Loops				
Number of Bases	1	5	10	20	30
Internal	--	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Hairpin	--	4.4	5.3	6.1	6.5



- ⦿ Given the energy tables, the free energy can be calculated for a structure
- ⦿ The score in dynamic programming is based on the free energy values
- ⦿ Gaps represent some form of a loop
- ⦿ The most widely used software that incorporates this minimum free energy algorithm is MFOLD/RNAfold

Drawbacks

- Only have one optimal solution

Features from secondary structure

- ◎ “Prediction of Mammalian MicroRNA Targets”, Benjamin P. Lewis etc.
- ◎ Assumption:
 - Perfect Watson-Crick complementarity to bases 2–8 of the miRNA
 - extend each seed match to a longer “target site”
- ◎ The miRNA/target site duplex stability was evaluated by assigning energy (ΔG) to the duplex
 - A candidate target site was rejected if the ΔG value was higher than a threshold

Features from secondary structure

- ◎ “Naïve Bayes for MicroRNA Target Predictions- Machine Learning for MicroRNA Targets”, Malik Yousef etc.
- ◎ Assumption:
 - A seed segment with weak complementarity can be compensated for by the out-seed sequence

- ⦿ partition the duplex into two parts, the seed (8nt of the miRNA) and out-seed
- ⦿ For each part, consider the following feature of the duplex
 - The number of paired bases
 - the number of bulges
 - the number of loops
 - the number of asymmetric loops
 - eight features, each representing the number of bulges of lengths 1-7 and those with lengths greater than 7

- Eight features, each representing the number of symmetric loops with lengths 1-7 and those with lengths greater than 7.
- Eight features each representing the number of asymmetric loops with lengths 1-7 and those with lengths greater than 7.
- the distance from the start of the seed (the 3' end) to the first paired base of the 5' start of the out-seed