RNA Nearest Neighbor Energy Model

and

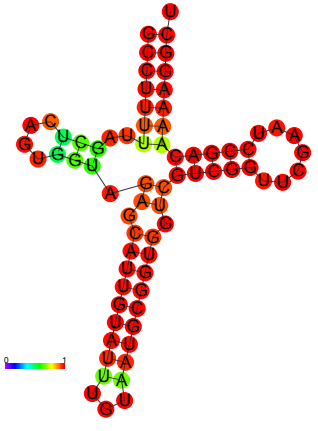
The Curse of Locality

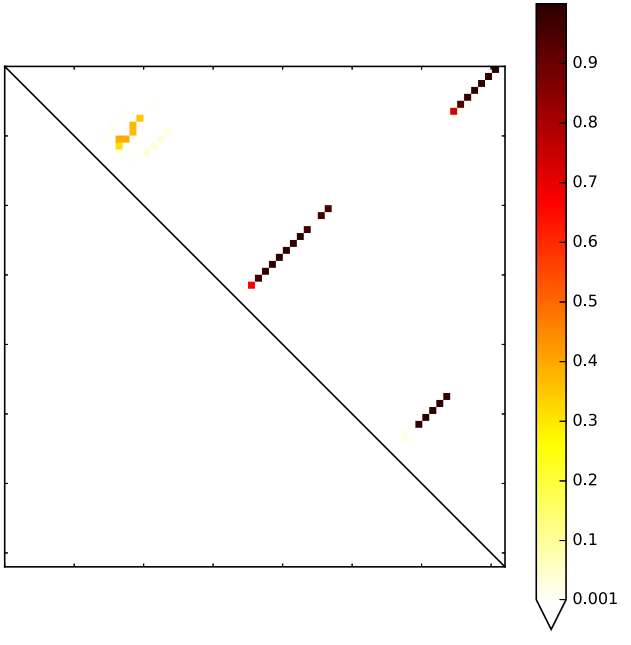
Milad Miladi

Herzogenhorn, April 2016

Target example

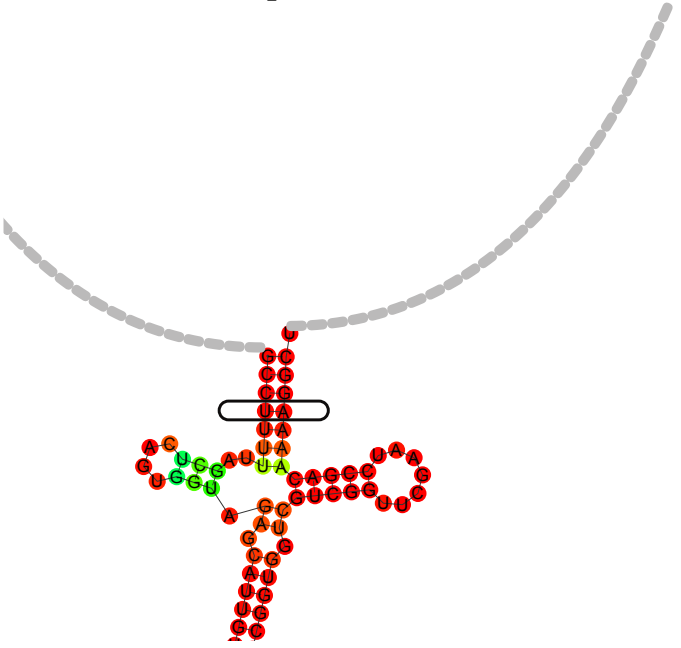
- A classic tRNA!





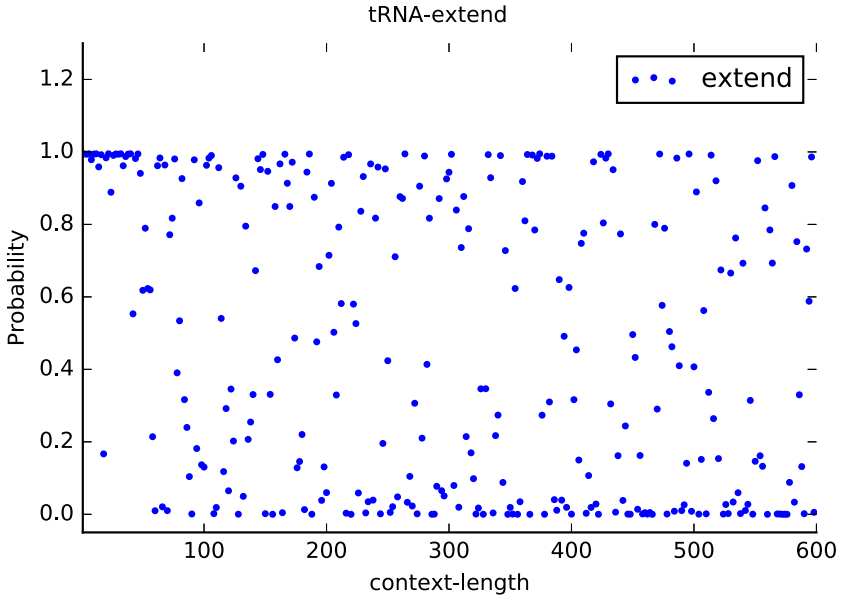
Test 1: Extension

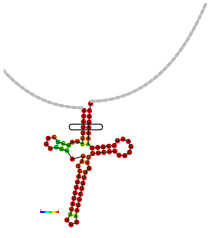
- Di-Nucleotide shuffled genomic context
- tRNA position:
  - close to the center of the extension
  - according to a normal distribution
- Target: a base-pair from *the acceptor stem*



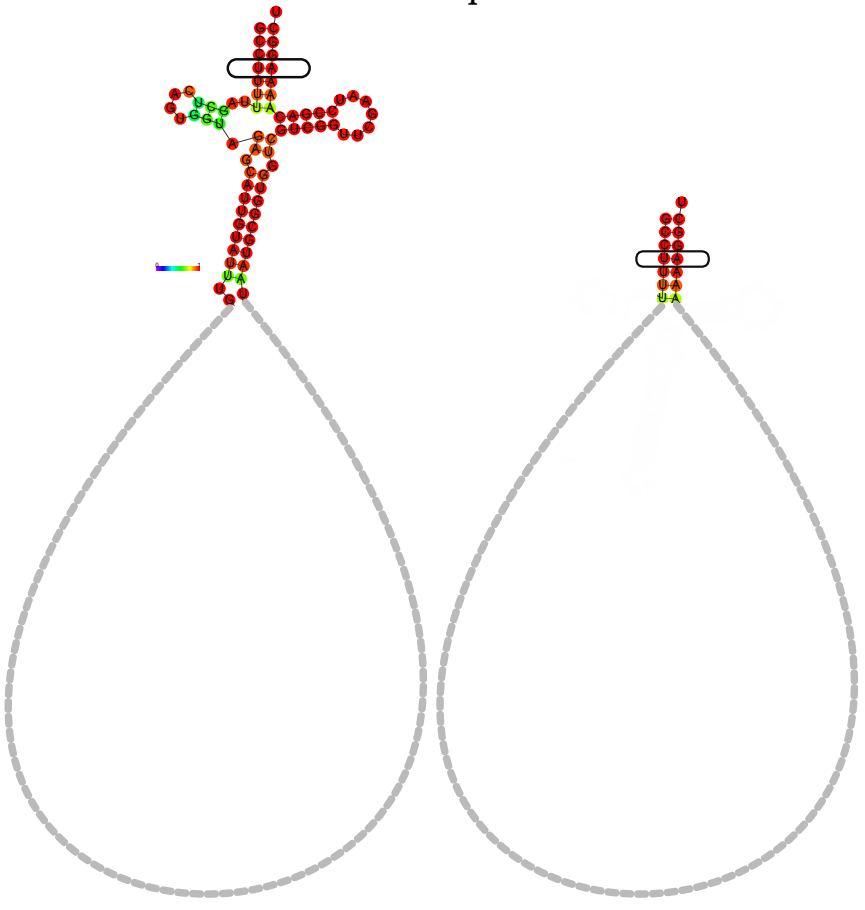
Probability of the selected base-pair (by global folding)

- Context-length:
  - Total length of the left and right extensions
- Each time the context is re-shuffled and re-sampled



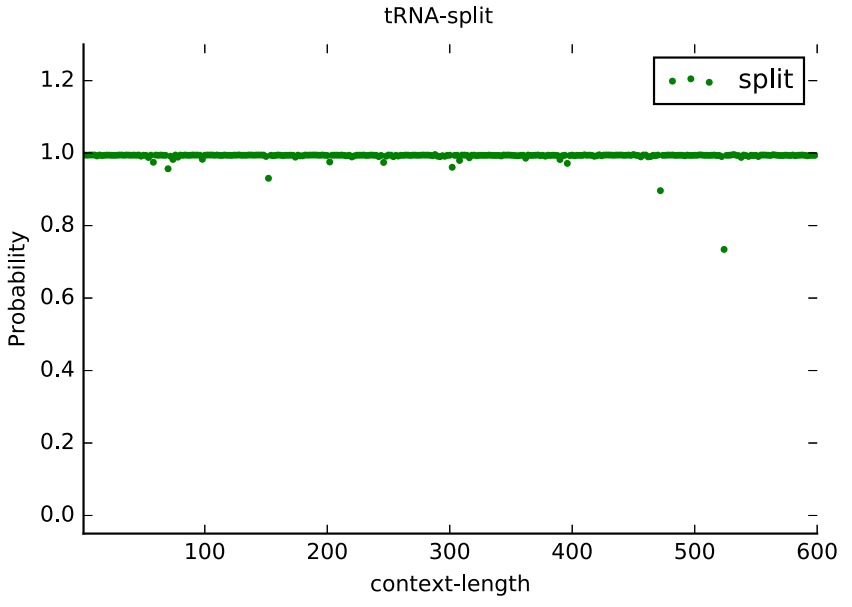


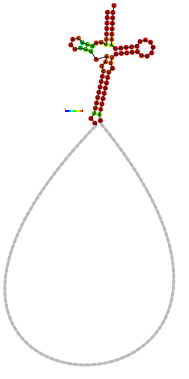
Test 2: Split



Probability of the selected base-pair (by global folding)

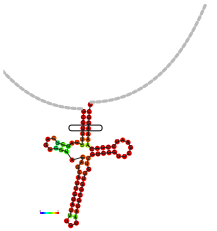
- Each time the context is re-shuffled and re-sampled





Problems

- **Locality** problem: *(extend)*
  - Desired base-pair probabilities easily distorted
  - Specially for the closing stems of multi-loops



- **Anti-locality** problem: *(split)*
  - No matter how long a sequence is ..
  - No matter what is inside ..
  - Few distant compatible base-pairs make an strong prediction!

What is missing?

Turner?

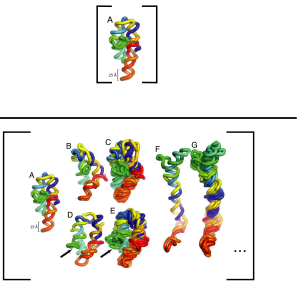
- Turner energy model should not be that much mad

McCaskill?

- McCaskill algorithm has no heuristics or simplification..

Probability of an structure in the ensemble

$$\frac{BW}{Z} = \frac{e^{\frac{-E(P)}{kT}}}{\sum_{P \text{ structure for } S} e^{\frac{-E(P)}{kT}}} =$$



- BW:
  - Boltzmann Weight
  - Exponential function => exponential scale behaviors!
- Z:
  - Partition function
  - Sum of the Boltzmann weights for the entire ensemble

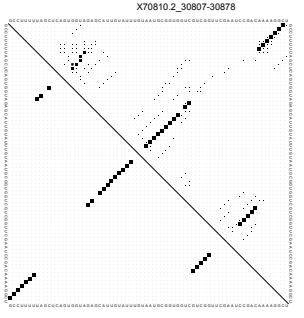
McCaskill, 1990,

- For a given sequence, efficient methods for:

1. **partition function (Z)**
    - Z(i,j)
    - For all sub-sequences
2. **probability of an individual base-pair** in ensemble
    - p(i,j)
    - For all possible pairs
3. Visualizing all base-pair probabilities as **dot plot**
    - Area(i,j) = p(i,j) . Unit-Area

Z(1,72)=-25.45 kcal/mol

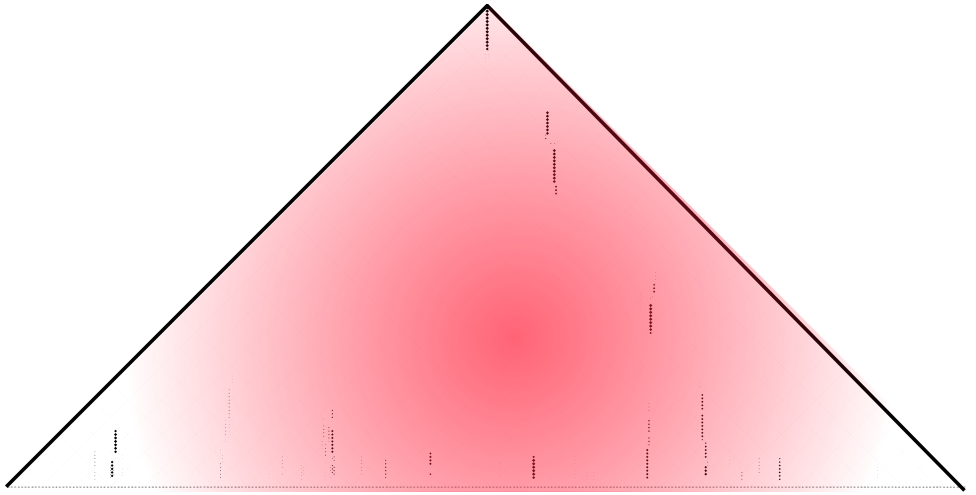
p(3,68)=0.9



What we have been missing?

The concept of noise and context in nowadays genomic biology

The fact that McCaskill's mindset was chemistry, not genome crawling



Can we solve it?

Calculating the base-pair probabilities with in inside algorithm

1. Base case: P_Hairpin(i,j)
2. Inner Loop: P_kl(ij| kl is closing ij)
3. Multiloop: coming soon..

Dot plot, 1

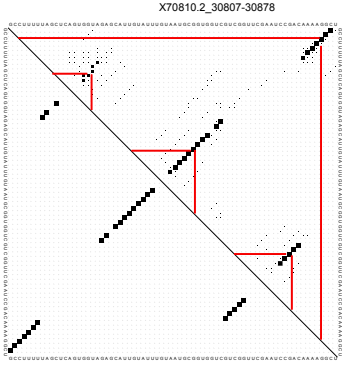
Advantages:

1. An excellent overview of **high** probable base-pairs
2. Great help to detect the **second** probable structure.
    - Ribo-switch/bistable RNAs for example

Dot plot, 2

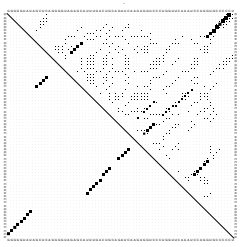
Advantages:

1. An excellent overview of **high** probable base-pairs
2. Great help to detect the **second** probable structure.
    - Ribo-switch/bistable RNAs for example
3. "Integration Test" 😎
    - For the new comers in the field of RNA-bioinf

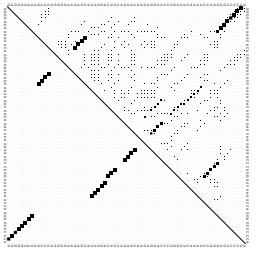


Dot plot: The evolution

1990



2016



*McCaskill's picture taken years are not exact :)

RNA Dotplots,

McCaskill

and the curse of Locality