RNA Nearest Neighbor Energy Model

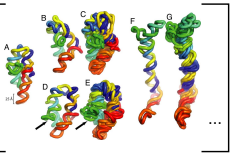
and

The Curse of Locality

Milad Miladi

Herzogenhorn, April 2016

Probability of an RNA structure in the ensemble

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

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

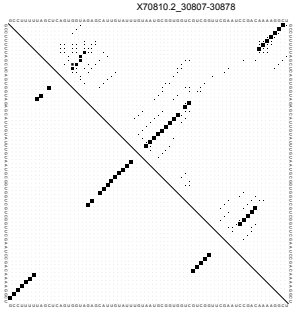
McCaskill, 1990,

- 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

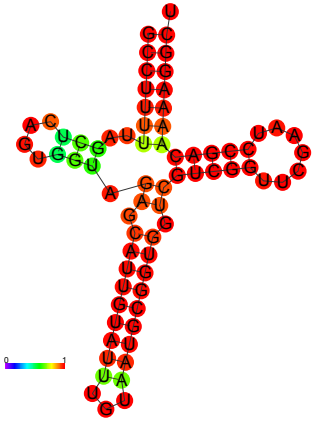
p(3,68)=0.9

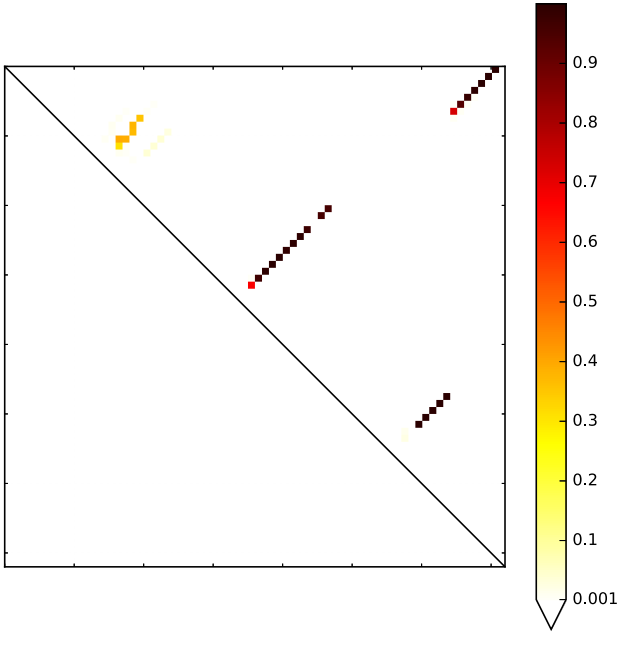


. Part 1: The problem

Target example

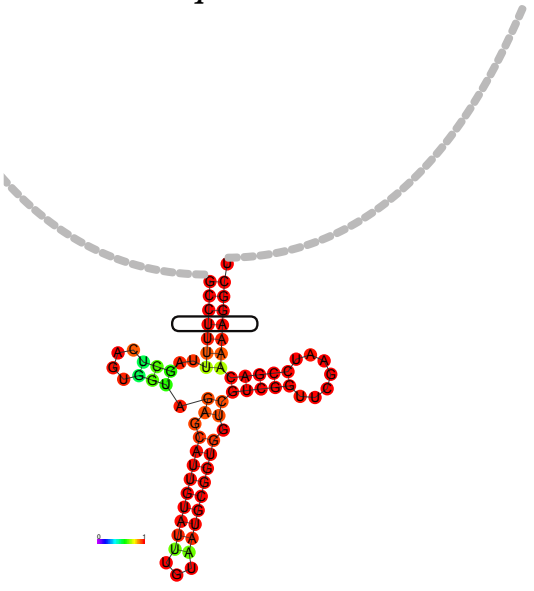
- A classic tRNA!





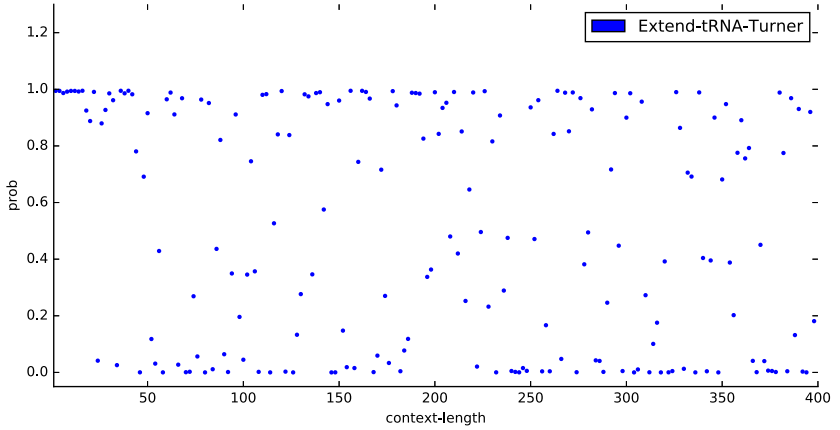
Test 1: Extension

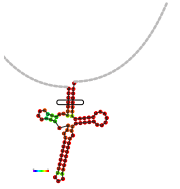
- Di-Nucleotide shuffled genomic context
- tRNA position:
  - close to the center of the extension
  - according to a normal distribution
- Target: a base-pair from *the acceptor stem*



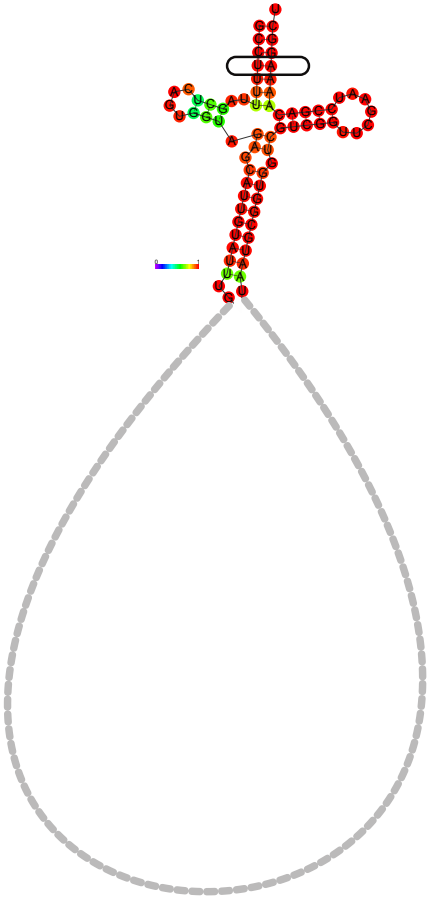
Probability of the selected base-pair (by global folding)

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



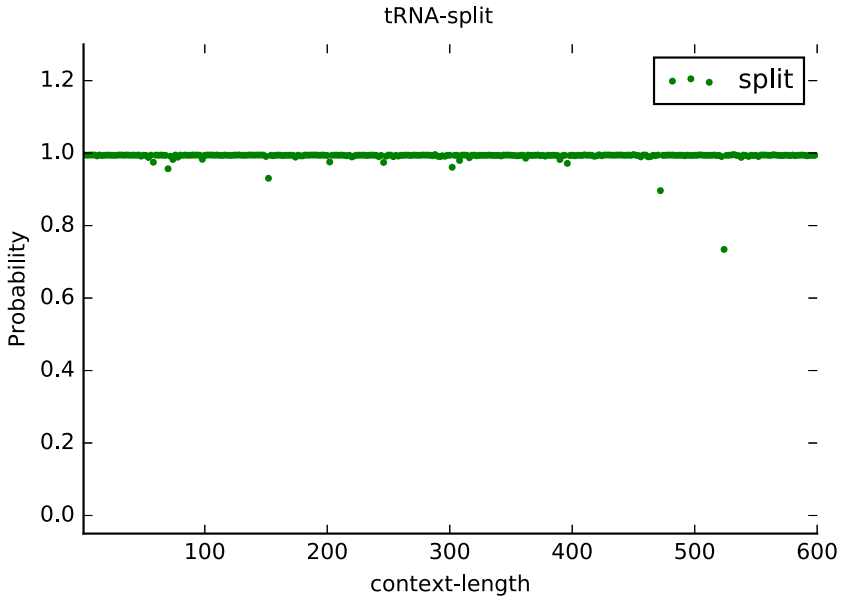


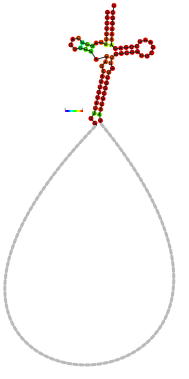
Test 2: Split



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

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



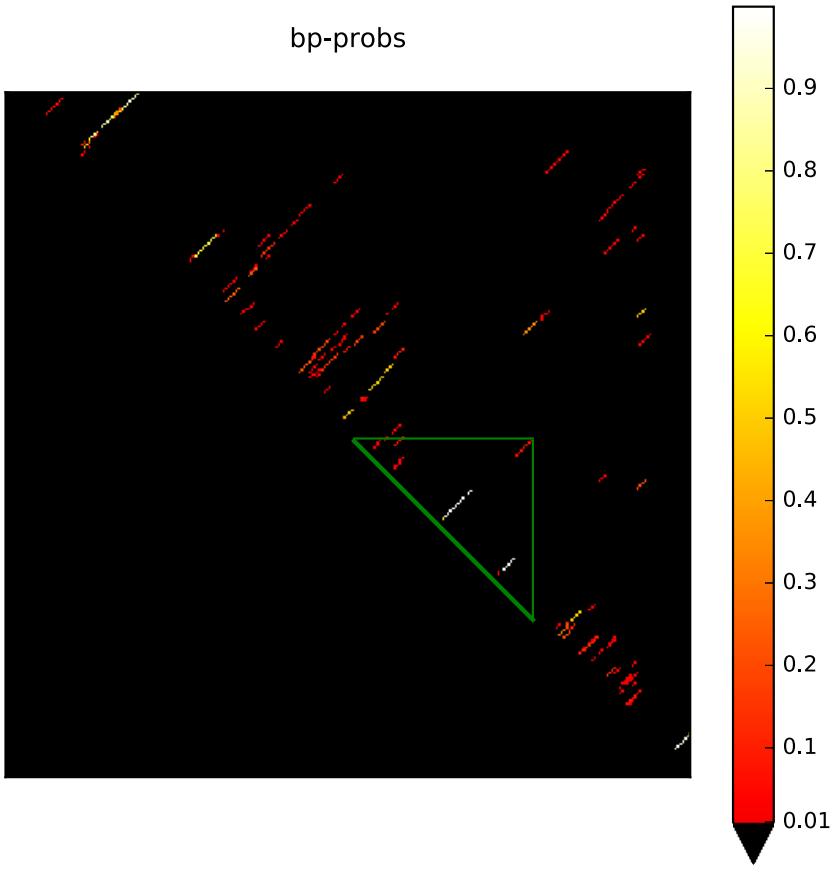


Problems

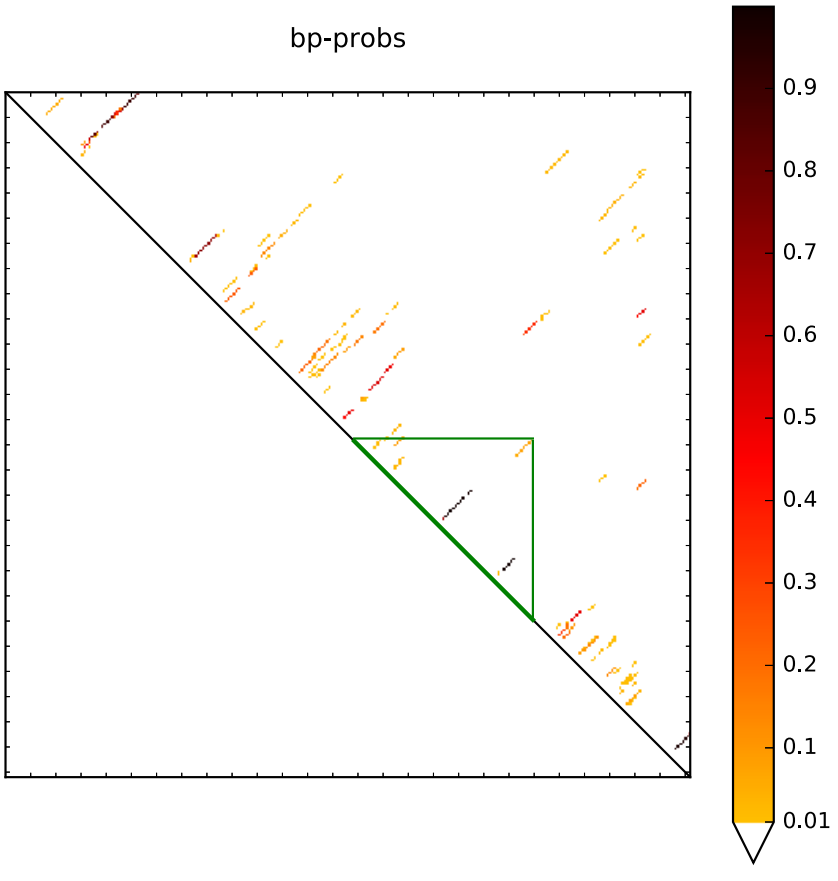
- **Locality** problem: *(extend test)*
  - Desired base-pair probabilities easily distorted
  - Specially for the closing stems of multi-loops

- **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!

Split example



Split example



(Slide from my talk last month)

What is missing?

Turner?

- Turner energy model should not be that much mad

McCaskill?

- McCaskill algorithm has no heuristics or simplification..

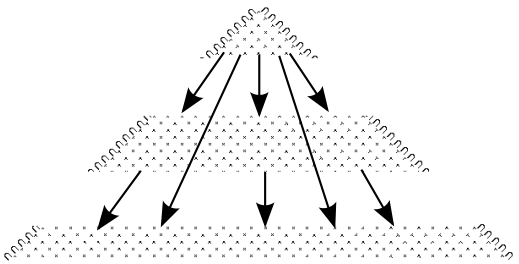
- **Part 1: The problem**

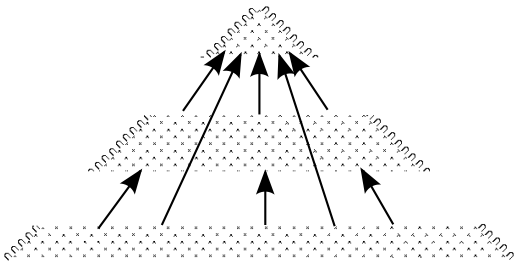
. Part 2: mmfold

Base pairing probability computation

Irreversibility hypothesis:

1. Markov chain of base-pair probabilities is not reversible
2. Computing the Markov chain with McCaskill's *outside* algorithm causes the locality problem (to some extend)

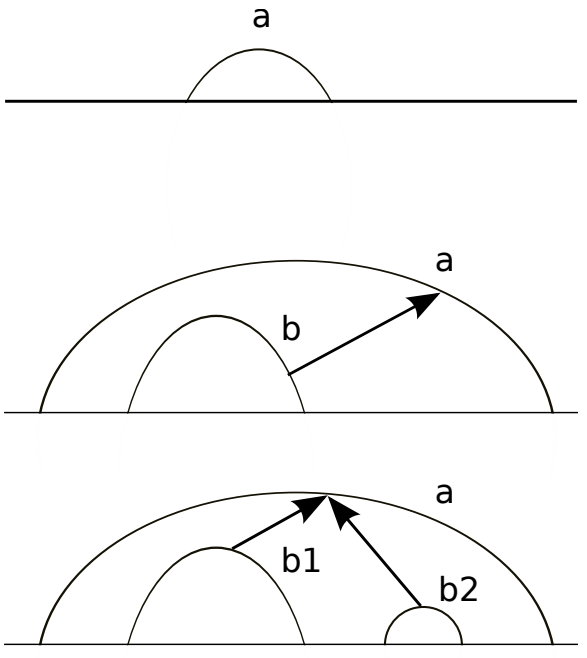




mmfold inside algorithm

- Calculating the base-pair probabilities with an inside algorithm

    1. Base case: P_Hairpin(a)
    2. Inner Loop: P(a| a is closing b)
    3. Multiloop: P(a| a is closing multiloop b1, b2, …)



mmfold implementation

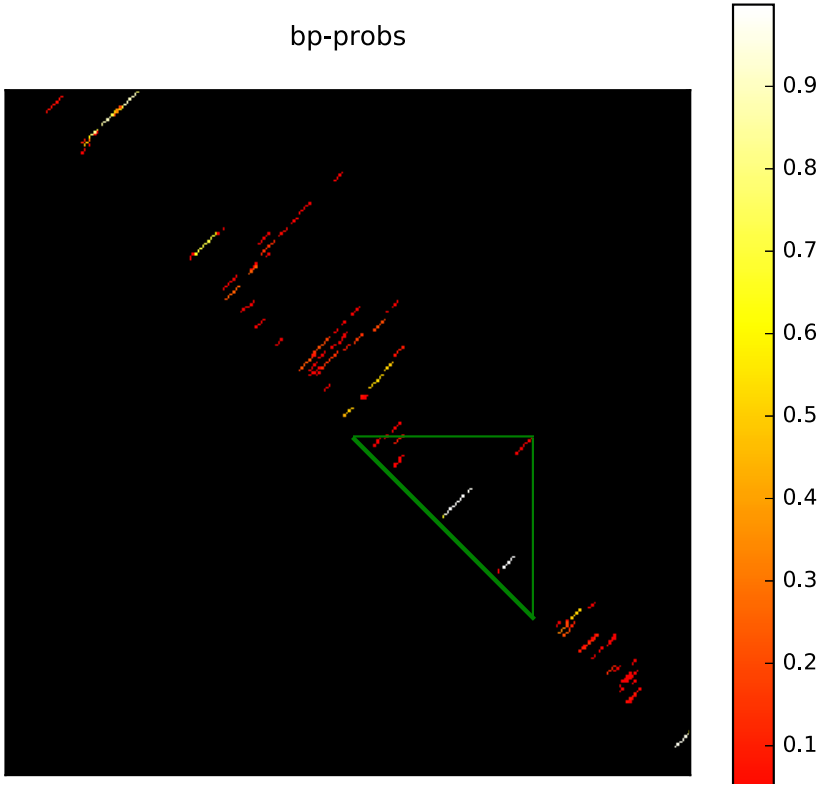
- Implemented in C with fun and pain! :D
- Directly inside cloned Vienna RNA package

```
PUBLIC void mm_pf_create_bppm( vrna_fold_compound_t *vc, char *structure){
    ...
    probs[ij] = mc_probs[ij]/
                qb[ij] *
                exp_E_Hairpin(j-i-1, type, S1[i+1], S1[j-1],
                    sequence+i-1 ,pf_params) * scale[j-i+1];
                    // TODO: Verify rescaling is correct!
    ...
    FLT_OR_DBL
    new_score = probs[ij] * (mc_probs[kl]/mc_probs[ij])
                * (qb[ij]/qb[kl]) * (scale[u1 + u2 + 2]
                * exp_E_IntLoop(u1, u2, type, type_2_r, S1[k+1], S1[l-1],
                *               S1[i-1], S1[j+1],  pf_params));

    ...

    tmp = qq1;  qq1 =qq;  qq =tmp;
    tmp = qqm1; qqm1=qqm; qqm=tmp;

}
```

- In my spare time (4 weekends + couple of afternoons)
- With a bunch of TODOS!

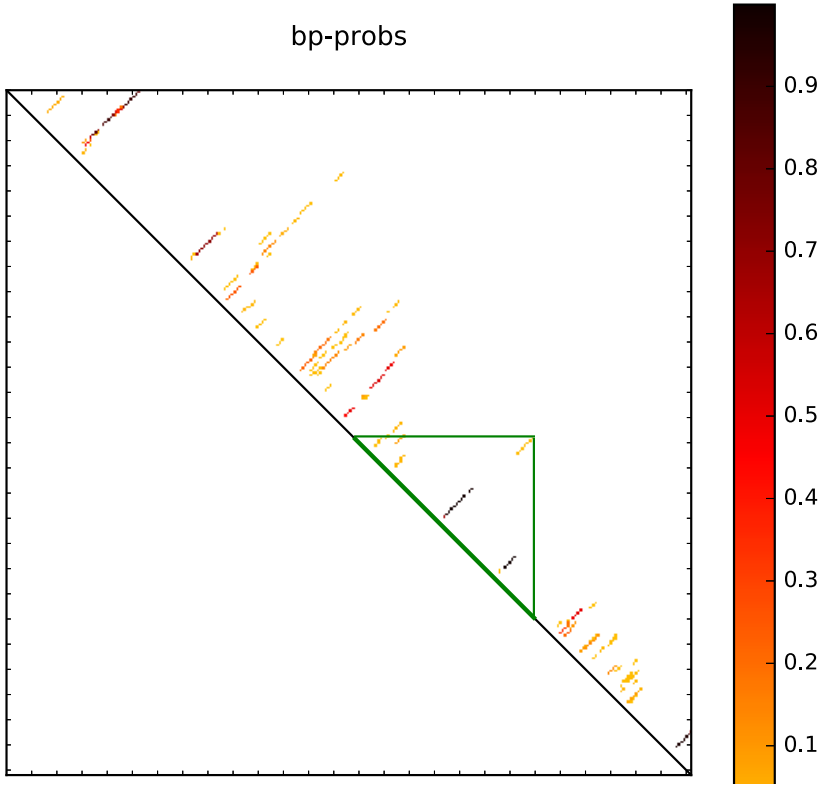
mmfold alpha: output

```
$ mmfold -p -P src/misc/rna_turner2004.par < trna2.fa
```



mmfold alpha: output

```
$ mmfold -p -P src/misc/rna_turner2004.par < trna2.fa
```



mmfold outcome

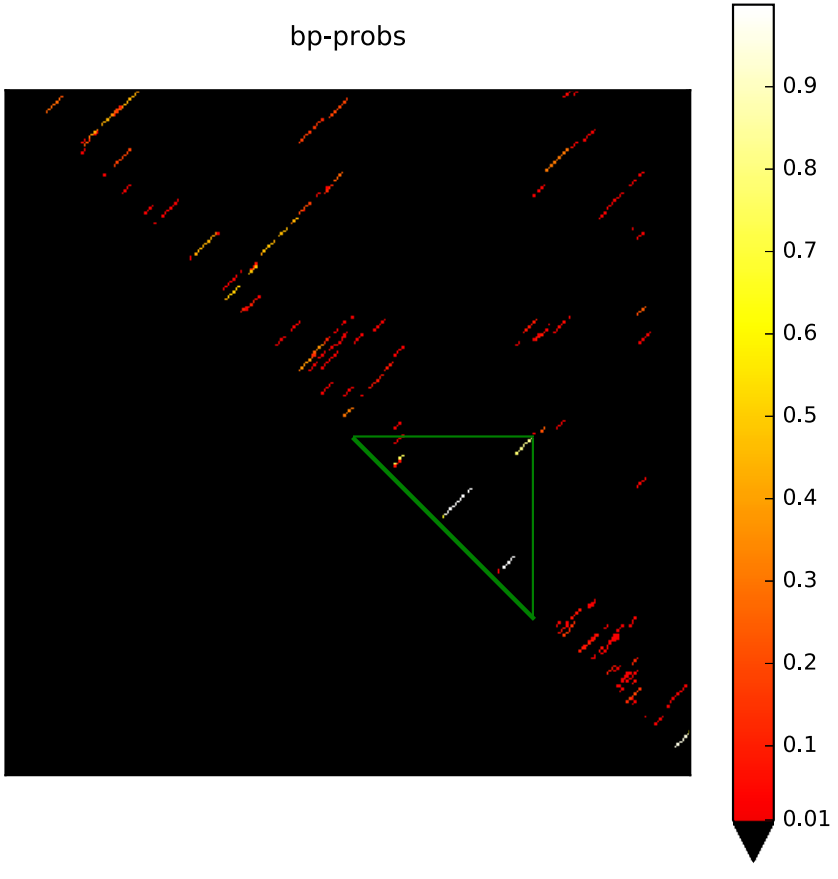
- My irreversibility hypothesis failed 👎
- Rolf and Martin were right ;-)
- But I also deep learned "Nearest Neighbor Energy Model"!

```
FLT_OR_DBL branch_energy = E_MLstem(type, S1[h-1], S1[k+1], vc->params);

FLT_OR_DBL branch_pf = exp_E_MLstem(rtype[type], S1[h-1], S1[k+1], pf_params);

FLT_OR_DBL qb_energy = (-log(qb[kh]) -(h-k+1)*log(pf_params->pf_scale))
                       *pf_params->kT/10.;
```

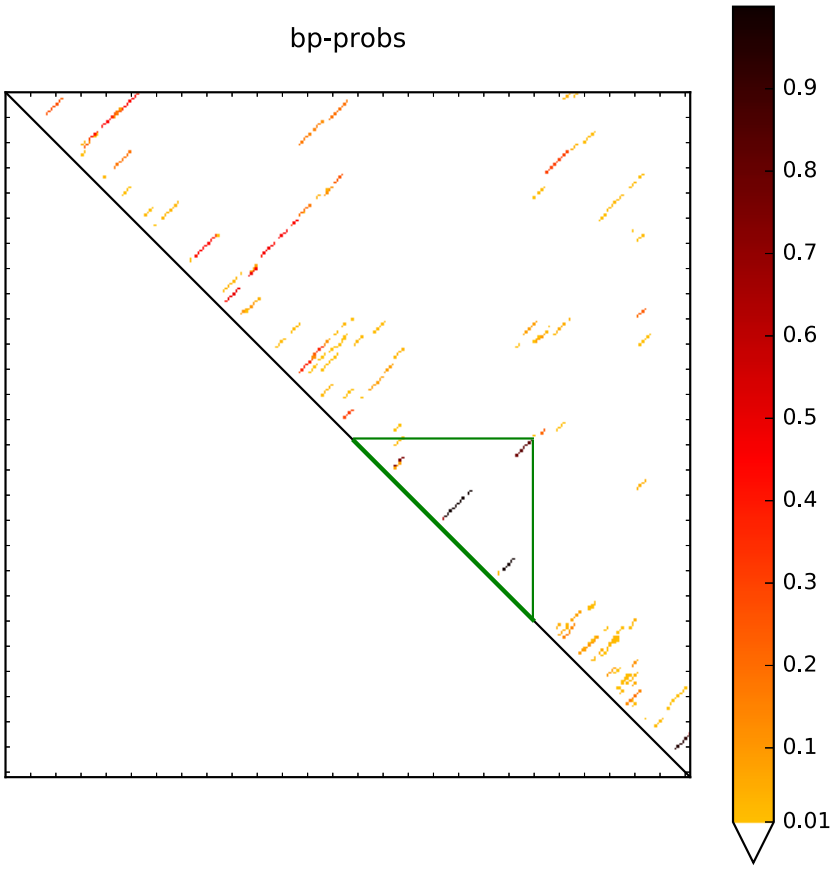
- **Part 1: The problem**
- **Part 2: mmfold**

. Part 3: Quake

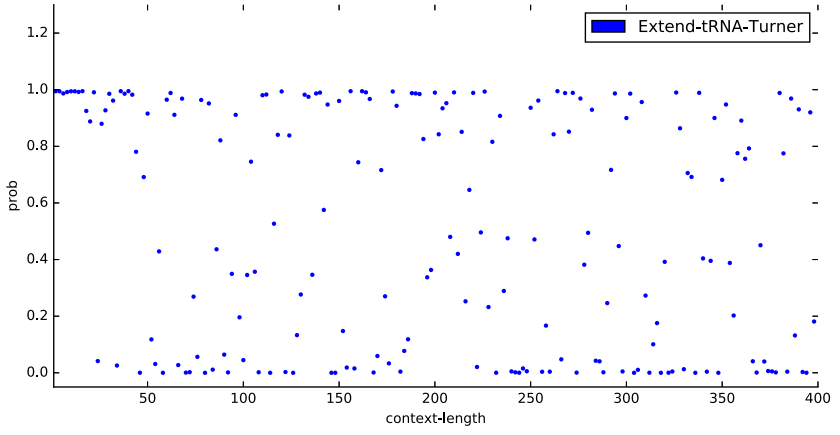
Quake example

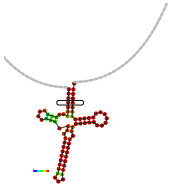


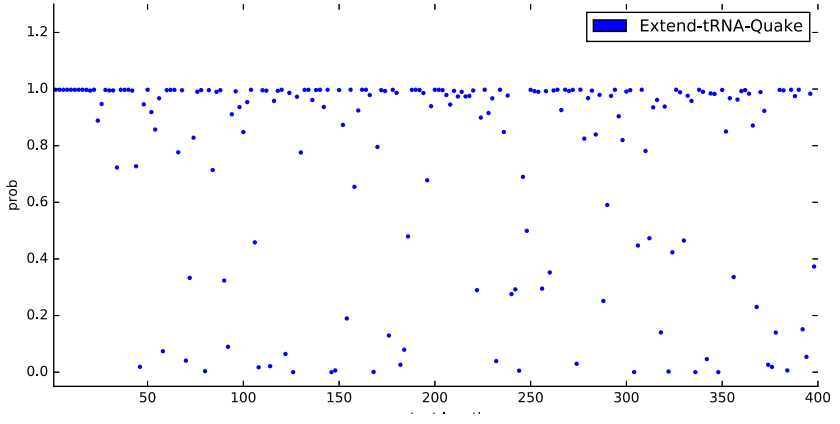
Quake example



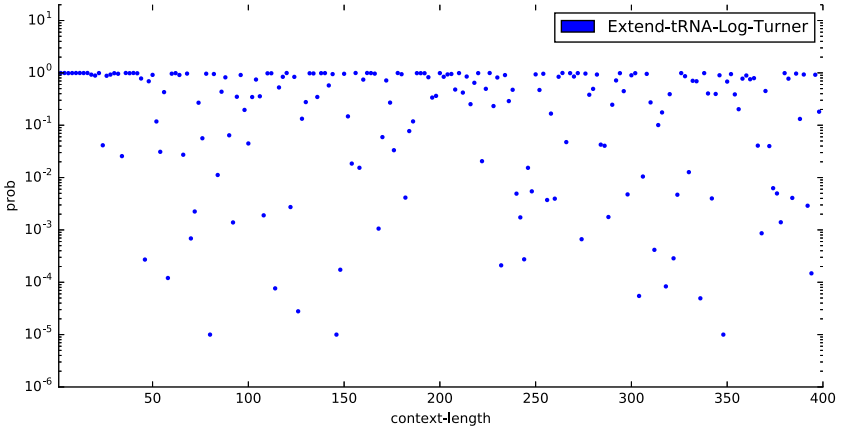
Extend: Turner vs Quake

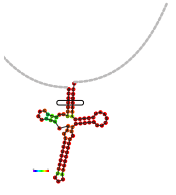


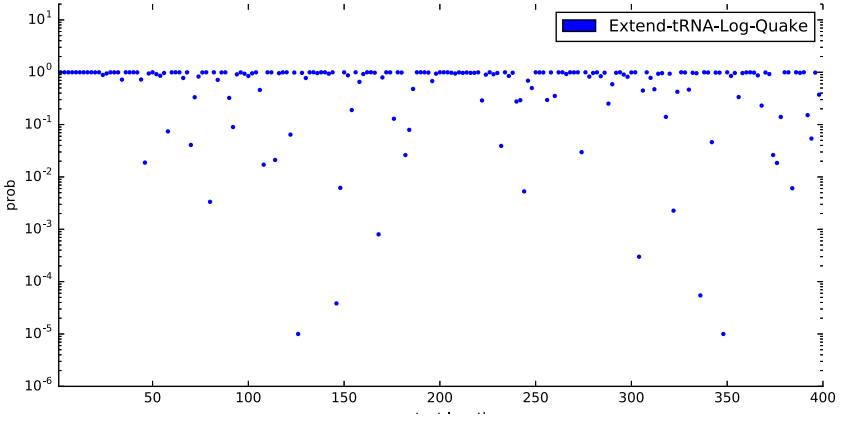




Extend: Turner vs Quake (Log scale)







What is Quake?

- RNAfold uses the famous Turner's energy parameters for free energy computations
- It is a new parameter set

```
RNAfold -p -P src/misc/quake.par
```

What is Quake?

- RNAfold uses the famous Turner's energy parameters for free energy computations
- It is a new parameter set

```
RNAfold -p -P src/misc/quake.par
```

- Not really!
- It is Turner's params except one param:
    - Unpaired nucleotide penalty of a multiloop region

Turner vs Quake

- Turner:

```
milad-Latitude:> ~/Downloads/ViennaRNA-2.2.4/misc
$ grep "ML" -A 3 rna_turner*
rna_turner1999.par:# ML_params
rna_turner1999.par-/* F = cu*n_unpaired + cc + ci*loop_degree (+TermAU) */
rna_turner1999.par-/*       cu   cu_dH      cc   cc_dH      ci   ci_dH  */
rna_turner1999.par-          0       0     340       0      40       0
--
rna_turner2004.par:# ML_params
rna_turner2004.par-      0       0     930    3000     -90    -220
rna_turner2004.par-
```

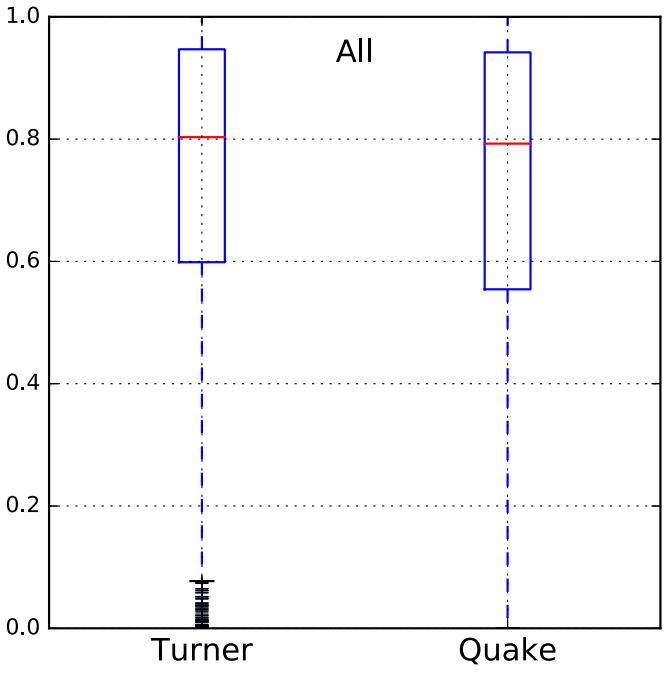
- Quake:

```
/* F = cu*n_unpaired + cc + ci*loop_degree (+TermAU) */
/*   cu      cu_dH   cc    cc_dH      ci     ci_dH  */
     50      0       930    3000     -190    -220
```

- **Part 1: The problem**
- **Part 2: mmfold**
- **Part 3: Quake**
- Part 4: Quake Evaluation

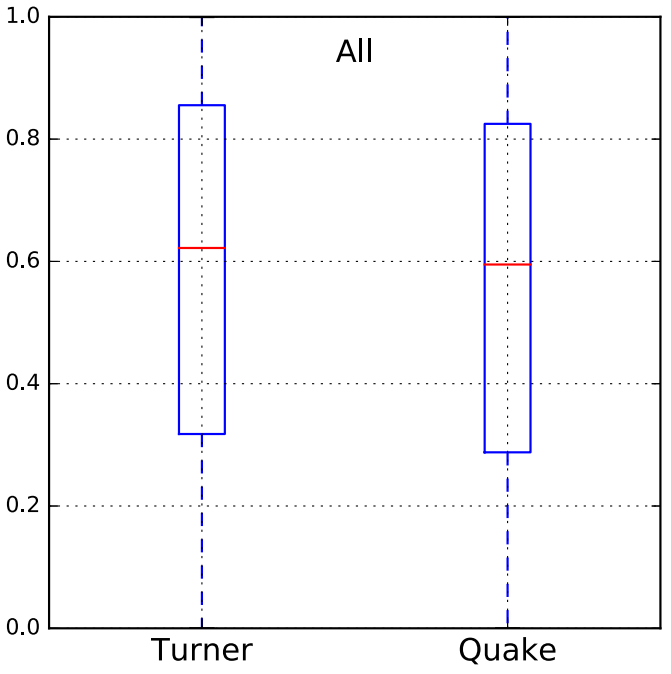
Localfold CisReg dataset, Context 0

Basepair accuracy (=expected sensitivity)



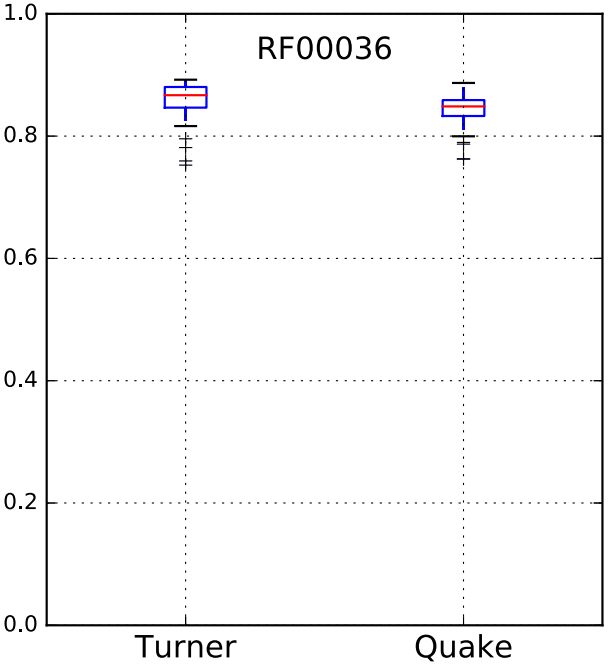
Localfold CisReg dataset, Context 200

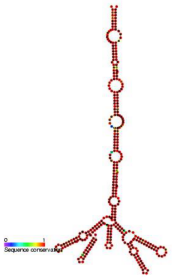
Basepair accuracy (=expected sensitivity)



Localfold CisReg dataset

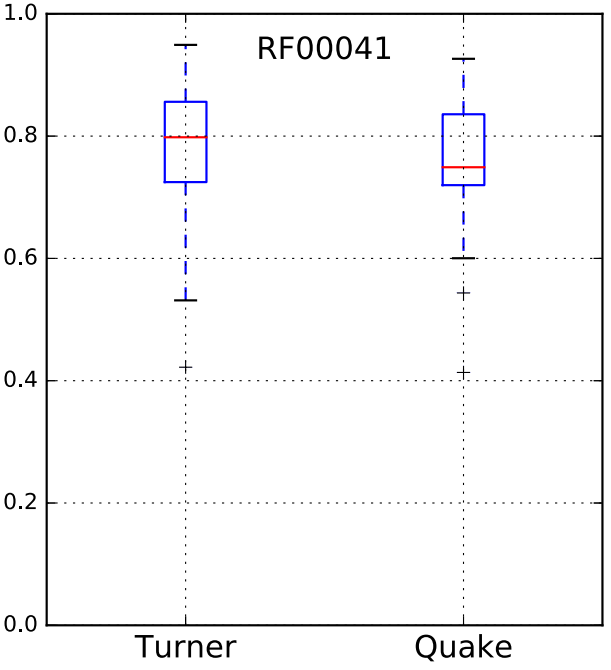
Basepair accuracy (=expected sensitivity)

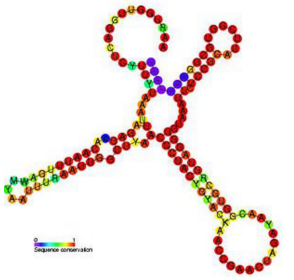




Localfold CisReg dataset

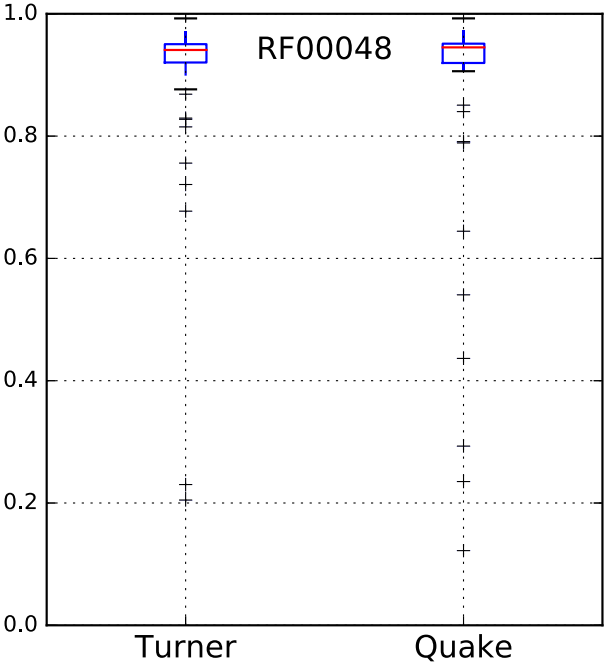
Basepair accuracy (=expected sensitivity)





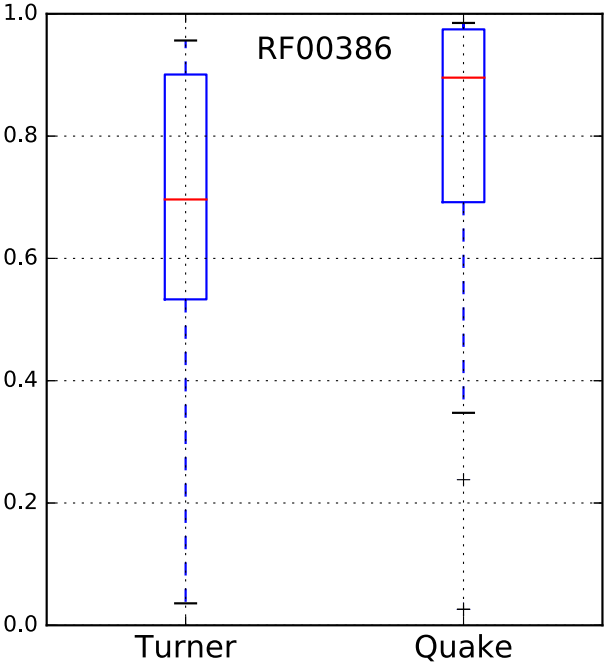
Localfold CisReg dataset

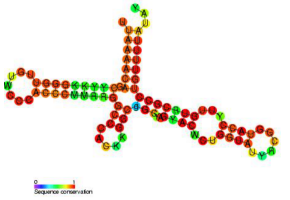
Basepair accuracy (=expected sensitivity)



Localfold CisReg dataset

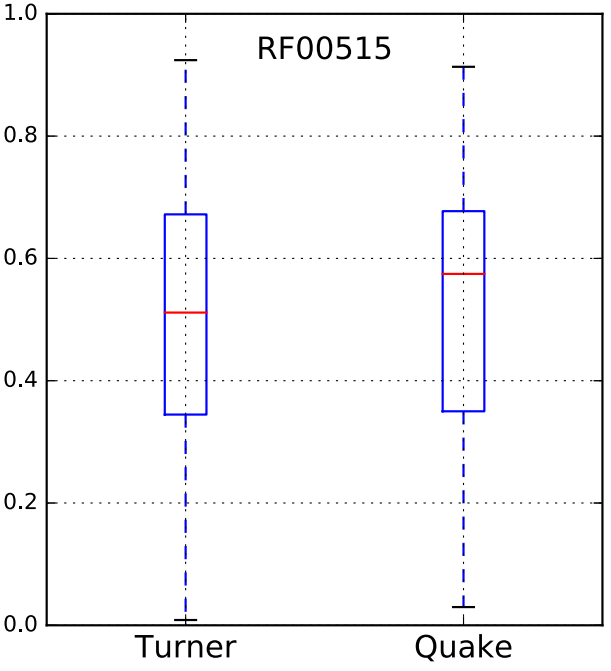
Basepair accuracy (=expected sensitivity)

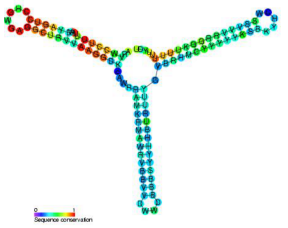




Localfold CisReg dataset

Basepair accuracy (=expected sensitivity)





(Slide from my talk last month)

What is missing?

Turner?

- Turner energy model should not be that much mad

McCaskill?

- McCaskill algorithm has no heuristics or simplification..

Update:

- Well the Turner energy model is not mad but highly overfitted to positive set of RNA strands, with nice boundaries
- For multiloop parameters (at least)
- More precisely the dynamic programming variation of Turner model is overfitted

RNA Dotplots,

McCaskill

and the curse of Locality