Locality Glitch

in

Established RNA Energy Models

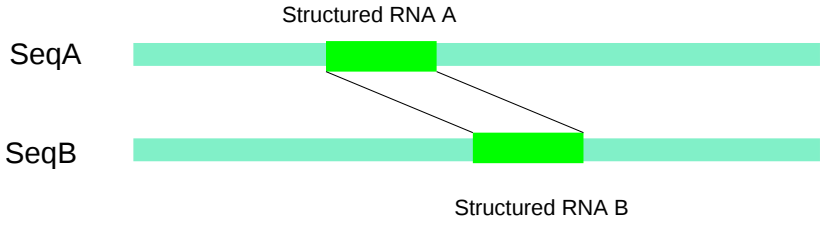
Milad Miladi

University of Freiburg

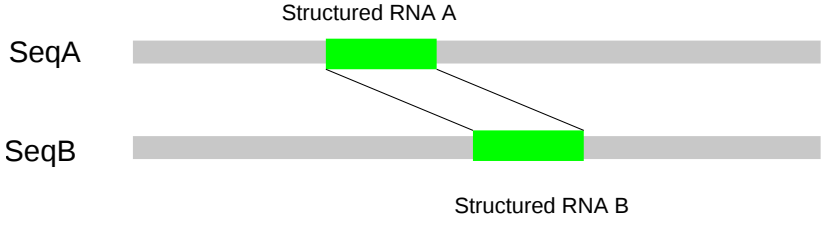
32nd TBI Winterseminar in Bled, F ebruary 2017

Motivation: Improving local alignment of RNAs

Genomic context

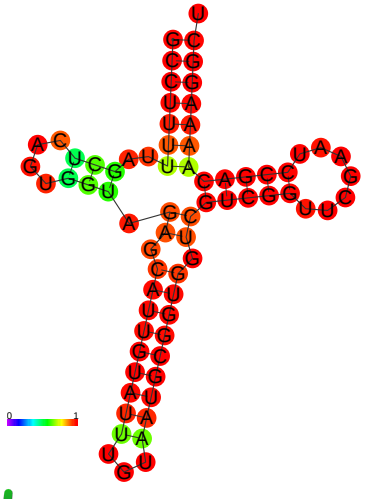


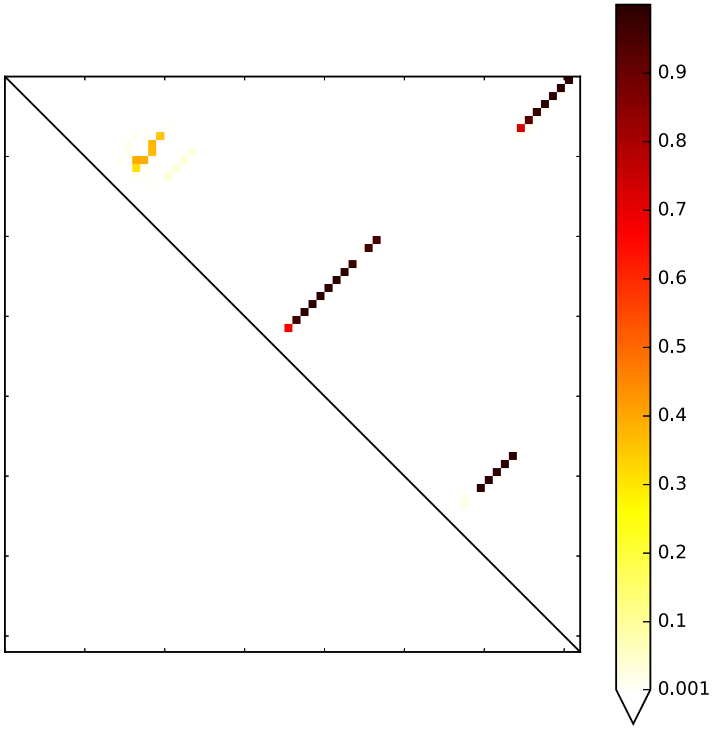
Shuffled genomic context



Inspected RNA

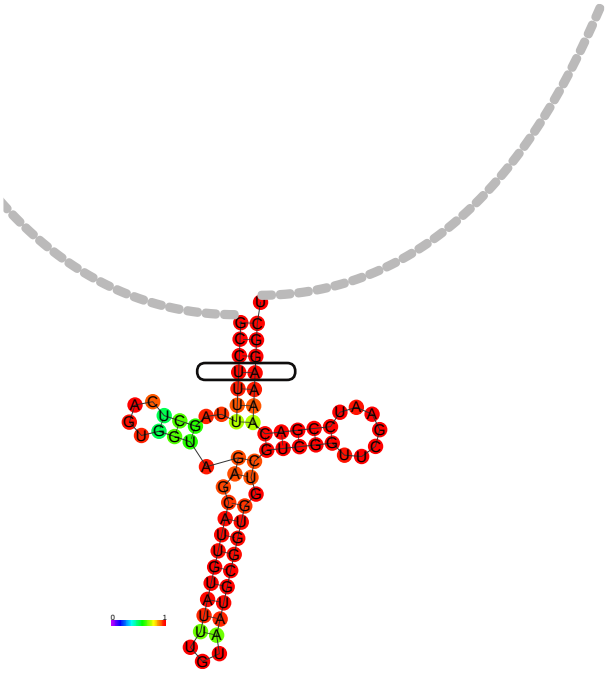
- A *tRNA* with its typical cloverleaf secondary structure
- Evaluate probability of two base-pairs from:
  - *Acceptor stem*
  - *Anticodon stem*



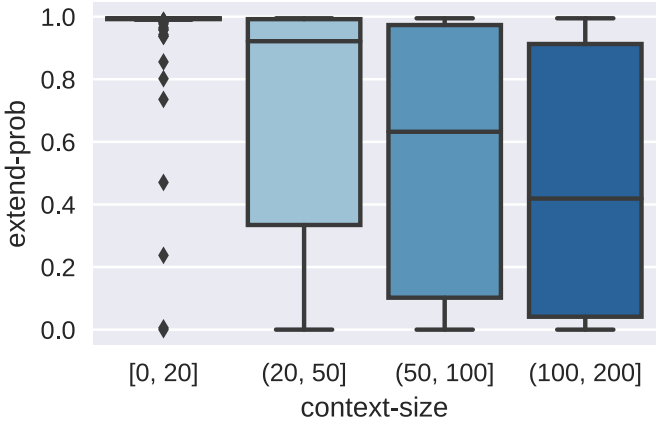


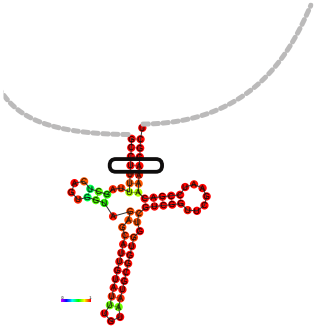
Experiment: Extension

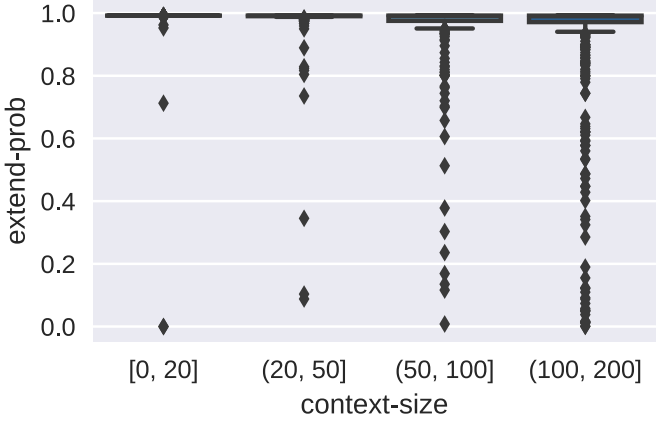
- Shuffled genomic context of the tRNA
- tRNA positioned
  - in the proximity of the midpoint
  - according to a normal distribution

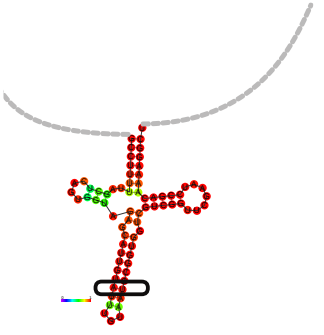


Probability of the selected base-pair (global folding)



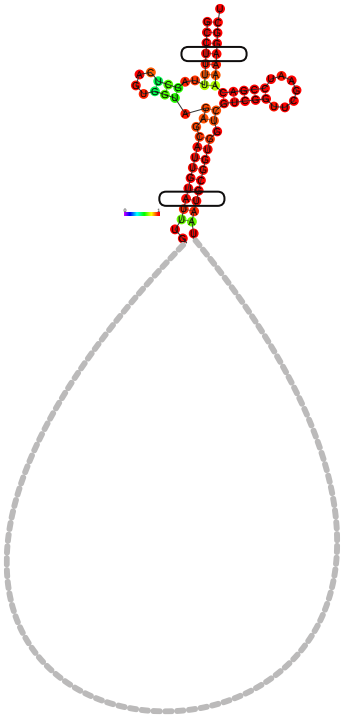




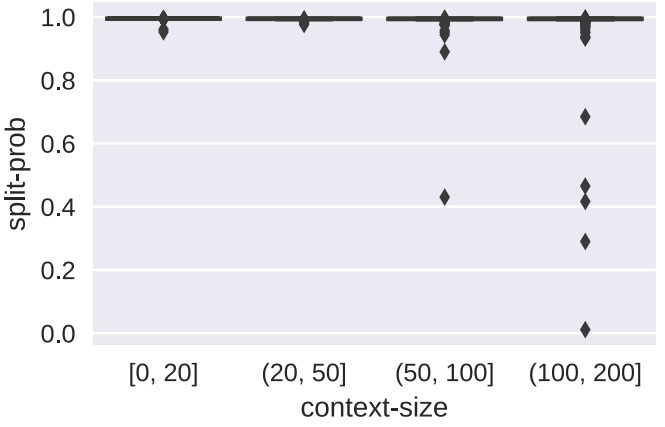


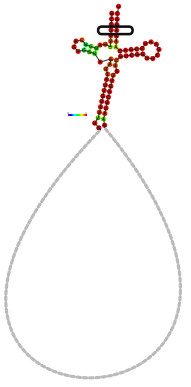
Experiment: Insertion

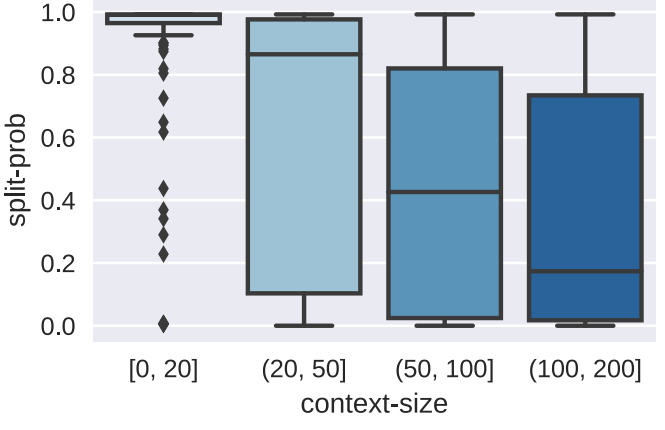
- Shuffled genomic context of the tRNA is inserted into the *Anticodon* loop

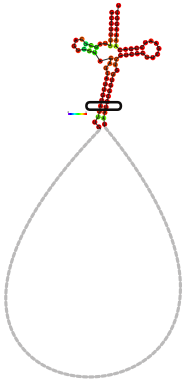


Probability of the selected base-pair (global folding)



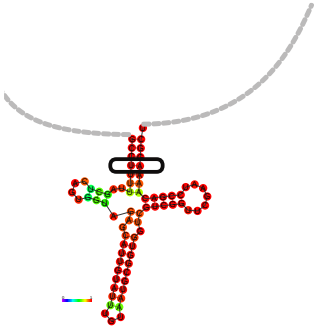






Observations

- **Locality**: *(extend test)*
  - A relatively short context can distort the acceptor signal
  - Specially for the closing stems of multi-loops



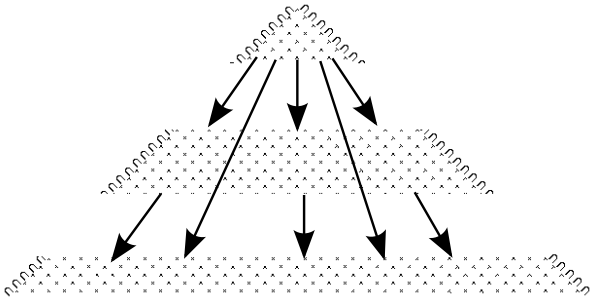
- **Anti-locality**: *(insert test)*
  - Independent of a sequence and content
  - Few distant compatible base-pairs make an strong prediction

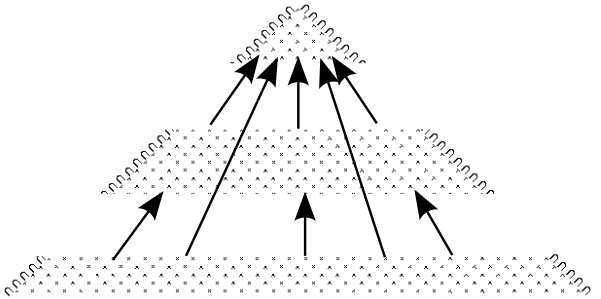
Irreversibility hypothesis:

Base pairing probability computation

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)

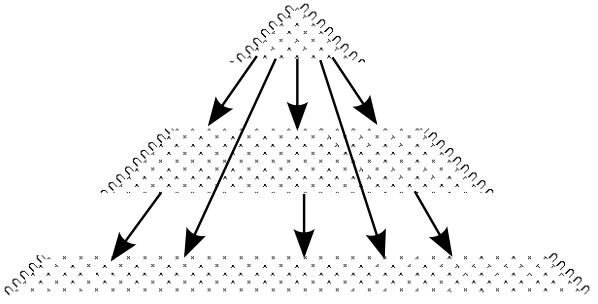


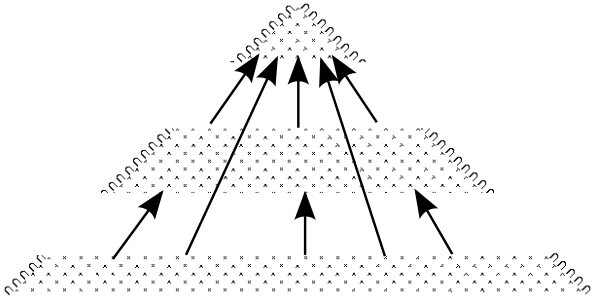


Irreversibility hypothesis:

Base pairing probability computation

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)





Shown to not be a valid hypothesis

Models multiloop parameters

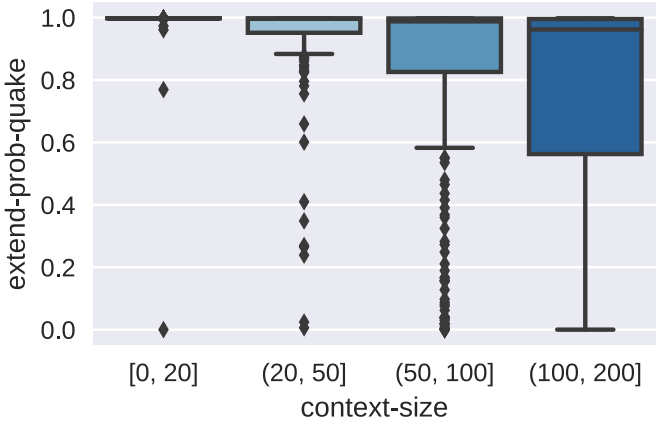
```
Multiloop free energy = cu*n_unpaired + cc + ci*loop_degree
```

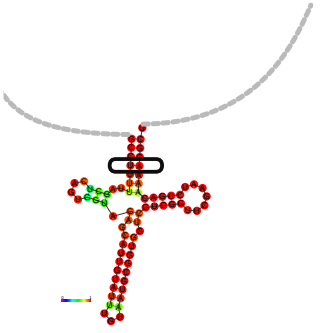
Model	cu	cc	ci
Turner-1999	**0**	340	40
Turner-2004	**0**	930	-90
Andronescu-2007	**4**	440	3
Quake (Patched Turner)	**50**	930	-190

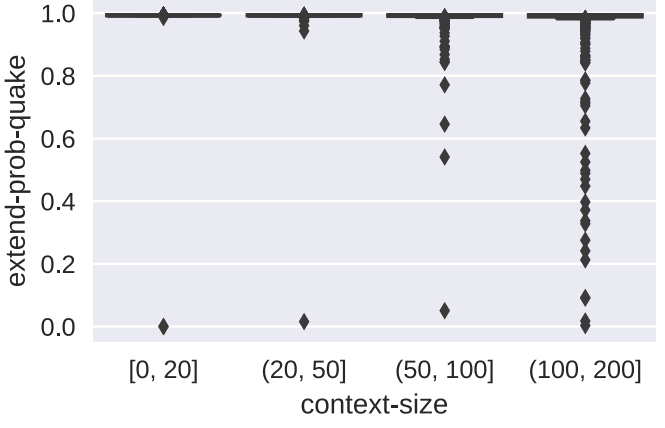
- (*) In fact Turner's lab proposed two versions of Multiloop scores:
  - Efficient version with a constant unpaired probability with value zero
  - Detailed version similar to inner-loop case
- More precisely, due to efficiency reasons, the dynamic programming variation of Turner model consider no penalty for unpaired region of Multiloops

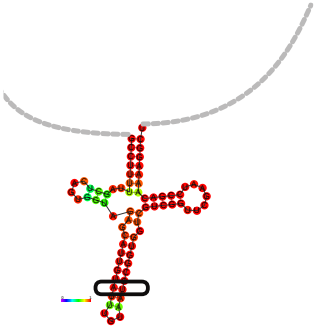
[Mathews et al. 1999, 2004]; [Andronescu et al.2007];

Probability of the selected base-pair (global folding, Quake)



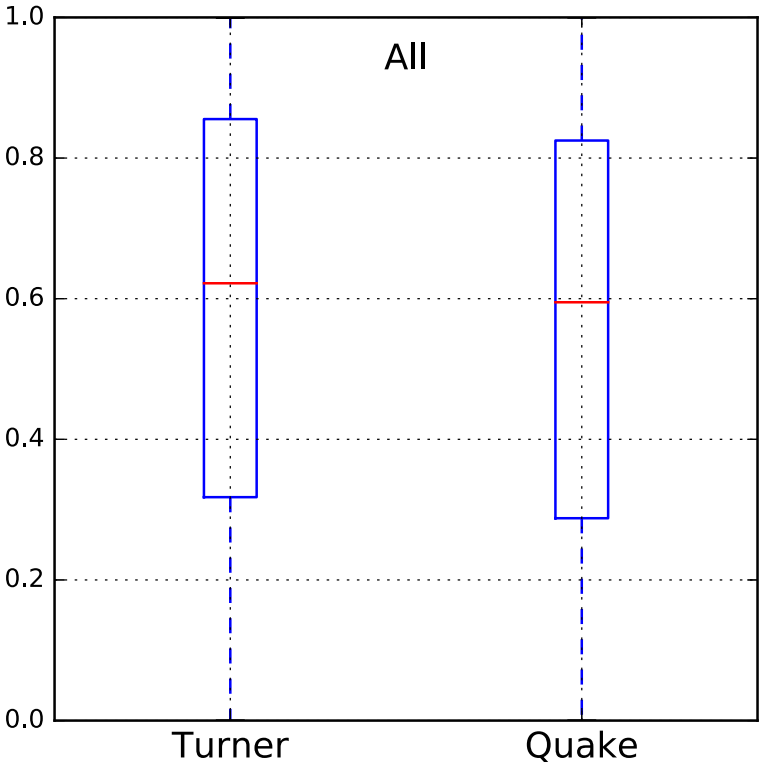






CisReg dataset*, Asymmetric Context 200

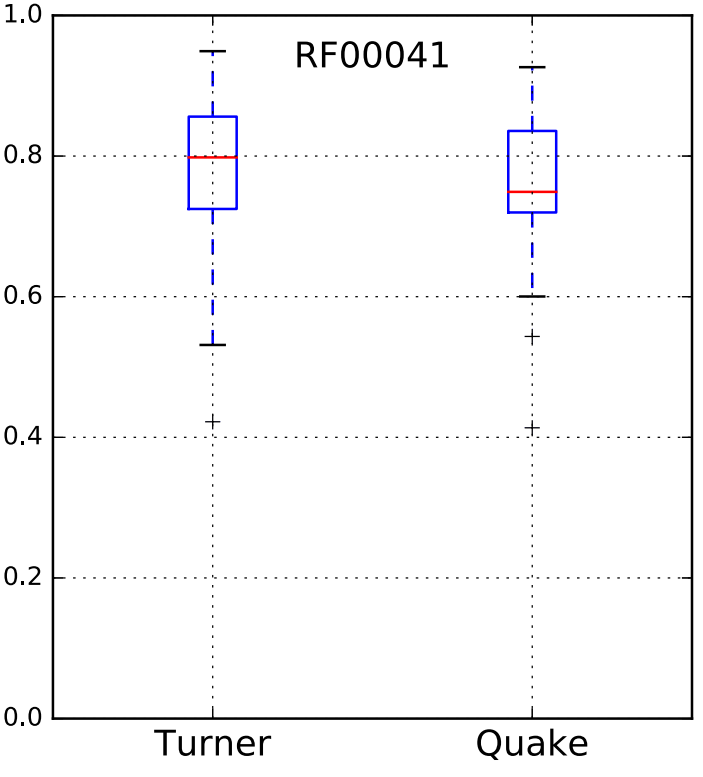
Basepair accuracy (=expected sensitivity)

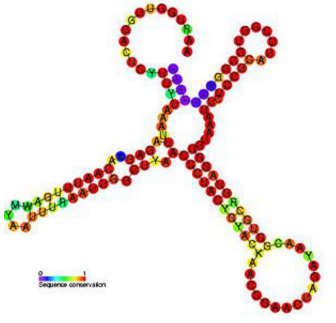


*Lange, Maticzka, Möhl et al. NAR-2012

CisReg dataset, Asymmetric Context 200

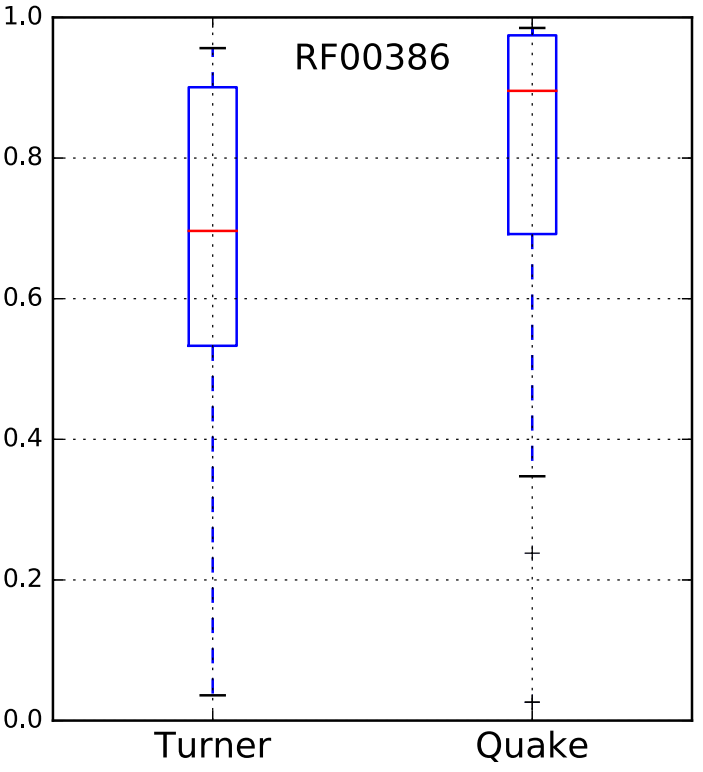
Basepair accuracy (=expected sensitivity)

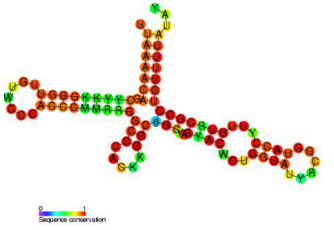




CisReg dataset, Asymmetric Context 200

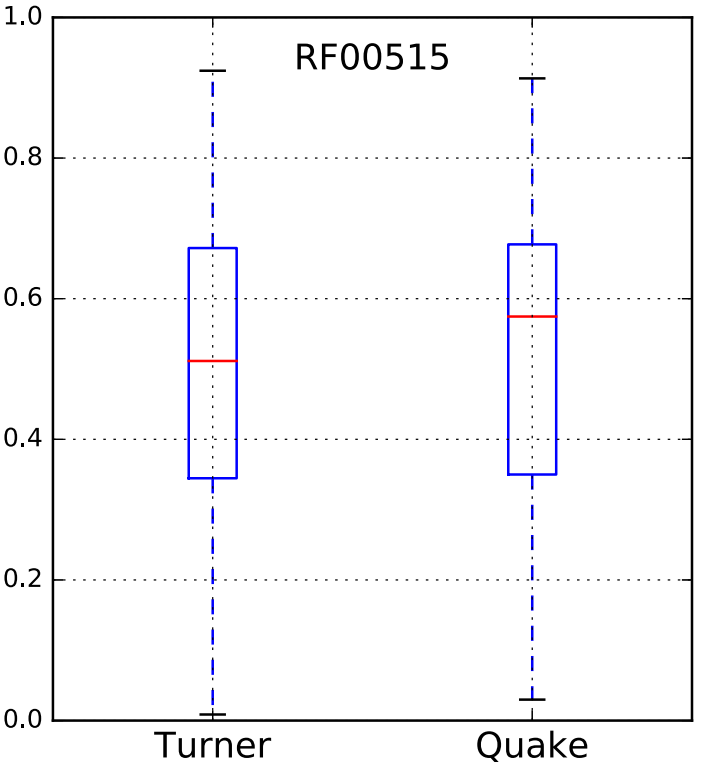
Basepair accuracy (=expected sensitivity)

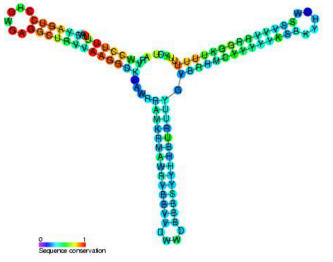




CisReg dataset, Asymmetric Context 200

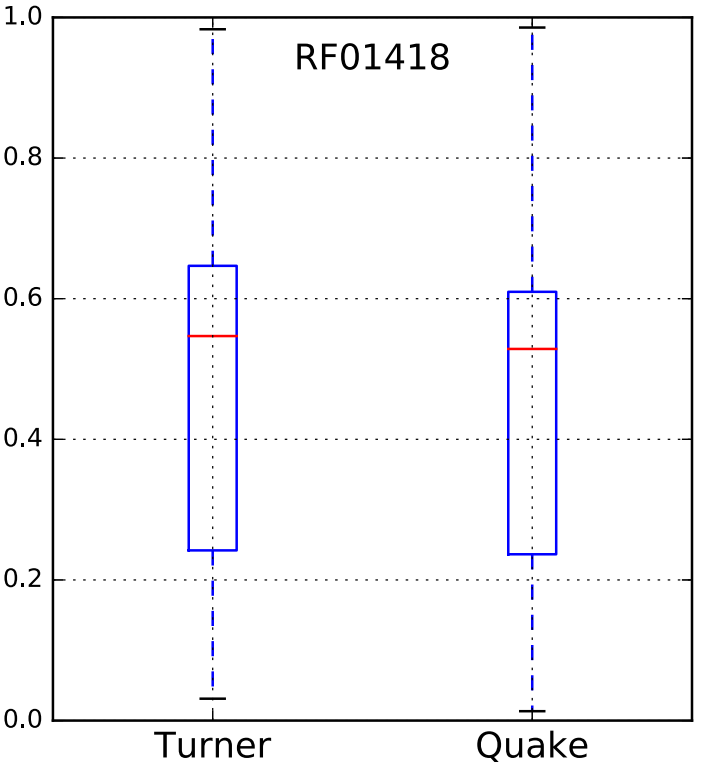
Basepair accuracy (=expected sensitivity)

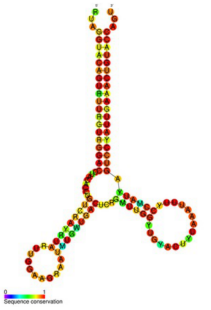




Localfold CisReg dataset

Basepair accuracy (=expected sensitivity)





Conclusion

- Well the established energy models seems to be leaned toward positive set of RNA strands, i.e. with nice boundaries
- Not considering a penalty for unpaired bases of Multiloops can result in favoring large multiloops and long base-pair interactions
- This yields into challenges for the local folding problem or probably in general for structure prediction of long RNA sequences

What can comes next?

- New Turner parameter set with less sensitivity to negative sequences?
- An implementation of folding algorithms supporting more exact Multiloop energy model?

Thanks for your attention!