DotCodeR: Alignment of RNA secondary structure dotplots

Jakob Hull Havgaard

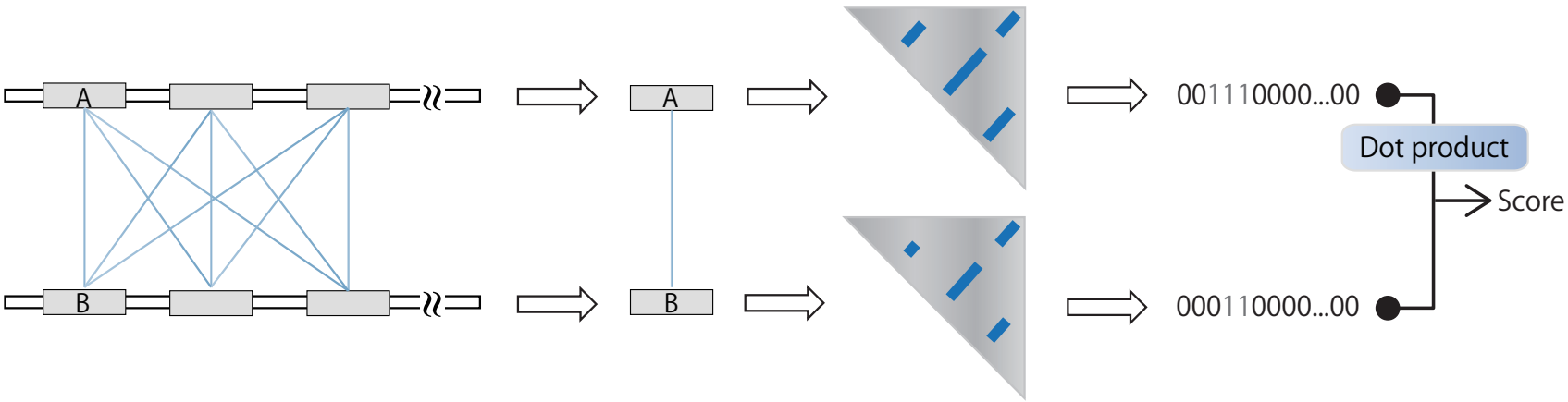
Objective:

- Make a pre-filter for comparative searches for structured RNA
- Scan chromosomal length sequences
- Throwing away unstructed regions and keeping potentially structured RNAs

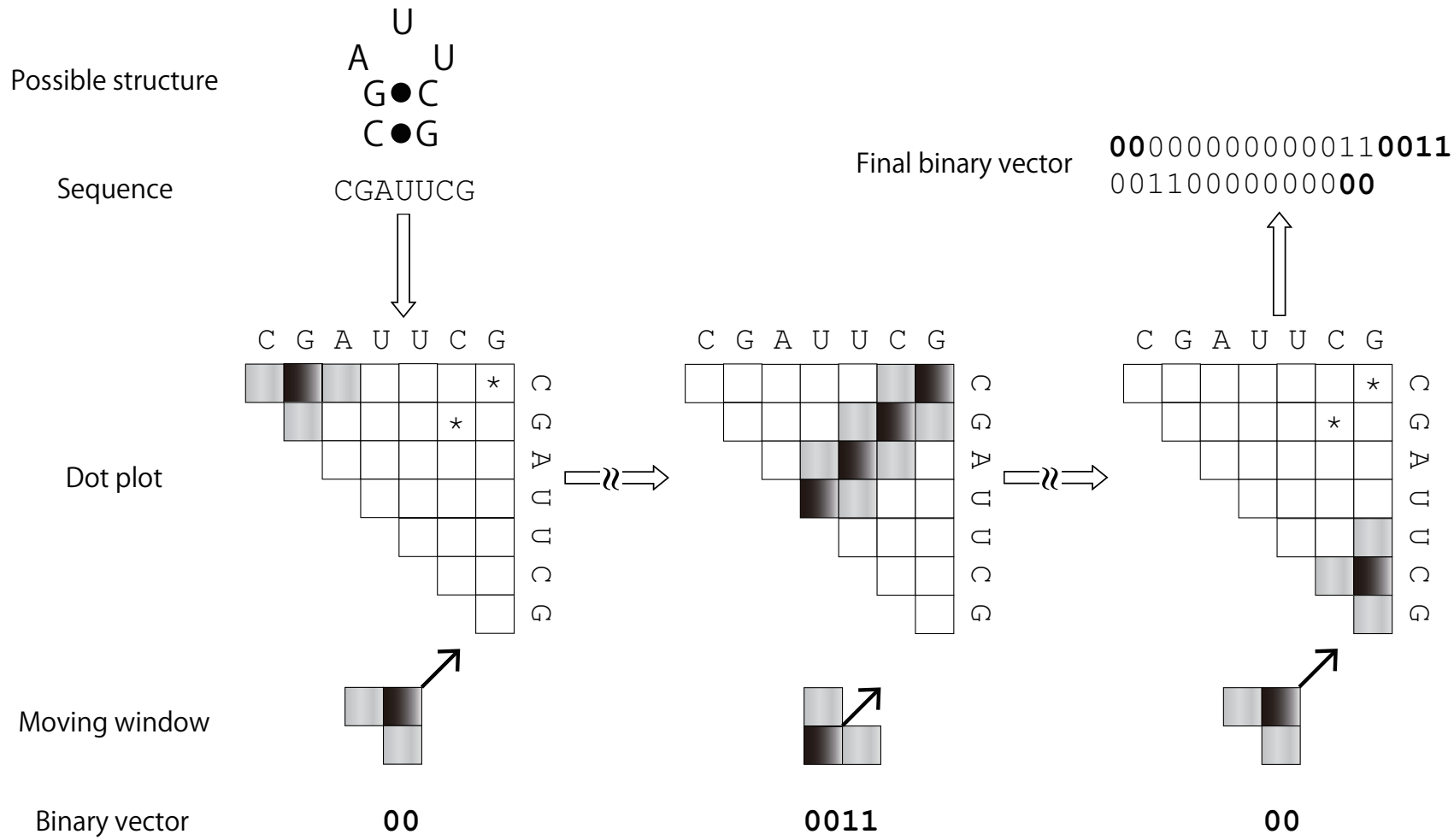
Idea:

- Quickly compare dot-plots by converting them it into binary vectors and calculating the dot product

Algorithmic overview



Algorithmic overview



Parameters

Parameters:

- d size of the neighborhood
- s window step size
- w window size
- c cutoff score

Data:

- Positive: Rfam 12.0[1] sequences
- Negative:
  - Gene: Shuffled Rfam sequences
  - Genomic: Shuffled genomic sequences

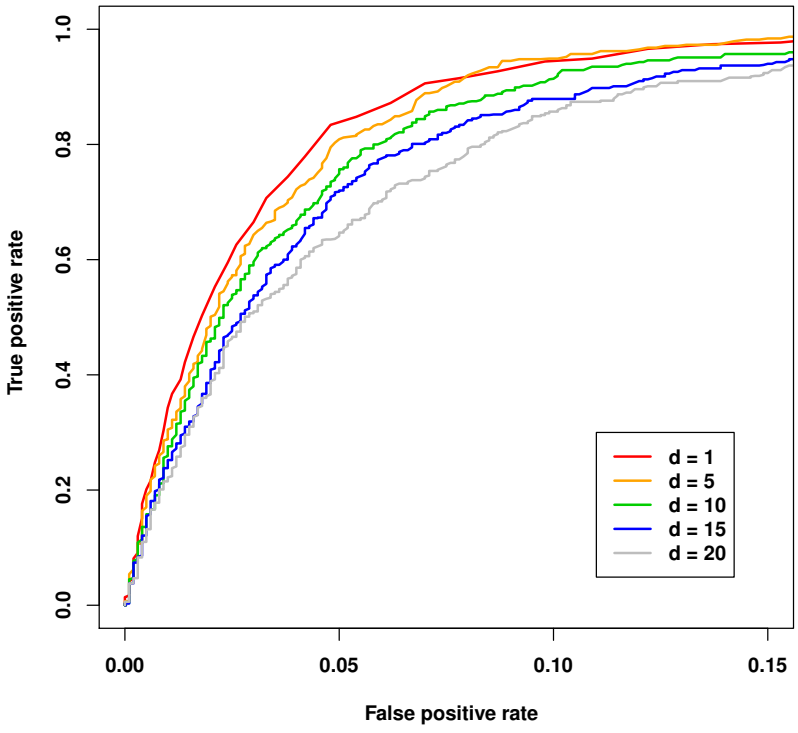
[1] Nawrocki et al. NAR 2015

RFAM dataset

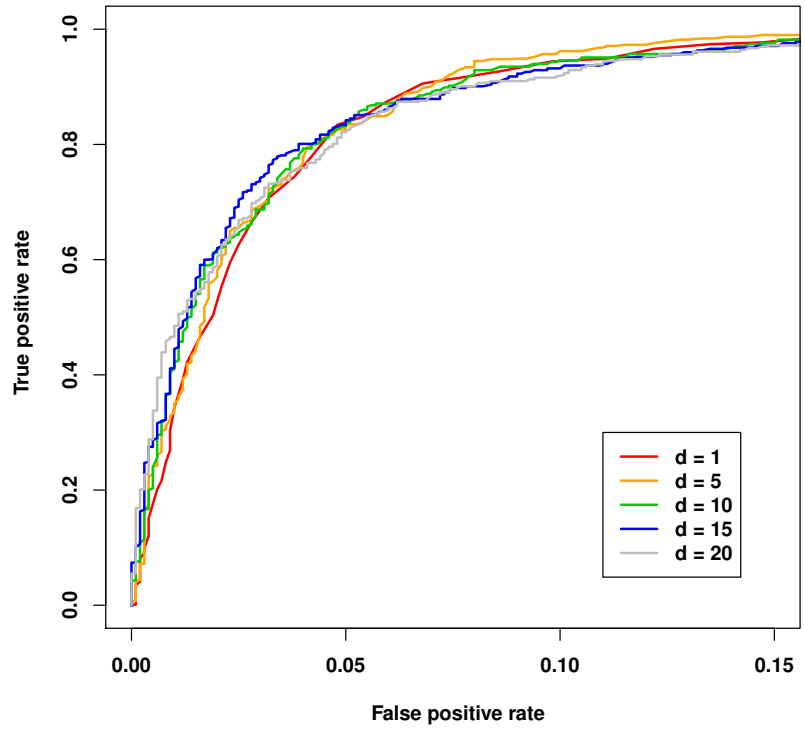
1. Remove sequences that have nucleotides in columns which are more than 80% gaps
2. Remove sequences that have more than 20% gaps
3. Remove families with less than 20 sequences
4. Redundancy reduce to at most 90% identiy
5. Ramdomly select five sequences from each clan
6. Split families into train or test set dependent on the first letter of the family name

d neighborhood size

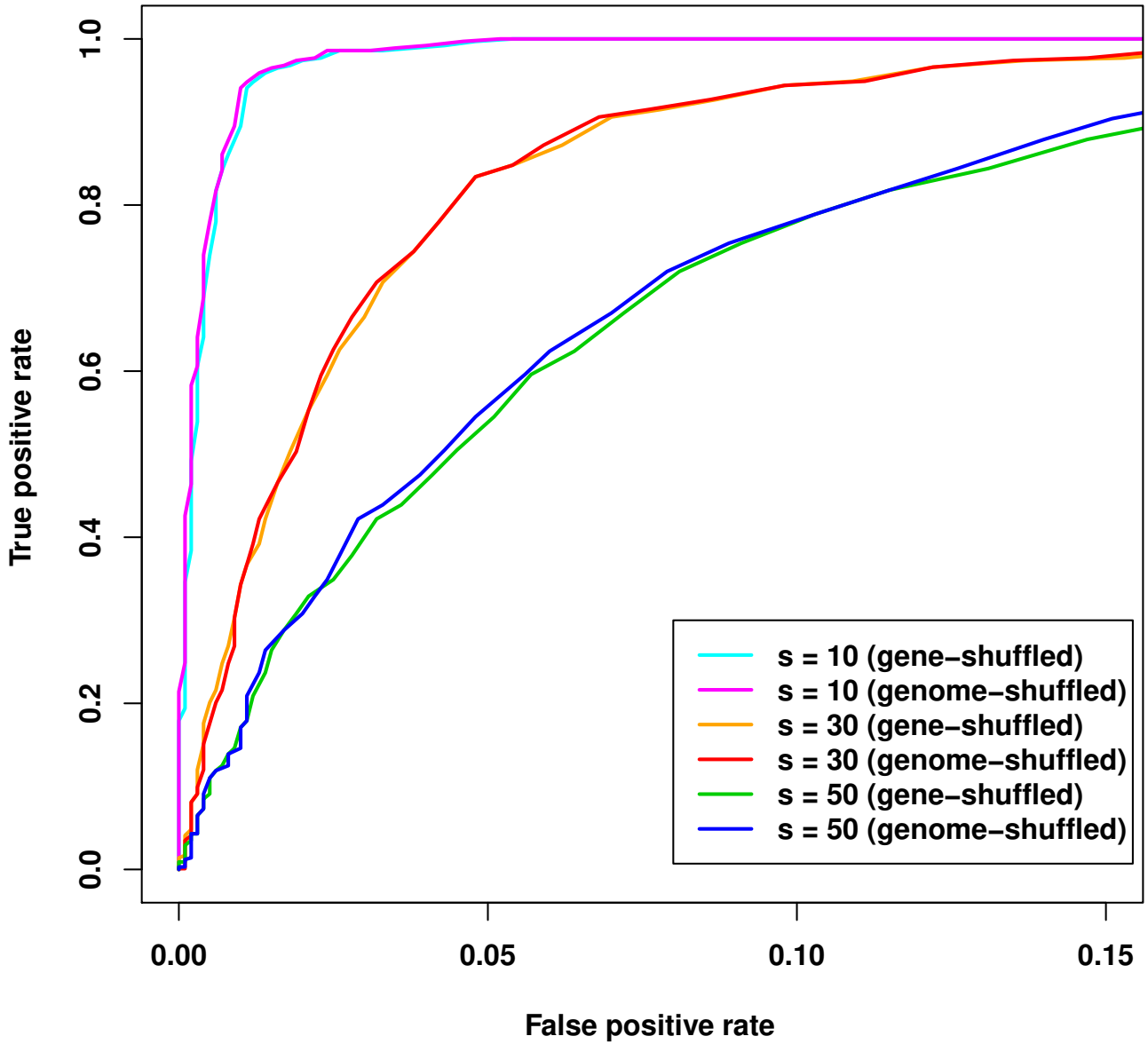
Gene-shuffled



Genomic-shuffled

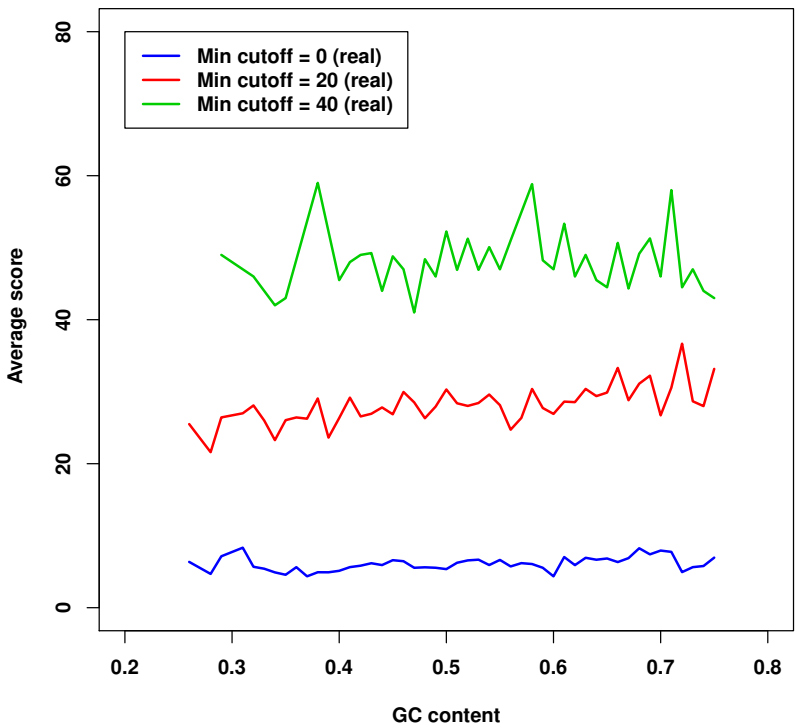


s step size

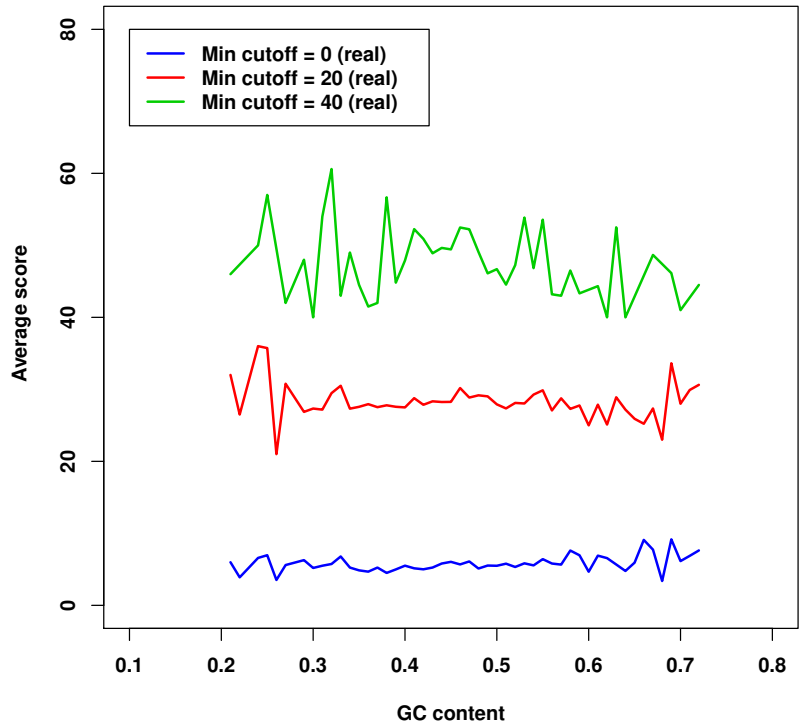

True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
False positive rate
s = 10 (gene−shuffled)
s = 10 (genome−shuffled)
s = 30 (gene−shuffled)
s = 30 (genome−shuffled)
s = 50 (gene−shuffled)
s = 50 (genome−shuffled)

Checking: Score as a function of GC contents

Training:

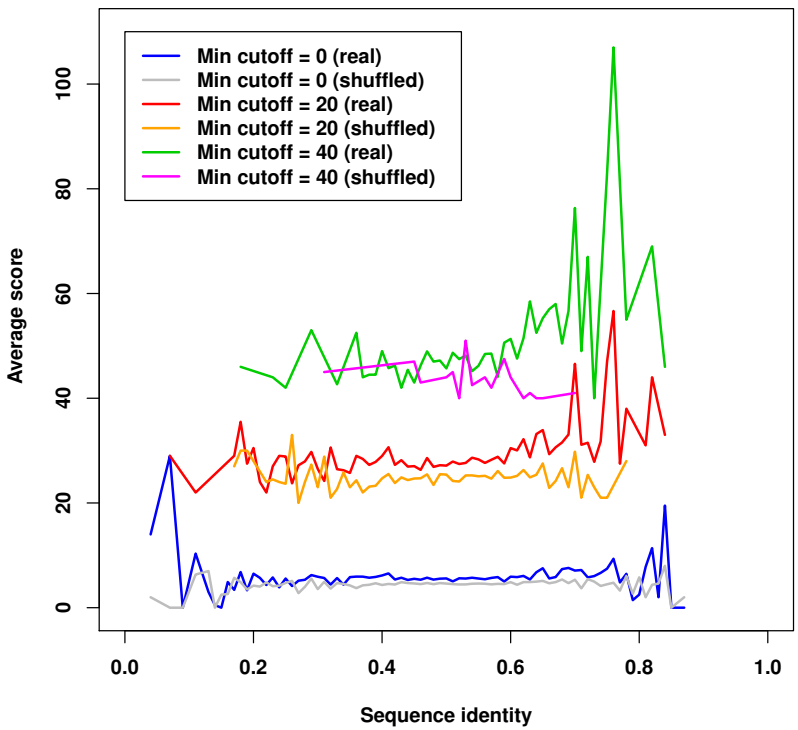


Test:

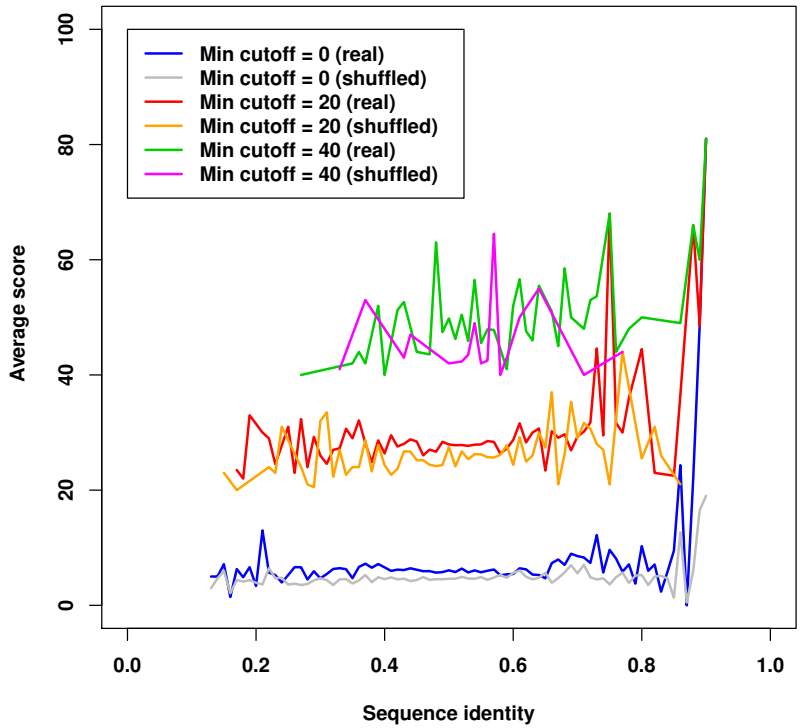


Checking: Score as a function of sequence identity

Training:



Test:

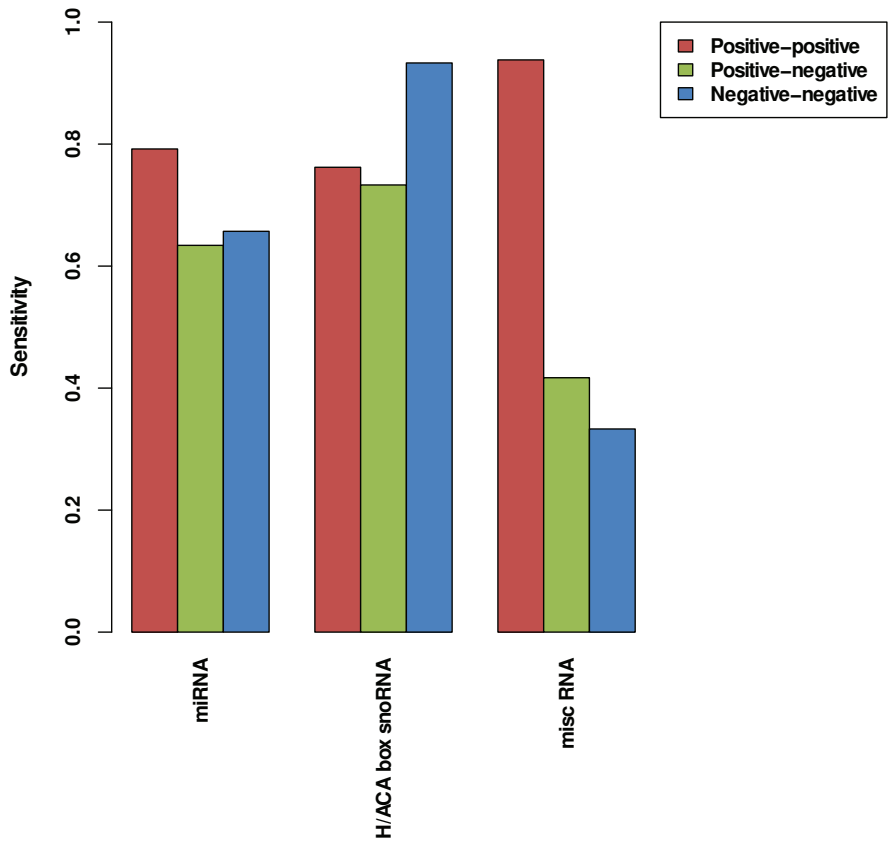


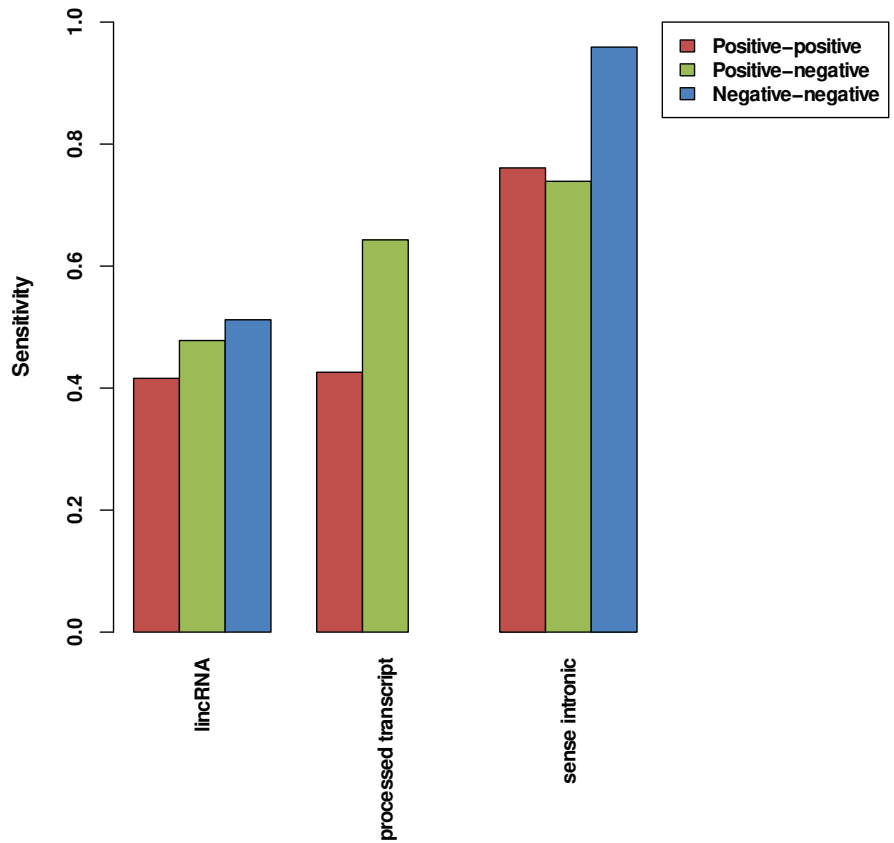
Testing with genomic sequences

- Hsa chr 21 vs Mmu chr 19
- Smallest human chromosome and the least syntenic mouse chromosome
- Remove repeats
- Remove aligned regions

Raw	Cleaned	Output	Reduction
3.18×10^{12}	2.30×10^{12}	6.75×10^{10}	97%

Structured RNAs in the chromosomal sequences





- Annotation from Ensembl[2]

[2] Cunningham et al. NAR 2015

Acknowledgement

- Yuki Kato (Osaka University)
- Jan Gorodkin (University of Copenhagen)

Kato, Y., Gorodkin, J. and Havgaard, JH.
Alignment-free comparative genomic screen for structured RNAs using coarse-grained secondary structure dot plots
BMC Genomics 2017 18:935 https://doi.org/10.1186/s12864-017-4309-y

Founding:

- JSPS KAKENHI
- Innovation fund Denmark
- Danish Center for Scientific Computing (DCSC, DeiC)