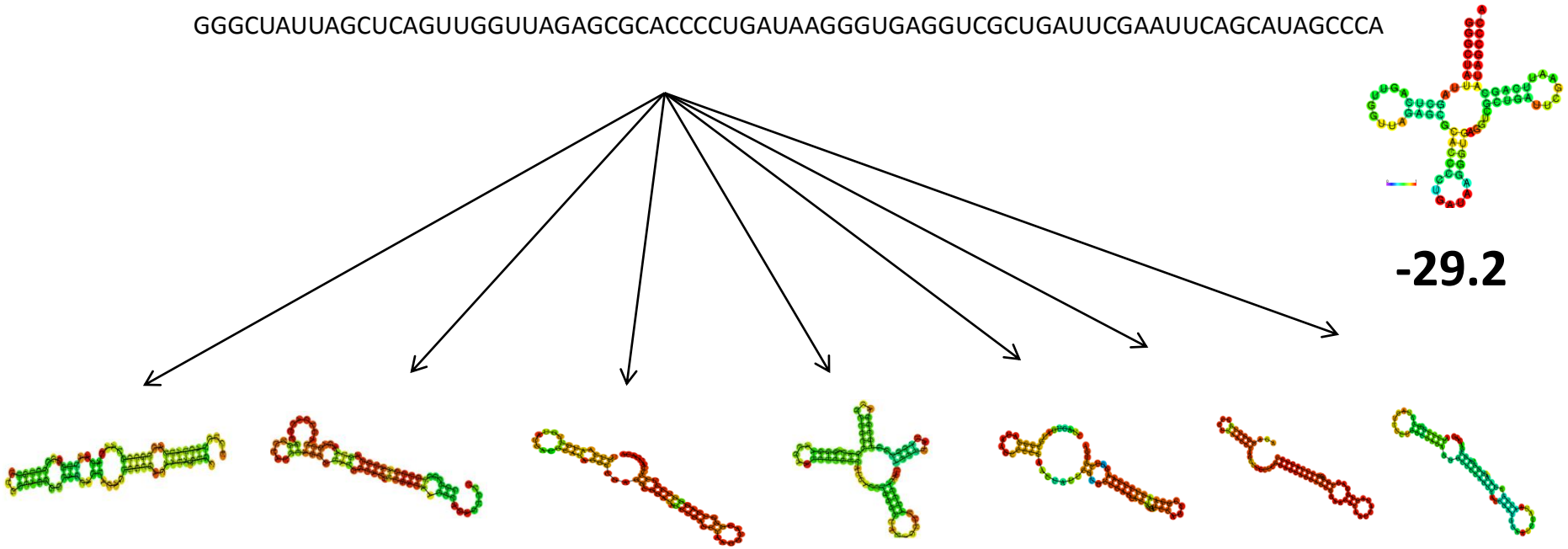
Detection of thermodynamically stable RNAs in long sequences

Ruslan Soldatov

Moscow State University

Thermodynamic stability of an RNA sequence

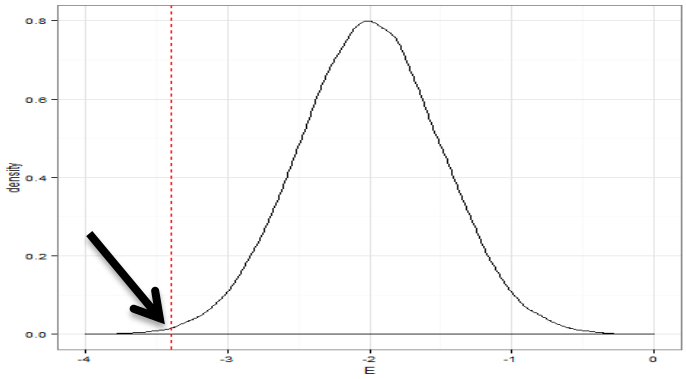
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA



-29.2

-17.5 **-17.2** **-19.7** **-18.4** **-17.9** **-21.3** **-17.9**

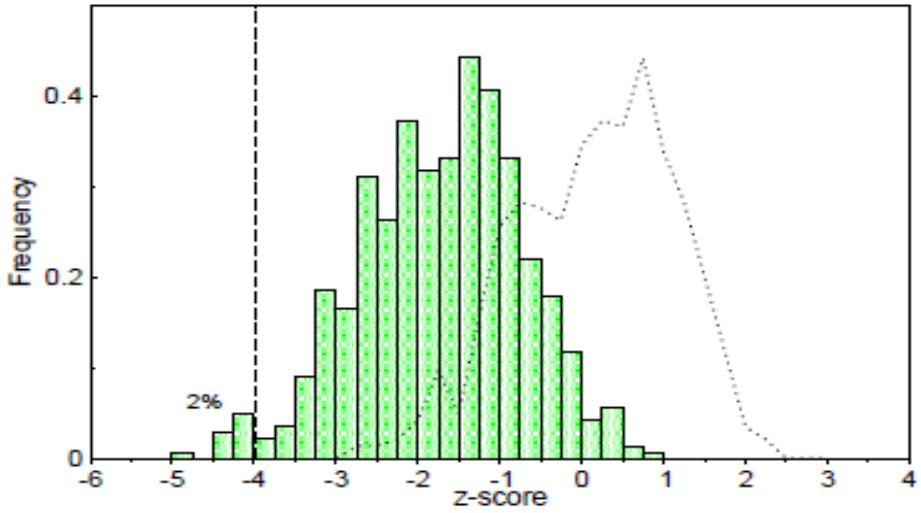
Thermodynamic stability of an RNA sequence



- length
- dinucleotide content

$$Z = (E - \mu) / \sigma$$

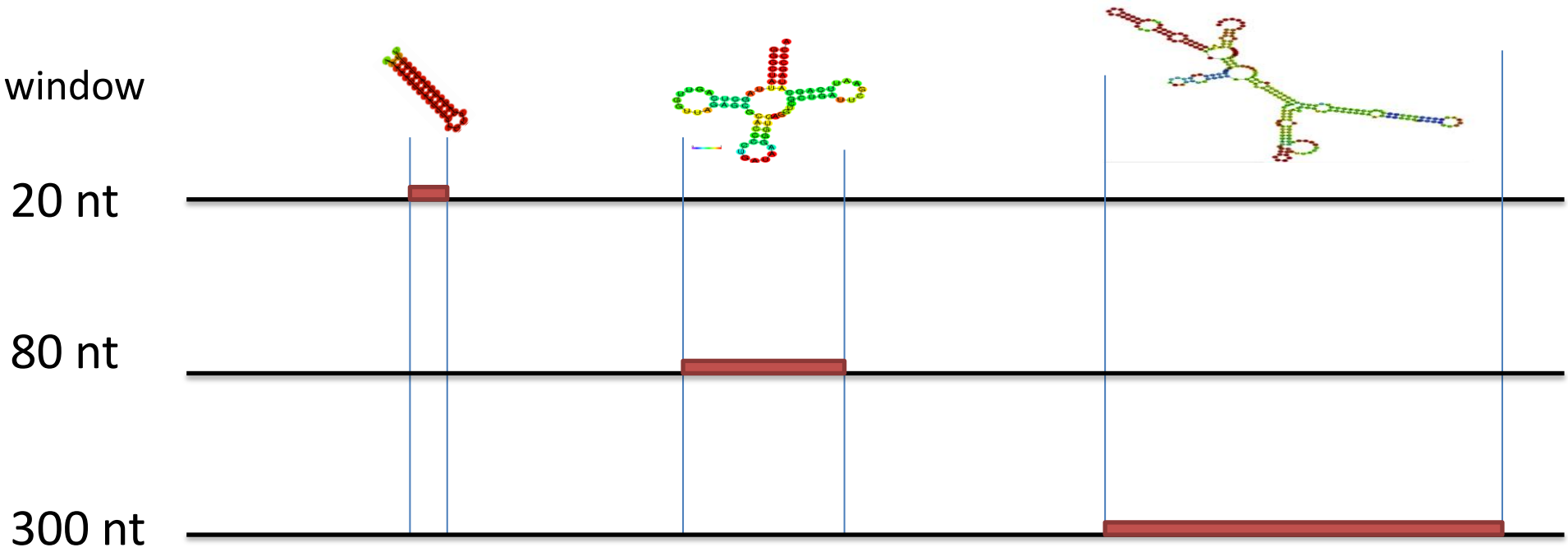
Thermodynamic stability of functional non-coding RNAs



ncRNA Type	No. of Seqs.	Mean z-score
tRNA	579	−1.84
5S rRNA	606	−1.62
Hammerhead ribozyme III	251	−3.08
Group II catalytic intron	116	−3.88
SRP RNA	73	−3.37
U5 spliceosomal RNA	199	−2.73

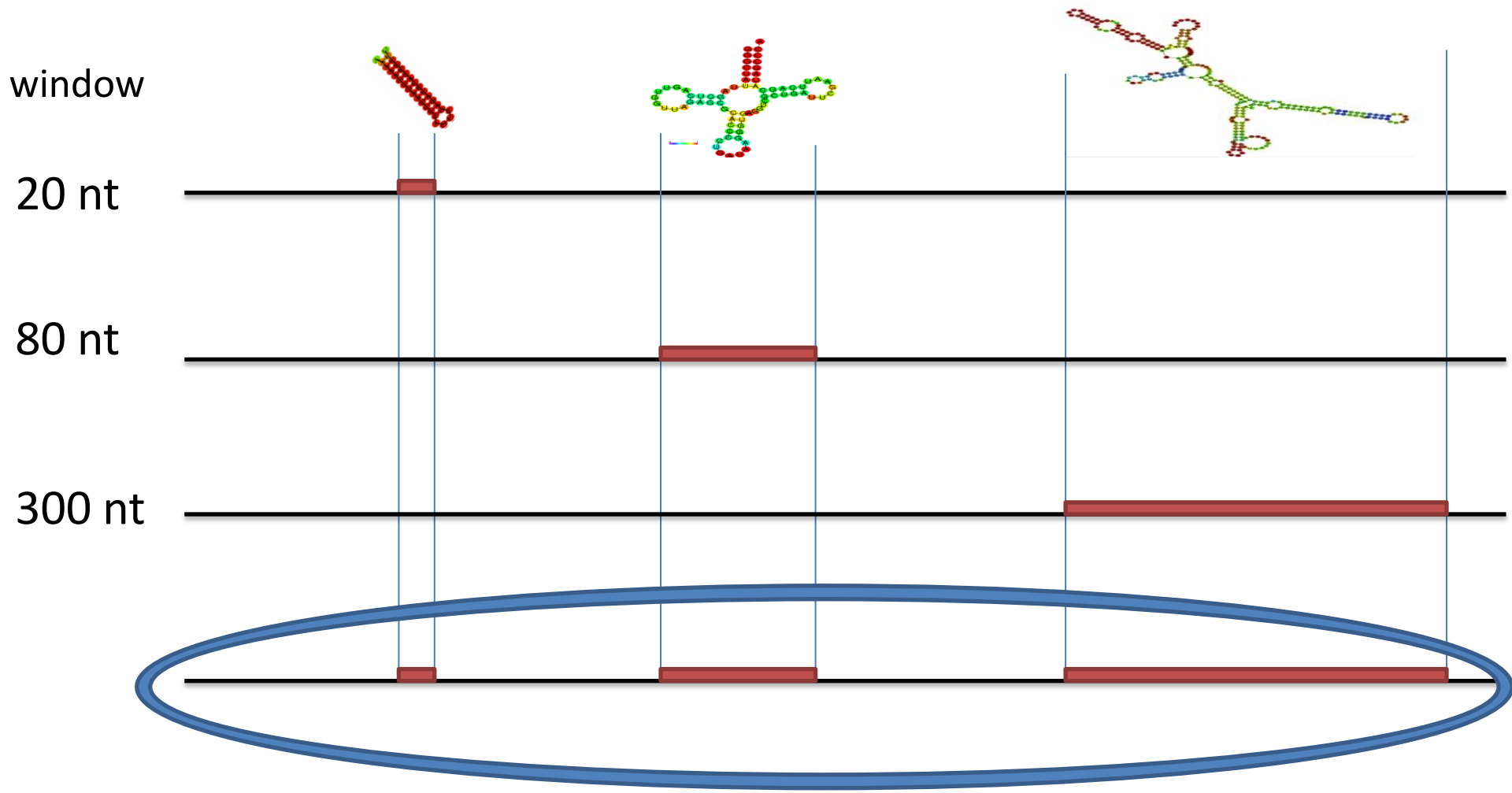
Washietl & Hofacker, *J. Mol. Biol.* (2004) 342:19

Detection of segments with low Z-score in sliding window



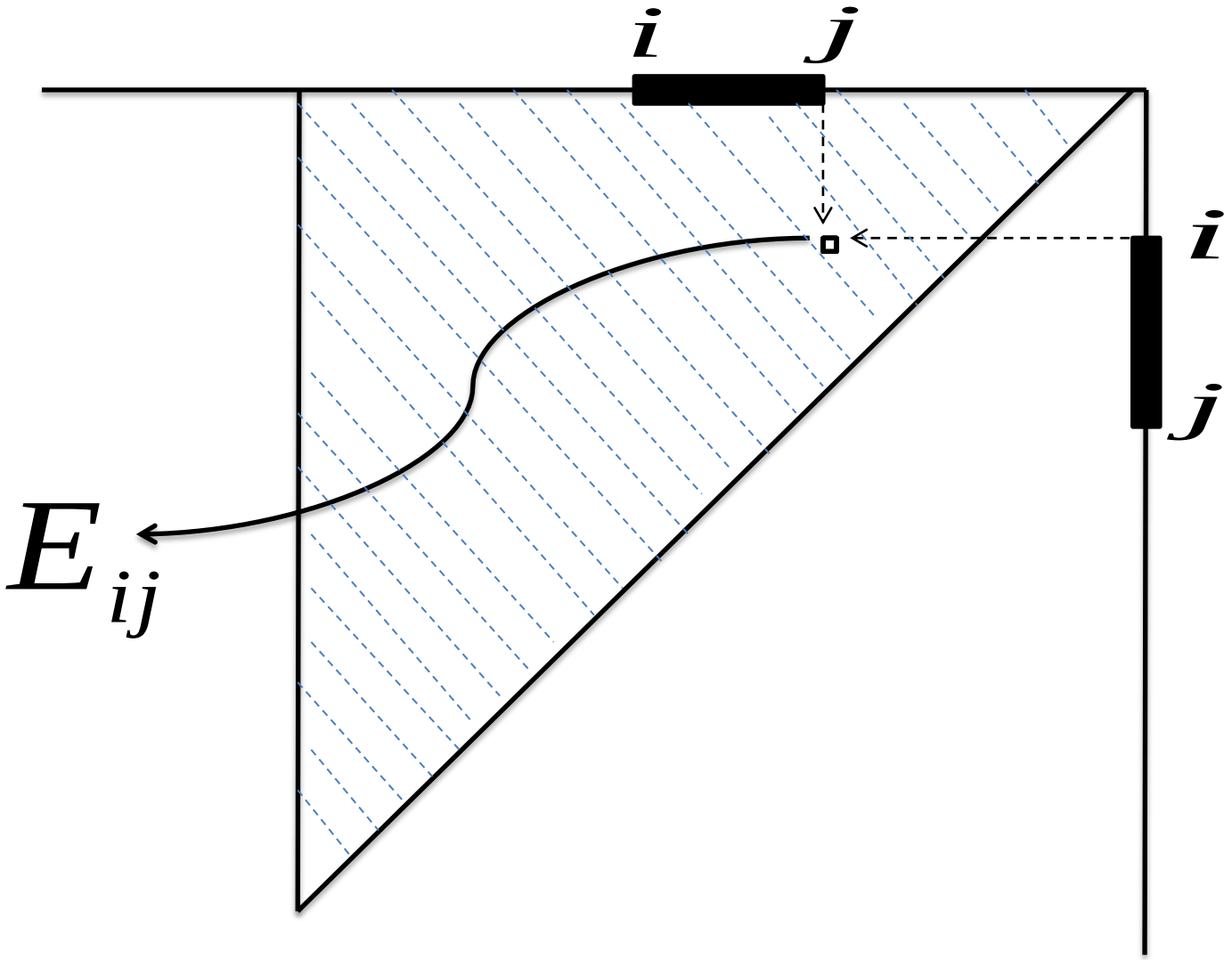
- RNAs have different complexities of a structure and sizes
- Detection is sensitive to the length of scanning window
- Combination of multiple windows is time-consuming and requires substantial post-processing

Detection of segments with low Z-score locally-optimal



Property of MFE matrix

- Calculate MFEs for each subsequence
- Use RNASlider with speed up techniques (sliding MFE recalculation, sparsification)

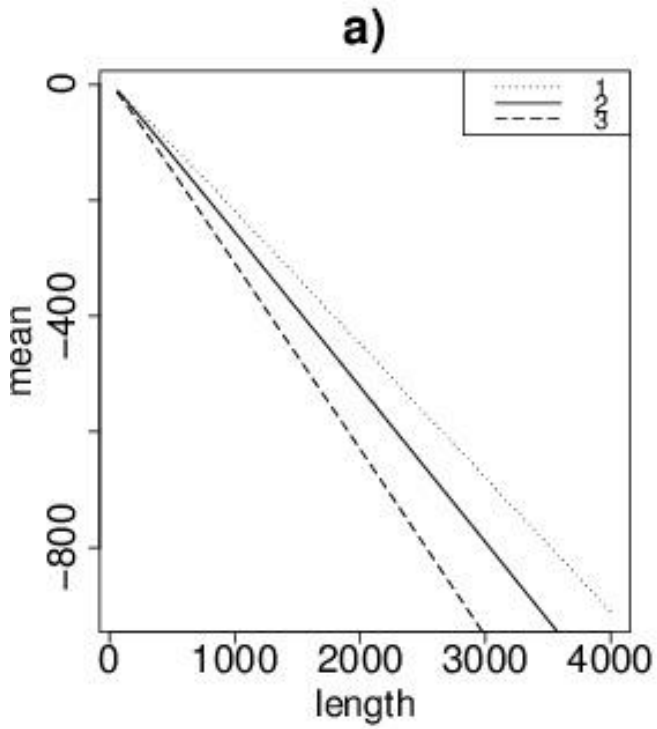


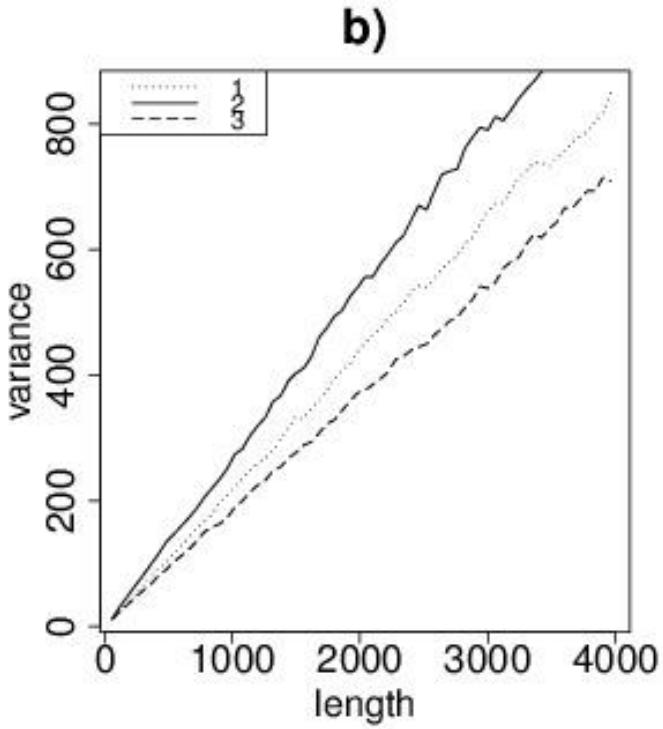
From MFE matrix to Z-score matrix

$$Z_{ij} = (E_{ij} - \mu_{ij}) / \sigma_{ij}$$

- Dependence on sequence length

$$\mu_{ij} = \mu(j - i + 1) \qquad \sigma^2{}_{ij} = \sigma^2(j - i + 1)$$





Dinucleotide content, %

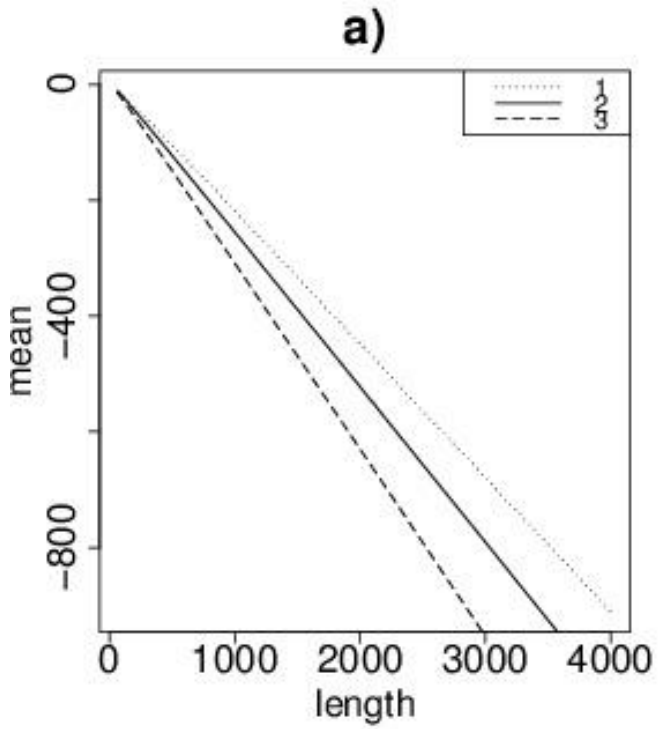
	1	2	3
aa	5.5	11	6.6
at	7.4	1.4	8.2
ag	5.5	11	1.6
ac	8.9	6.9	8.2
ta	6	5.5	1.6
tt	5.5	4.1	4.9
tg	6.4	2.7	9.8
tc	7.8	4.1	8.2
ga	9.2	8.2	14.8
gt	5.5	8.2	9.8
gg	3.3	11	6.6
gc	3.2	4.1	3.3
ca	6.4	5.5	3.3
ct	7.4	2.7	1.6
cg	6	6.9	9.9
cc	6	6.9	1.6

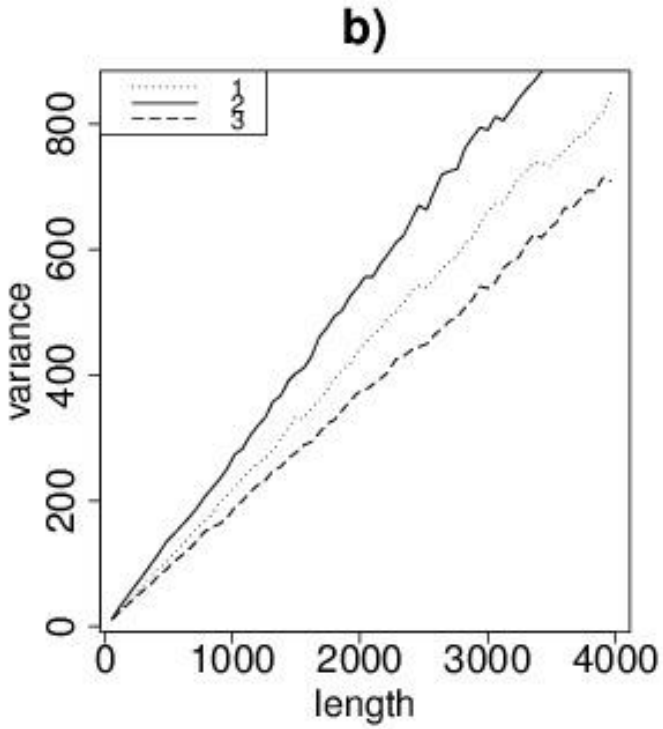
From MFE matrix to Z-score matrix

$$Z_{ij} = (E_{ij} - \mu_{ij}) / \sigma_{ij}$$

- Dependence on sequence length

$$\mu_{ij} = \mu(j - i + 1) \qquad \sigma^2{}_{ij} = \sigma^2(j - i + 1)$$





Dinucleotide content, %

	1	2	3
aa	5.5	11	6.6
at	7.4	1.4	8.2
ag	5.5	11	1.6
ac	8.9	6.9	8.2
ta	6	5.5	1.6
tt	5.5	4.1	4.9
tg	6.4	2.7	9.8
tc	7.8	4.1	8.2
ga	9.2	8.2	14.8
gt	5.5	8.2	9.8
gg	3.3	11	6.6
gc	3.2	4.1	3.3
ca	6.4	5.5	3.3
ct	7.4	2.7	1.6
cg	6	6.9	9.9
cc	6	6.9	1.6

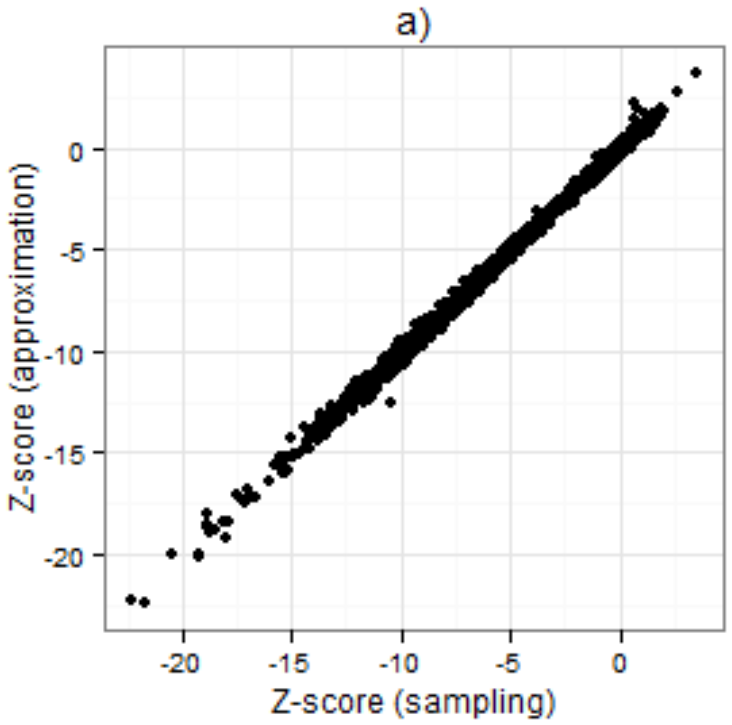
- Dependence on dinucleotide content

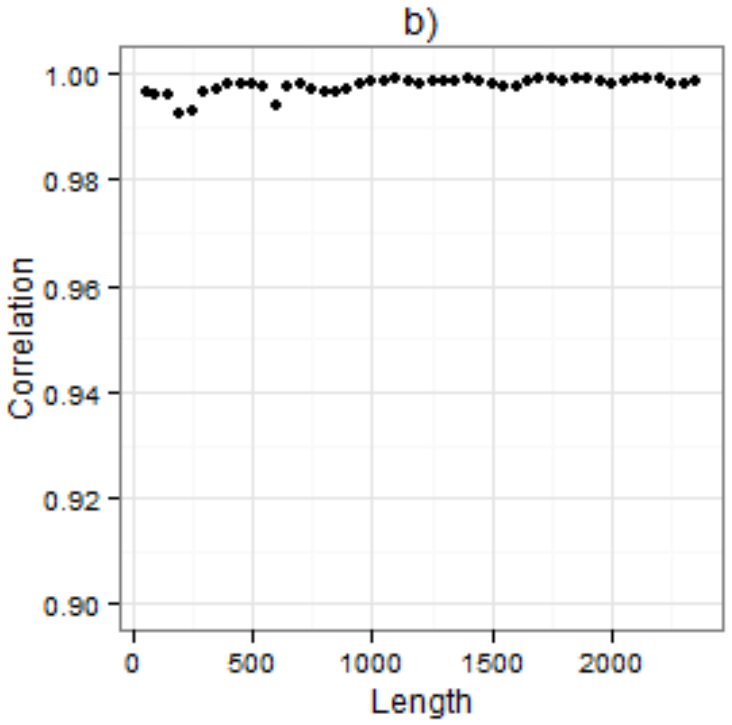
$$\mu(f_1, \ldots, f_{16} \mid l) = \sum_{1 \le k \le l \le 16} a_{kl} f_k f_l$$

From MFE matrix to Z-score matrix

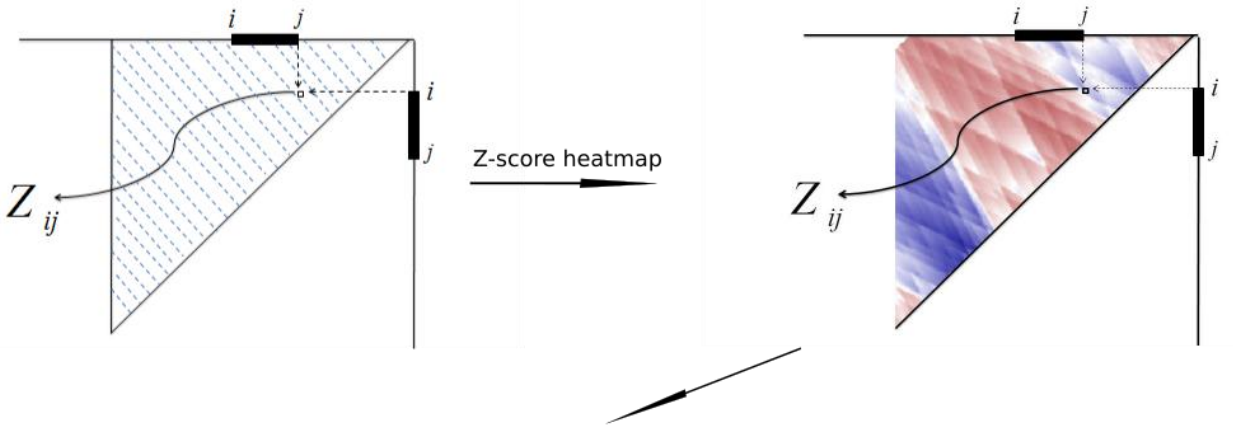
- Estimation of regression parameters
  - 27 quadratic regressions were fitted for each selected length
  - 20’000 learning parameters were used to estimate parameters of each quadratic regression

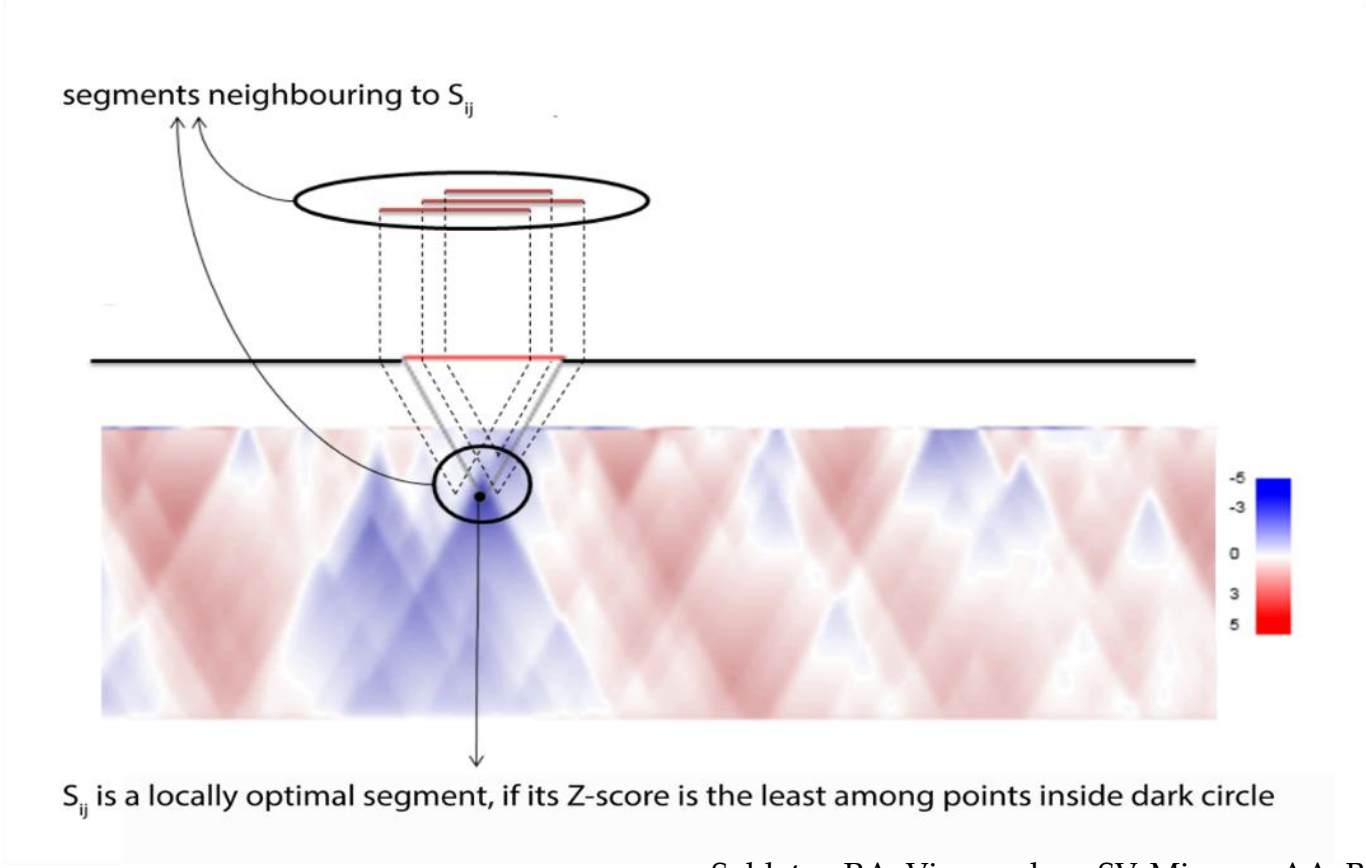
- High quality of approximation





RNASurface

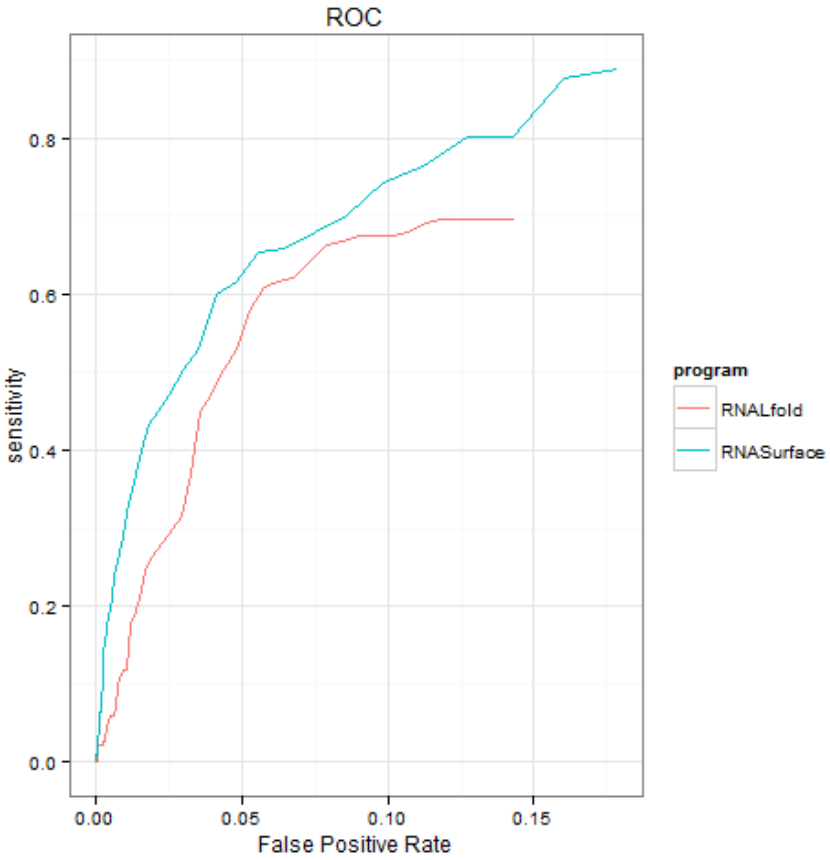




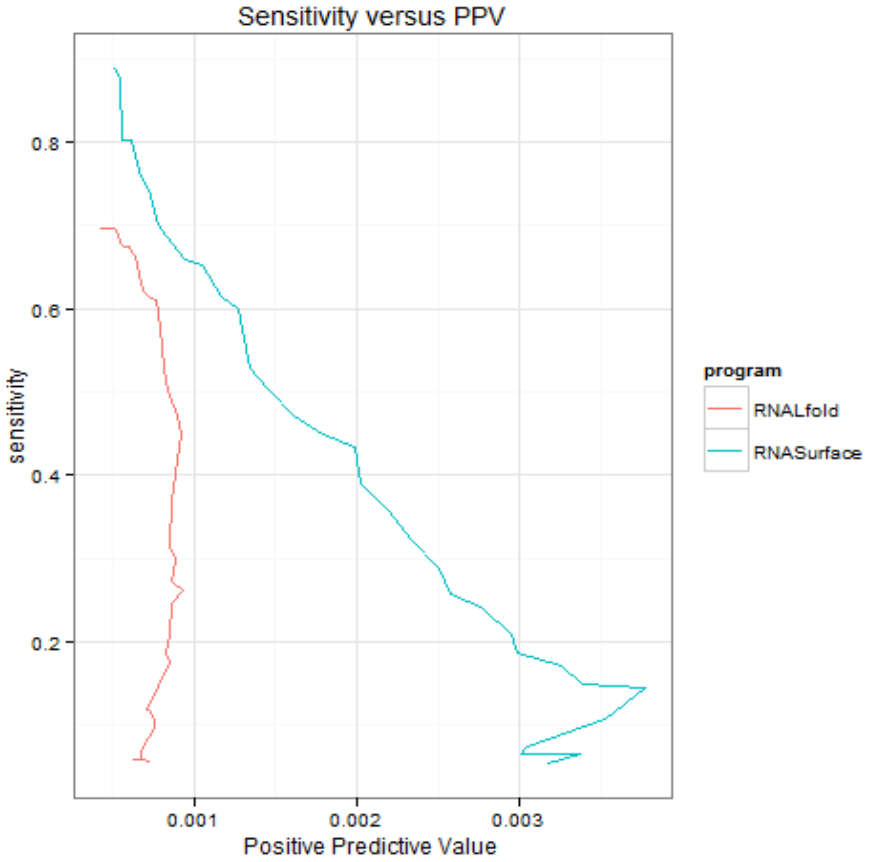
Soldatov RA, Vinogradova SV, Mironov AA. Bioinformatics. 2014

Benchmark using *Bacillus subtilis*

ROC curve



Sensitivity versus PPV



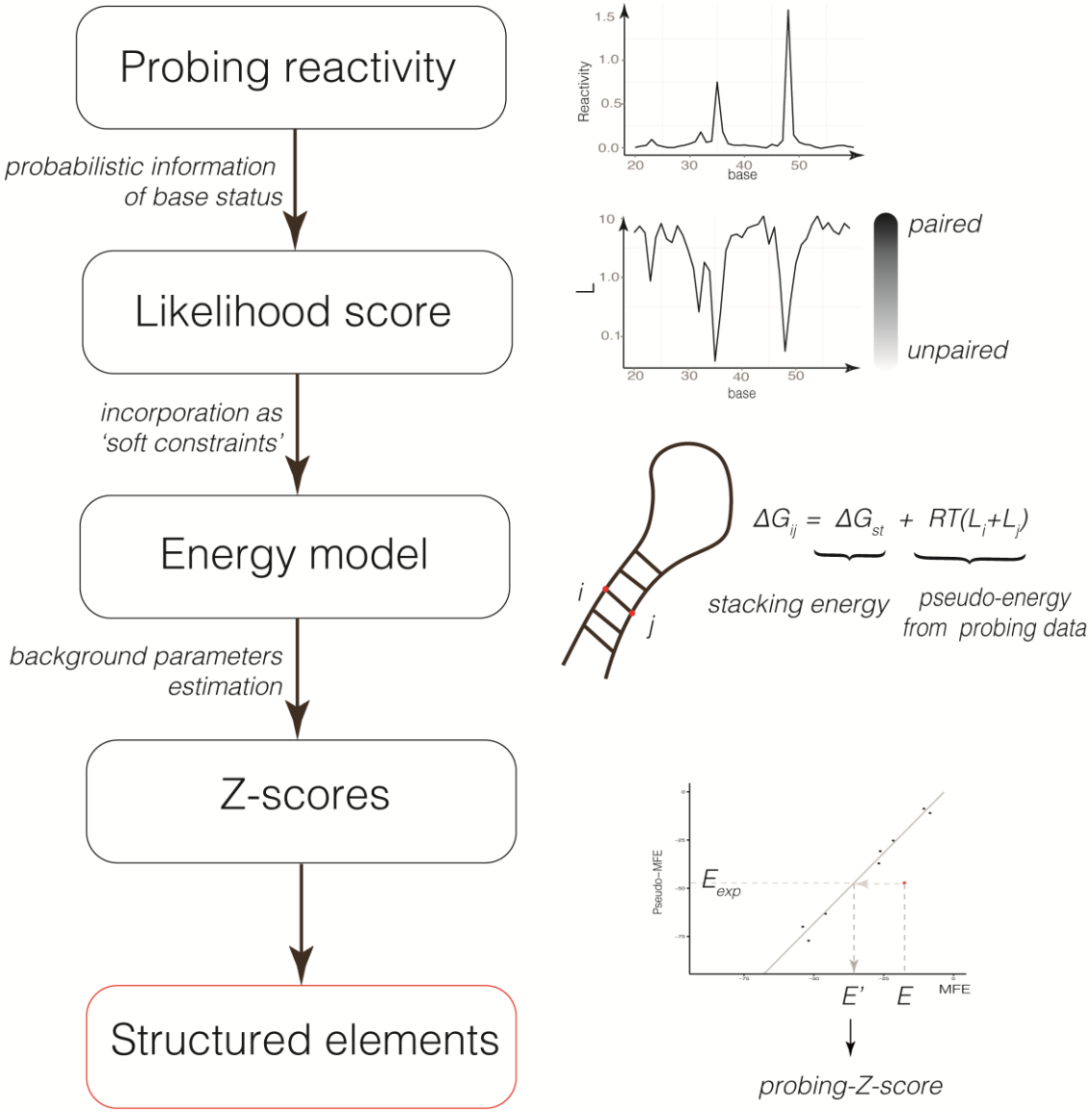
Applications

- Preprocessing in detection of functional structured RNAs
- Large-scale correlations with other genomic tracks (e.g. cds boundaries, ribosome profiling, RNA-seq etc)

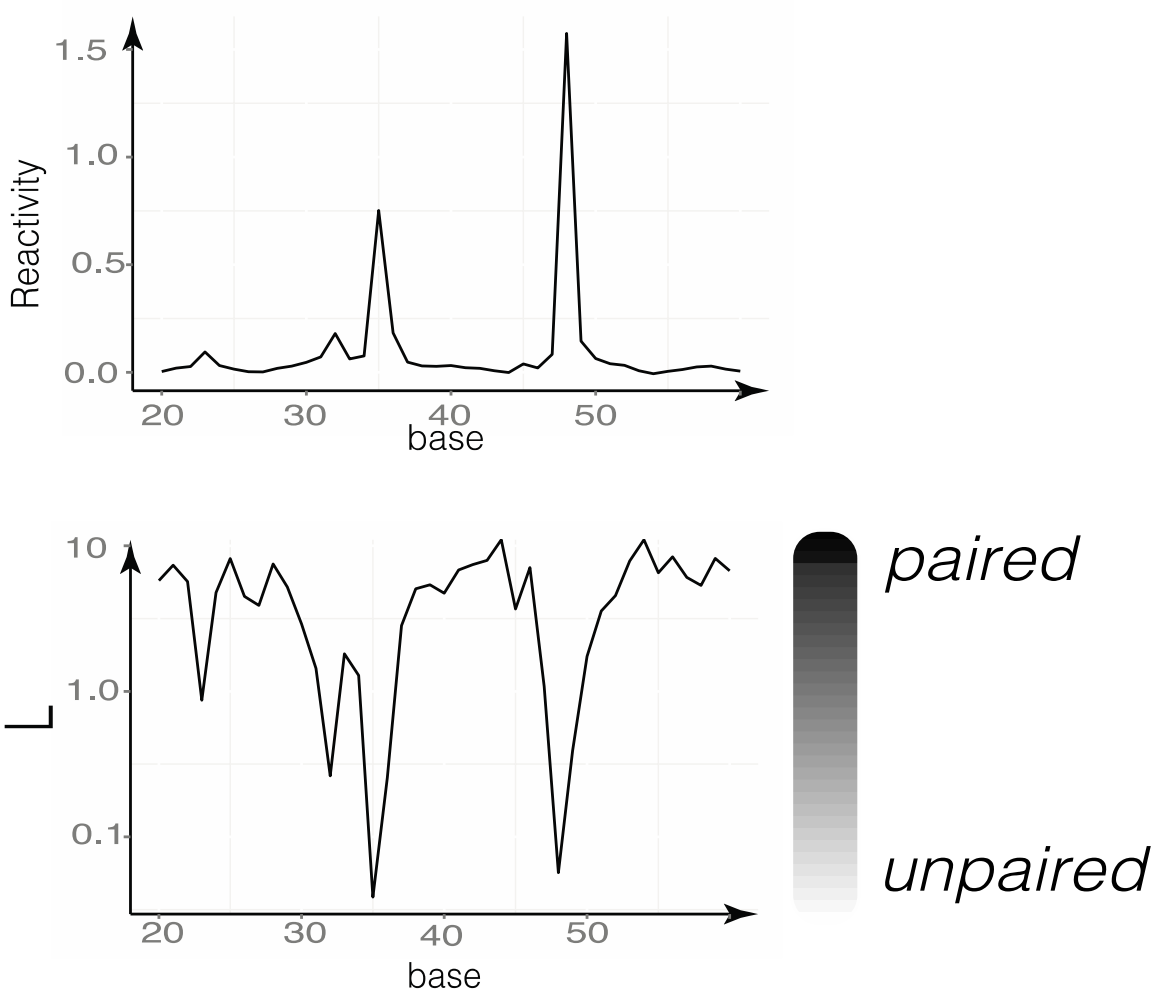
RNASurface
+
Probing data

- Probing data increases quality of the RNA secondary structure prediction
- Whether and how probing data contributes to the **detection** of structured RNAs?

Outline of the approach

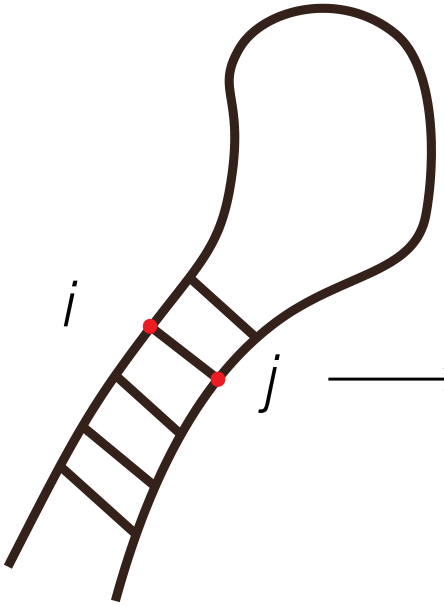


From reactivity to likelihood



Reactivity distribution of paired/unpaired bases is inferred from high-confidence nucleotides according to partition function

Energy model

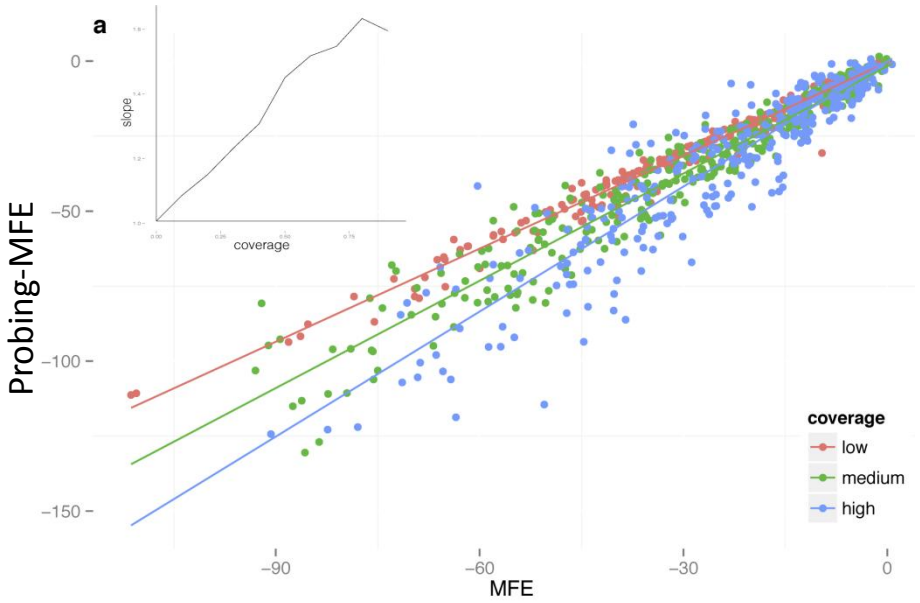


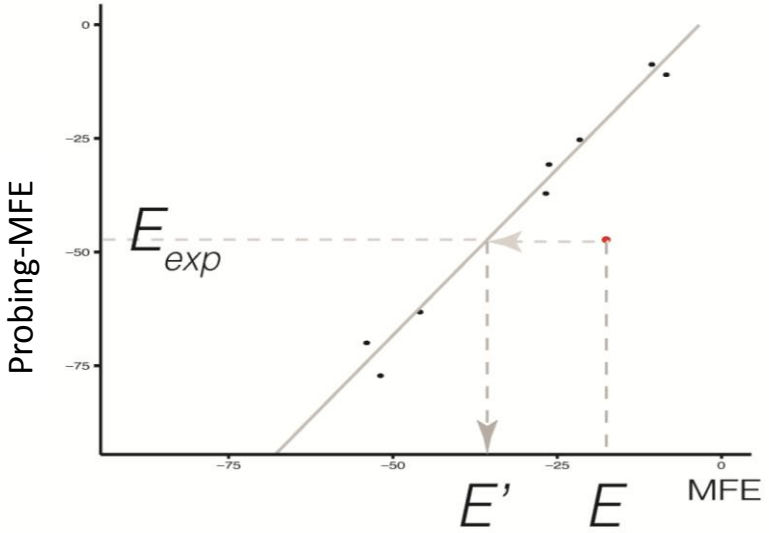
$$\Delta G_{ij} = \overbrace{\Delta G_{st}}^{\text{free energy}} + \overbrace{RT(L_i+L_j)}^{\text{pseudo-energy from probing data}}$$

L_i is the probing log-likelihood of being paired for position i in the RNA sequence

How to estimate background of probing-directed MFE?

mRNAs as a set of sequences with low fraction of functional secondary structures





$$Z = (E^{'} - \mu) / \sigma \quad \Leftarrow$$

Effect of probing data

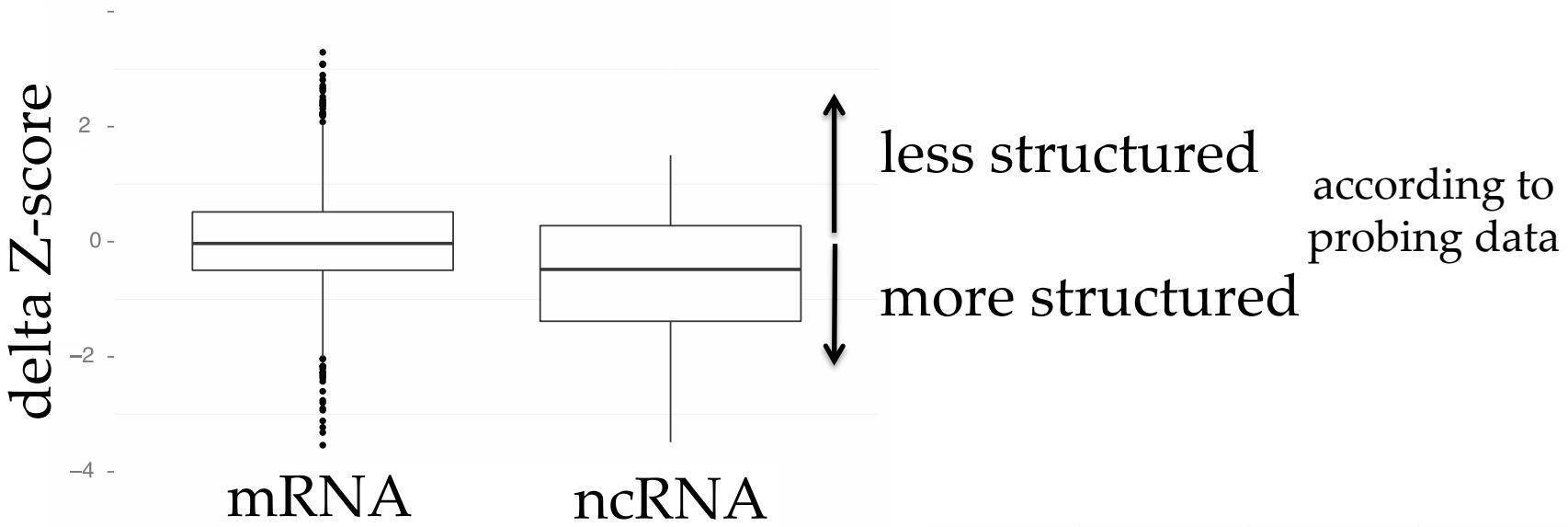
$E^{'} > E$ destabilizing

$E^{'} \approx E$ background

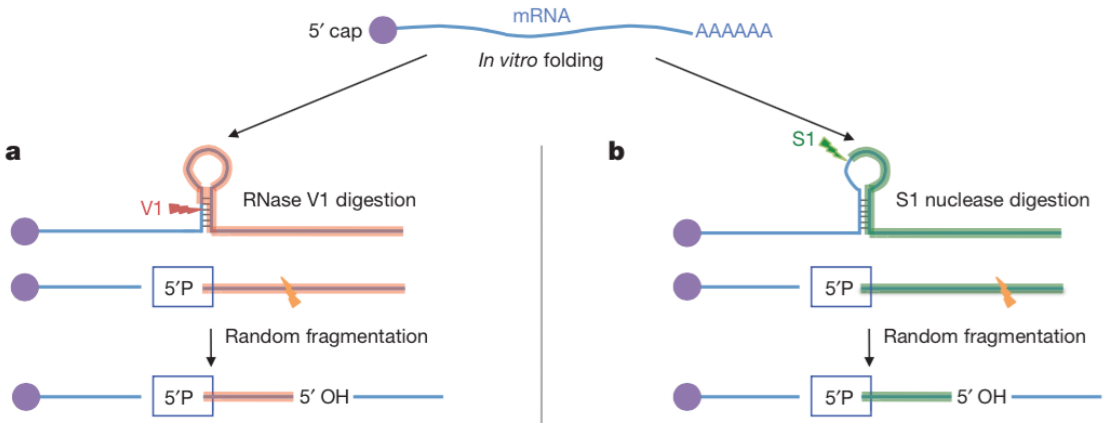
$E^{'} < E$ stabilizing

Probing-directed Z-score of mRNAs and ncRNAs

delta Z-score = Z-score - Probing Z-score



Transcriptome-wide screen with PARS data



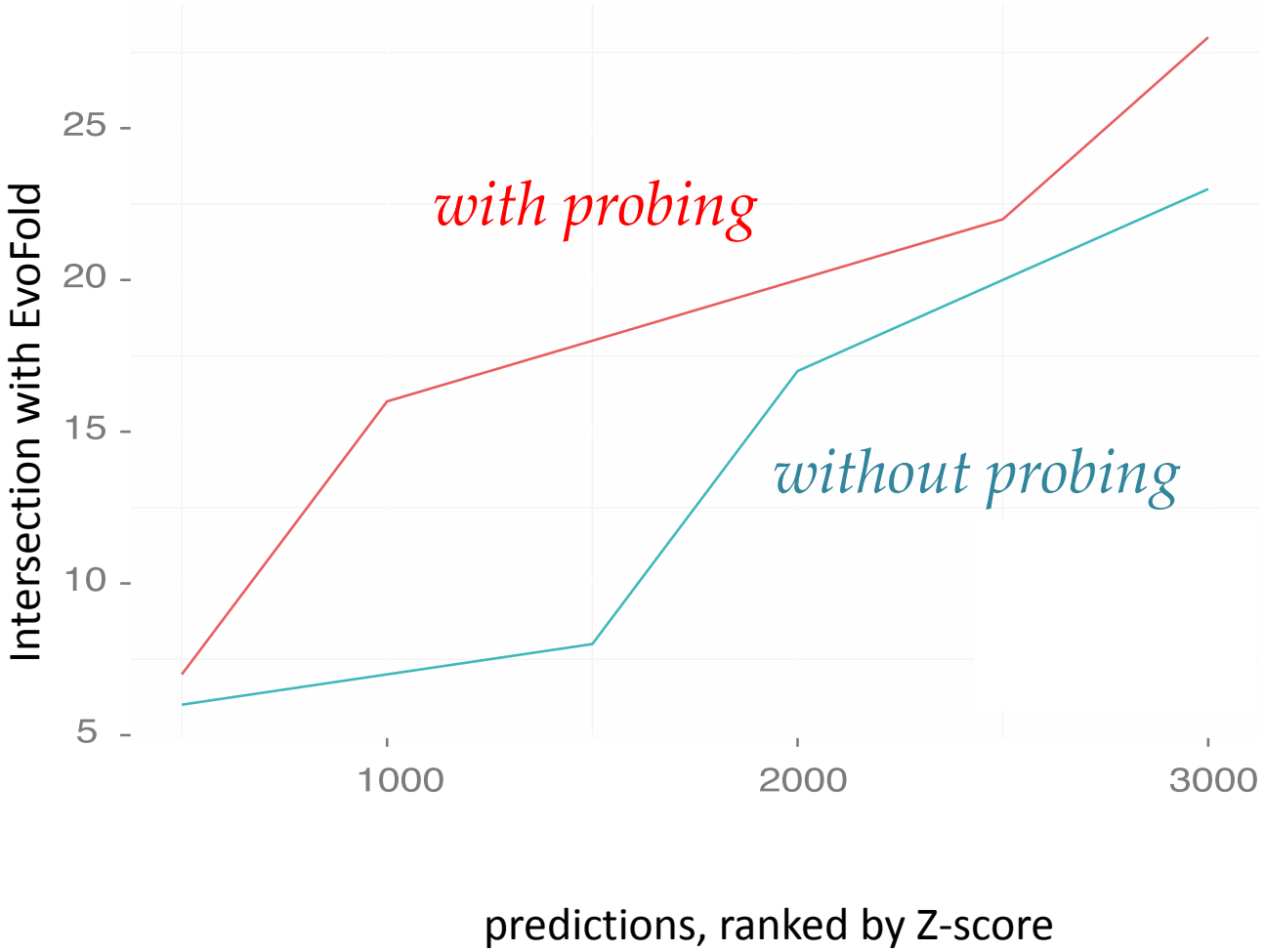
Wan Y et al. Nature. 2014

Two runs:

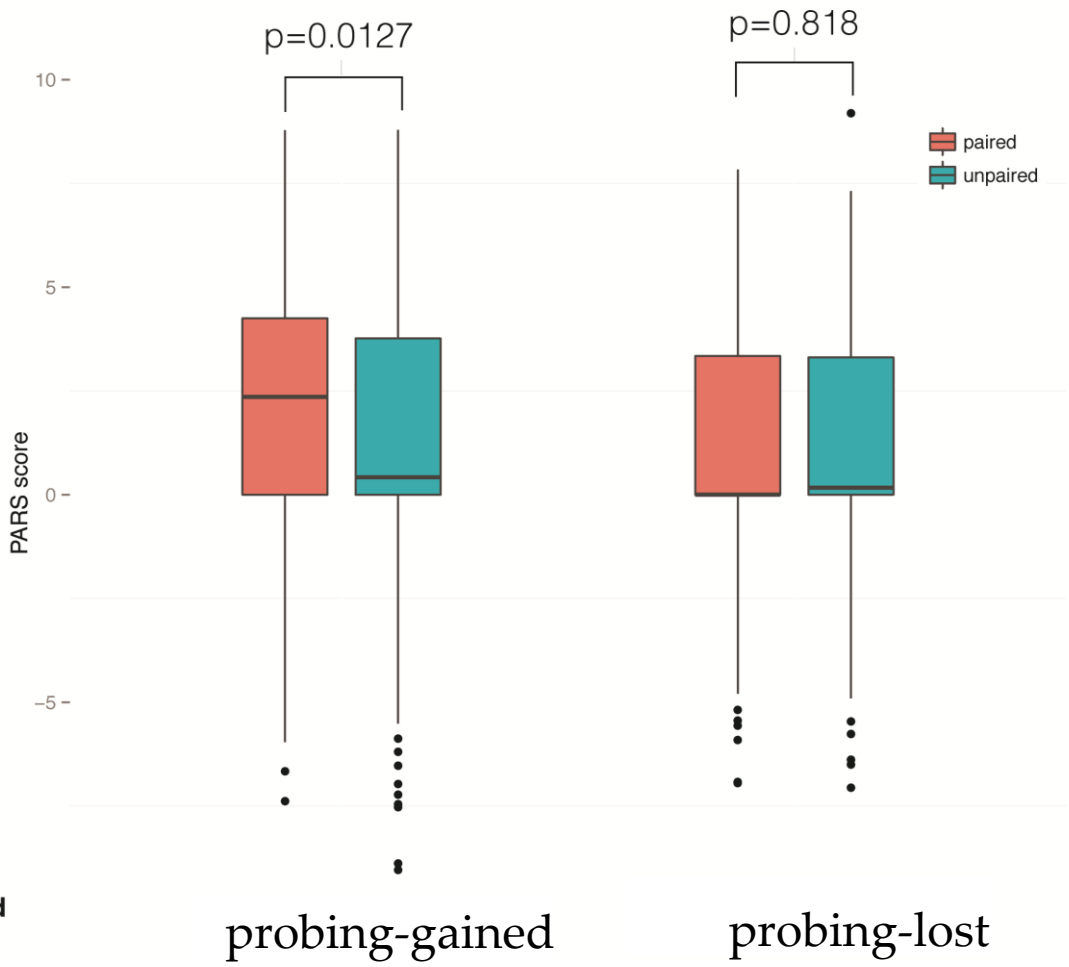
- 3587 elements in probing-constrained RNASurface run
- 3201 elements in RNASurface run

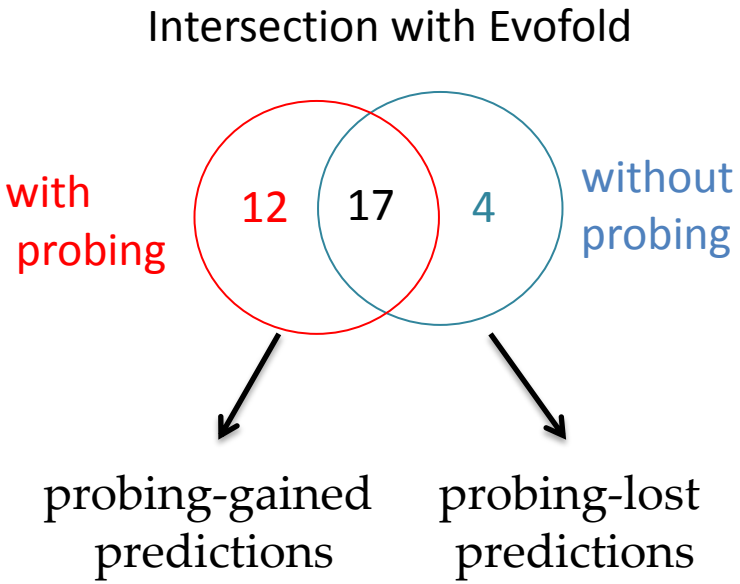
Z-score < -3

Results with/without probing data are compared with Evofold prediction



Consistency of probing data with evolutionary conserved RNA secondary structures



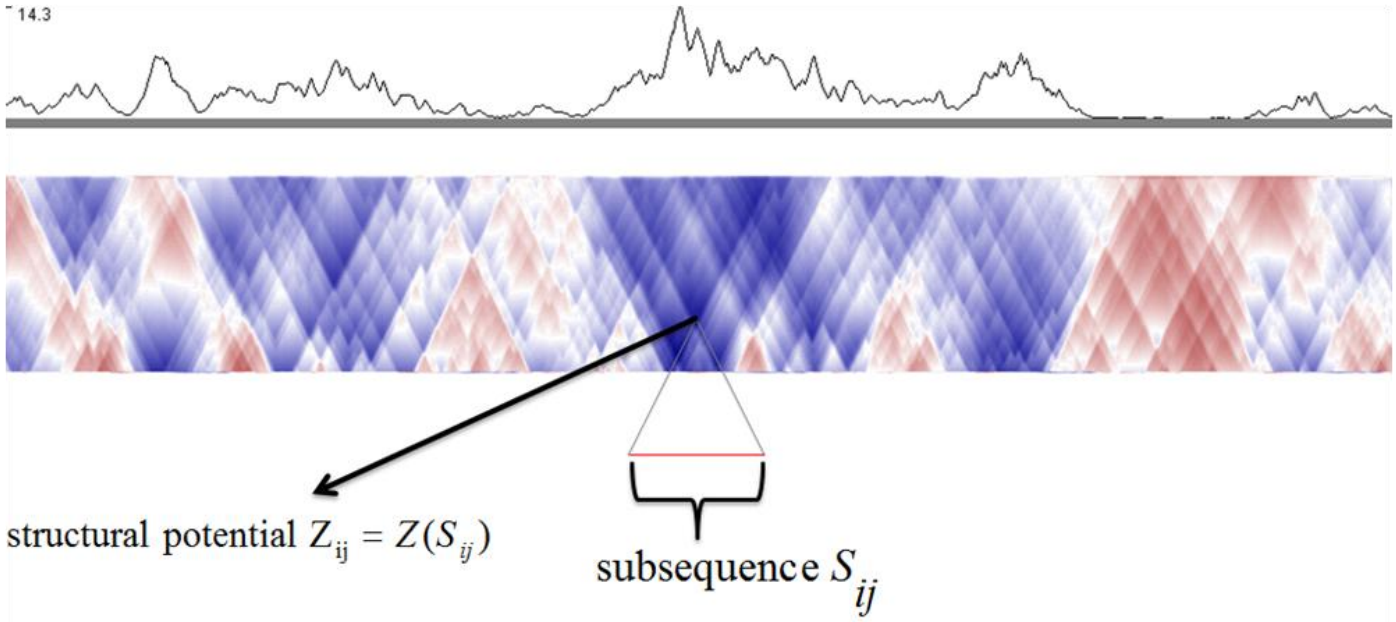


Conclusions

- Program RNASurface using a set of regressions efficiently detects locally-optimal segments with low Z-score in long sequences

- Integration of RNA probing data with RNASurface allows increased prediction quality
- Web-server http://bioinf.fbb.msu.ru/RNASurface/

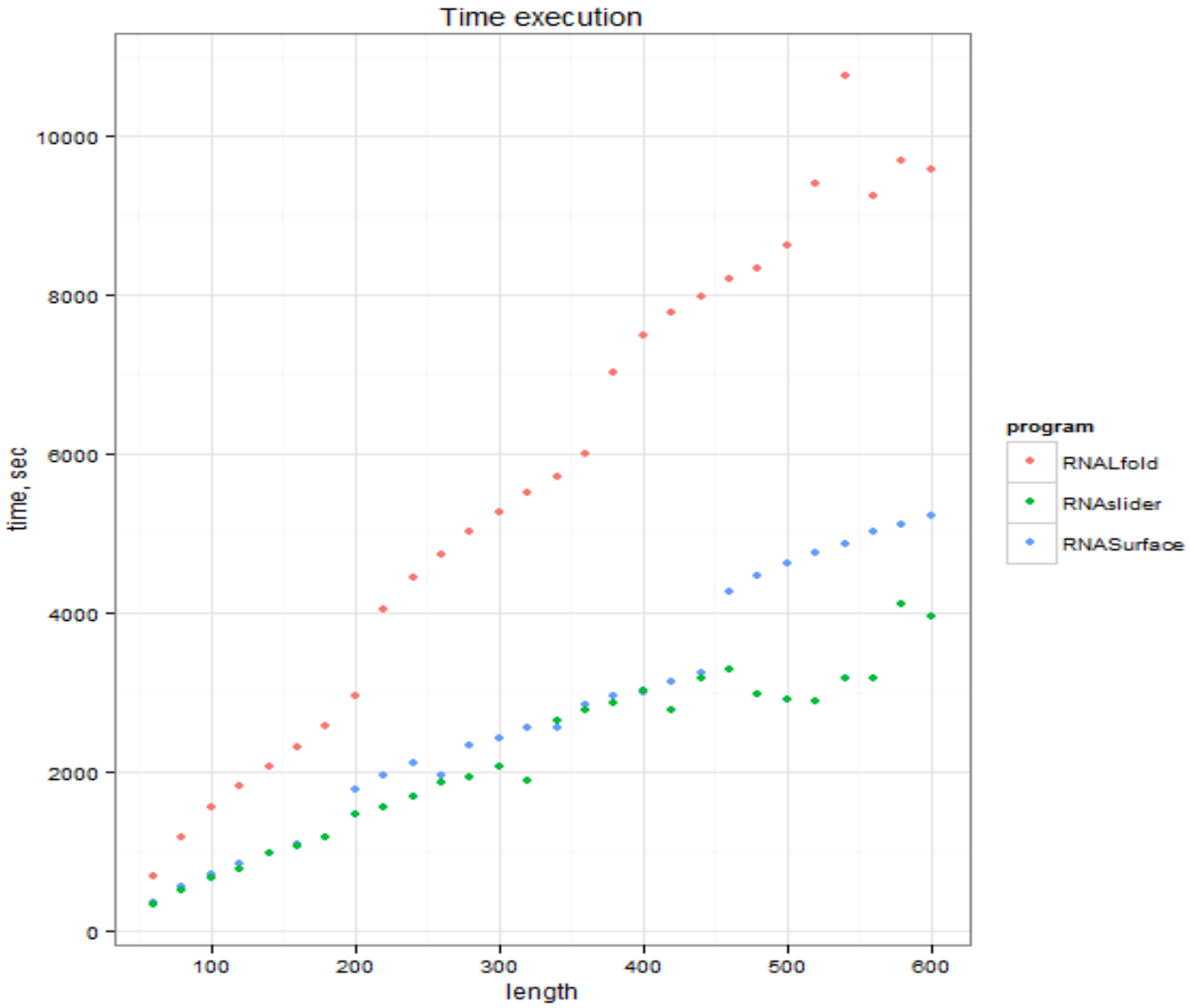
One-dimensional tracks



$$MZ(i) = \max_{i-l=r-i, r-l+1 \leqslant L} Z_{kl}^2 I\{Z_{kl} \leqslant 0\}$$

$$\rho_w(i) = \frac{1}{w} \sum_{k,l} Z_{kl}^2 \cdot I\{S_{kl} \in \text{locally optimal output}\} \cdot I\{i - w \leqslant \frac{k+1}{2} \leqslant i + w\},$$

Time requirement



Calculations were performed on Intel Xeon Processor E5506

Distribution of structured predictions along different types of regions in *Bacillus subtilis*

Table 2. Relative abundance of structured regions in various functional parts of the *Bacillus subtilis* genome

Z-score	-2	-3	-4	-5
coding regions	0.91 (148441/162599)	0.68 (21050/30963)	0.37 (2010/5399)	0.15 (153/1017)
upstream regions	0.98 (6442/6601)	1.41 (1809/1280)	2.09 (480/230)	3.05 (131/43)
downstream regions	2.55 (6166/2420)	5.68 (2793/492)	10.11 (950/94)	12.5 (225/18)
intercoding regions	1.34 (16448/12249)	2.41 (5753/2387)	4.41 (1901/431)	12.52 (551/44)
intercoding regions in operons	1.71 (274/160)	3.18 (105/33)	5.86 (41/7)	13 (13/1)

First and second numbers in parentheses are observed and expected number of predictions in selected region. Abundance is the ratio of these numbers.

Detection of different ncRNA classes in *Bacillus subtilis*

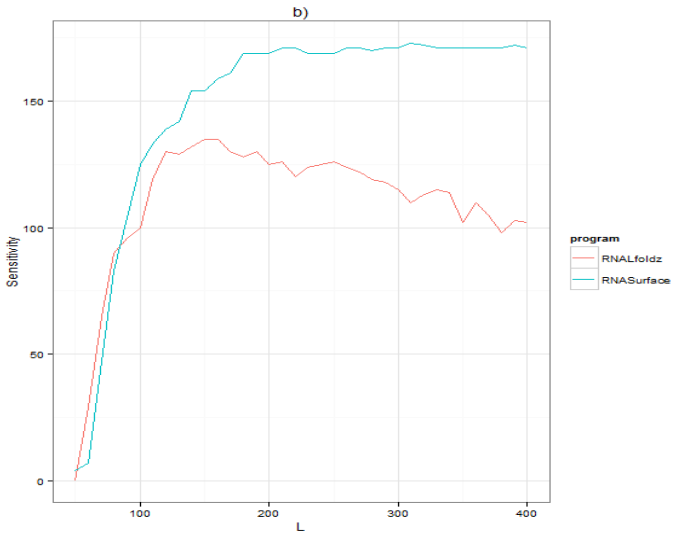
Table 1. Percent (number) of predictions for different types of RNA for three Z-score thresholds

Z-score	Riboswitch	T-box	L-leader	sRNA	tRNA	5S rRNA	FPR, %	PPV, %
-1	79 (34)	92 (12)	67 (4)	75 (15)	95 (81)	100 (20)	18	0.05
-2	65 (28)	85 (11)	50 (3)	65 (13)	62 (53)	35 (7)	5	0.1
-3	44 (19)	69 (9)	33 (2)	35 (7)	16 (14)	15 (3)	1	0.25

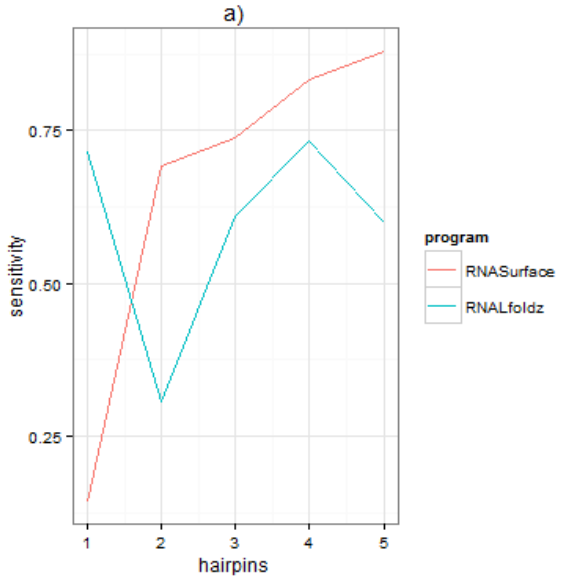
FPR - false positive rate, PPV - positive predictive value;

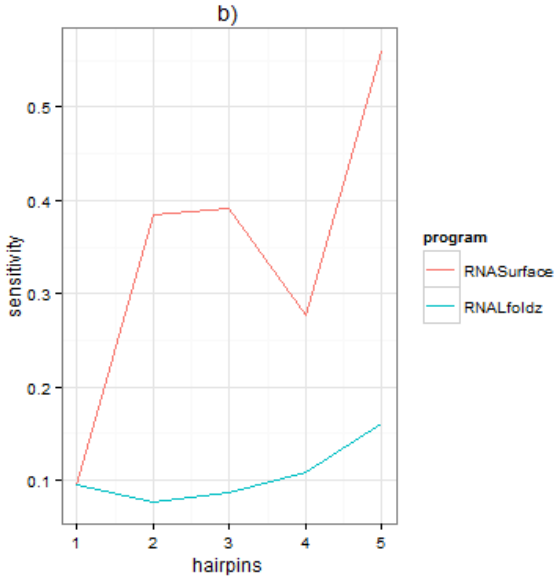
Prediction features

- Robustness of predictions to variable window length.

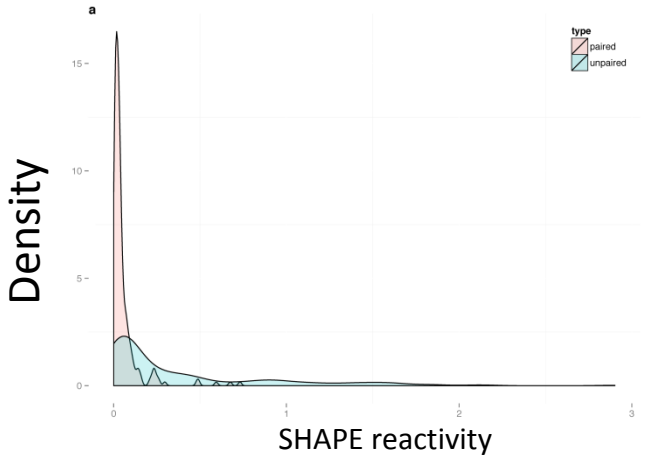


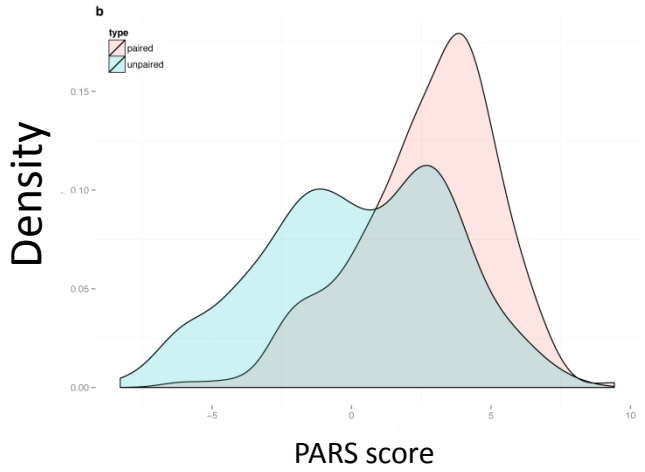
- Impact of the structure complexity on RNASurface and RNALfoldz performance

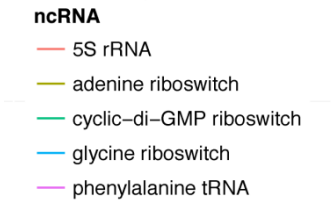


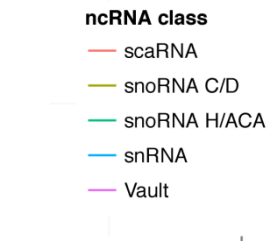


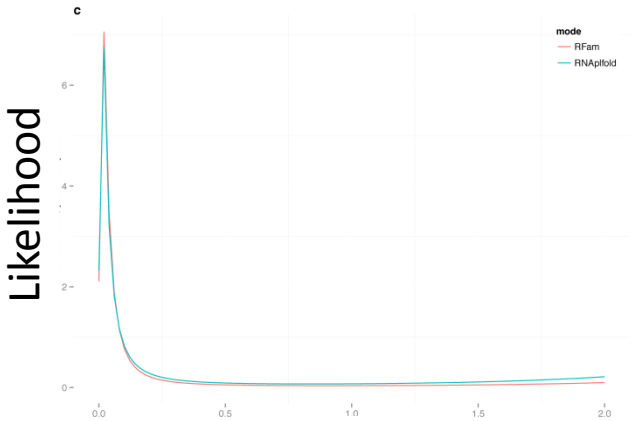
From reactivity to likelihood

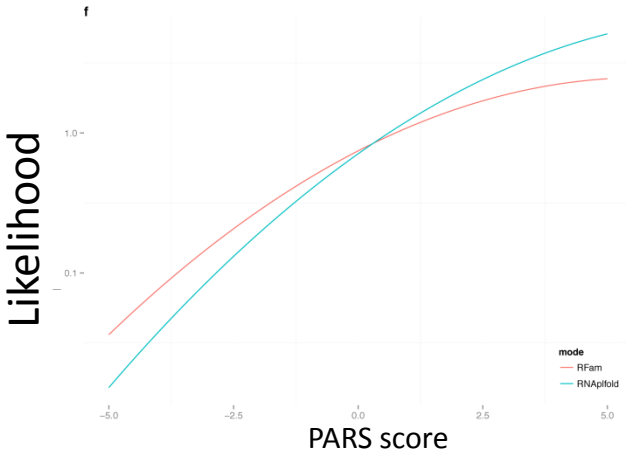












Reactivity r → high-confidence bases from partition function

$$L(r) = \log \frac{P(r|paired)}{P(r|unpaired)}$$

Distribution across mRNAs