Global or local?

Predicting secondary structure in mRNAs

Sita Lange

Bioinformatics Freiburg

objective

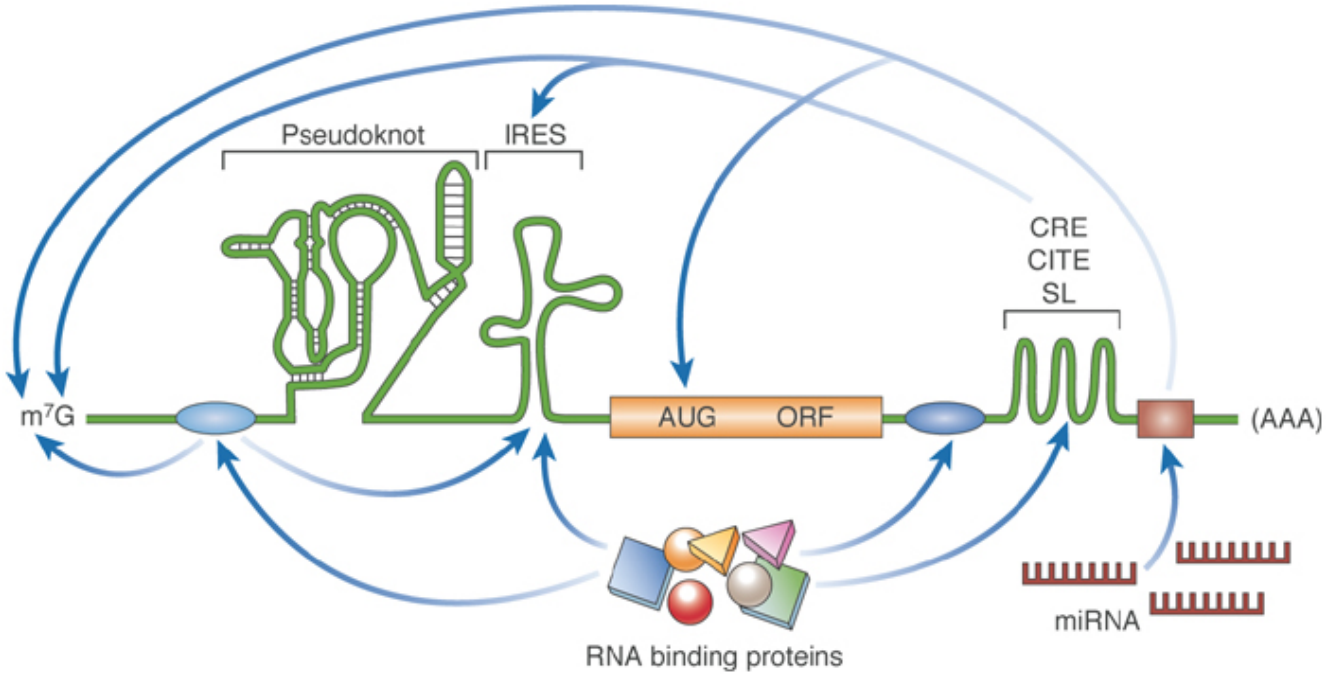
WINDOW-BASED LOCAL FOLDING...

...achieves the best structure prediction accuracy in application to mRNAs

and possibly other long RNA molecules,
i.e. long non-coding RNA

mRNA structure involved in regulation

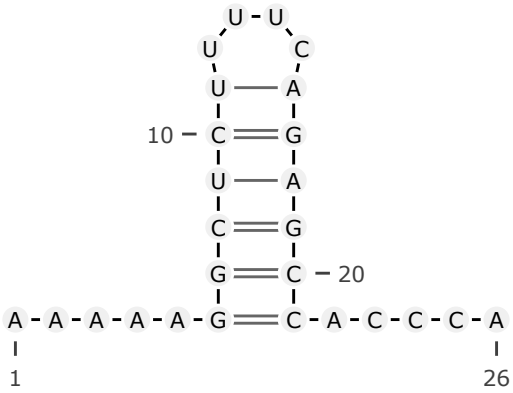
- *Cis-acting*: *Cis*-regulatory elements located primarily on untranslated regions (UTRs)
- *Trans-acting*: Structural accessibility of binding sites



Roberts & Holcik, EMBO reports (2009)

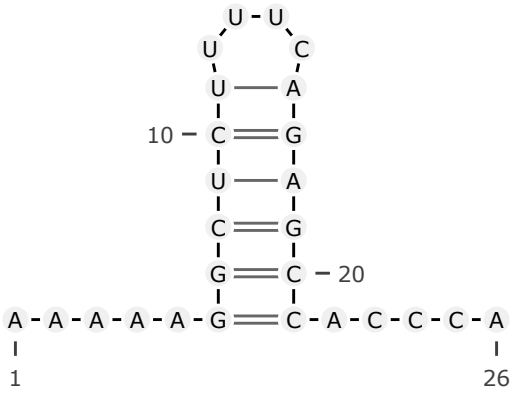
local structure

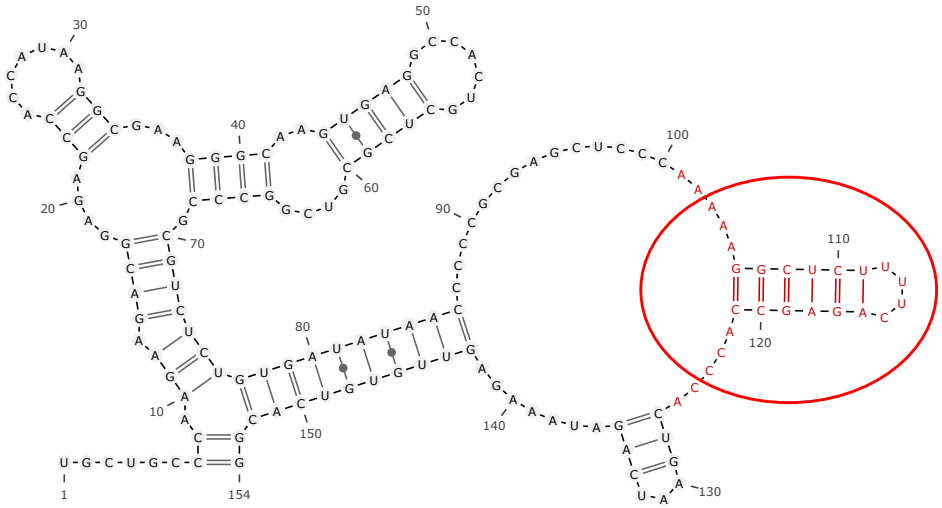
e.g. histone 3' UTR stem-loop



local structure

e.g. histone 3' UTR stem-loop

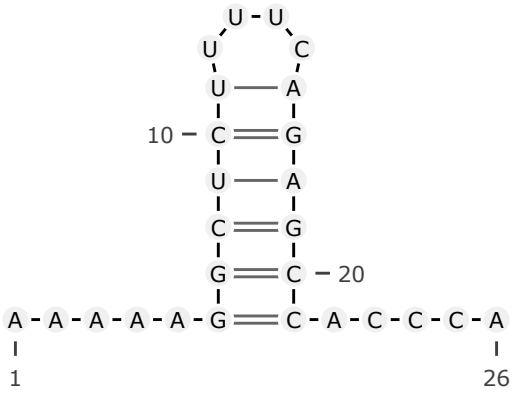


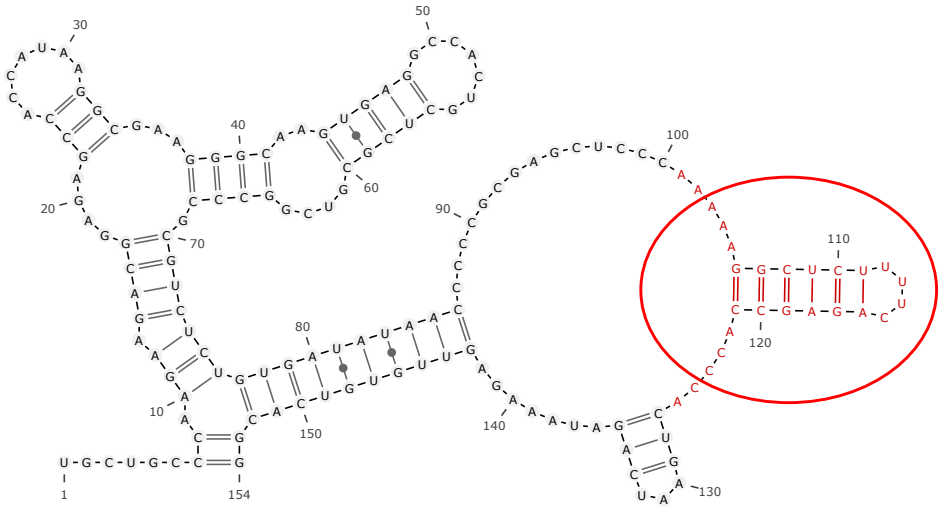


a substructure in its natural context

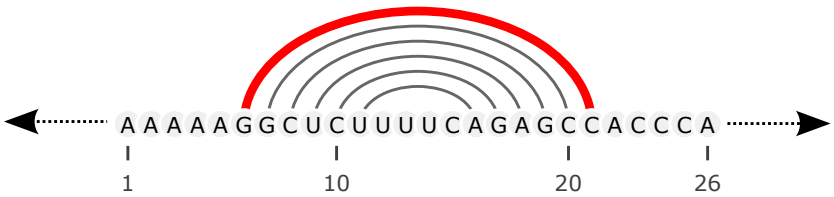
local structure

e.g. histone 3' UTR stem-loop



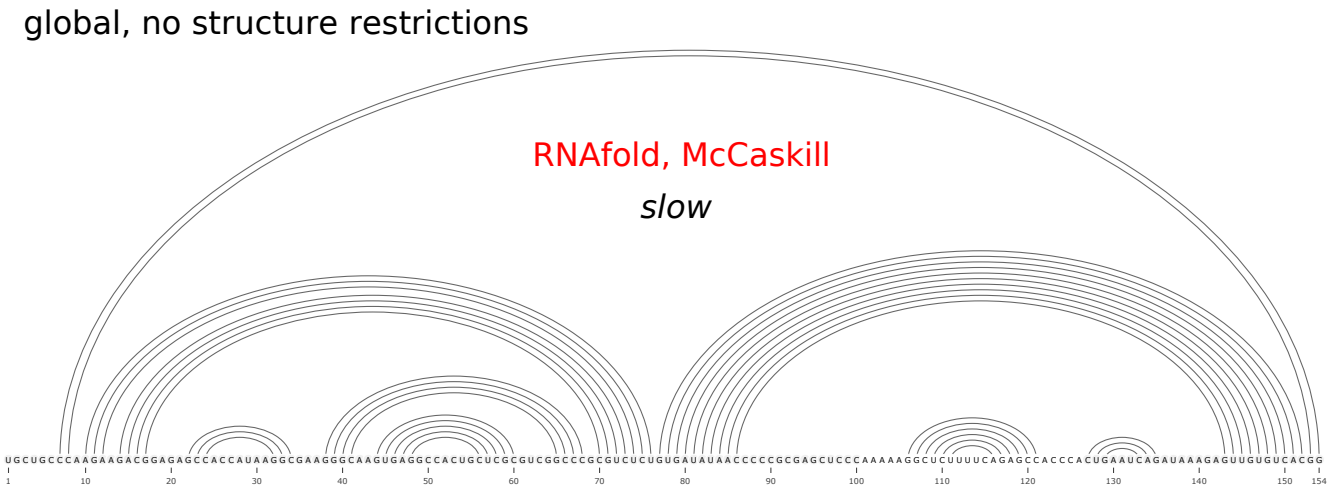


a substructure in its natural context



define a max. base-pair span (L)

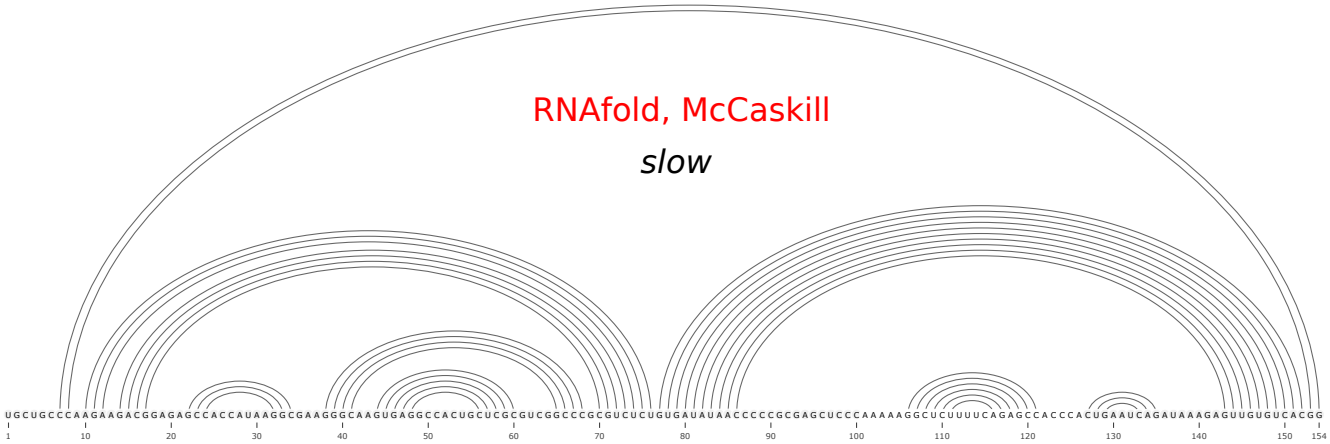
predicting local structure



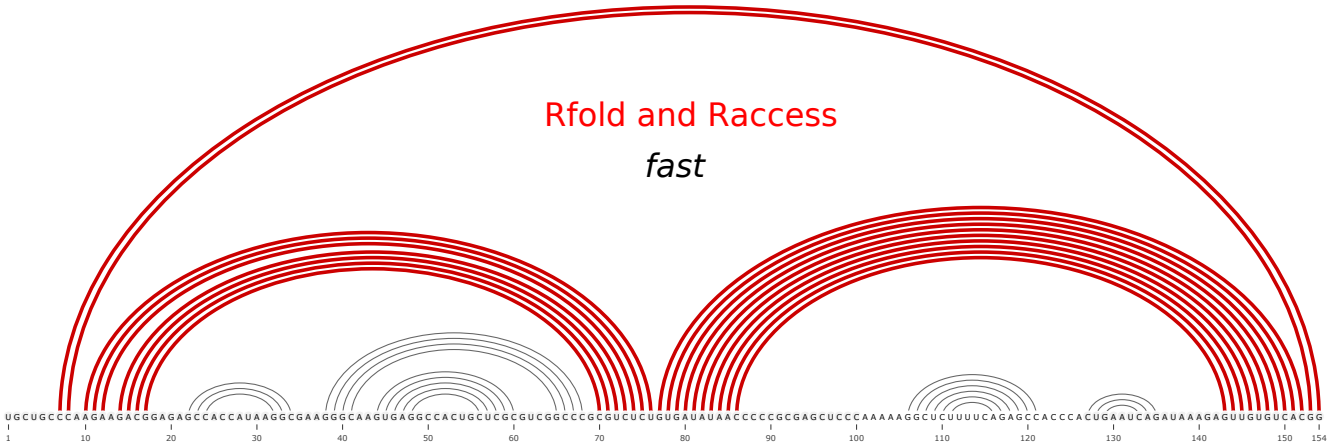
RNAfold: Hofacker *et al.* Monatshefte für Chemie (1994),
UNAfold/mfold: Markham & Zuker Methods Mol. Biol. (2008)

predicting local structure

global, no structure restrictions



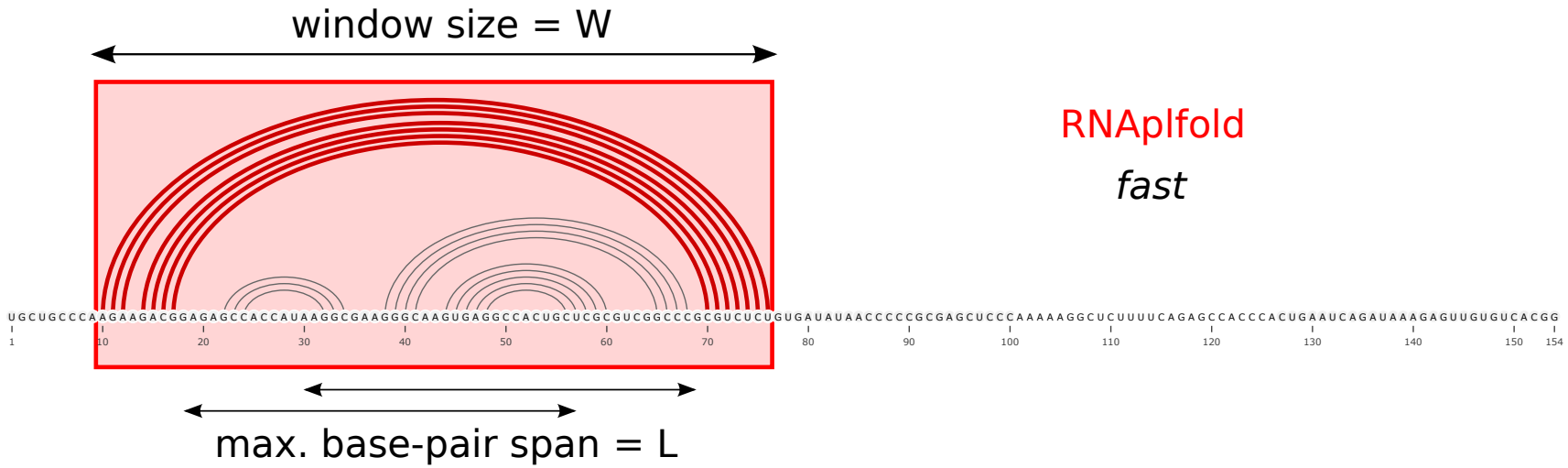
semi-local, restricting max. base-pair span, parameter L



Rfold: Kiryu *et al.* Bioinformatics (2008),
Raccess: Kiryu *et al.* Bioinformatics (2011)

predicting local structure

local, sliding local windows ignoring larger context, parameters W, L



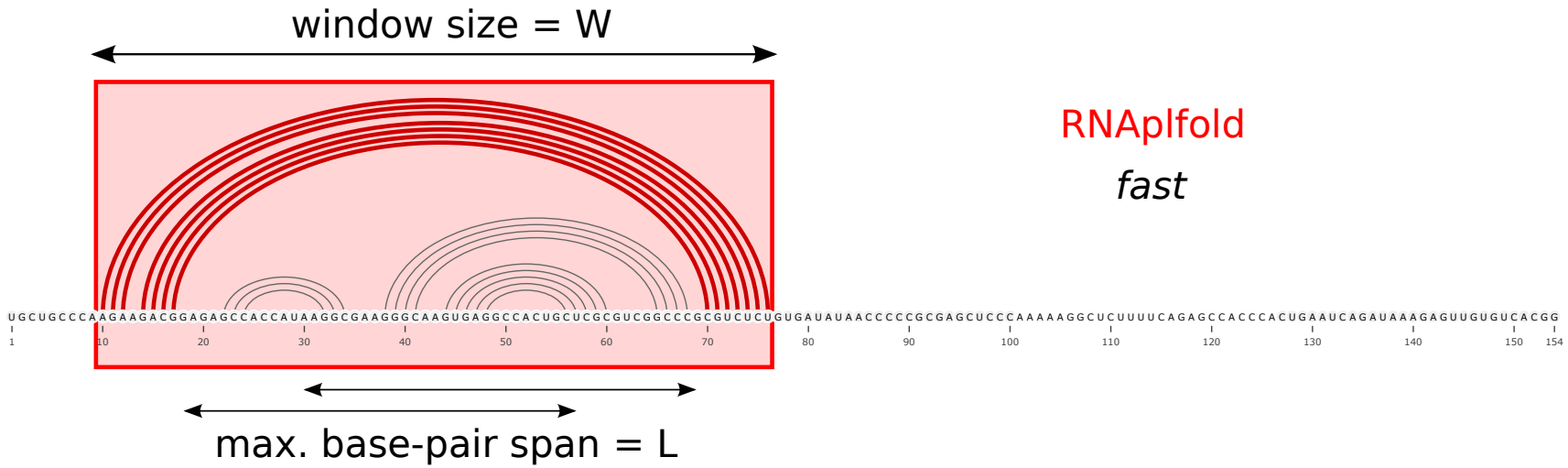
RNAplfold

fast

Bernhart *et al.* Bioinformatics (2006)

predicting local structure

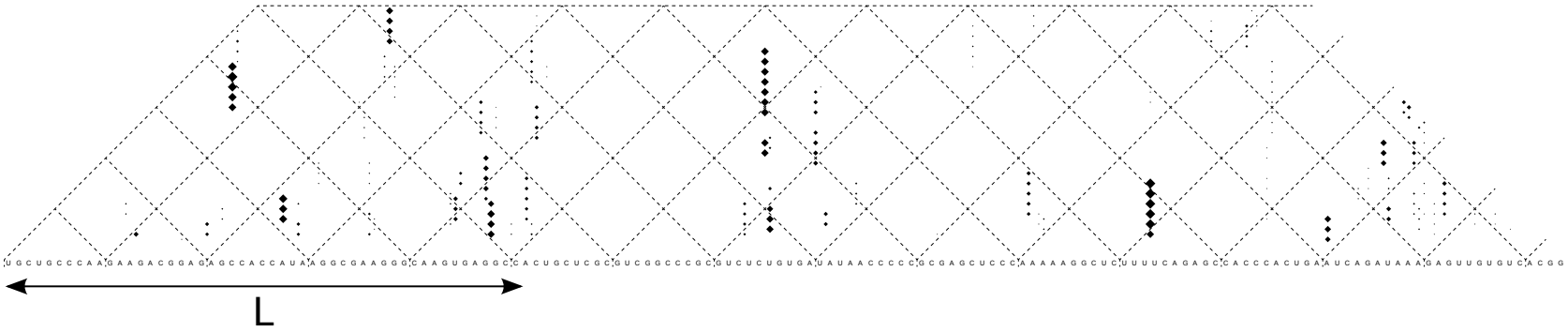
local, sliding local windows ignoring larger context, parameters W, L



RNAplfold

fast

averaged base-pair probabilities



Bernhart *et al.* Bioinformatics (2006)

three unresolved questions

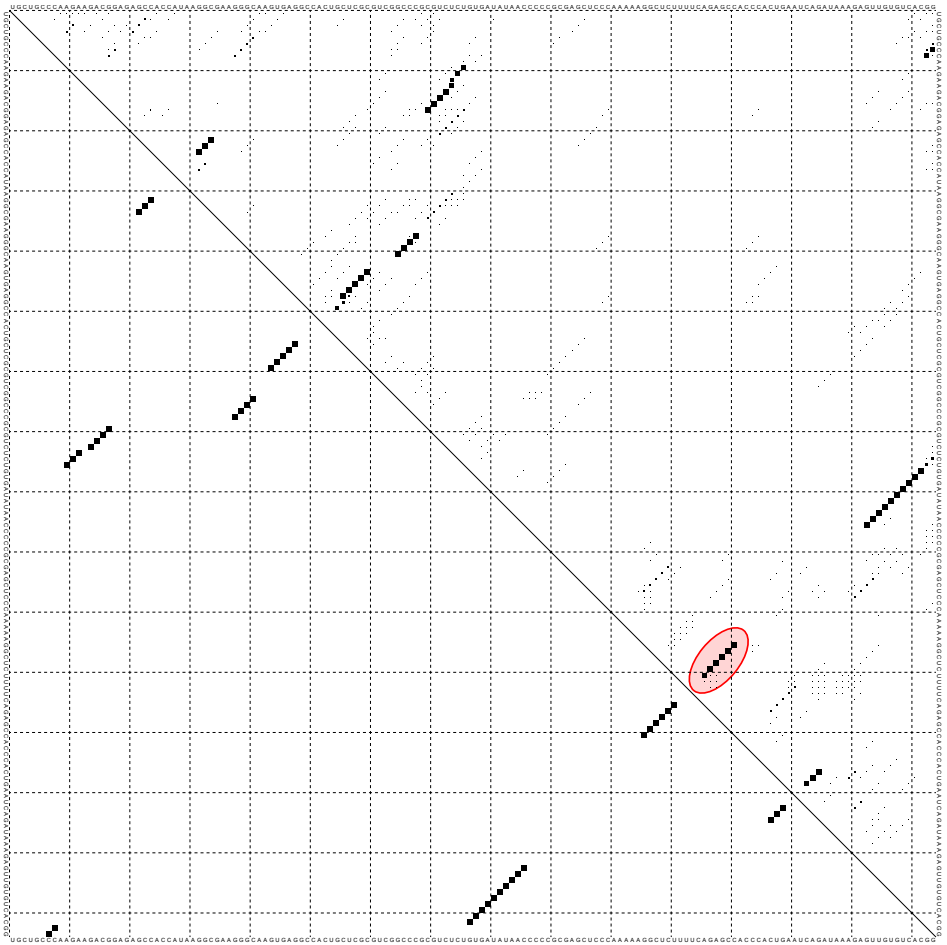
- Which method is more accurate for mRNA? Global or local?
- How to set the parameters for local folding?
- How to evaluate the quality of structure prediction methods on local structure? Data? Performance measures?

our new database: CisReg mRNA

- 95 families of *cis*-regulatory elements, hand-selected from Rfam
- 2500 individual structures, mapped, extended and filtered from Seed alignments
- $>$ 85, 000 base-pairs, many spieces/kingdoms

http://lancelot.otago.ac.nz/CisRegRNA/

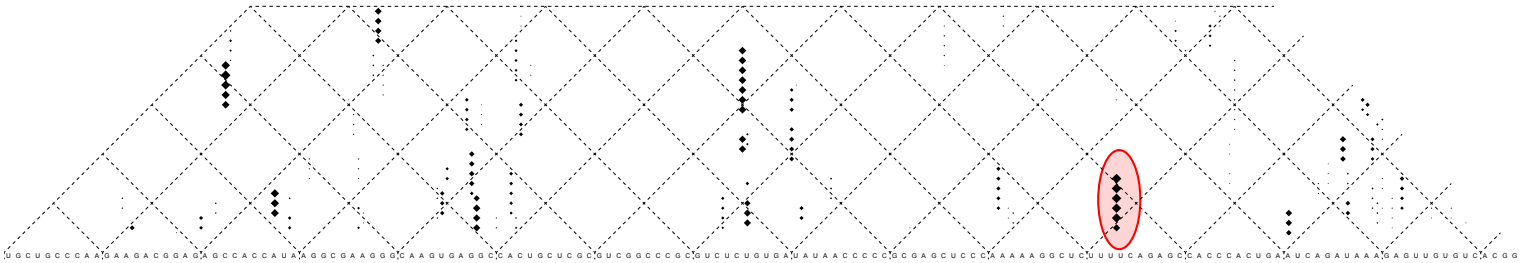
performance measure: structure accuracy



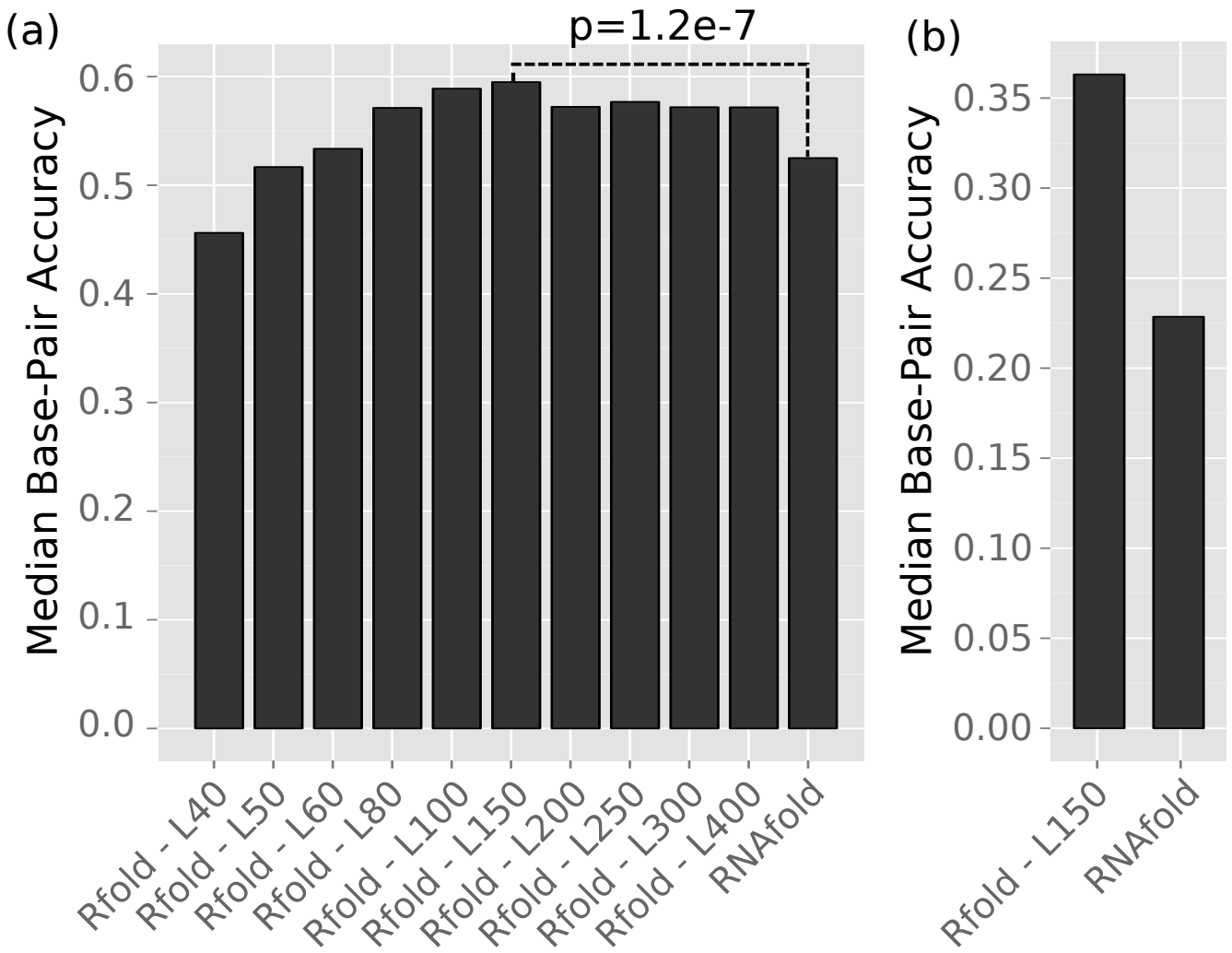
for global predictions

ACCURACY:
average base-pair probability

for local predictions

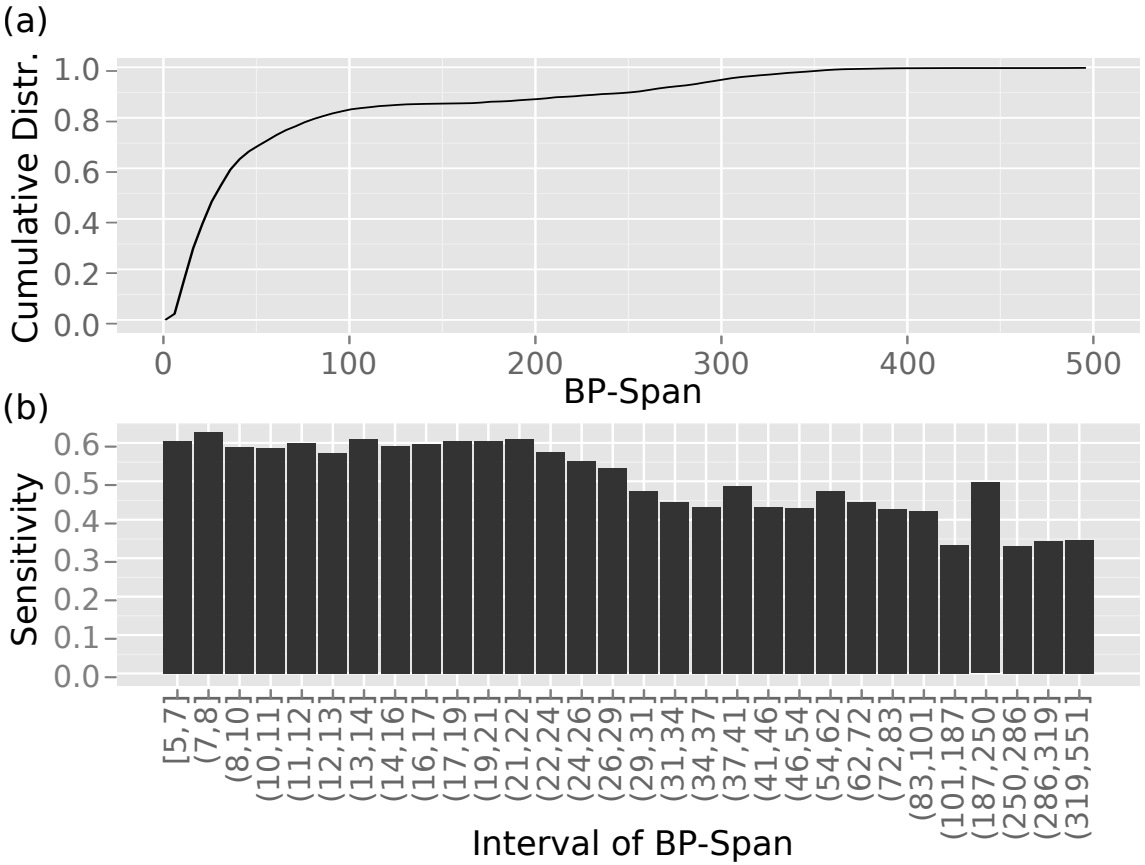


restricting the max. base-pair span L



- best *L* at 150 nt
- structures are locally stable

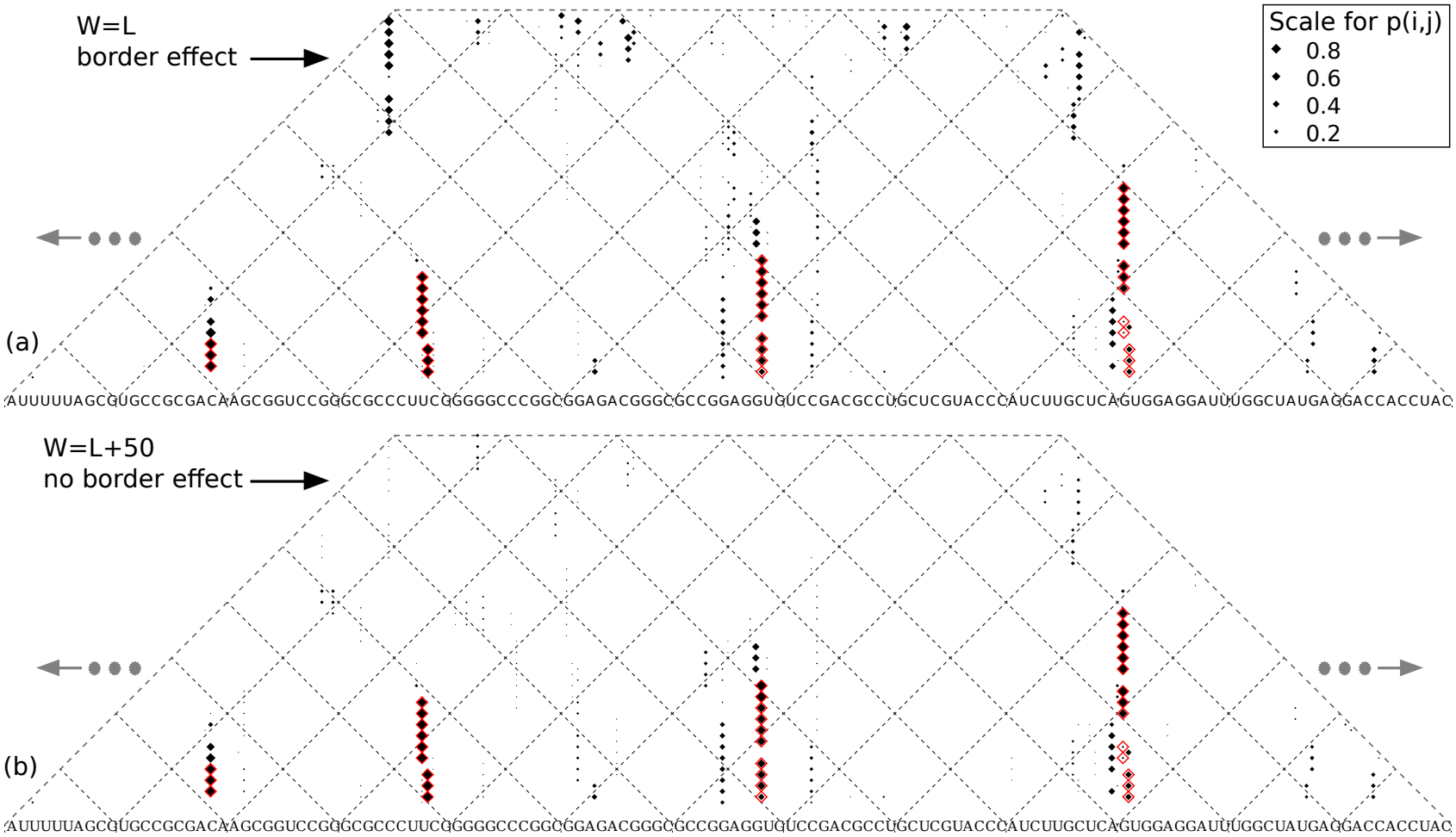
restricting the max. base-pair span L



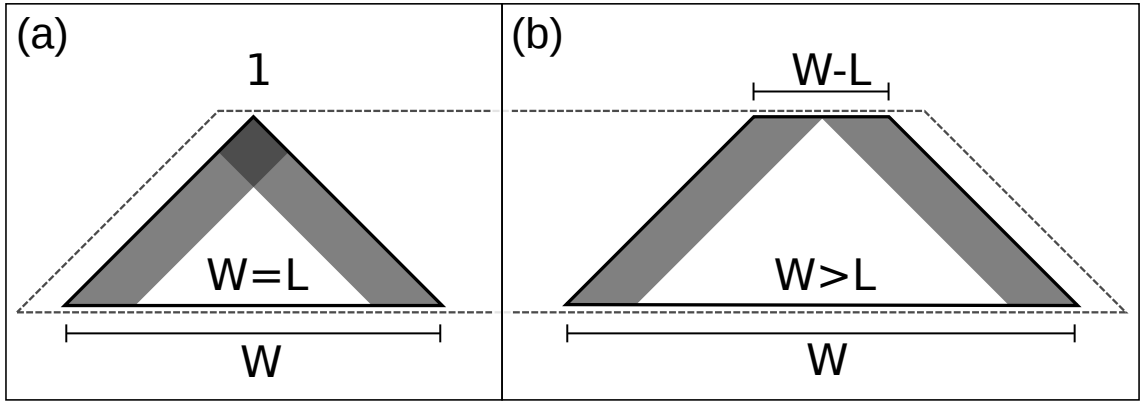
- most base-pairs have short spans: 83% $\leq$ 100 nt (exponential)
  - 75% $\leq$ 100 nt for rRNA, Doshi *et al.* BMC Bioinformatics 2004
  - 85% $\leq$ 100 nt for Rfam, Kiryu *et al.* Bioinformatics 2011
- $L \approx 150$ maximises included base-pairs and minimises incorrect long-range predictions

indroducing artificial windows: border bias

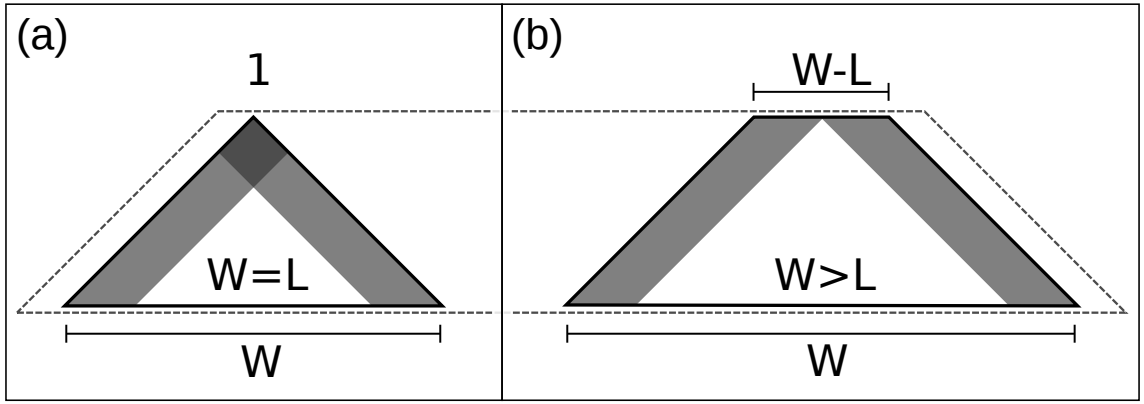
- ROSE element predicted with an accuracy of 0.65
- strong border effect when $W = L$

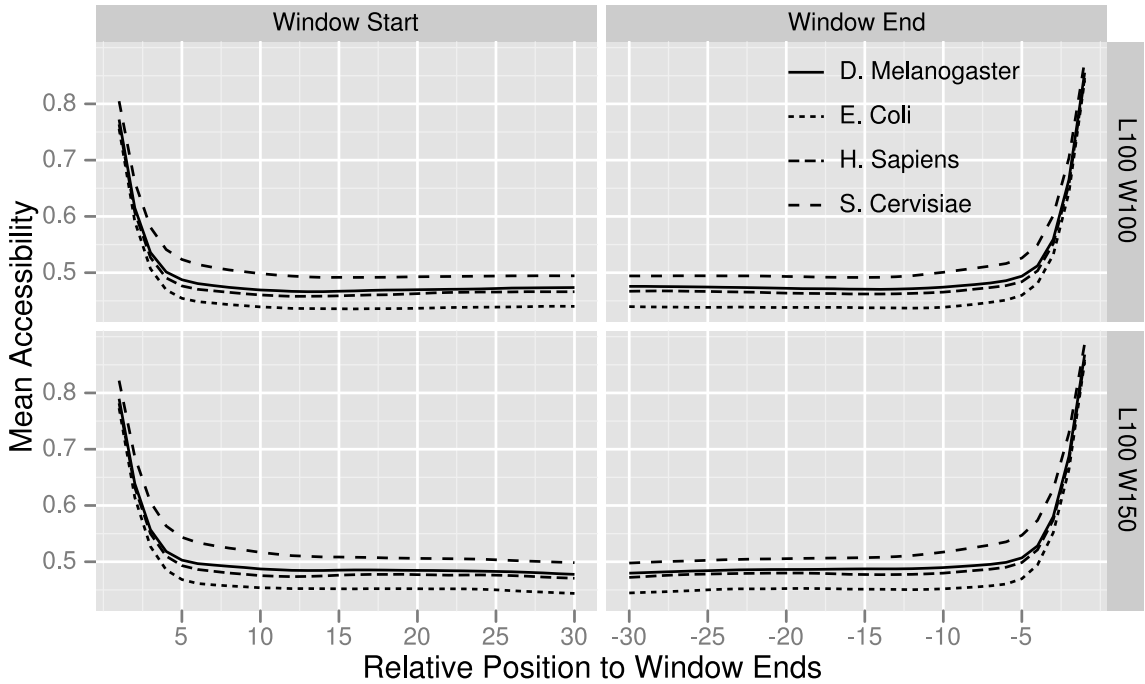


introducing artificial windows: border effects

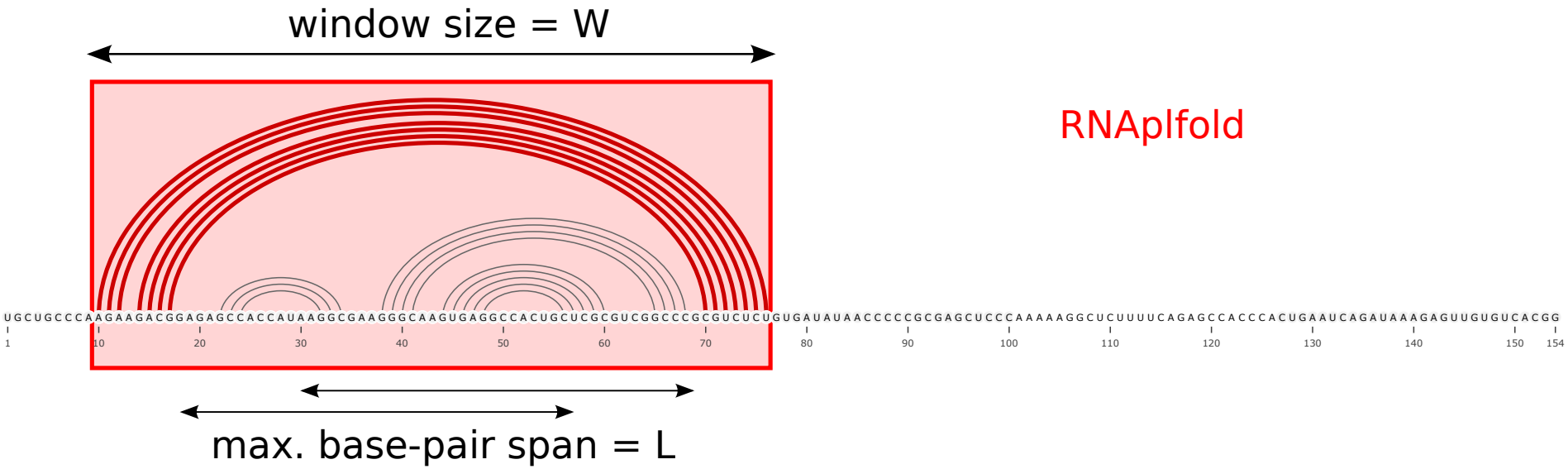


introducing artificial windows: border effects

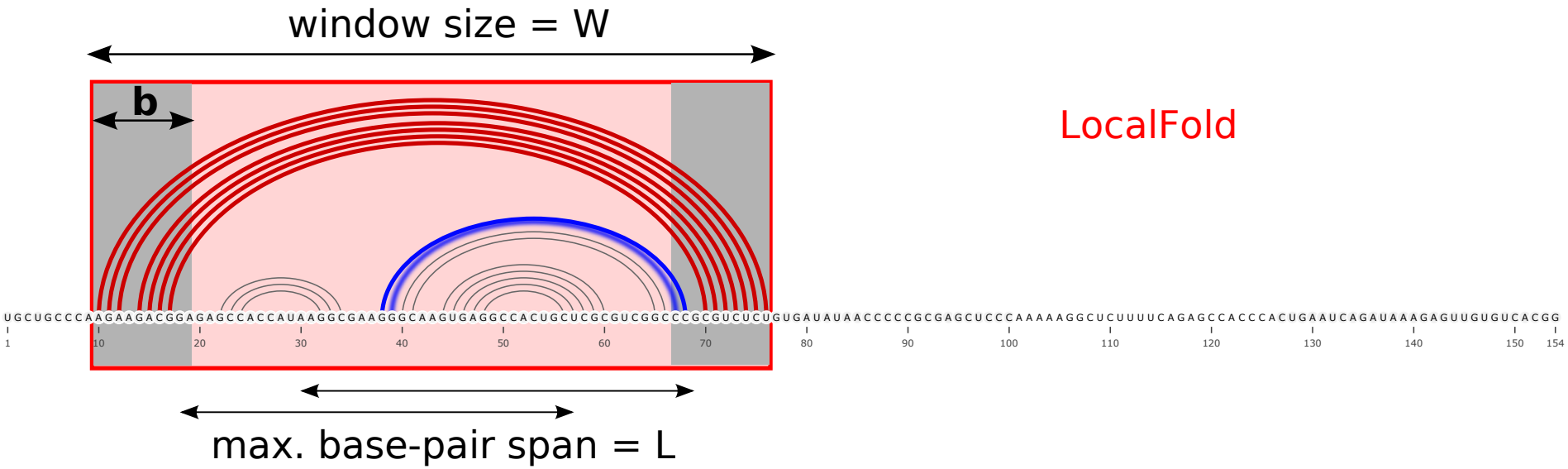




LocalFold: diminish border effects

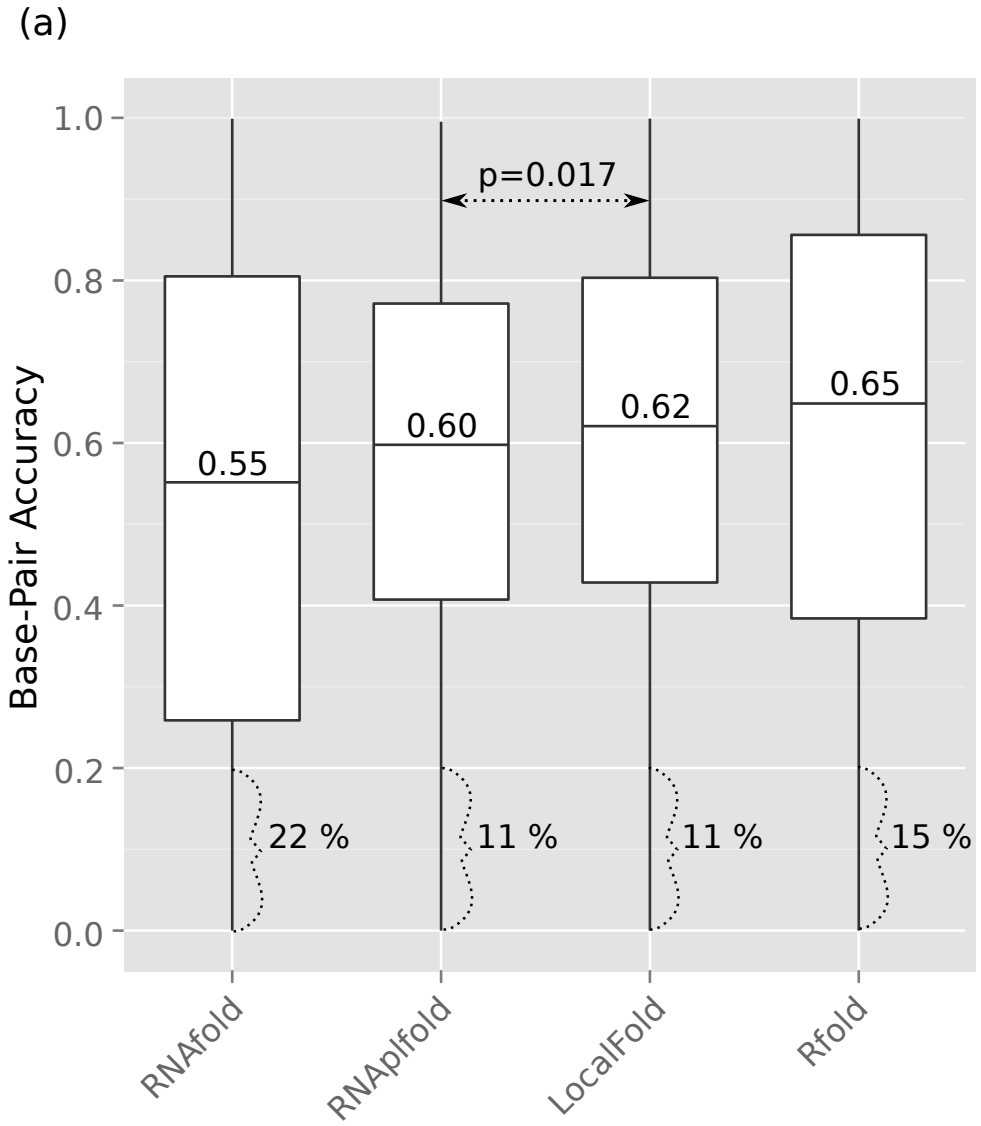


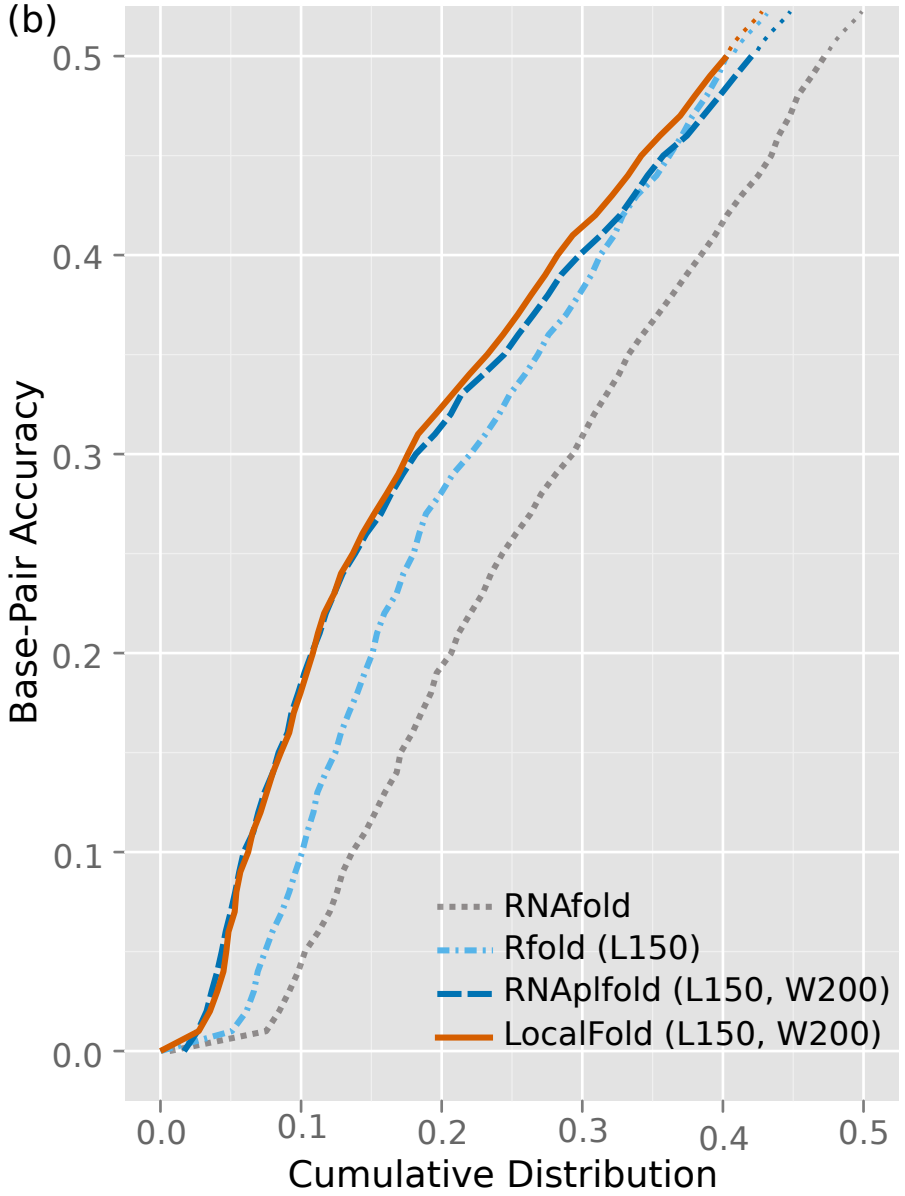
RNAplfold



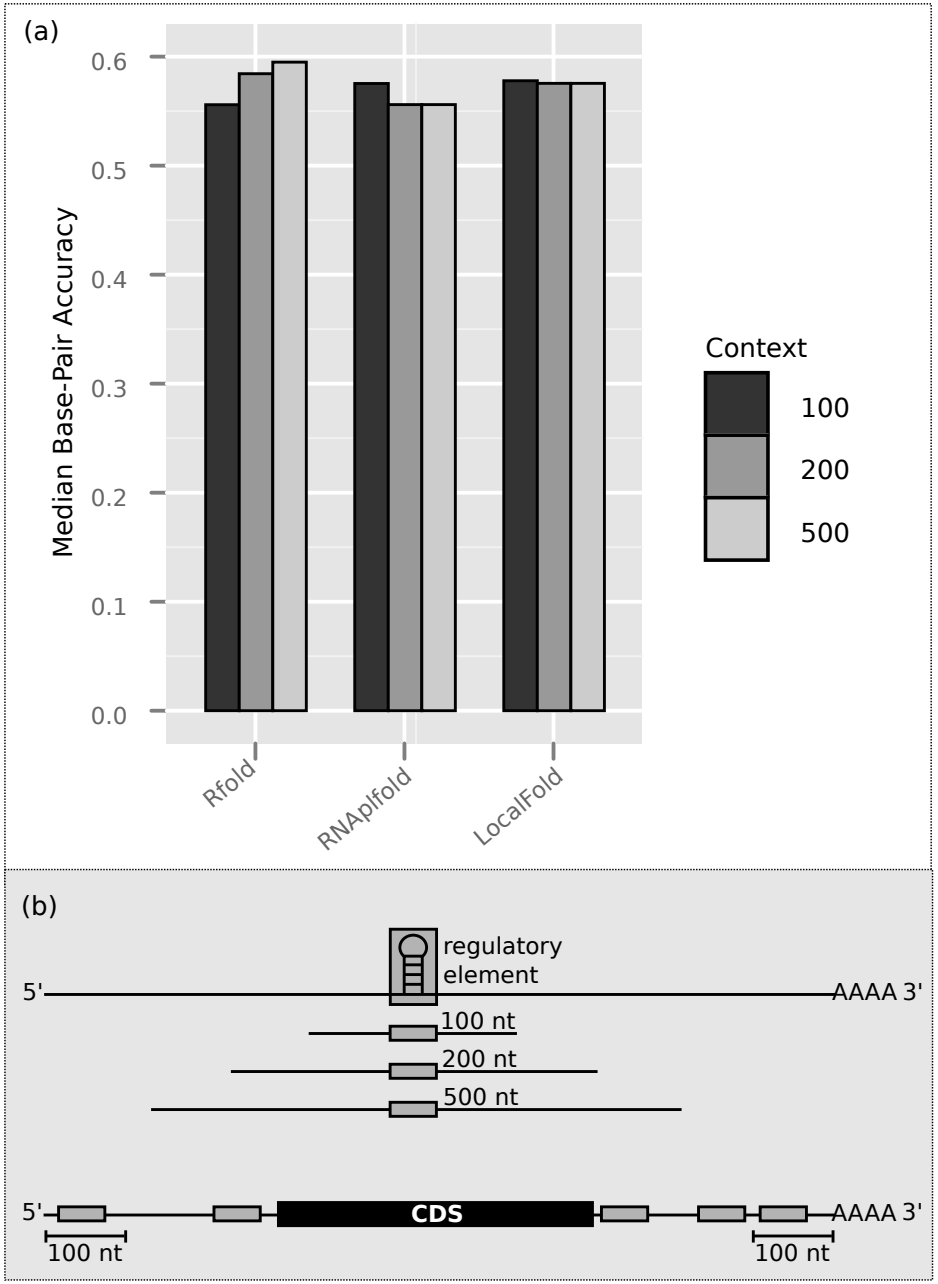
LocalFold

performance comparison on CisReg





Rfold has problems at sequence ends



evaluating accessibility

ACCESSIBILITY

unpaired probability for one or more nucleotide (we use 1nt)

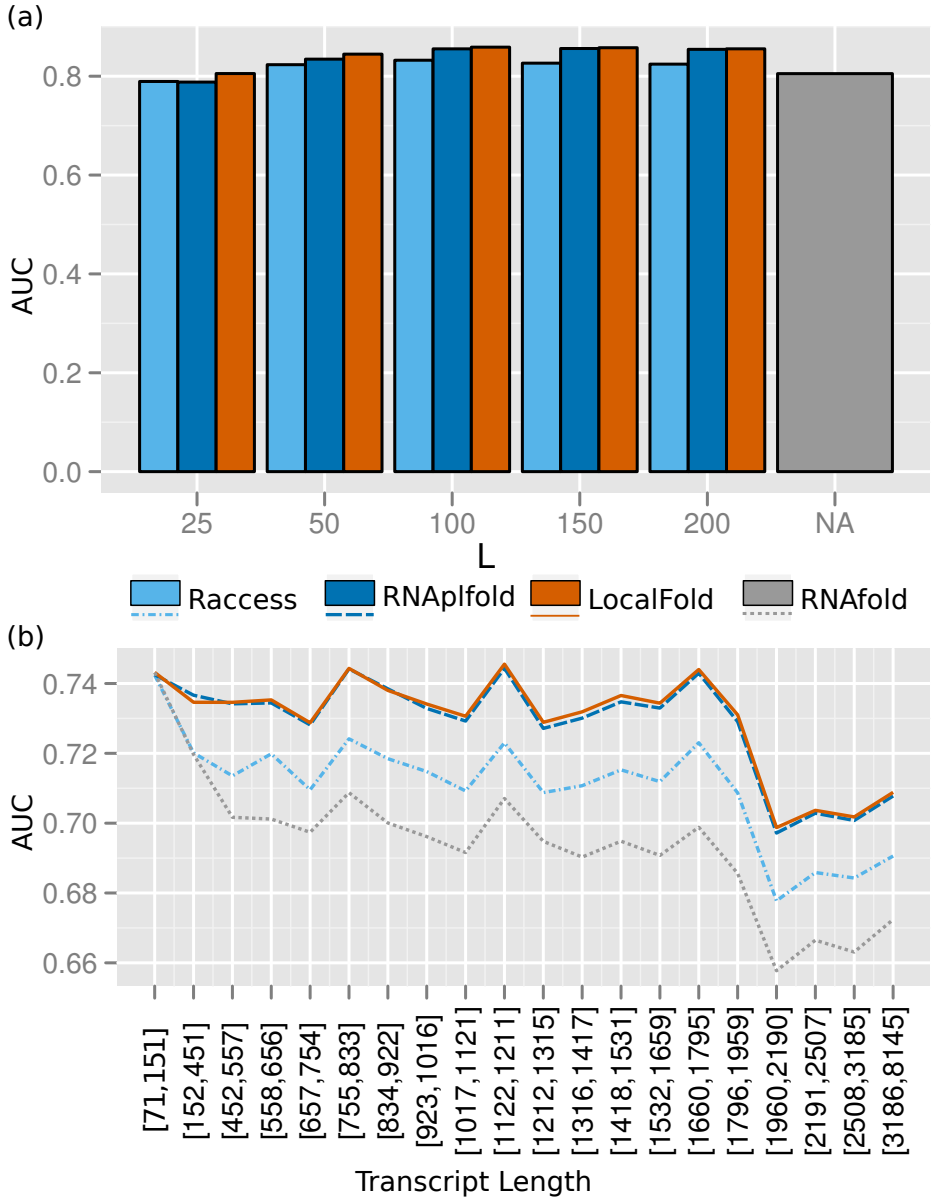
DATA

- Kertesz et al. Nature 2010, *Genome-wide measurements of mRNA secondary structure in yeast*
- restriction enzyme decay, double- and single-stranded cleavage $\rightarrow$ PARS score (position-wise single-strandedness)
- evaluated positions with highest (paired) and lowest (unpaired) score ($\geq 80,000$ nt)

PERFORMANCE MEASURE

- performance measure: Area Under the ROC Curve (AUC)
- implicitly reflects base-pairing distributions

introducing windows: border effects



the last slide

TAKE-HOME

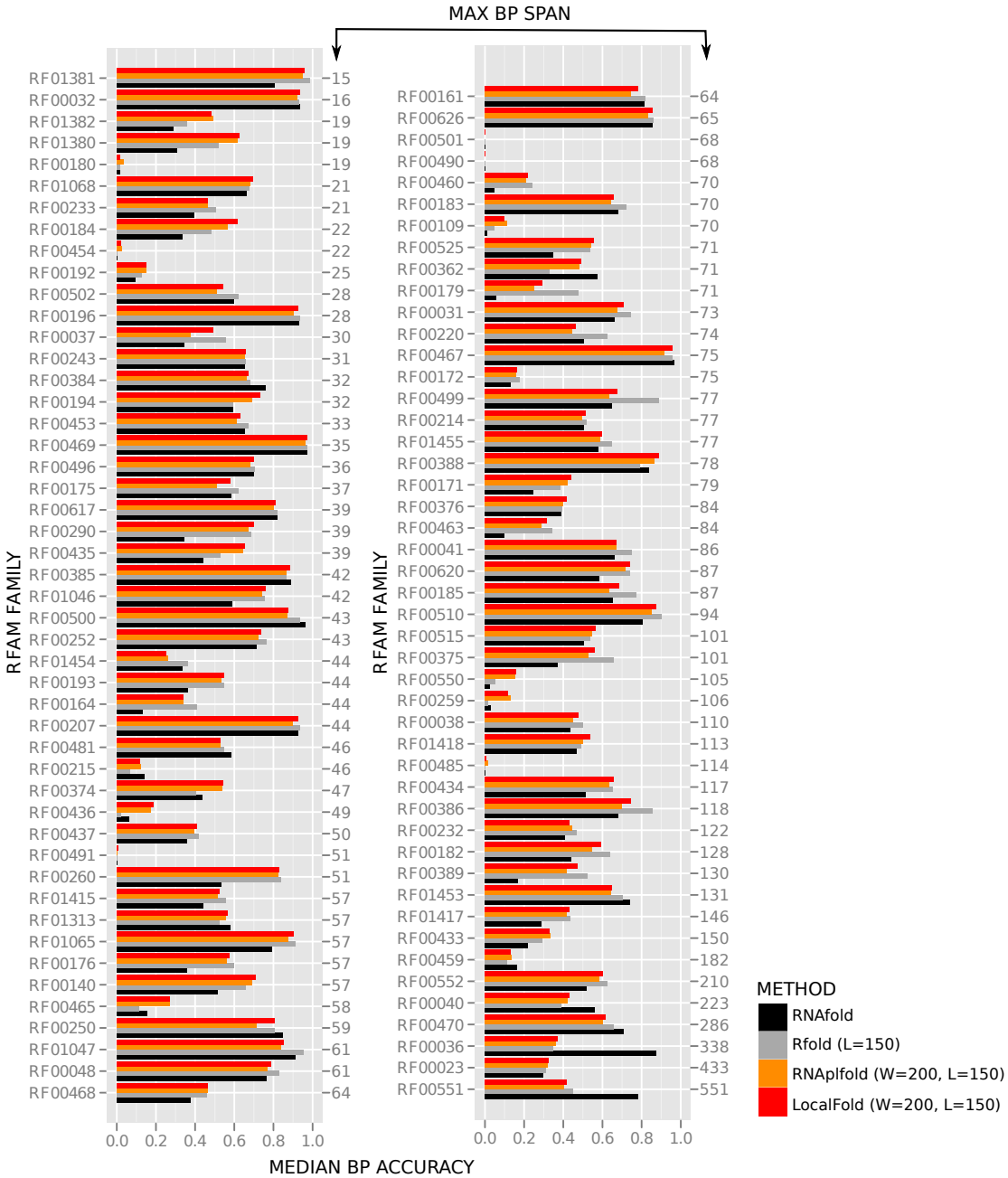
- window-based folding is the most accurate (LocalFold or RNAplfold)
- bias at artificial window borders
- $W = L + 50$ & exclude window termini

ACCEPTED BY NAR FEB 2012

I acknowledge and thank my co-first author Daniel Maticzka, and Mathias Möhl, Josh Gagnon, Chris Brown, Rolf Backofen

THANK YOU for your attention!

similar trend for most families



window size is a stable parameter