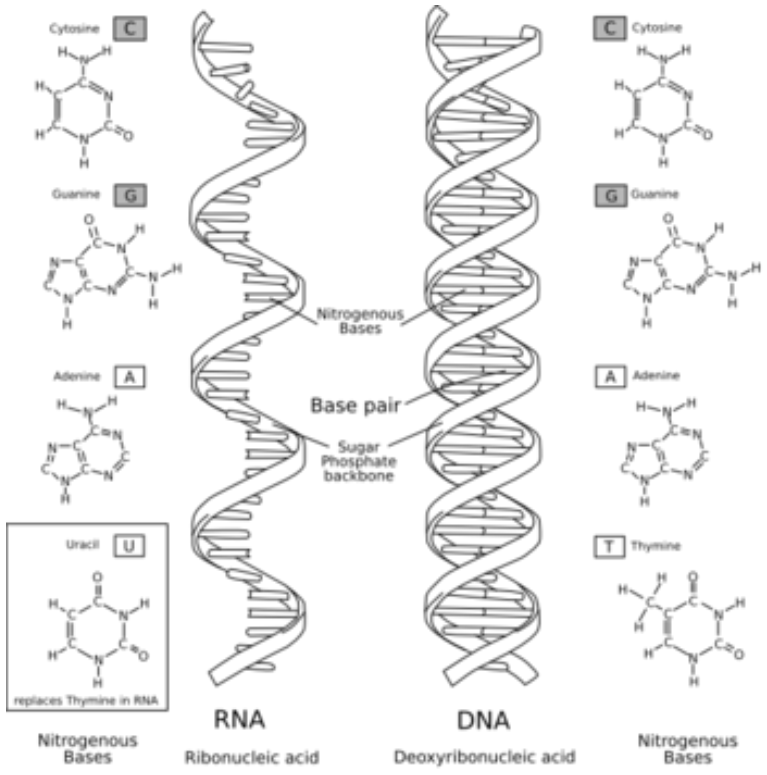
LocARNA-P: Accurate Boundary Prediction and Improved Detection of Structured RNAs

Sebastian Will

CSAIL, MIT

RNA

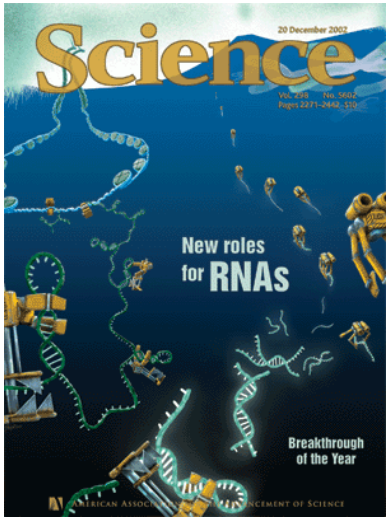


chain of building blocks A,C,G,U

once upon a time . . .

RNA was considered only an intermediate in protein synthesis

Changed Picture: RNA plays central role



Multitude of Non-coding RNAs
with various functions

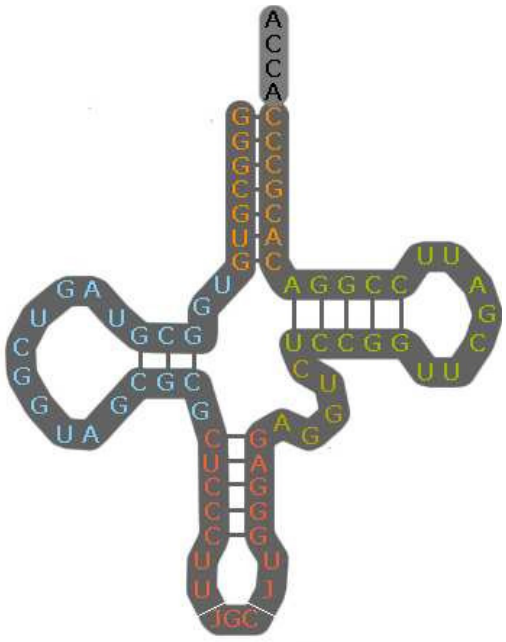
- gene regulation
- catalysis of reactions
  ("ribozymes")

Versatility due to RNA Structure

primary structure = sequence (of bases A,C,G,U)

GGGCGUGUGGCGUAGUCGGUA . . . GUUCGAUUCCGGACACGCCCACCA

secondary structure



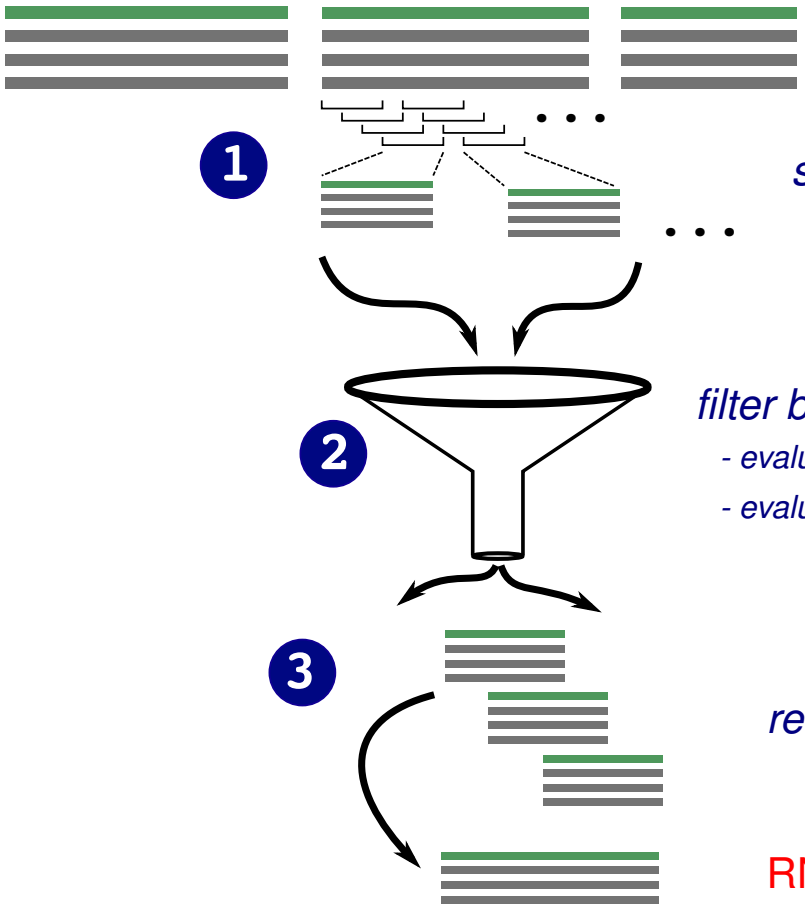
base pairs: C-G,A-U,G-U

tertiary structure

all atoms 3D

De-novo Prediction of Structural RNAs

RNAz Washietl *et al*, Fast and reliable prediction of noncoding RNAs, *PNAS*, 2005



RNAz detects $\geq$ 40,000 ncRNA candidate loci in Fly

Current Limits of De-novo ncRNA Prediction

Main Problems:

- coarse RNA window boundaries
- high false discovery rate (FDR) $\approx 50\%$
  experimental analysis expensive

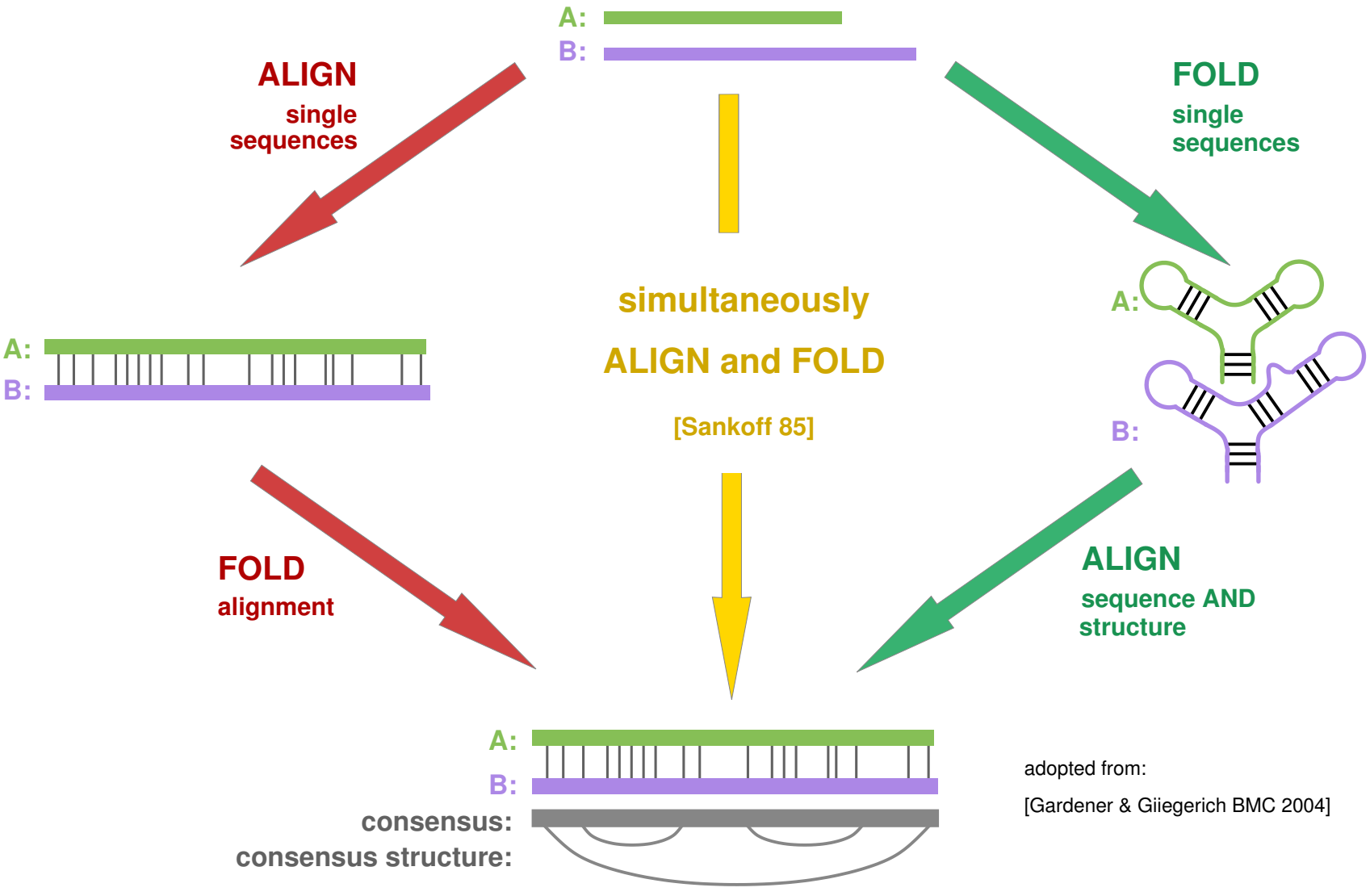
Main Limitations:

- no structural re-alignment (WGA sequence-based)
- no information about alignment reliability
- no information about alignment space

Recall talk title

"LocARNA-P: Accurate Boundary Prediction and Improved Detection of Structured RNAs for Genome-wide Screens"

Comparative RNA Analysis



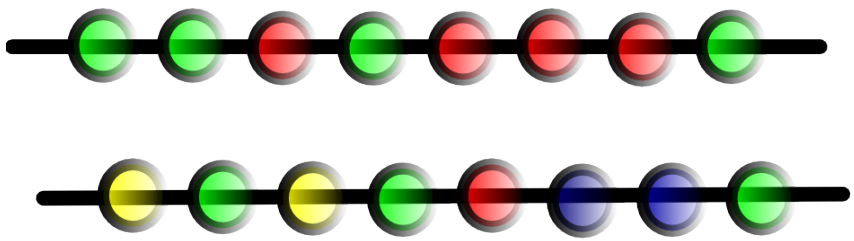
adopted from:

[Gardener & Giiegerich BMC 2004]

LocARNA: more efficient Alignment and Folding

- **LocARNA** does a simplified variant of Simultaneous and Folding due to Hofacker

Sankoff	**Hofacker**
Zuker × Needleman-Wunsch	Nussinov × Smith-Waterman + McCaskill
full energy model	use full energy model via base pair probabilities

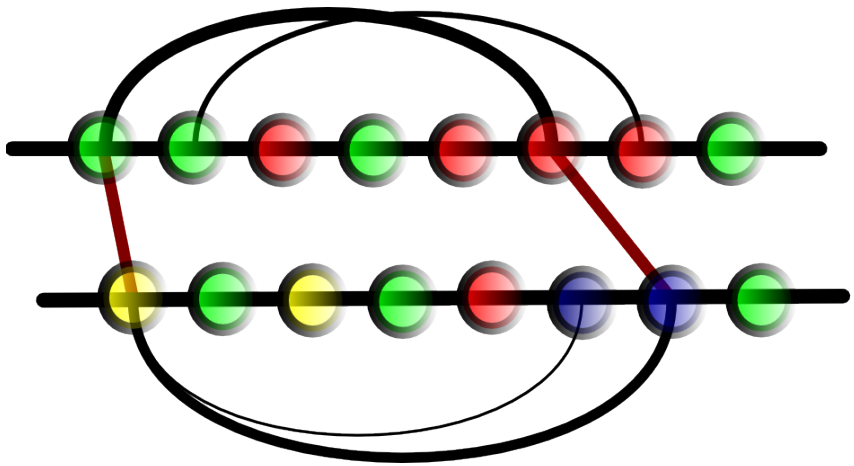


Hofacker et al., PMcomp, *Bioinformatics*, 2004.

LocARNA: more efficient Alignment and Folding

- **LocARNA** does a simplified variant of Simultaneous and Folding due to Hofacker

Sankoff	**Hofacker**
Zuker × Needleman-Wunsch	Nussinov × Smith-Waterman + McCaskill
full energy model	use full energy model via base pair probabilities



Hofacker et al., PMcomp, *Bioinformatics*, 2004.

LocARNA: more efficient Alignment and Folding

- **LocARNA** does a simplified variant of Simultaneous and Folding due to Hofacker

Sankoff	**Hofacker**
Zuker $\times$ Needleman-Wunsch	Nussinov $\times$ Smith-Waterman + McCaskill
full energy model	use full energy model via base pair probabilities

- **LocARNA** is fast and robust
- **LocARNA** improves scoring and adds features (e.g. locality)
- **LocARNA** reduces space complexity to $O(n^2)$

Hofacker et al., PMcomp, *Bioinformatics*, 2004.

LocARNA-P adds Probabilities

- match probabilities for sequence **and** structure

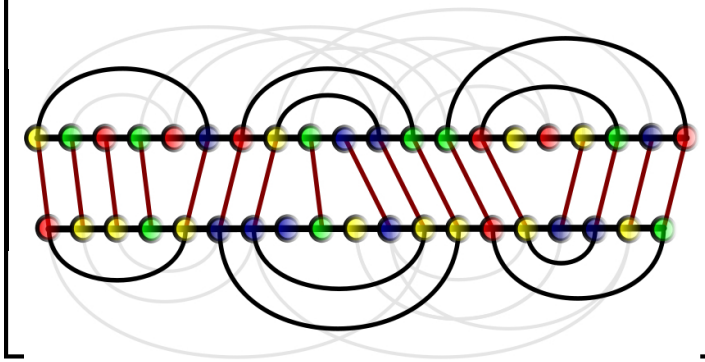
- reduced space complexity: $O(n^2)$ "as in LocARNA"
- make use of very accurate LocARNA scoring

Alignment Probabilities

LocARNA scoring function: $\text{score}(\mathcal{A}, \mathcal{S})$

Given sequences A, B: define **probability**

$$\Pr\left[\ \right]$$



i.e.: Pr[pair of alignment $\mathcal{A}$ **and** consensus structure $\mathcal{S}$]

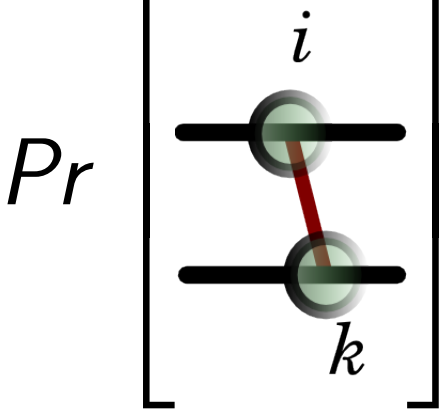
as probability in **Boltzmann ensemble**

$$Pr[(\mathcal{A}, \mathcal{S})] \sim exp\big(\beta\, \text{score}(\mathcal{A}, \mathcal{S})\big)$$

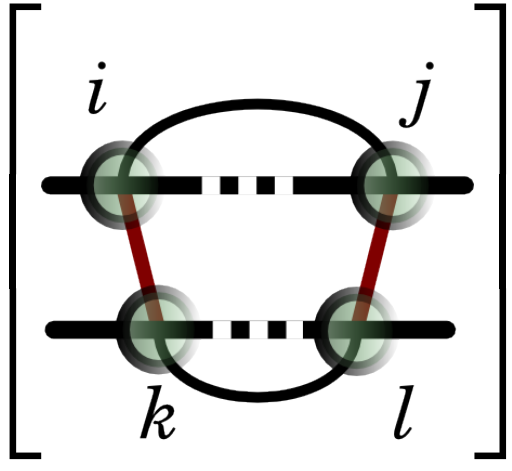
needs **partition function** $Z = \sum_{(\mathcal{A},\mathcal{S})} exp\big(\beta\, \text{score}(\mathcal{A}, \mathcal{S})\big)$

Match Probabilities

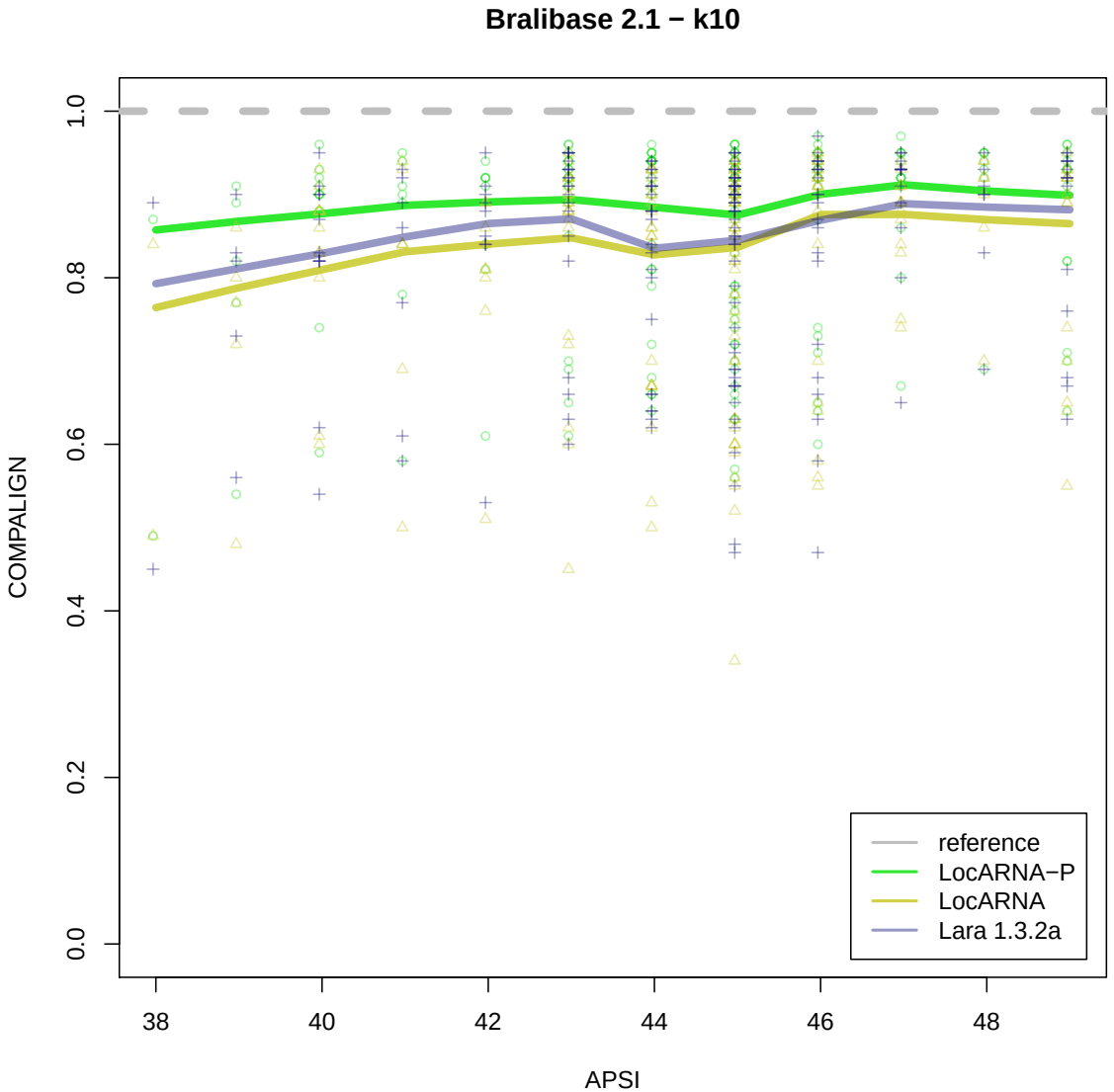
$$Pr\left[\;\right] = \sum_{\substack{(\mathcal{A},\mathcal{S}) \text{ with} \\ i \sim k \in \mathcal{A}_s}} Pr[(\mathcal{A},\mathcal{S})]$$



$$Pr\left[\;\right] = \sum_{\substack{(\mathcal{A},\mathcal{S}) \text{ with} \\ (i,j) \sim (k,l) \in \mathcal{S}}} Pr[(\mathcal{A},\mathcal{S})]$$



LocARNA-P Improves Multiple Alignment



10-fold alignment, low sequence identity $\leq 50\%$

Local Alignment Quality: Reliability Profiles

Goal measure reliability of alignment columns

Method sum pairwise match probabilities

Result Reliability profile (structure # + sequence *)

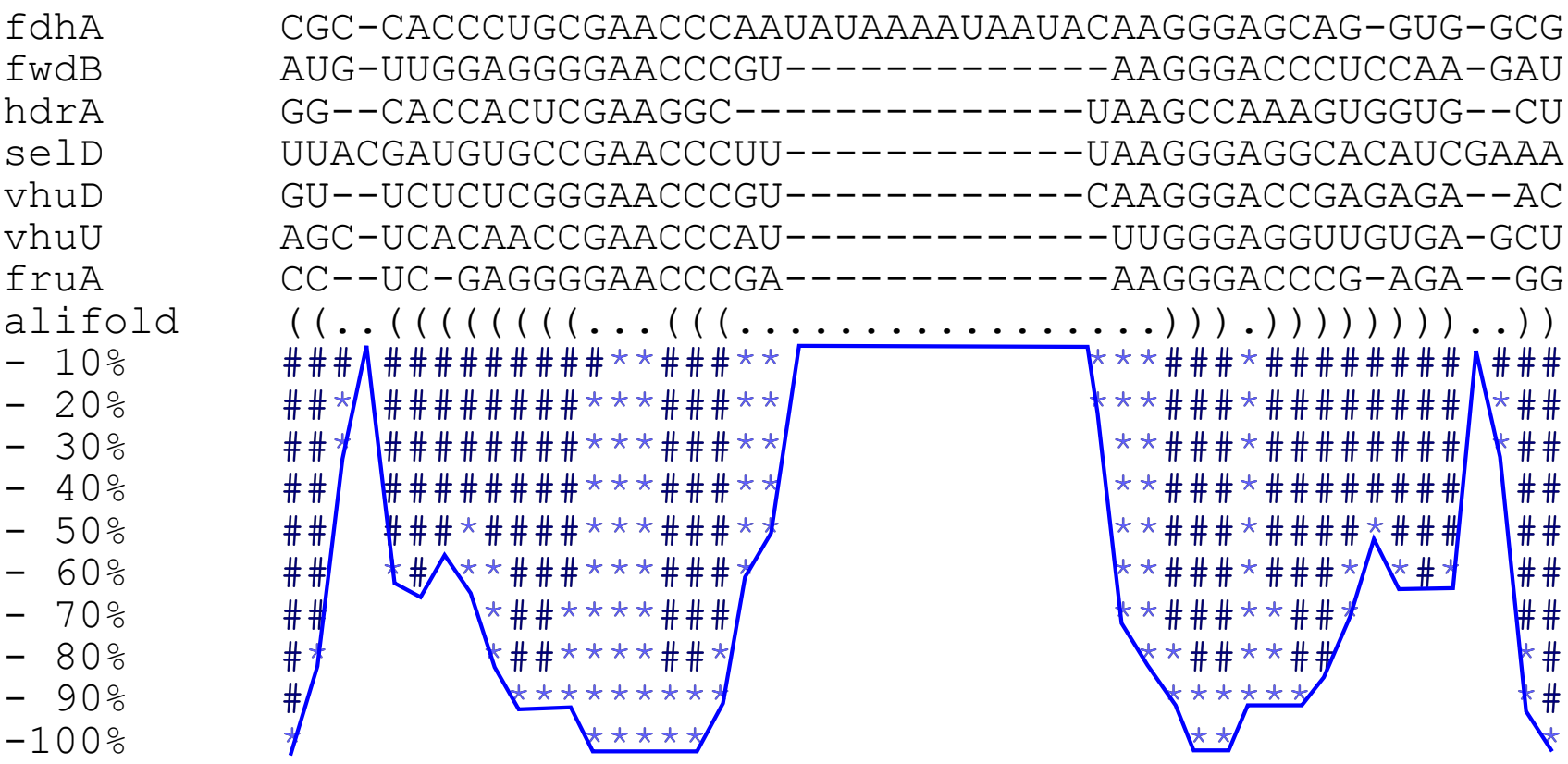
```
fdhA      CGC-CACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-GUG-GCG
fwdB      AUG-UUGGAGGGGAACCCGU-------------AAGGGACCCUCCAA-GAU
hdrA      GG--CACCACUCGAAGGC--------------UAAGCCAAAGUGGUG--CU
selD      UUACGAUGUGCCGAACCCUU------------UAAGGGAGGCACAUCGAAA
vhuD      GU--UCUCUCGGGAACCCGU------------CAAGGGACCGAGAGA--AC
vhuU      AGC-UCACAACCGAACCCAU------------UUGGGAGGUUGUGA-GCU
fruA      CC--UC-GAGGGGAACCCGA------------AAGGGACCCG-AGA--GG
alifold   ((..(((((((((...(((.................))).))))))))).))
- 10%      ### ##########*****             ******#######  ###
- 20%      ##* ##########******            ******#######  *##
- 30%      ##* ##########******              *****#######  *##
- 40%      ##  ##########******              *****#######   ##
- 50%      ##  ###*#########******           *****#####*##   ##
- 60%      ##  *# **#########*              *******###* **   ##
- 70%      ##      *########                *********       ##
- 80%      #*      *########*                ******         *#
- 90%      #        *********                 ******        *#
-100%      *           *****                   **           *
```

Local Alignment Quality: Reliability Profiles

Goal measure reliability of alignment columns

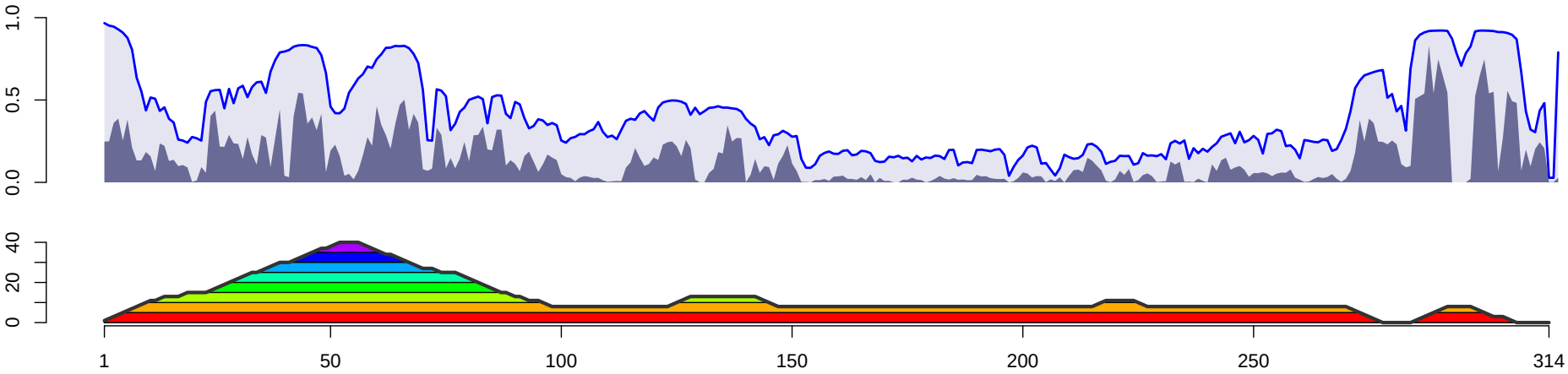
Method sum pairwise match probabilities

Result Reliability profile (structure # + sequence *)



Reliability Profile: Case Study 7SK

- Here: realign hand curated alignment
- Reliability profile fits with experience from manual alignment



dark blue structure reliability

light blue sequence reliability

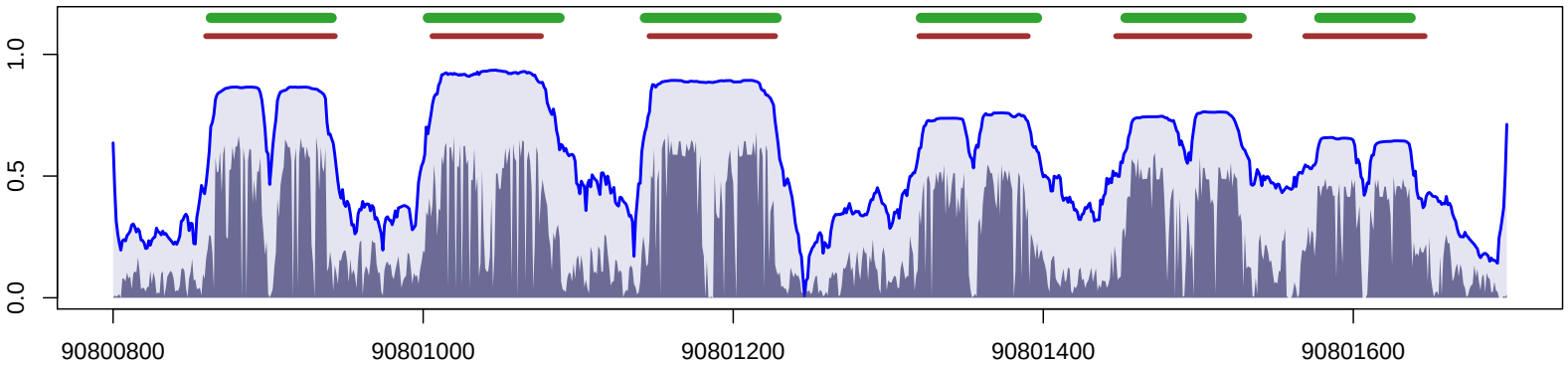
blue line total reliability = structure + sequence

rainbow colored plot mountain plot, shows RNA structure

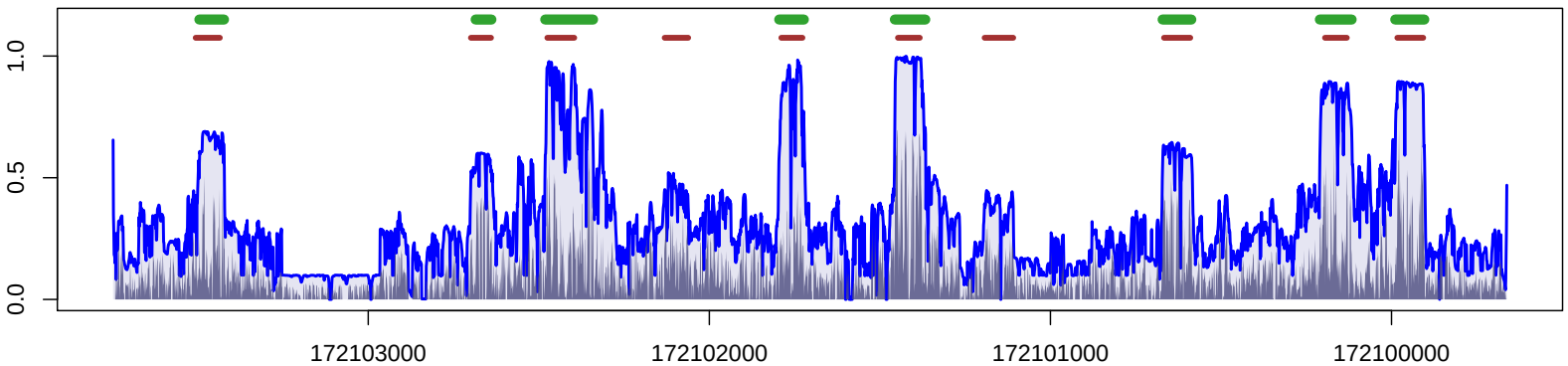
Reliability Profiles

- genomic cluster with known ncRNAs
- align corresponding regions in 10/5 vertebrates
- show reliability profile for human DNA

cluster of 6 micro RNAs, length ≈900

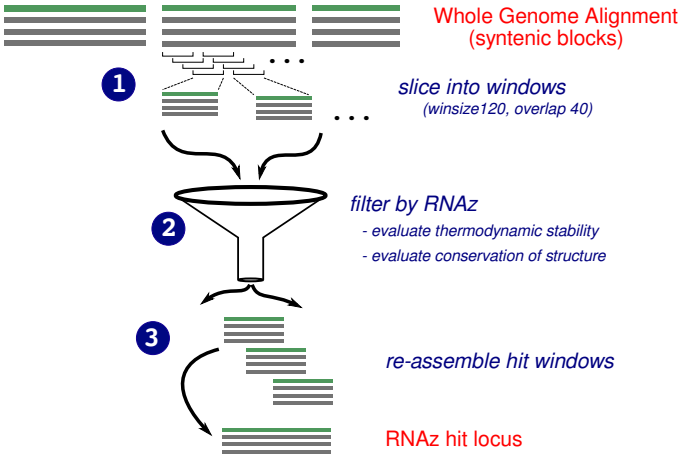


cluster of 10 CD-Box snoRNAs 'GAS5', length ≈4000



green = LocARNA-P prediction; red = ncRNA annotation

Reliabilities for Refining the Drosophilids ncRNA Screen



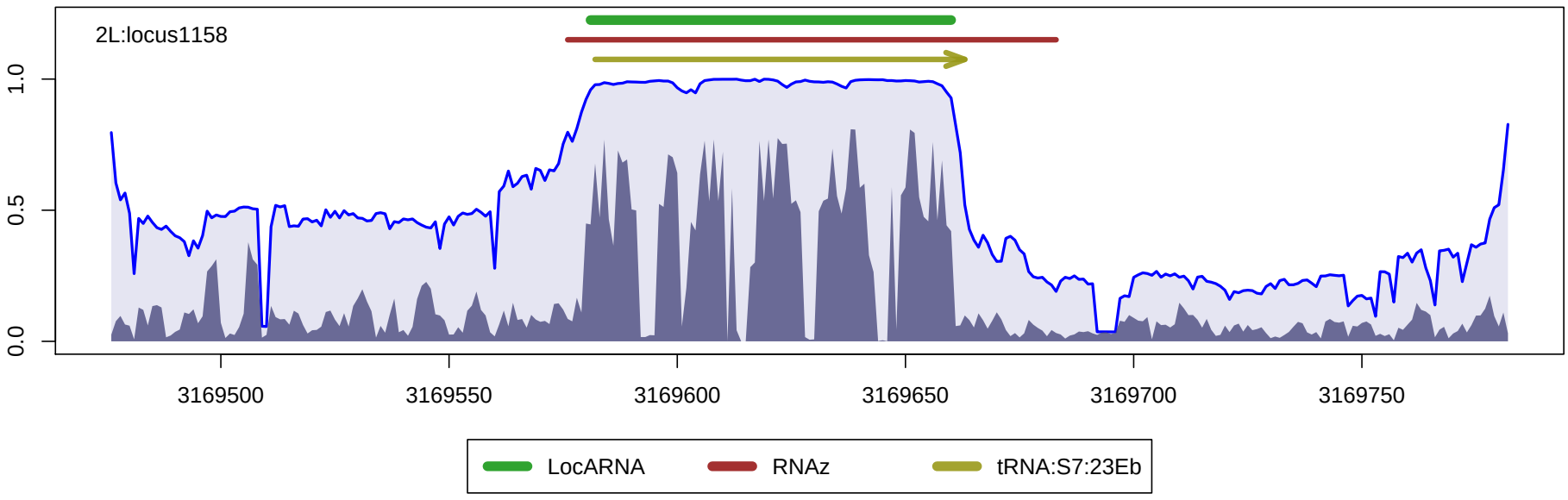
Rose *et al.* Computational RNomics of Drosophilids, BMC Genomics, 2007.

12 Drosophilid genomes alignment, RNAz: 120nt windows in 40nt, combine windows into loci

NEW: Analyze hits using reliabilities

- Realign with context (100 up, 100 down) and reliability profile
- Predict boundaries of ncRNA $\Rightarrow$ **exact location**
- Compute reliability score $\Rightarrow$ **improve predictive power**
- Benchmark: 301 RNAz loci annotated as ncRNA

LocARNA-P Refines Drosophilids ncRNA Screen

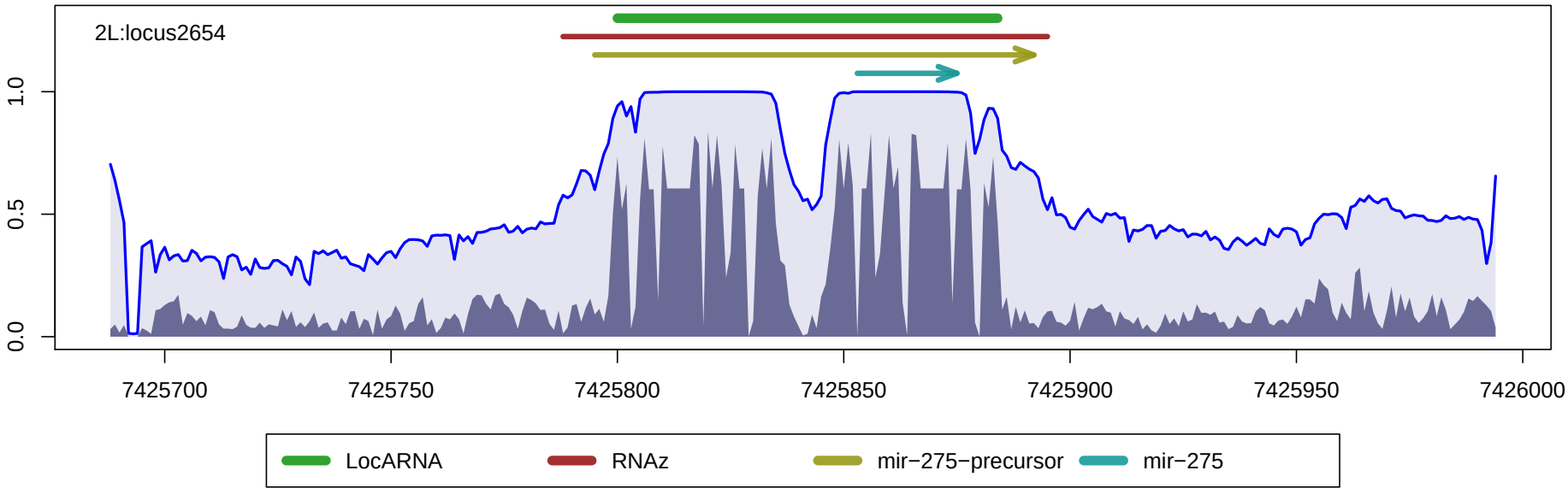


dark blue structure reliability

light blue sequence reliability

blue line total reliability = structure + sequence

LocARNA-P Refines Drosophilids ncRNA Screen

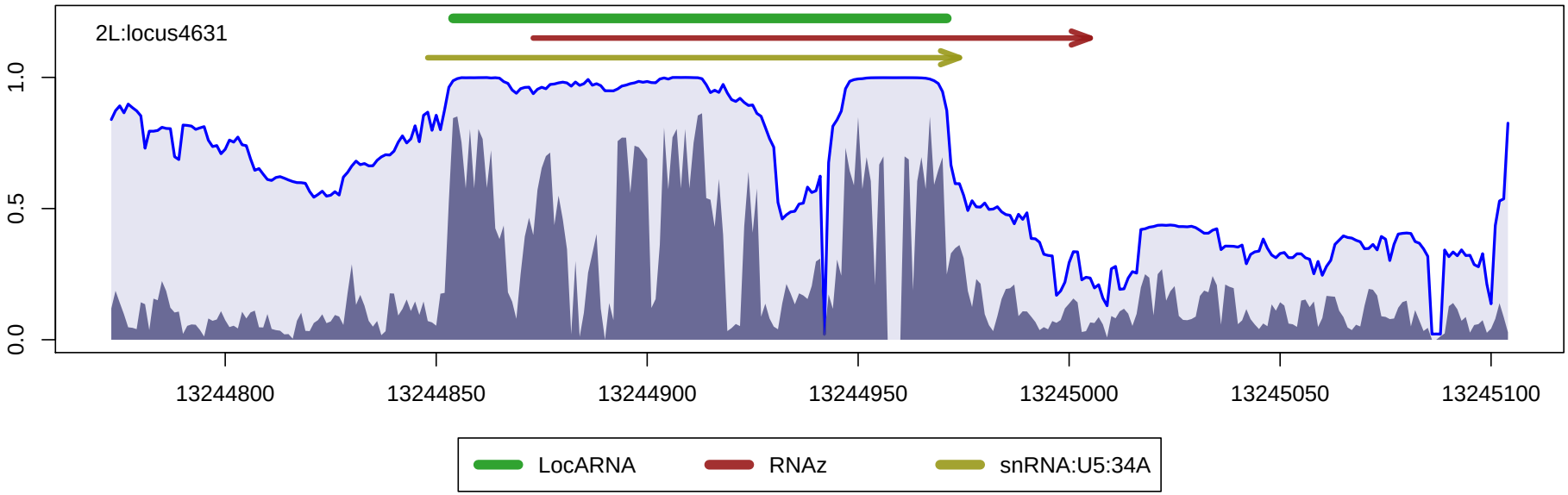


dark blue structure reliability

light blue sequence reliability

blue line total reliability = structure + sequence

LocARNA-P Refines Drosophilids ncRNA Screen

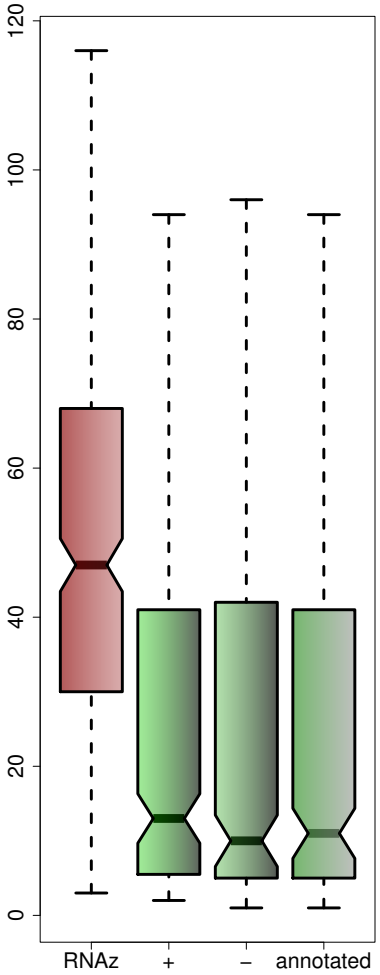


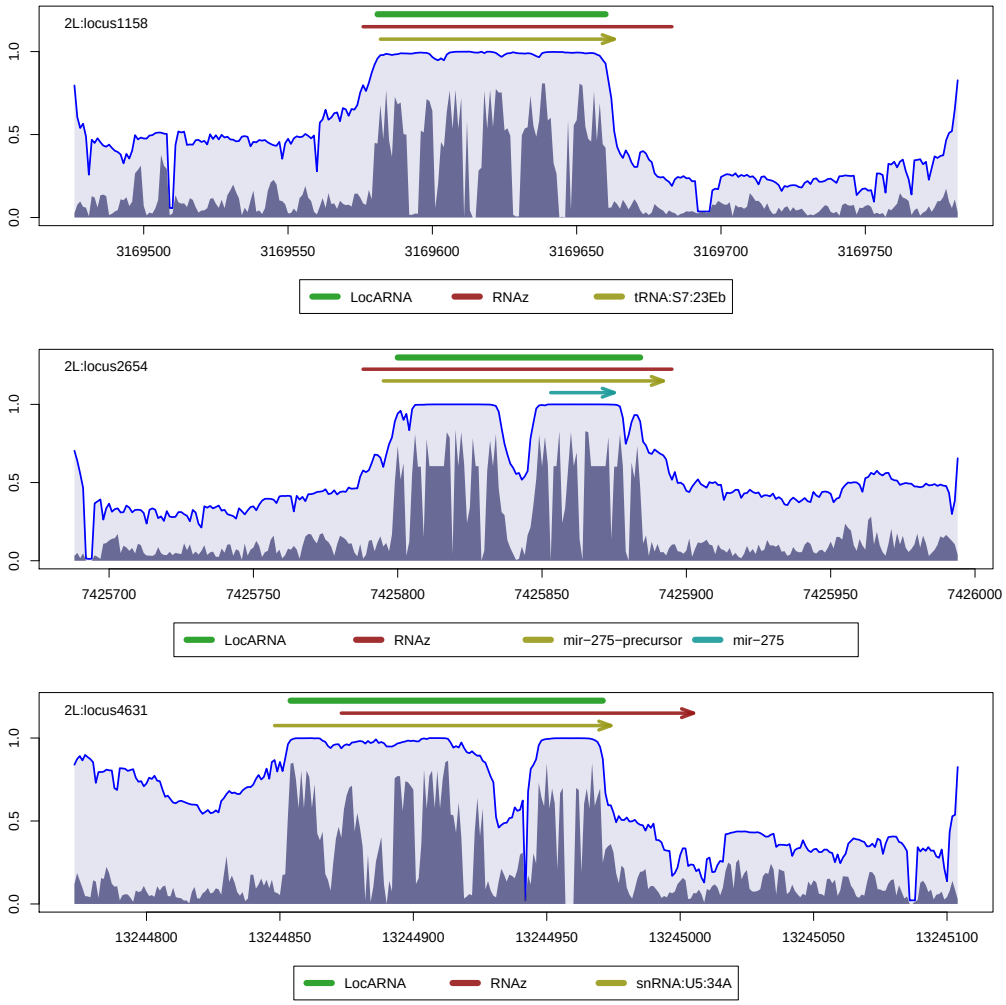
dark blue structure reliability

light blue sequence reliability

blue line total reliability = structure + sequence

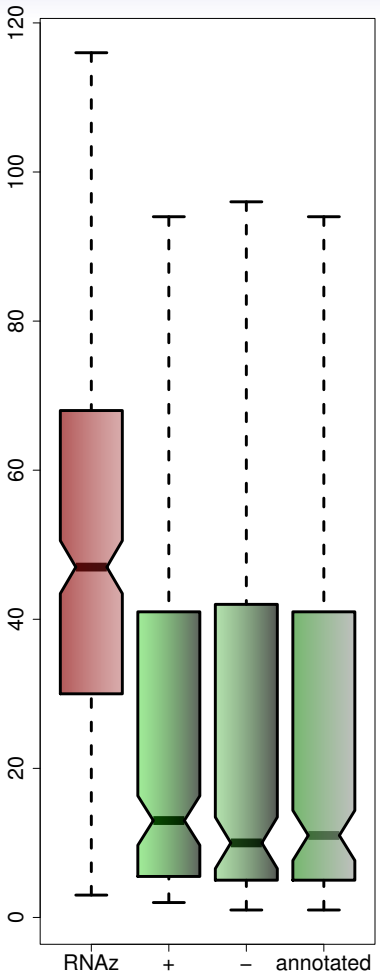
Significant Improvement of Boundaries

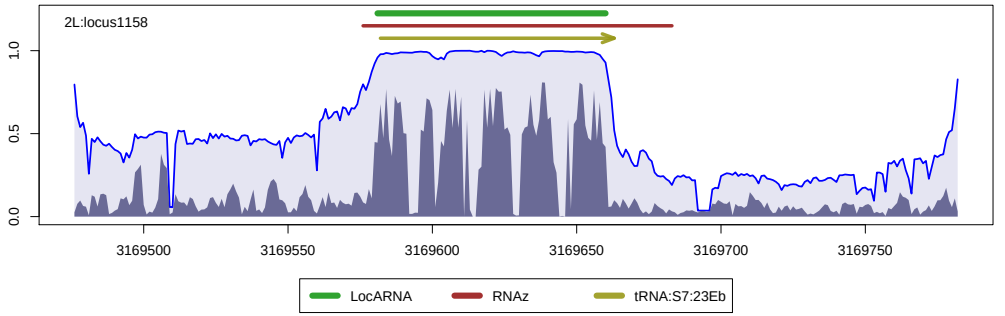


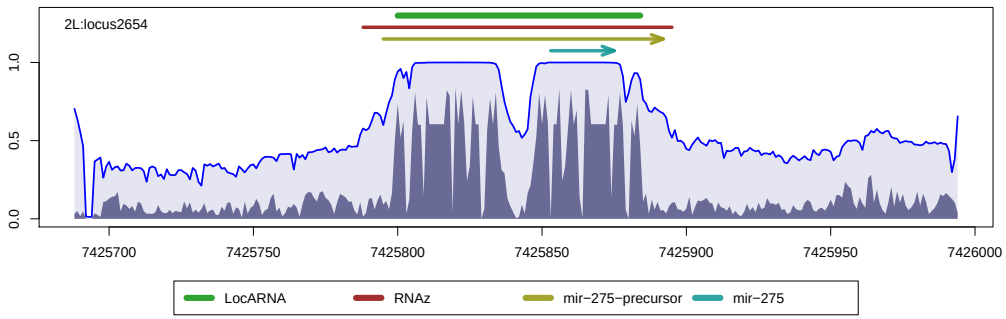


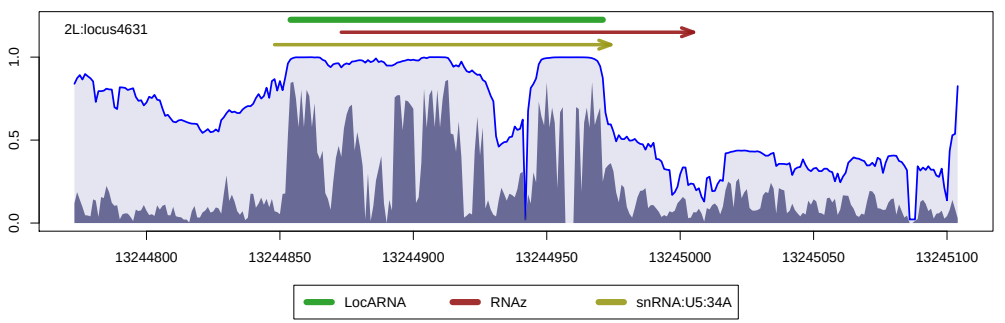
S.Will

Significant Improvement of Boundaries



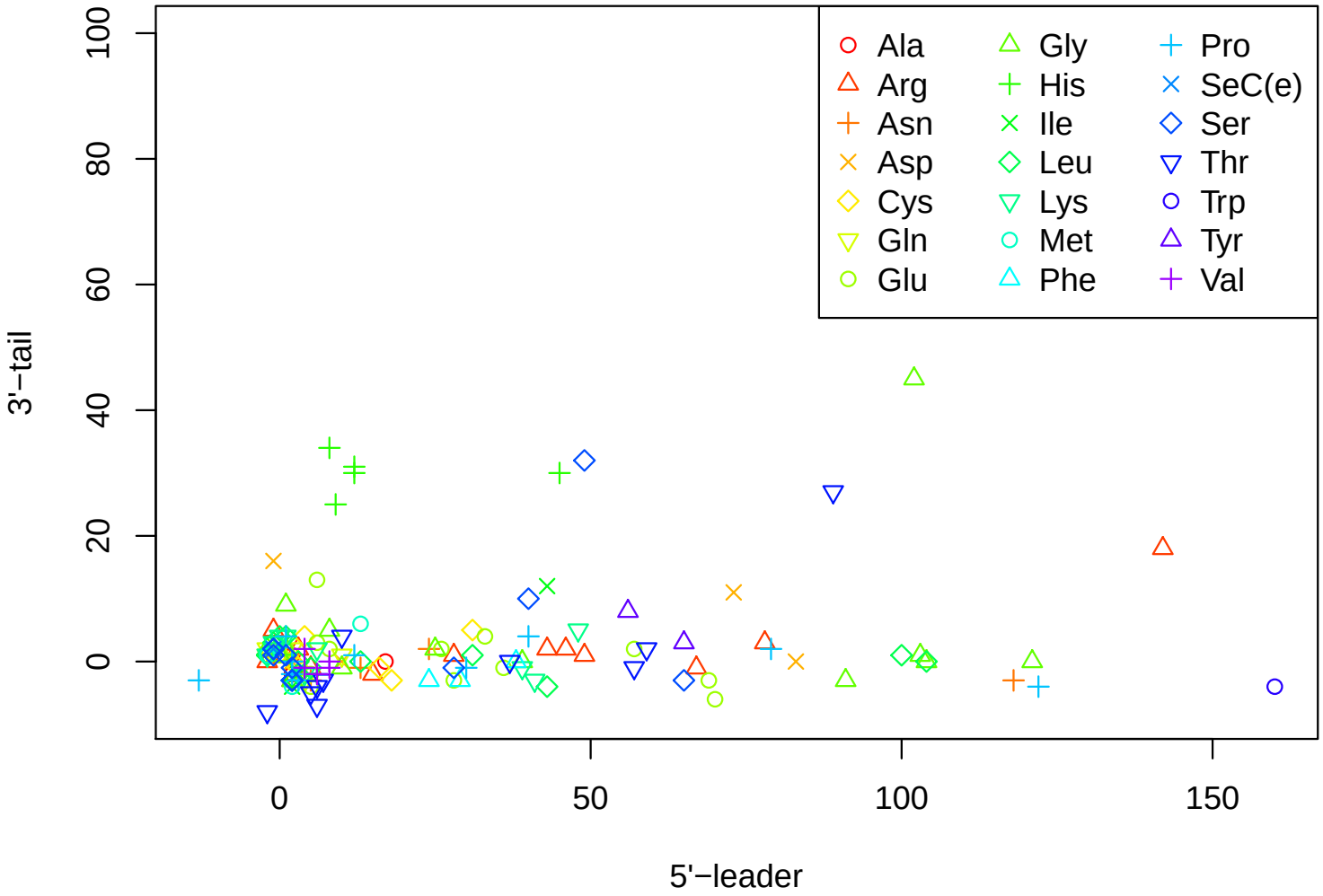






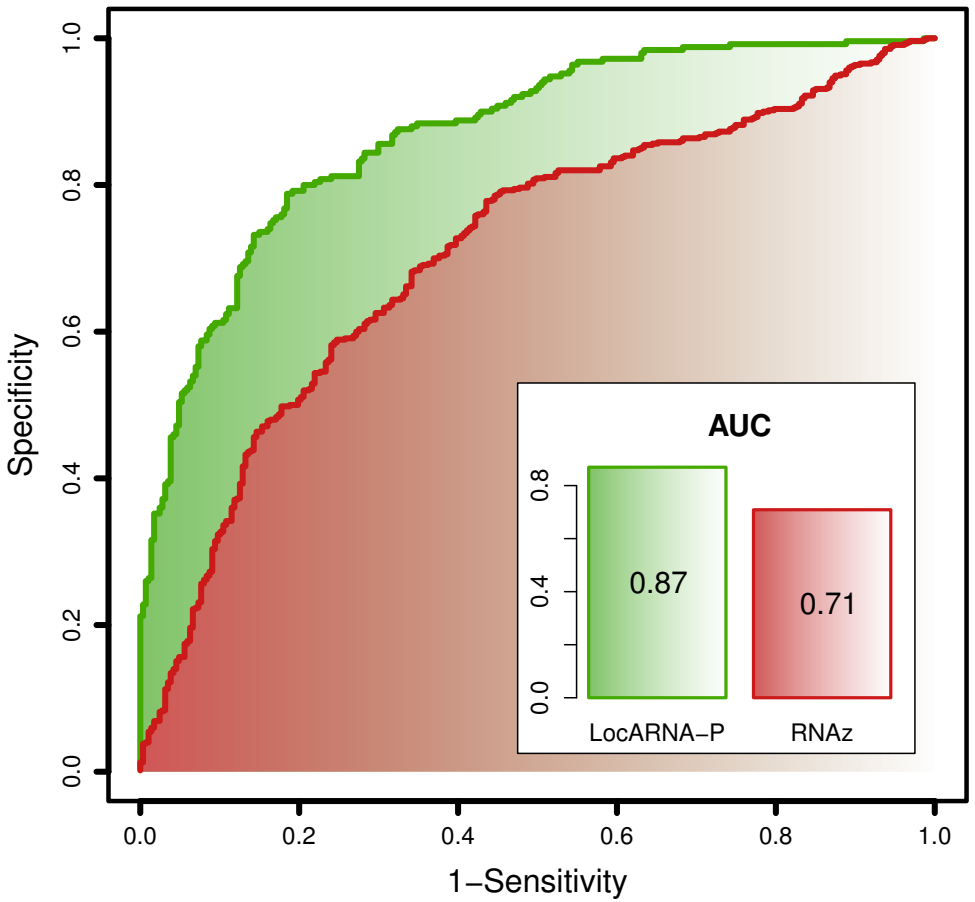
Moreover: many deviations have a biological reason
for example: tRNA

Deviation Reveals 5' Signal for tRNAs



Non-random difference of deviation distributions in header and tail indicates signal!

LocARNA-P Improves Predictive Power of Predictions



- positive set = Rfam annotated RNAz hits
- negative set by shuffling of RNAz hits
- LocARNA-P: reliability score vs. RNAz max. P score

Partition Functions by Inside Algo

Partition Function $Z = \sum \exp(\beta \, \text{score}(\mathcal{A}, \mathcal{S}))$

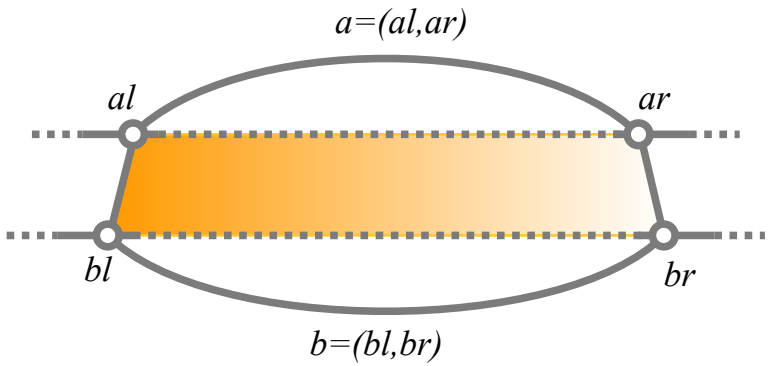
: $\mathcal{A}$ alignment, $\mathcal{S}$ structure

Inside PFs $Z(a, b) =$ Partition function, where

$a = (al, ar) \quad \sim \quad b = (bl, br)$

inside $[al..ar]$ and $[bl..br]$

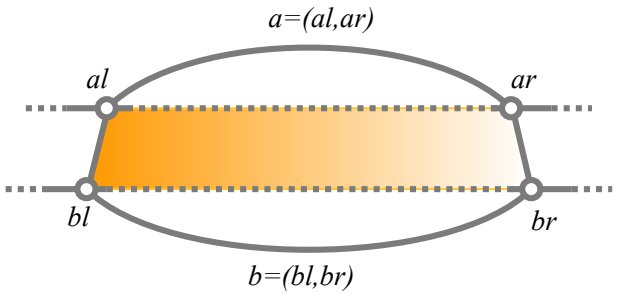
$Z(a, b) =$



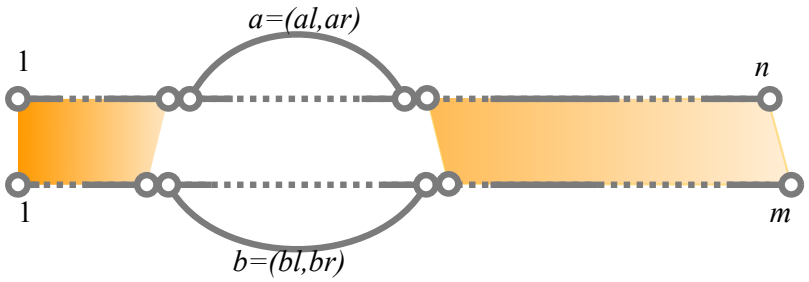
Computation of Match Probabilities

Requires inside and outside partition functions

Inside $Z(a, b)$:



Outside $Z'(a, b)$:

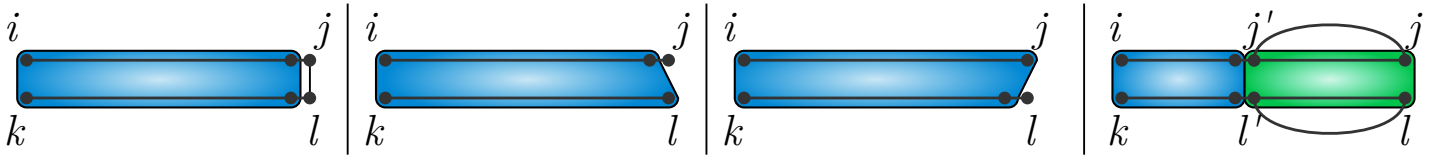


Structural Match probability $Pr[a \sim b] = Z(a, b) \cdot Z'(a, b)/Z$

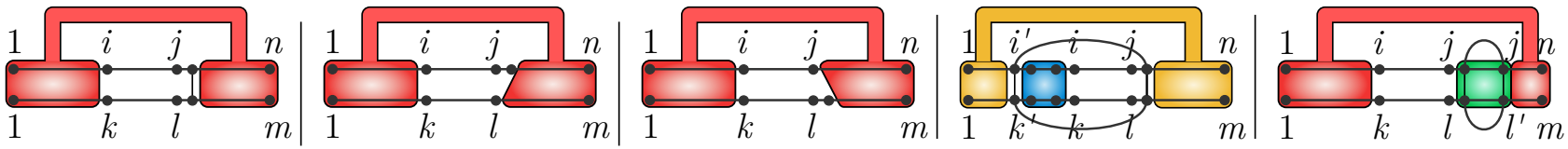
Sequence Match probability slightly more complex
(sum over lots of cases)

Inside and Outside by DP

Inside Decomposition



Outside Decomposition



Complexity

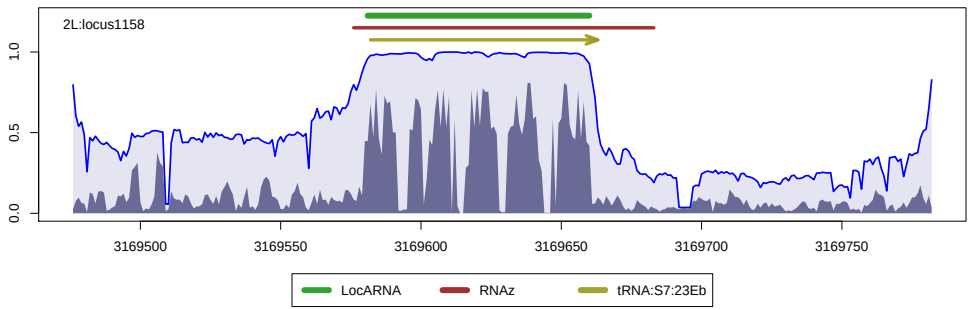
naïve:	$O(n^6)$ time, **$O(n^4)$ space**
LocARNA-P:	$O(n^4)$ time, **$O(n^2)$ space**

Take home: LocARNA-P

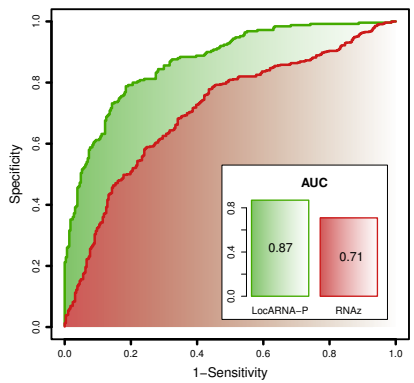
- Reliability Profiles due to sequence-structure alignment

$$Pr\left[\begin{array}{c} i \\ \\ k \end{array}\right] \text{ and } Pr\left[\begin{array}{cc} i & j \\ k & l \end{array}\right]$$

- Refines de-novo prediction of ncRNA
  - Boundaries



  - Predictive Power



- Potential for broad, general application to RNA analysis

Thank you for your attention

and thanks to these people for working with me on LocARNA(-P)

Freiburg	**Leipzig**	**Wien**
Rolf Backofen	Peter Stadler	Ivo Hofacker
Steffen Heyne	Kristin Reiche	
Tejal Joshi	Wolfgang Otto	
	Michael Siebauer	

More Info+Download:

`http://www.bioinf.uni-freiburg.de/Software/LocARNA/`

Web Server:

`http://rna.informatik.uni-freiburg.de`