TARGET PREDICTION FOR BOX H/ACA SNORNAS

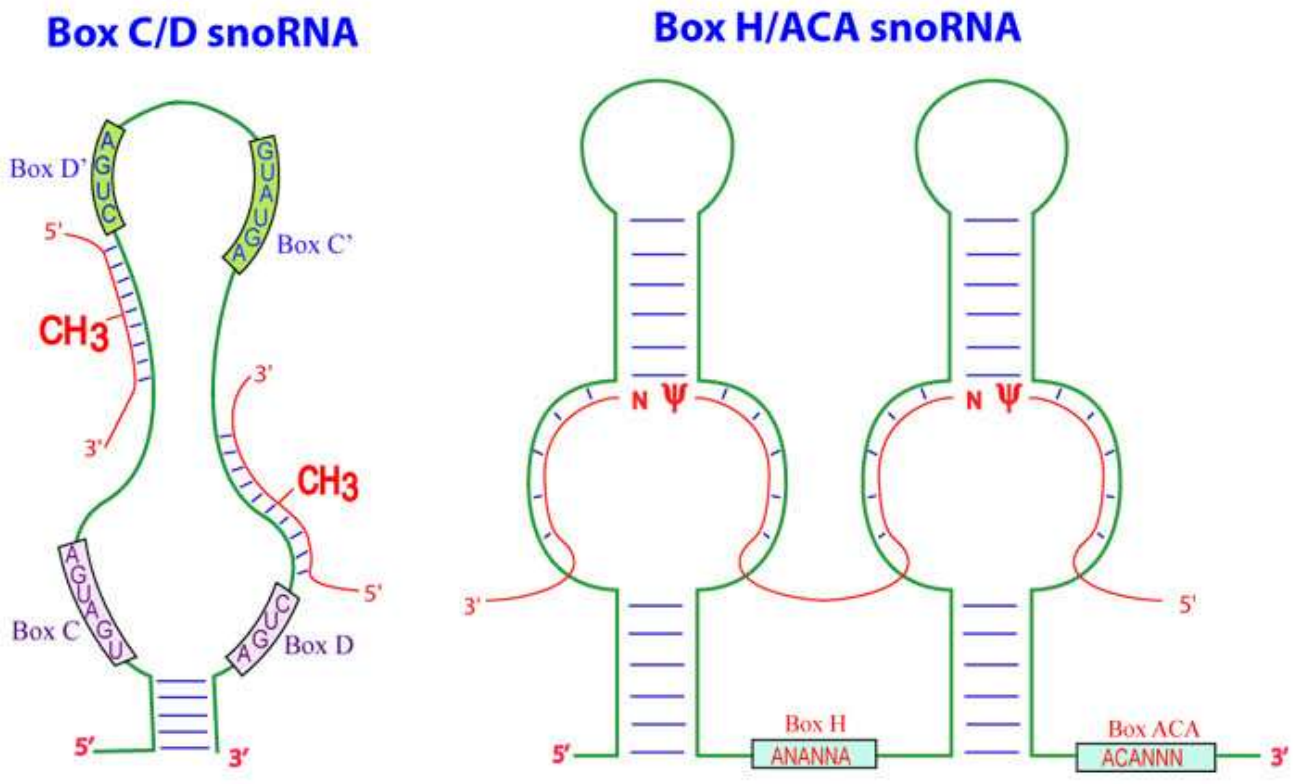
Stephanie Kehr, Hakim Tafer

TBI
University of Vienna
Bioinformatics
University of Leipzig

Bled, 2009

SNORNAS

- CD-box: guide methylation of target RNA
- H/ACA-box: guide pseudouridylation of target RNA

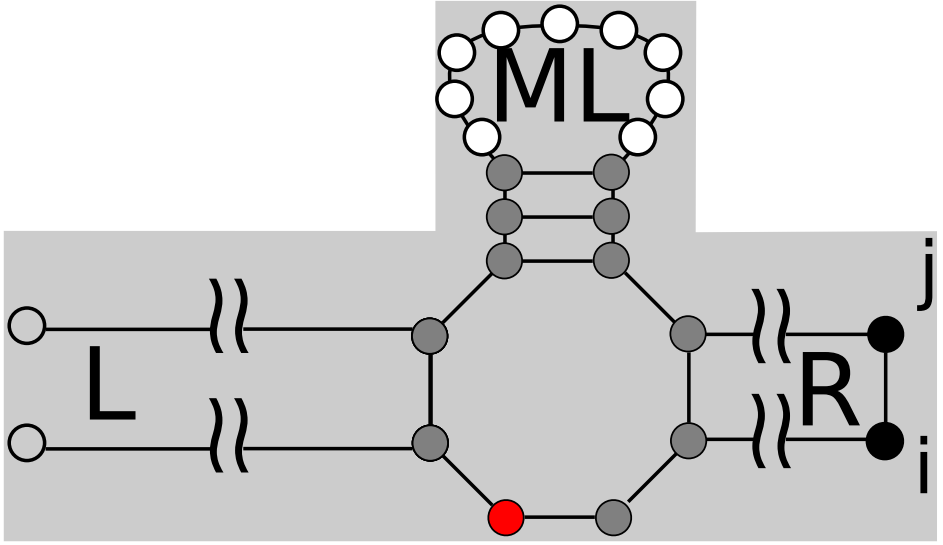


MOTIVATION

- snoRNA with microRNA most often newly detected ncRNA
- pseudouridylation one of most abundant post-transcriptional modifications
- many snoRNAs with unknown targets
- also uridylated-sites with unknown guide snoRNAs
- targets in rRNA, snRNA, tRNA and maybe more???
- until now no sufficent tools for target prediction of HACA-snoRNAs $0(n^3 * m^3)$
- genome-wide search for targets was impossible

THE RNASNOOP ALGORITHM

- we can devide the problem into three parts
  - 1. compute matrix for snoRNA structure
  - 2. compute matrices for left binding site
  - 3. compute matrices for right binding site



THE RNASNOOP ALGORITHM

1. compute matrix for snoRNA structure

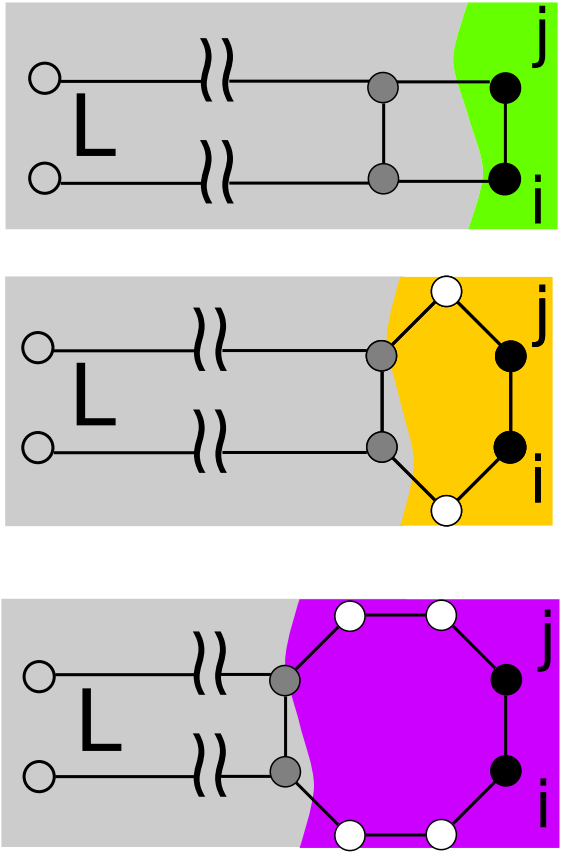
- one fixed matrix for snoRNA structure
- not updated in following steps
- runtime advantage

$$
\begin{aligned}
F_{p,q} &= \min \begin{cases} \mathcal{H}(p,q) \\ F_{p-k,q+l} + \mathcal{I}(p;q;p-k,q+l) \end{cases} \\
C_{i,j} &= \min_{0<k=l<3} \left\{ C_{i-k,j+l} + \mathcal{I}(i,j;i-k,j+l) \right. \\
R_{i,j} &= \min_{0<k=l<3} \begin{cases} R_{i-k,j+l} + \mathcal{I}(i,j;i-k,j+l) \\ \mathit{if}\,(x_{i-2} = U)\, \min_{1\le q\le s} C_{i-3,j+q} + F_{j+1,j+q-1} \end{cases}
\end{aligned}
$$

THE RNASNOOP ALGORITHM

2. compute matrices for left binding site

- (i,j) paired
- (i,j) paired & i-1 , j-1 unpaired
- (i,j) paired & i-1 , j-1 & i-2 , j-2 unpaired
- simple RNAduplex with restriction, only symetrical unpaired regions with maximum length 2

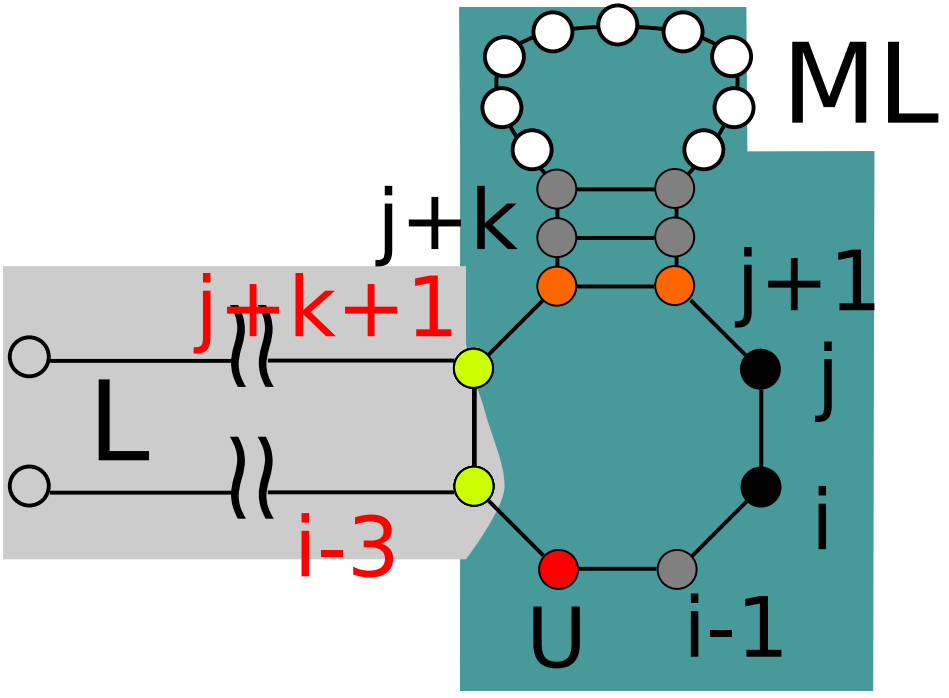


THE RNASNOOP ALGORITHM

3. compute matrix for right binding site

3.1. find possible starting point

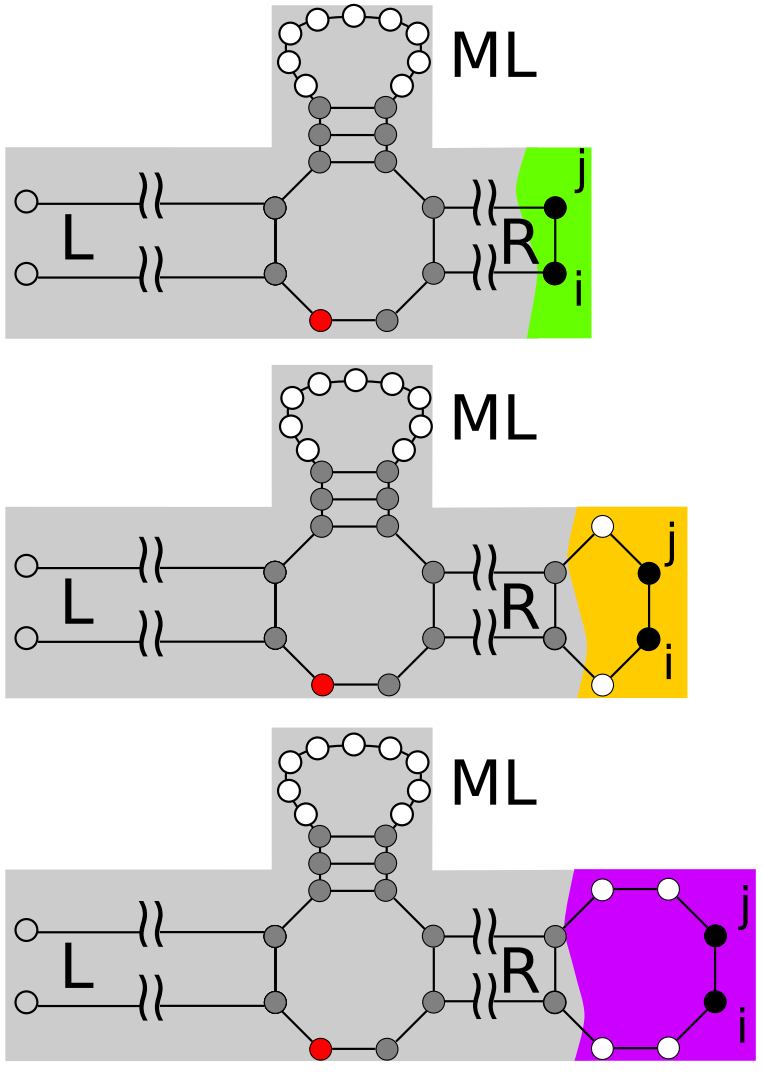
- conditions to start right binding
  - (i,j) must be paired (begin of right binding)
  - i-2 must be Uracil (pseudouridylation)
  - (j+1,j+k) paired (stem)
  - (j+k+1,i-3) paired (end of left binding)



THE RNASNOOP ALGORITHM

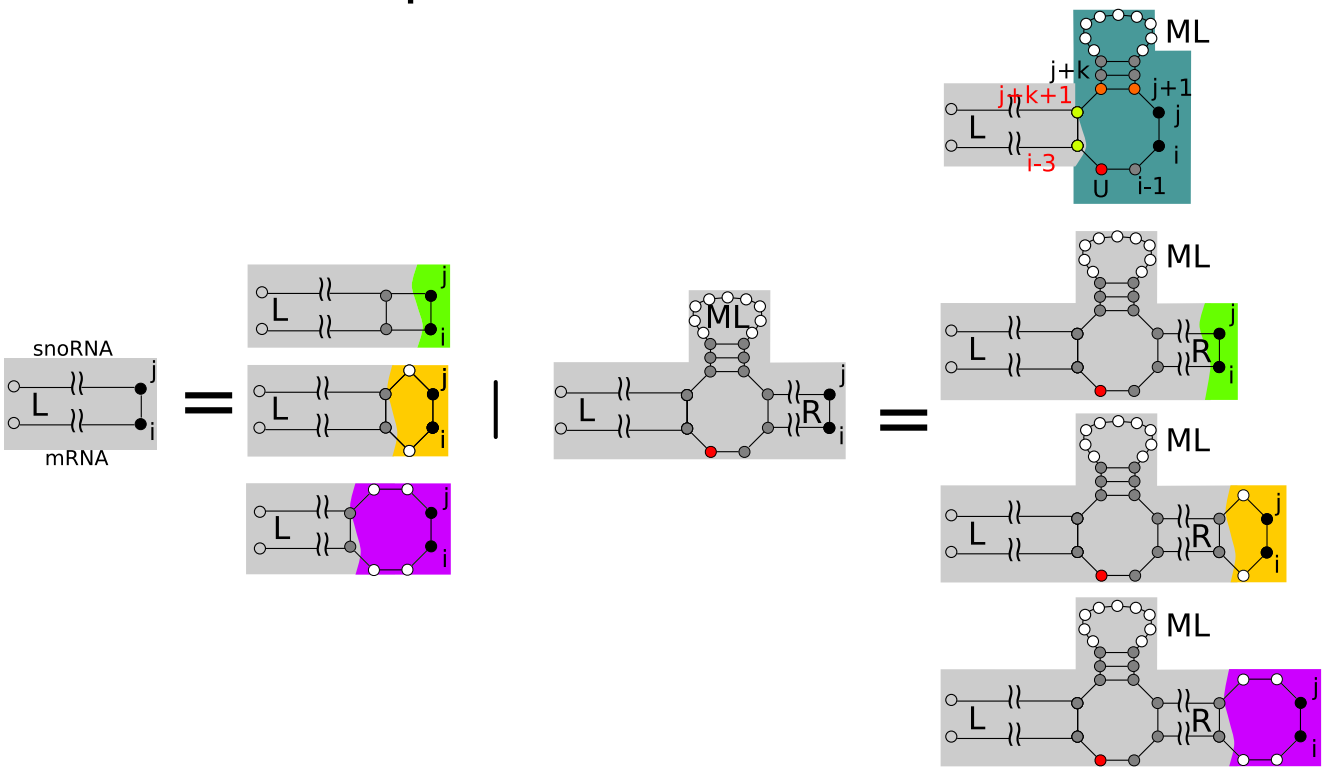
3. compute matrices for right binding site

- (i,j) paired
- (i,j) paired & i-1 , j-1 unpaired
- (i,j) paired & i-1 , j-1 & i-2 , j-2 unpaired



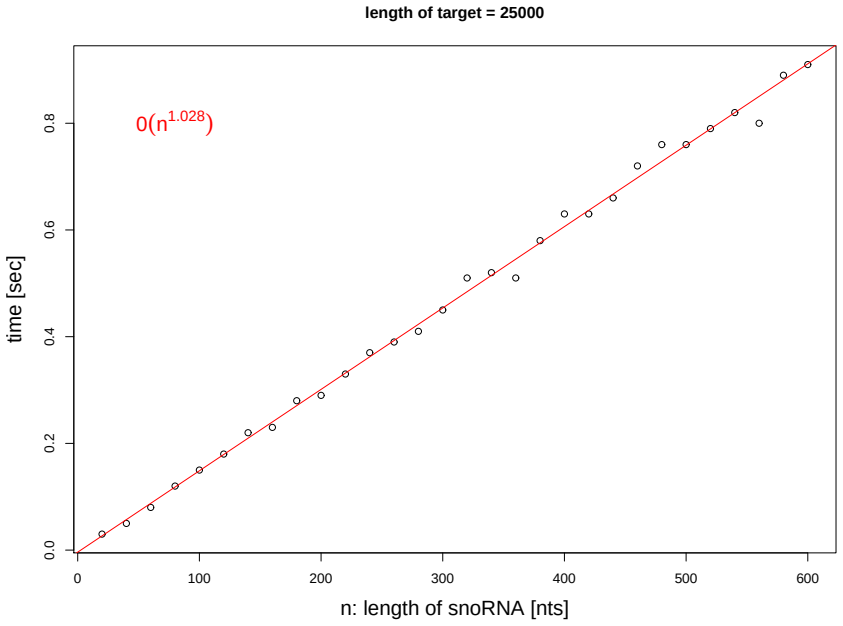
THE RNASNOOP ALGORITHM

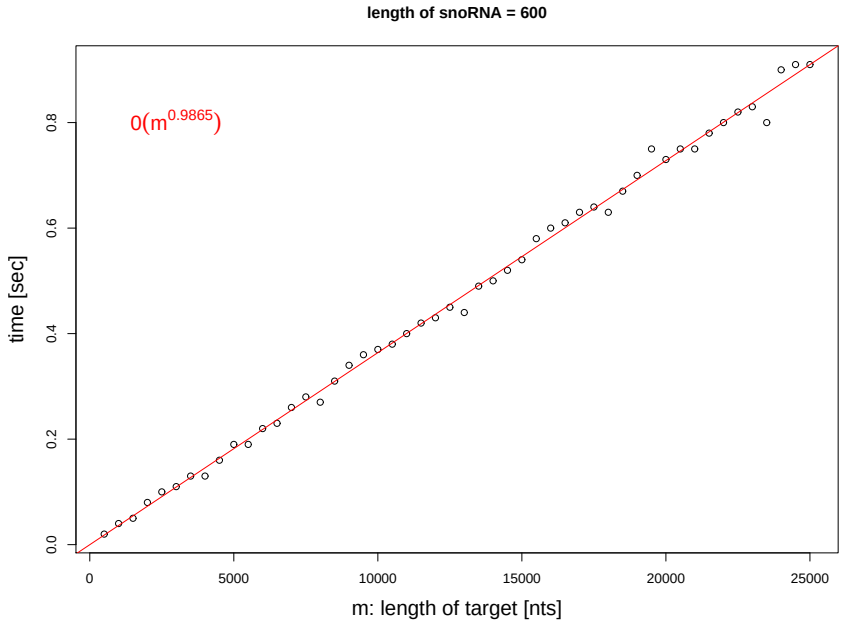
- 3 matrices (ML, L, R)
- use lowest energy value in matrix R
- get structure by backtracking
- suboptimal solutions possible



PERFORMANCE

RUNTIME:





- proportional to length of snoRNA and length of target!! $O(n * m)$

PERFORMANCE

MEMORY:

- expected $O(n * m)$
- better $O(n)$ because you need at most 5 nts of target RNA (start of right binding)
- 5xn matrix can be kept in cache which means another runtime advantage
- best energy value of each column is stored in a list to recover best structure
- 20% faster with this option

- genome-wide target search is possible

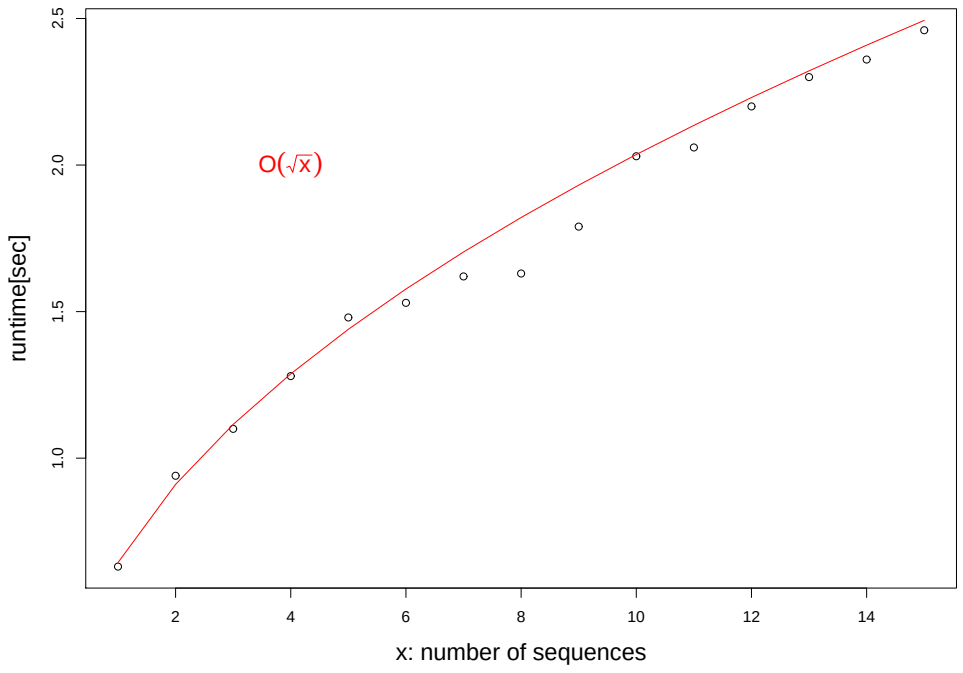
EXTENSION FOR ALIGNMENTS

- interaction length is very short (12 -20 nts) so possiblity of finding a random hit in genome-wide search is quite high
- interaction is well conserved over species it is an indication for correctness of prediction
- especially if we find compensatory mutations

- computation similar to RNAalifold

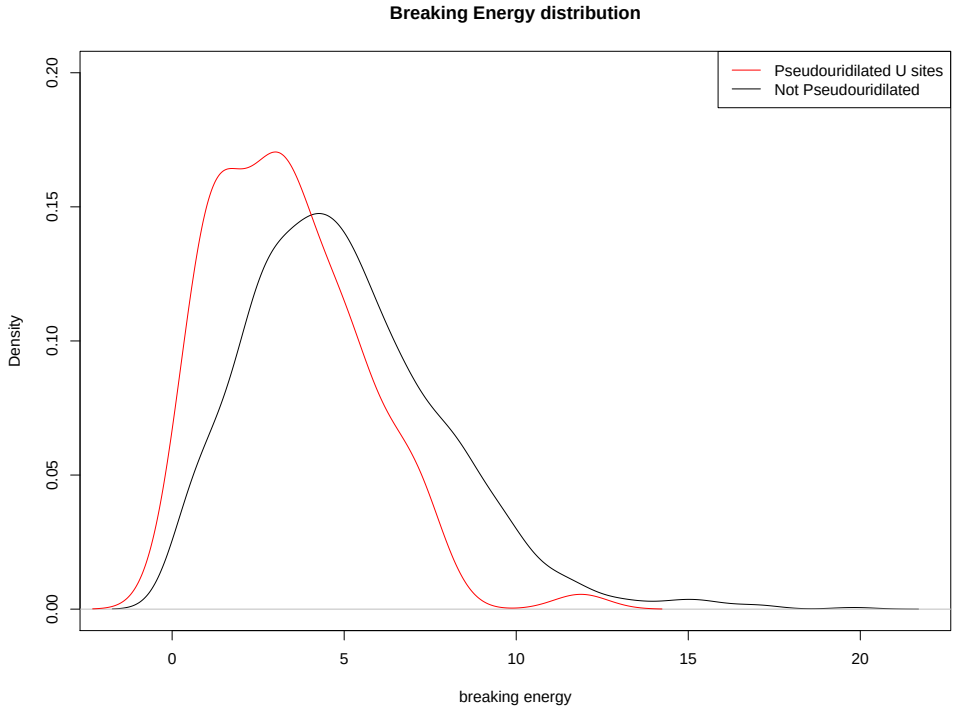
EXTENSION FOR ALIGNMENTS

RUNTIME:



- length of target-sequences = 25000 nts
- length of snoRNA-stem-sequences = 61 nts
- runtime about $0(\sqrt{x})$

ACCESSIBILITY



- 6nts in front of U and 6nts after U in rRNA
- in general openig energy for regions around pseudouridilated U is lower

PREDICTIONS FOR YEAST

- all experimentally verified predictions in yeast

snoRNA	PseudoU	snoGPS	RNAsnoop -a	RNAsnoop
snR32	25S-2191	1	1	1
snR82	25S-2349	1	1	1
snR33	25S-1042	1	1	1
snR43	25S-966	1	1	1
snR35	18S-1191	1	1	1
snR5	25S-1124	1	1	1
snR85	18S-1181	1	1	1
snR82	25S-2351	1	1	1
snR46	25S-2865	1	1	1
snR3	25S-2133	1	1	1
snR83	18S-1290	1	1	1
snR189	25S-2735	1	1	1
snR34	25S-2880	1	1	2
snR49	18S-302	1	1	5
snR44	25S-1056	2	1	1
snR34	25S-2826	2	1	1
snR189	18S-466	2	1	1
snR49	18S-120	3	1	1
snR5	25S-1004	3	1	1
snR83	18S-1415	4	1	1
snR3	25S-2129	4	1	1
snR36	18S-1187	12	1	6

snoRNA	PseudoU	snoGPS	RNAsnoop -a	RNAsnoop
snR8	25S-986	55	1	3
snR81	25S-1052	57	1	2
snR82	25S-1109		1	1
snR80	18S-775		1	2
snR84	25S-2266	1	2	1
snR31	18S-999	1	2	1
snR44	18S-106	1	2	2
snR191	25S-2258	1	2	4
snR42	25S-2975	1	2	5
snR86	25S-2314	13	2	4
snR80	18S-759		2	2
snR3	25S-2264	2	3	3
snR49	18S-211	2	3	5
snR161	18S-632	6	3	7
snR49	25S-990	4	4	10
snR8	25S-960	68	4	6
snR161	18S-766	1	6	8
snR10	25S-2923	2	11	7
snR11	25S-2416	3	11	12
snR9	25S-2340	33	32	20
snR191	25S-2260	1		
snR37	25S-2944	1		

PREDICTIONS FOR HUMAN

- all experimentally verified interactions in human

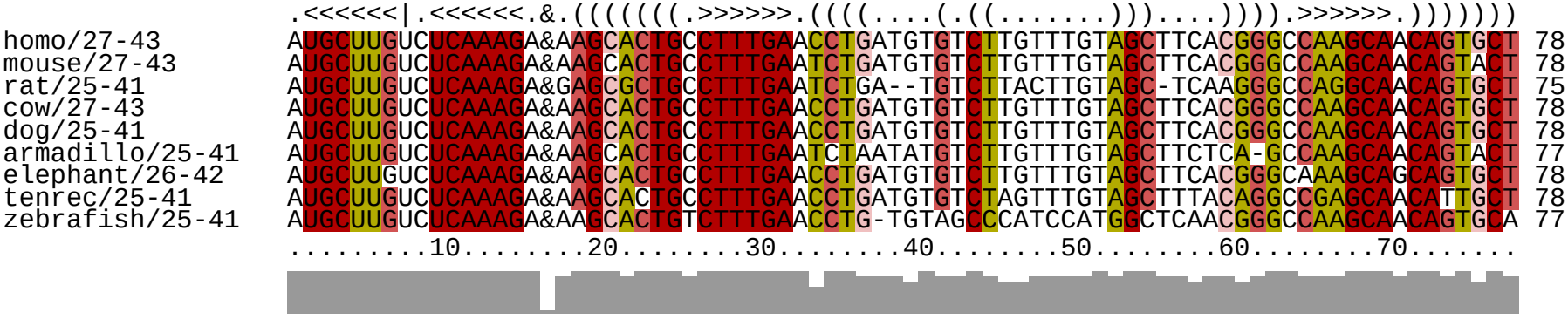
snoRNA	PseudoU	snoGPS	RNAsnoop -a	RNAsnoop	RNAsnoop -A	SVM	posneg
ACA19_1	28S-3709	1	1	1	1	1	+
ACA19_2	28S-3618	25	1	1	2	1	+
ACA24_1	18S-863		1	1	1	0	+
ACA28_1	18S-815	1	1	3	1	1	+
ACA28_2	18S-866		2	2	1	1	+
ACA42_2	18S-109	1	1	1	5	1	+
ACA50_1	18S-34	1	1	1	1	0	+
ACA50_2	18S-105	2	2	2	1	1	+
ACA62_1	18S-34	3	5	21	2	3	+
ACA62_2	18S-105	2	1	1	4	1	+
ACA67_1	18S-572	2	3	1	5	1	+
ACA67_2	18S-109	1	1	1	1	1	+
ACA19_1	18S-863	10	5	2	5	0	-
ACA19_1	18S-863	10	5	2	5	0	-
ACA24_2	18S-612	86	3	3		0	-
ACA42_1	18S-572	3	4	4		0	-

- SVM trained with 144 datapoints in yeast and tested on human

PREDICTIONS FOR HUMAN

ACA50:

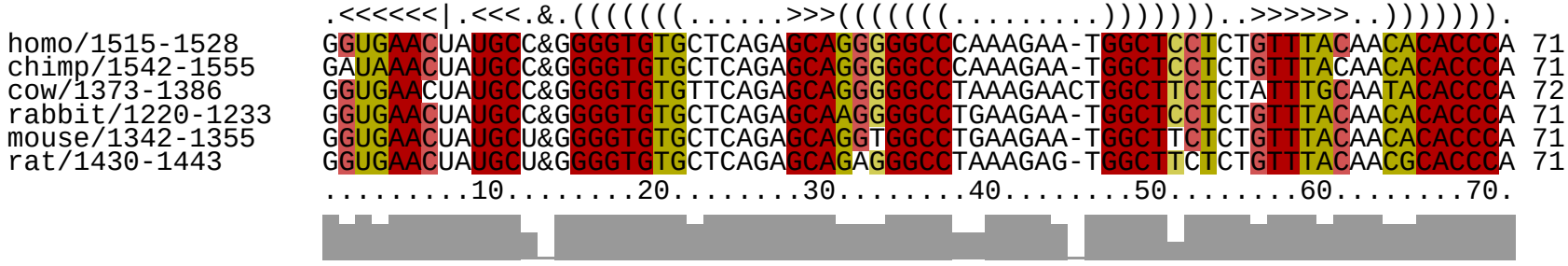
- **18S-34**
- best hit in human
- best hit with alignment-option



PREDICTIONS FOR ORPHAN TARGETS

- pseudouridylated positions in human rRNA with unknown guiding snoRNA

- **28S-1523**
  - only possible solution in human:
    - **U107** (orphan snoRNA)
    - 6 species aligned
    - compensatory mutations in binding-site and in the stem



THANK YOU

Hakim, Jana, Peter, Ivo