SNOBOARD

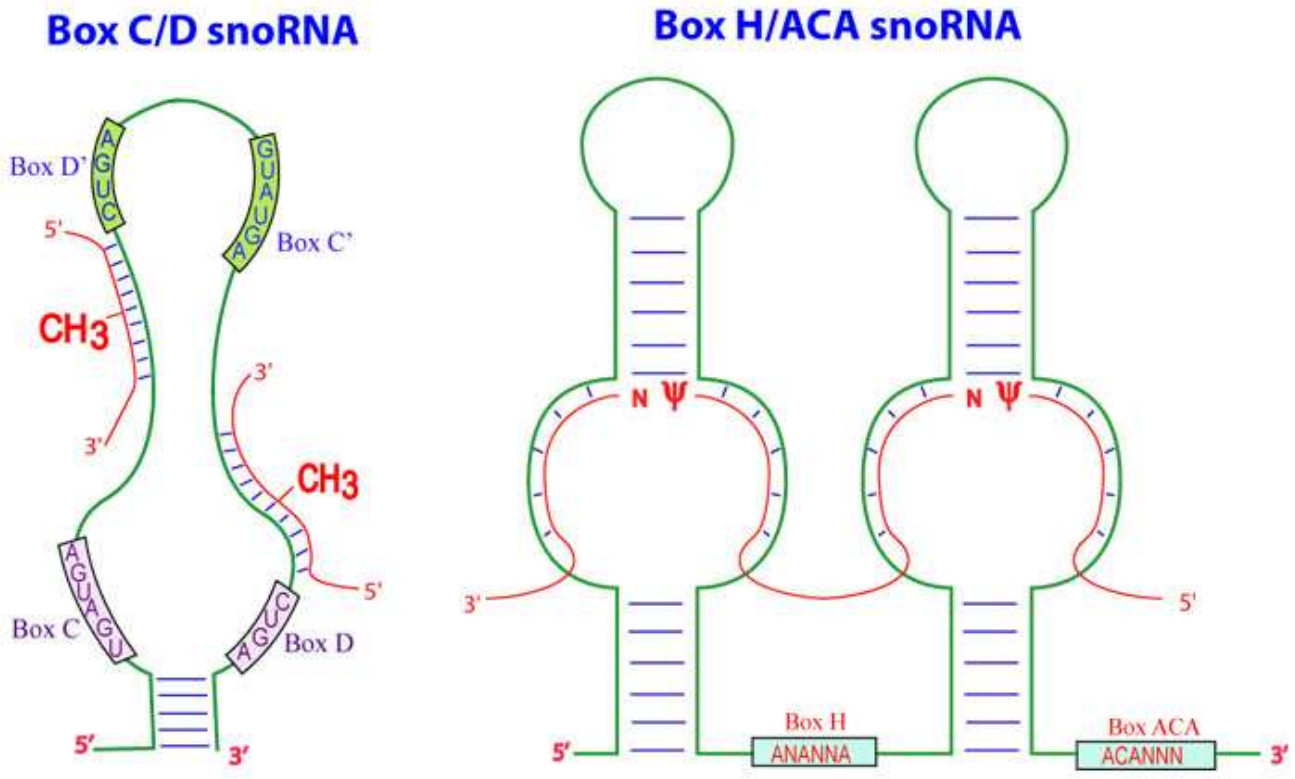
Stephanie Kehr, Sebastian Bartschat

Bioinformatics
University of Leipzig

Bled, Slovenia, 2010

SNORNAS

- CD-box: guide methylation of target RNA
- H/ACA-box: guide pseudouridylation of target RNA
- scaRNAs: hybrid of H/ACA and C/D domains



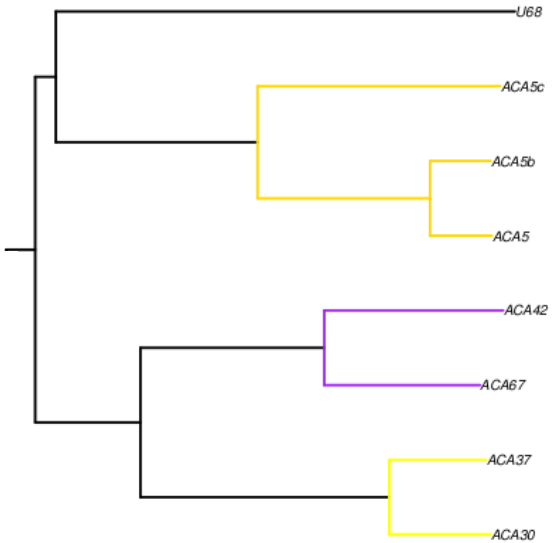
Motivation

- snoRNA with microRNA most often newly detected ncRNA
- pseudouridylation and methylation are highly abundant post-transcriptional modifications
- guide modifications in rRNA, snRNA, tRNA and some mRNAs
- precursors of microRNAs
- orphan snoRNAs with unknown function

⇒ *snoRNA set which is as complete as possible to get further insights into snoRNAs and their evolution*

STARTINGPOINT

- human, chicken, platypus, rhesus, C.elegans
- try to find homologs and paralogs among metazoan
- name conflicts
- paralogs and families??



PROPERTIES

Table ▲	
CD-boxes	
CD-targets	
genome-source	
HACA-boxes	
HACA-targets	
homology	
snoRNA-source	
snoRNAs	
target-source	
9 table(s)	

PROPERTIES

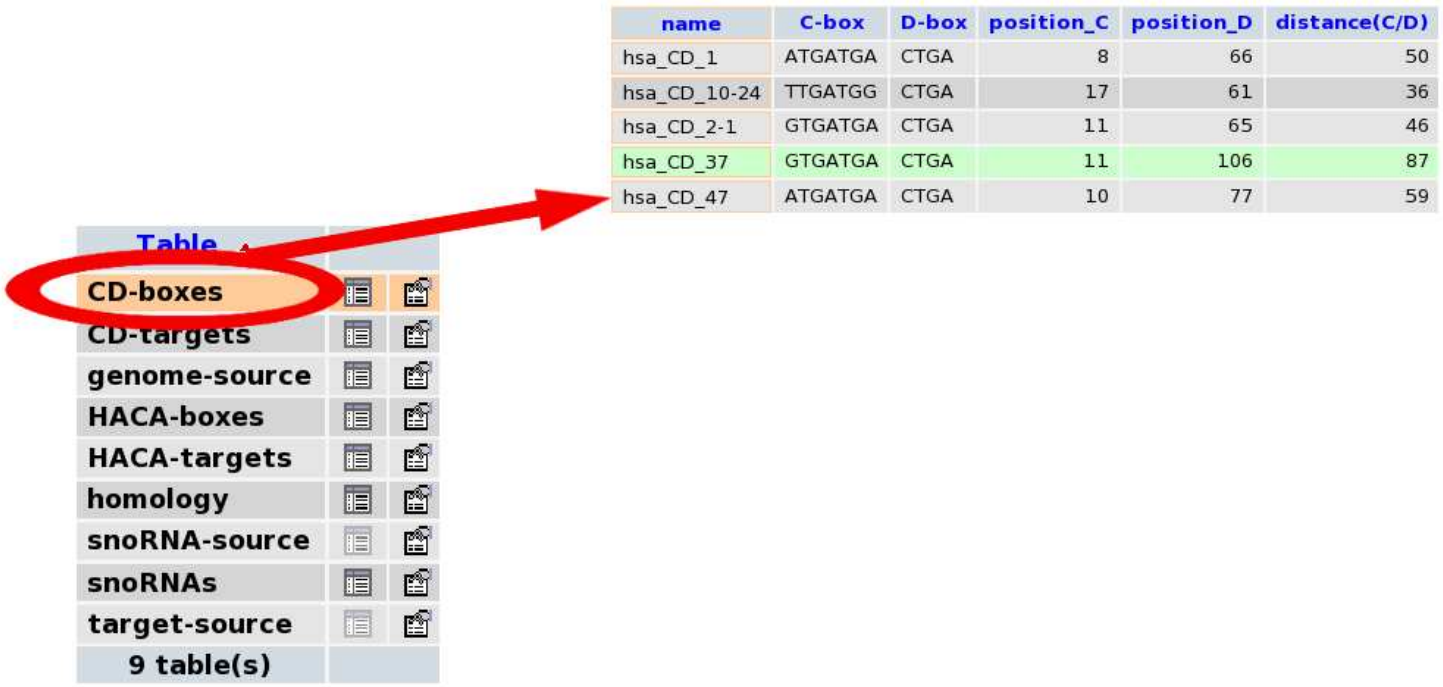
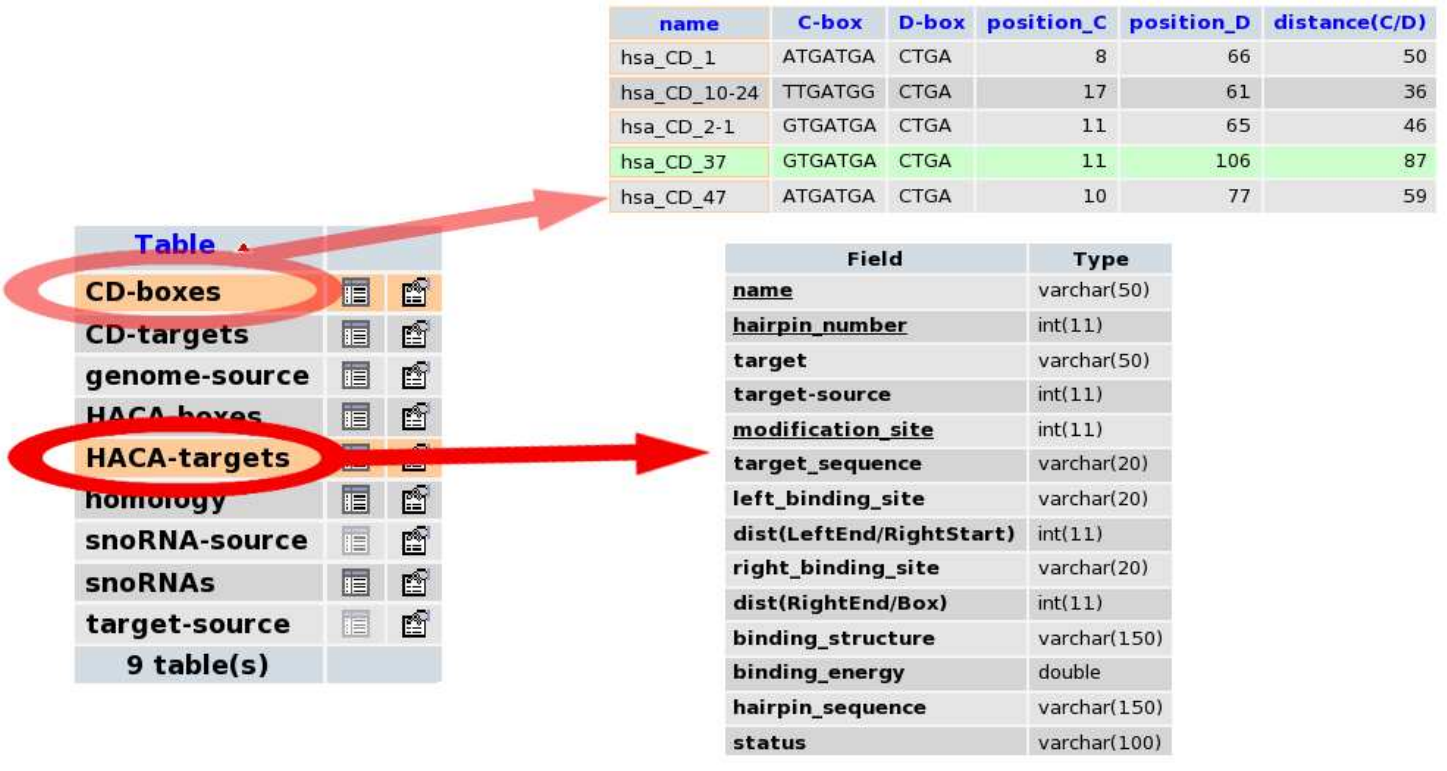


Table
CD-boxes
CD-targets
genome-source
HACA-boxes
HACA-targets
homology
snoRNA-source
snoRNAs
target-source
9 table(s)

name	C-box	D-box	position_C	position_D	distance(C/D)
hsa_CD_1	ATGATGA	CTGA	8	66	50
hsa_CD_10-24	TTGATGG	CTGA	17	61	36
hsa_CD_2-1	GTGATGA	CTGA	11	65	46
hsa_CD_37	GTGATGA	CTGA	11	106	87
hsa_CD_47	ATGATGA	CTGA	10	77	59

PROPERTIES



name	C-box	D-box	position_C	position_D	distance(C/D)
hsa_CD_1	ATGATGA	CTGA	8	66	50
hsa_CD_10-24	TTGATGG	CTGA	17	61	36
hsa_CD_2-1	GTGATGA	CTGA	11	65	46
hsa_CD_37	GTGATGA	CTGA	11	106	87
hsa_CD_47	ATGATGA	CTGA	10	77	59

Table
CD-boxes
CD-targets
genome-source
HACA-boxes
HACA-targets
homology
snoRNA-source
snoRNAs
target-source
9 table(s)

Field	Type
name	varchar(50)
hairpin_number	int(11)
target	varchar(50)
target-source	int(11)
modification_site	int(11)
target_sequence	varchar(20)
left_binding_site	varchar(20)
dist(LeftEnd/RightStart)	int(11)
right_binding_site	varchar(20)
dist(RightEnd/Box)	int(11)
binding_structure	varchar(150)
binding_energy	double
hairpin_sequence	varchar(150)
status	varchar(100)

PROPERTIES

name	C-box	D-box	position_C	position_D	distance(C/D)
hsa_CD_1	ATGATGA	CTGA	8	66	50
hsa_CD_10-24	TTGATGG	CTGA	17	61	36
hsa_CD_2-1	GTGATGA	CTGA	11	65	46
hsa_CD_37	GTGATGA	CTGA	11	106	87
hsa_CD_47	ATGATGA	CTGA	10	77	59

Table
CD-boxes
CD-targets
genome-source
HACA-boxes
HACA-targets
homology
snoRNA-source
snoRNAs
target-sourc
9 table(s)

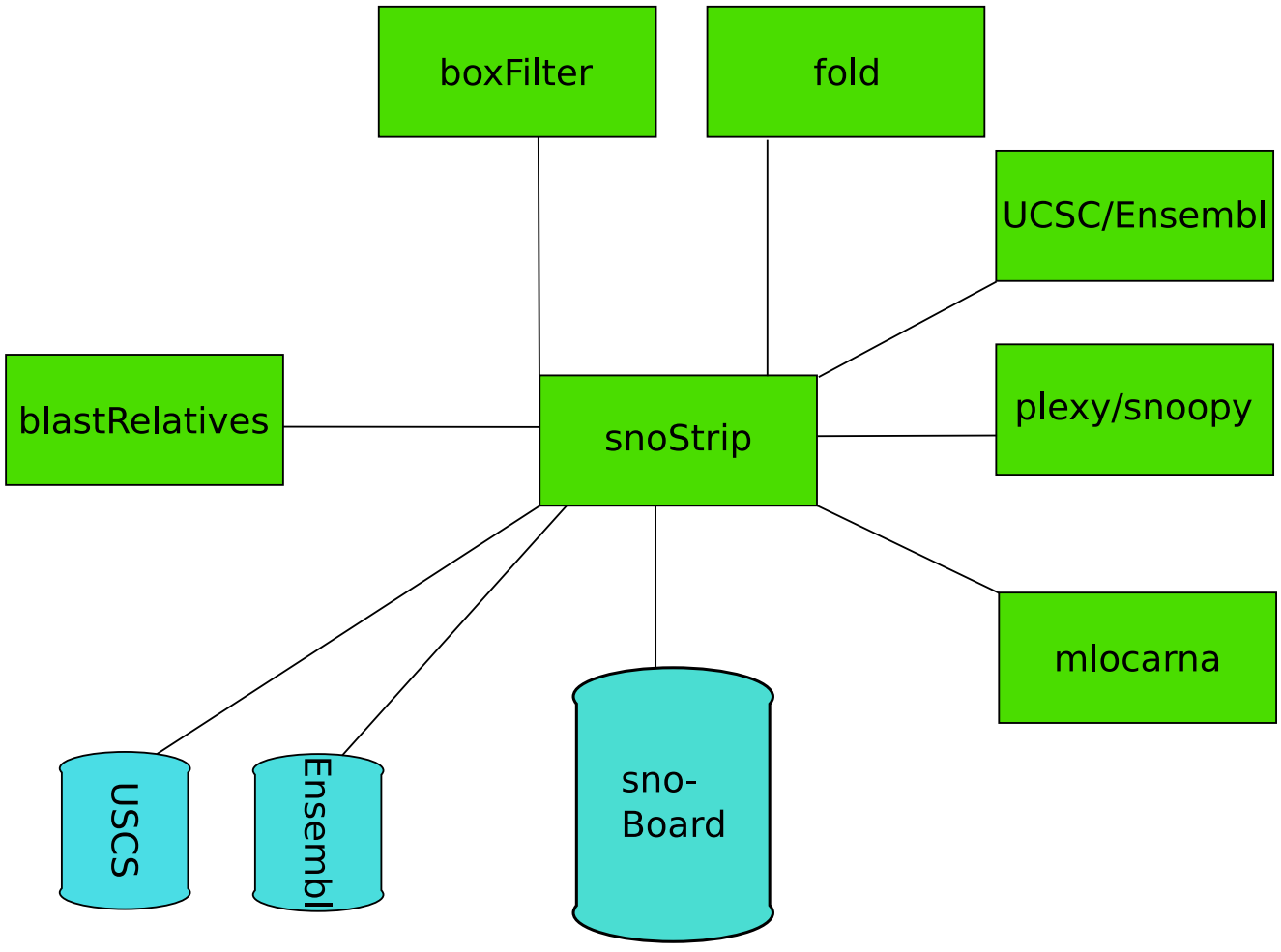
Field	Type
name	varchar(50)
hairpin_number	int(11)
target	varchar(50)
target-source	int(11)
modification_site	int(11)
target_sequence	varchar(20)
left_binding_site	varchar(20)
dist(LeftEnd/RightStart)	int(11)
right_binding_site	varchar(20)
dist(RightEnd/Box)	int(11)
binding_structure	varchar(150)
binding_energy	double
hairpin_sequence	varchar(150)
status	varchar(100)

name	sequence	chromosome	position	strand	length
ptr_HACA_43	GCTGTCCTGGACCTGTTGGCACCACAGACAGTTGCTCTGCTGTGCCTGTG...	chr9	136970559,136970694	2	136

hostgene	localisation	intronnumber	transcript ID	structure	minimum_free_energy	found_with
SNHG7 ()	intronic	2,1,1	NR_024543,NR_024542,NR_003672	..(((((((((..(..(.(((((.. (((((....).)))))..)))))).).....	-51.80	hsa_HACA_43

HOMOLOGY

name	H.sapiens	P.troglodytes	P.pygmaeus	M.mulatta
HACA_1	hsa_HACA_1	ptr_HACA_1-1,ptr_HACA_1-2	ppy_HACA_1-1,ppy_HACA_1-2,ppy_HACA_1-3	mml_HACA_1-1,mml_HACA_1-2
HACA_10	hsa_HACA_10	ptr_HACA_10	ppy_HACA_10	mml_HACA_10
HACA_2	hsa_HACA_2-1,hsa_HACA_2-2	ptr_HACA_2	ppy_HACA_2	mml_HACA_2-1,mml_HACA_2-2
HACA_23	hsa_HACA_23	ptr_HACA_23	ppy_HACA_23	mml_HACA_23
HACA_38	hsa_HACA_38-1,hsa_HACA_38-2	ptr_HACA_38	ppy_HACA_38	mml_HACA_38
HACA_39	hsa_HACA_39	ptr_HACA_39	ppy_HACA_39	mml_HACA_39
HACA_64	hsa_HACA_64	ptr_HACA_64-1,ptr_HACA_64-2,ptr_HACA_64-3,ptr_HACA...	ppy_HACA_64	mml_HACA_64
HACA_68	hsa_HACA_68-1,hsa_HACA_68-2	ptr_HACA_68-1,ptr_HACA_68-2	ppy_HACA_68	mml_HACA_68-1,mml_HACA_68-2


boxFilter
fold
UCSC/Ensembl
blastRelatives
snoStrip
plexy/snoopy
mlocarna
USCS
Ensembl
sno-
Board

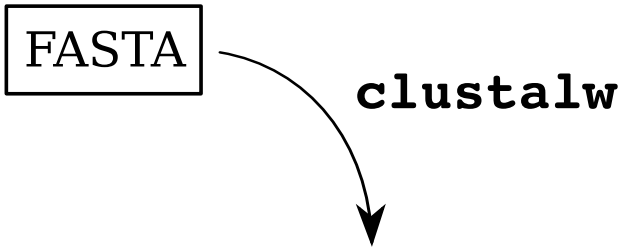
BLAST-SEARCH

- stepwise approach using all paralogs
- best blasthit as anchor for cutoff values
- decrease penalties for mismatches, gap opening and gap extensions

Alternatives:

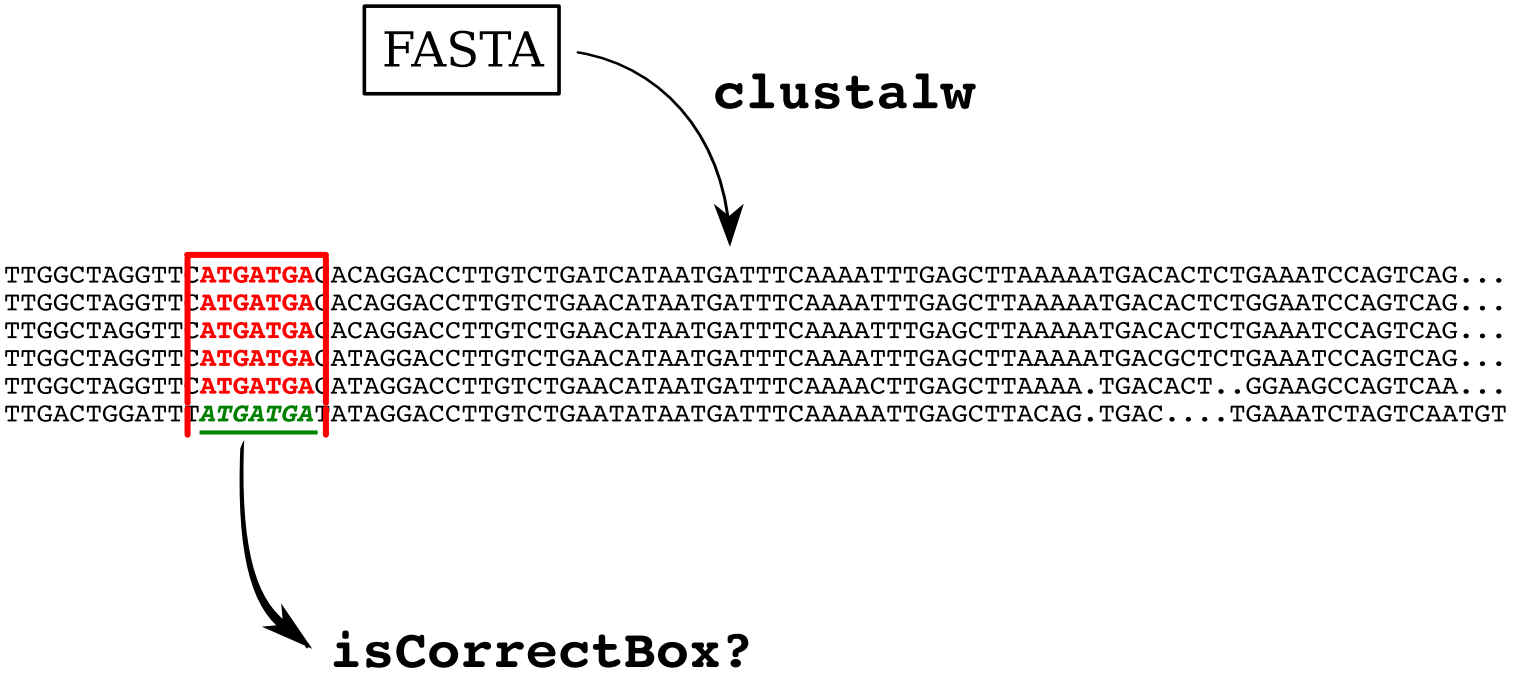
- christian’s chainer
- infernal search runs

BOX-FILTER

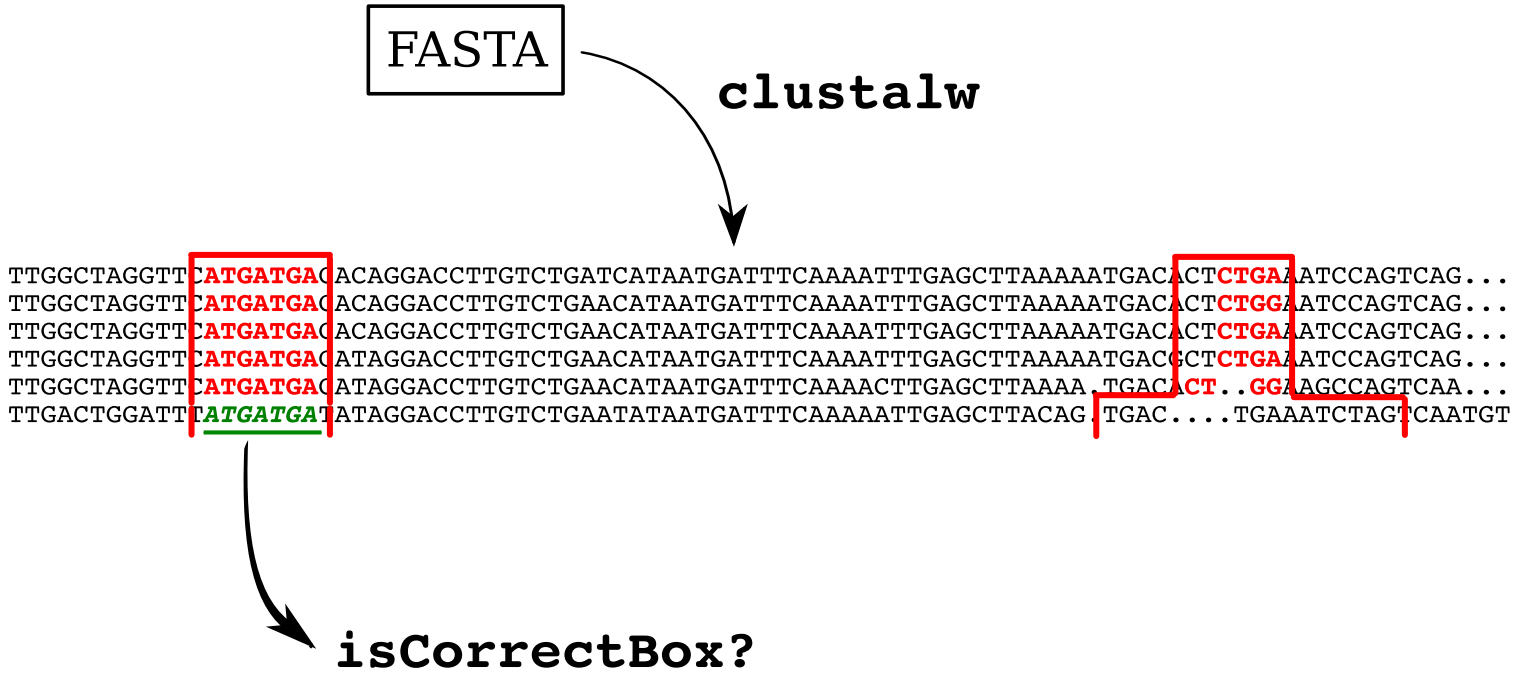



```
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGATCATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGGAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACGCTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAACTTGAGCTTAAAA.TGACACT..GGAAGCCAGTCAA...
TTGACTGGATTTATGATGATATAGGACCTTGTCTGAATATAATGATTTCAAAAATTGAGCTTACAG.TGAC....TGAAATCTAGTCAATGT
```

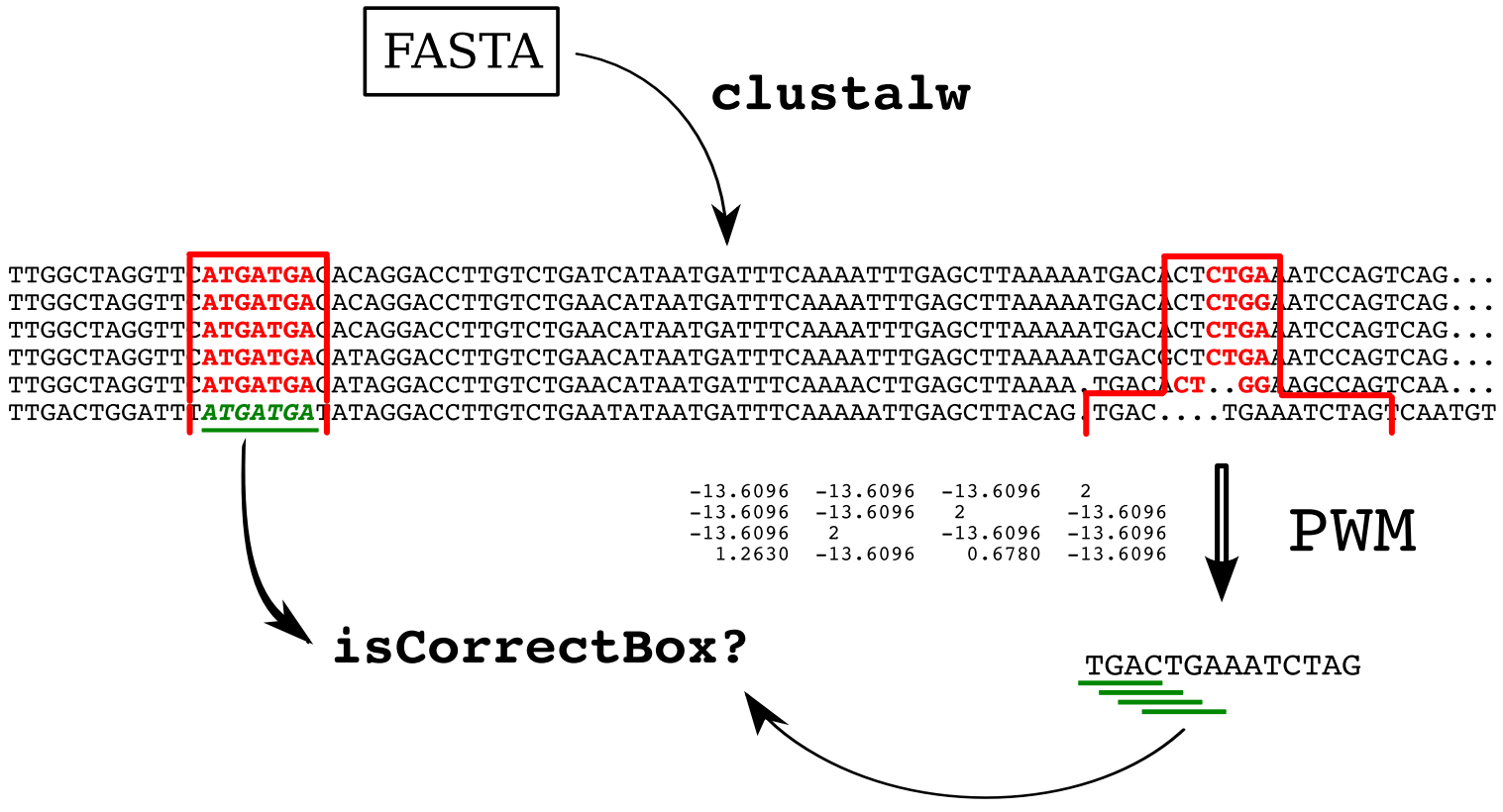
BOX-FILTER


FASTA
clustalw
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGATCATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGGAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACGCTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAACTTGAGCTTAAAA.TGACACT..GGAAGCCAGTCAA...
TTGACTGGATTTATGATGATATAGGACCTTGTCTGAATATAATGATTTCAAAAATTGAGCTTACAG.TGAC....TGAAATCTAGTCAATGT
isCorrectBox?

BOX-FILTER


FASTA
clustalw
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGATCATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGGAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACGCTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAACTTGAGCTTAAAA.TGACACT..GGAAGCCAGTCAA...
TTGACTGGATTTATGATGATATAGGACCTTGTCTGAATATAATGATTTCAAAAATTGAGCTTACAG.TGAC....TGAAATCTAGTCAATGT
isCorrectBox?

BOX-FILTER

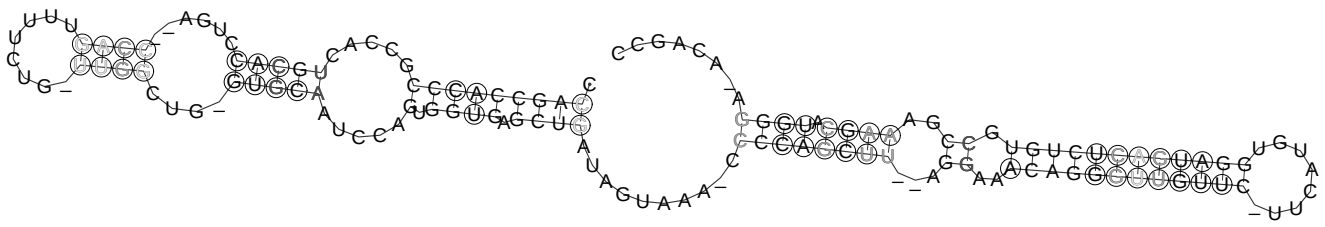

FASTA
clustalw
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGATCATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGGAATCCAGTCAG...
TTGGCTAGGTTCATGATGACACAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACACTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAATTTGAGCTTAAAAATGACGCTCTGAAATCCAGTCAG...
TTGGCTAGGTTCATGATGACATAGGACCTTGTCTGAACATAATGATTTCAAAACTTGAGCTTAAAA.TGACACT..GGAAGCCAGTCAA...
TTGACTGGATTTATGATGATATAGGACCTTGTCTGAATATAATGATTTCAAAAATTGAGCTTACAG.TGAC....TGAAATCTAGTCAATGT
-13.6096 -13.6096 -13.6096 2
-13.6096 -13.6096 2 -13.6096
-13.6096 2 -13.6096 -13.6096
1.2630 -13.6096 0.6780 -13.6096
PWM
isCorrectBox?
TGACTGAAATCTAG

STRUCTURE

- fold single sequences with RNAsubopt

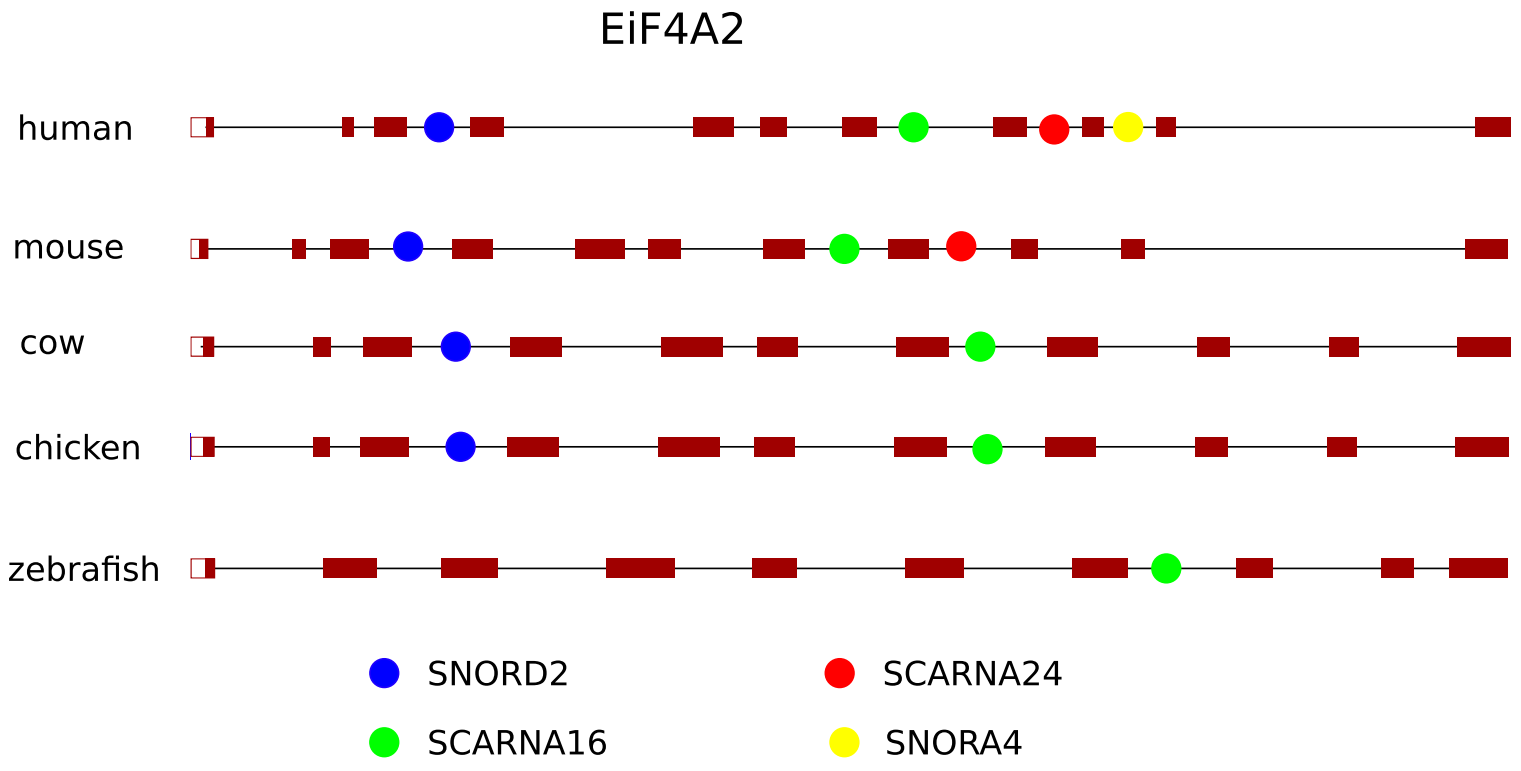
```
GUGCCCUUUUAAGGUUGACCCAGUGCUUUAAGAGGCUAACACAGAAGGGUAAAGUAAGUCUCCAUAAAACCCAGAGAAGAGACUGGAAAGCUCCUCUUUGGAUCCUGUCUGGAGUCACAACU  -3524
..(((((((((..((....)).....(((.....)))).....)))))))).......(.(((((......(((((((.(((.((....))))).)))))))........)))))..)..... -35.24
..(((((((((..((.....)).....(((.....)))).....)))))))).......(.(((((......(((((((.(((.((....))))).)))))))........)))))..)..... -35.24
..(((((((((...(((..(((..........).)).))).)))))))).......(.(((((......(((((((.(((.((....))))).)))))))........)))))..)..... -35.24
..(((((((((...((((.(((..........).))).))))..)))))))).......(.(((((.(....(((((((.(((.((....))))).)))))))......).)))))..)..... -35.20
```

- fold alignments with RNAalifold



GENOMIC ORGANISATION & HOSTGENE

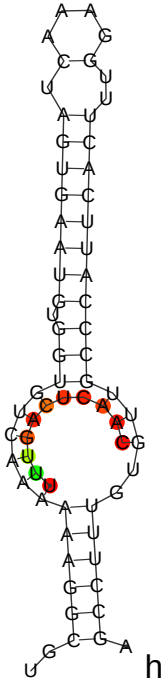
- known gene annotation from UCSC and ENSEMBL
- in vertebrata mostly located in introns
- recognize evolutionary events like intron or hostgene switches
- problem alternative splicing



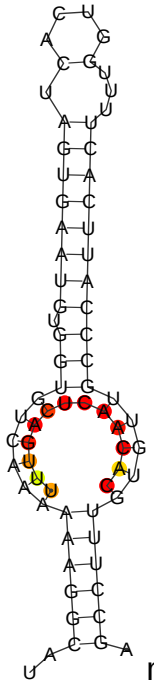
TARGET PREDICTION

- HACA: RNAsnoop for each hairpin
- apply svm to evaluate results
- CD: compute RNAduplex for 15nts upstream of D and D’-box
- use acessibility-profiles of target RNAs by RNAup

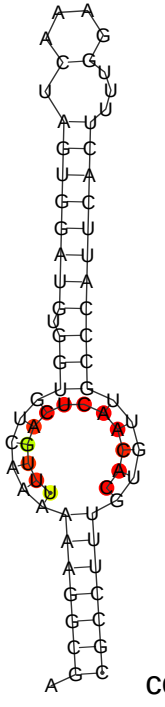
SNORA13-hairpin1



human



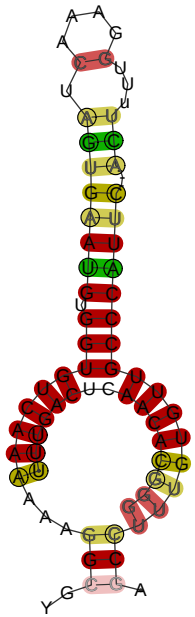
mouse

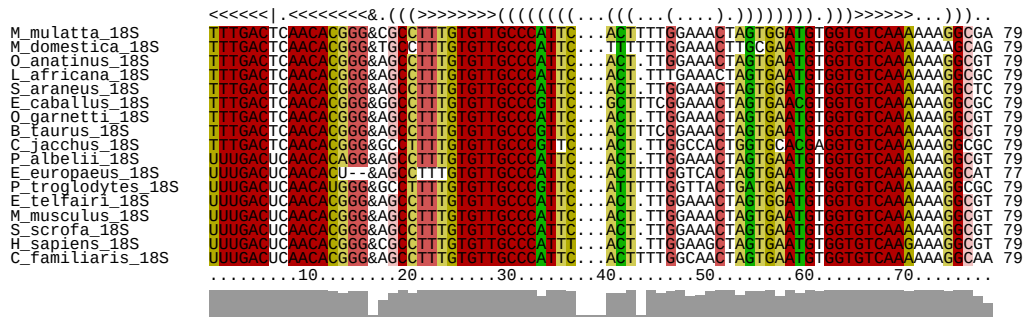


cow

TARGET PREDICTION

- RNAsnoop for the alignment
- identify recombination and separation





SNORA13-hairpin1 consensus target 18S-1248

ALIGNMENT CONSTRUCTION

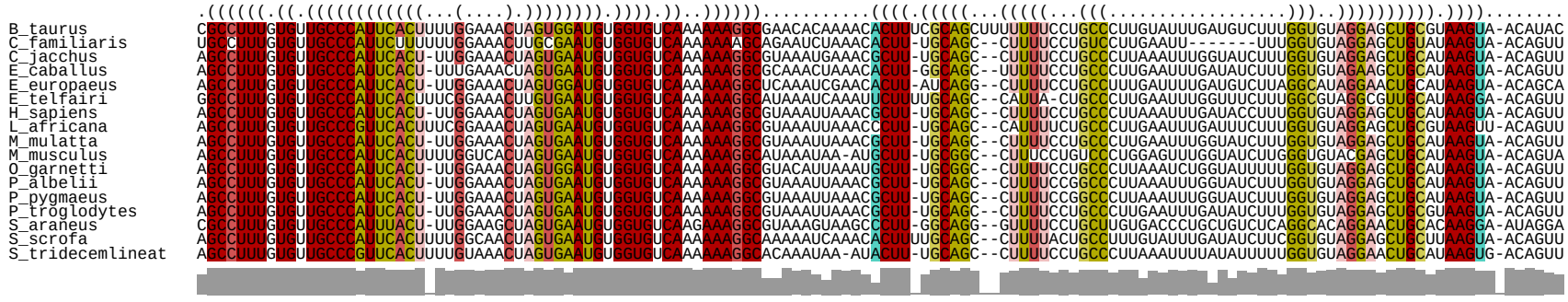
- use mlocarna to compute final alignments
- structural constraints
- anchors constraints to align boxes

```
>H.sapiens
AGCCTTTGTGTTGCCCATTCACTTTGGAAACTAGTGAATGTGGTGTCAAAAAAGGCGTAAATTAAACGCTTTGCAGCCTTTTCCTGCCCTTAAATTTGATACCTTTGGTGTAGGAGCTGCATAAGTAACAGTT
.....................................................xxxxxx.....................................................xxx... #S
.....................................................AAAAAA.....................................................BBB... #1
.....................................................123456.....................................................123... #2
>M.musculus
AGCCTTTGTGTTGCCCATTCACTTTTGGTCACTAGTGAATGTGGTGTCAAAAAAGGCATAAATAAATGCTTTGCGGCCTTTCCTGTCCCTGGAGTTTGGTATCTTGGGTGTACGAGCTGCATAAGTAACAGTA
.....................................................xxxxxx.....................................................xxx... #S
.....................................................AAAAAA.....................................................BBB... #1
.....................................................123456.....................................................123... #2
```

```
>H.sapiens
AAGTGAAATGATGGCAATCATCTTTCGGGACTGACCTGAAATGAAGAGAATACTCATTGCTGATCACTT
.......xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx...... #S
.......AAAAAAA.....................................BBBB...... #1
.......1234567.....................................1234...... #2
>M.musculus
AAGTGAAATGATGGCAAATCATCTTTCGGGACTGACCTGAAATGAAGAGAATACTCTTGCTGATCACTTA
.......xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx....... #S
.......AAAAAAA.....................................BBBB....... #1
.......1234567.....................................1234....... #2
```

EXAMPLE

- SNORA_13



THANK YOU

Hakim, Jana, Peter, Christian