20th TBI Winterseminar, Bled 2005 Slovenia
"Computational Mathematics and Theoretical Biology"

Homology based approaches to detect non-coding RNAs in the genomic sequence of Ciona intestinalis

Dominic Rose
Undergraduate seminar, winter term 2004/05

Supervision: K. Missal and P. Stadler

Introduction

- Topic:
  - Computational genomics, RNA detection
- Objective target:
  - Identification and annotation of functional non-coding RNAs in given genomes by homology based methods
- Tasks:
  - Setting up a database to handle
    - Source data (genome sequences of target organisms)
    - Resulting data from analyses
    - Additional: Documentation and website
  - Process homology based analyses
  - Fill database and extract knowledge

Motivation – Why RNA?

- RNA sequence analyses answers evolutionary and phylogenetic questions
- Cellular activity without protein influence:
  - Self-splicing Introns
  - miRNAs

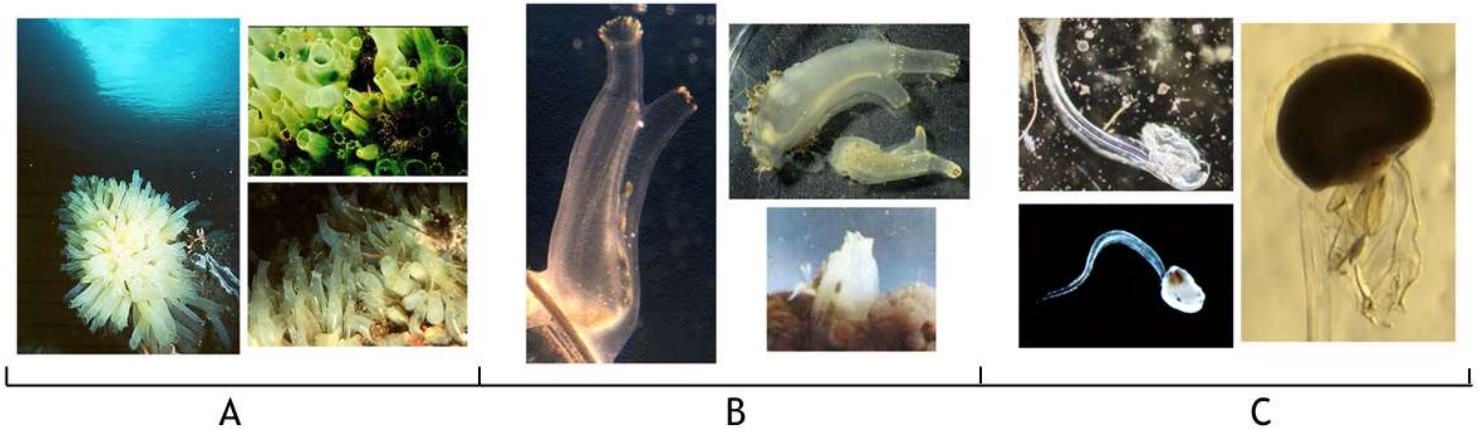
➡ RNA regulates and is catalytic active

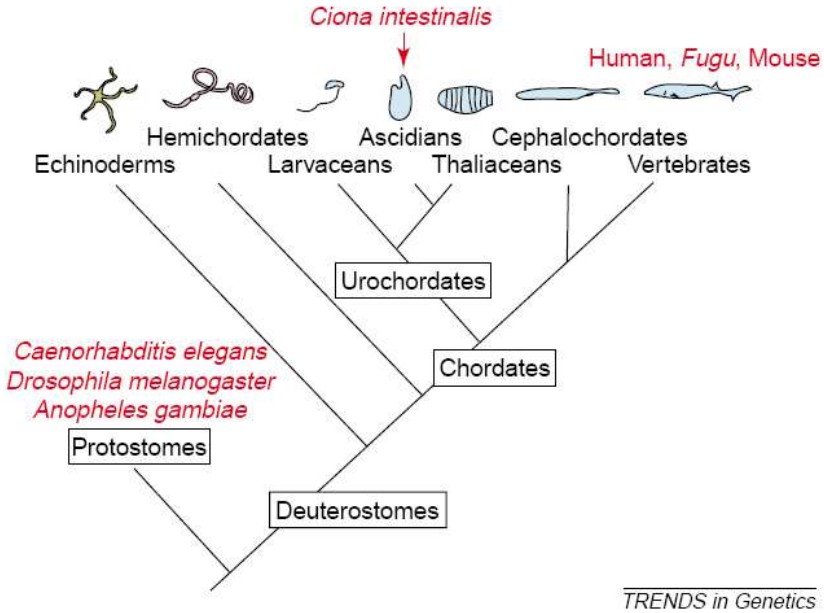
Motivation – Why non-coding RNA?

- Field of active research
- Major part of ncRNA functionality is not understood
- There is verified relation and specifity of ncRNA to
  - Diseases
  - Sex
  - Species
  - More: NONCODE
    - `http://bioinfo.org.cn/NONCODE/index.htm`

➡ Before you think about an RNAs function, you should have one...

Objetcs of research





A: Ciona intestinalis (Ci)

B: Ciona savignyi (Cs)

C: Oikopleura dioica (Od)

Procedures

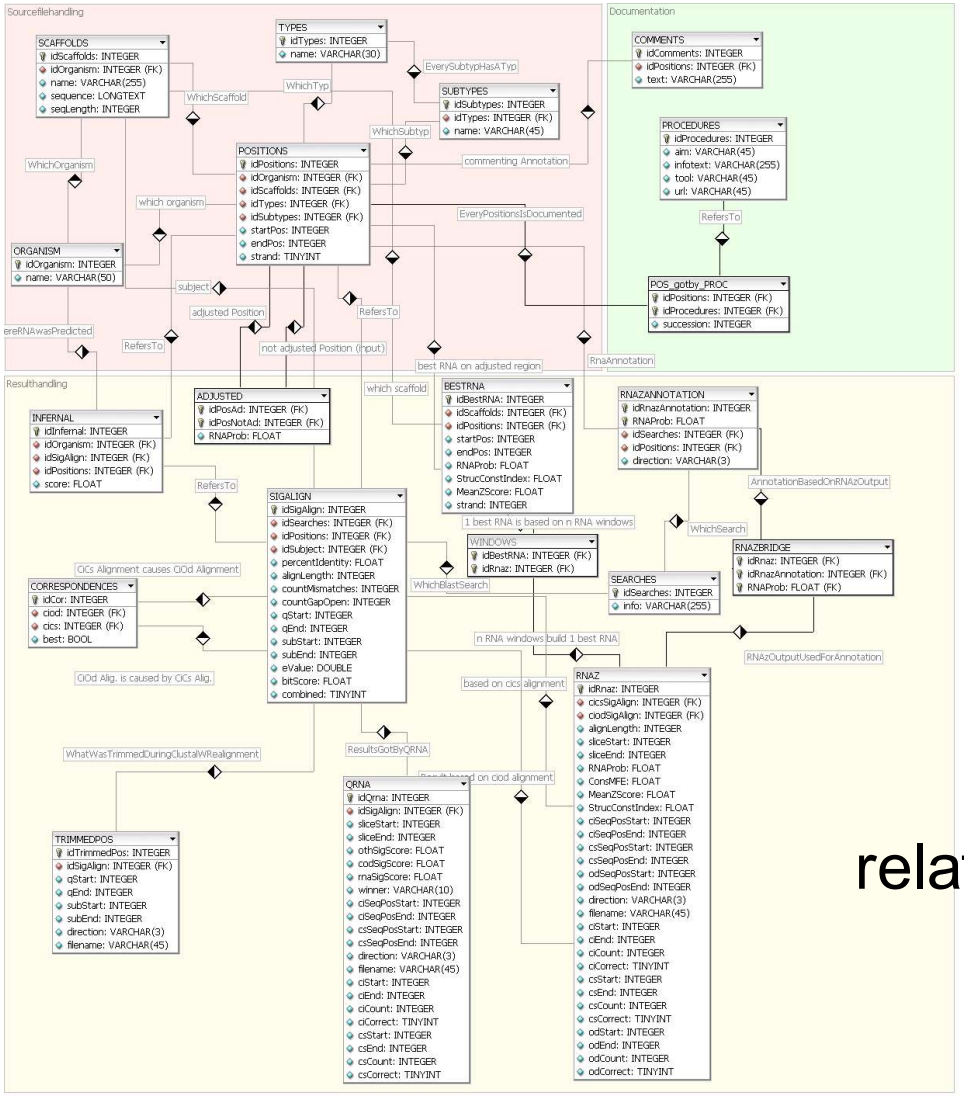
1. Setting up the database
2. Homology based analyses producing data
3. Perform annotation
4. Evaluation, statistic analyses
5. Publication, building website to view results

1. Setting up the database

- Which DB-System?
- Requirements
  - For free, OpenSource ;-)
  - Efficient operations for string manipulation
  - Support of these functions for large objects (BLOBs, CLOBs)
  - Good documentation
  - Platform: Linux, Fedora Core 2
- MySQL or PostgreSQL?

⇨ MySQL 4.1

1. Setting up the database



relational model

1. Setting up the database

- Database state after initial loading of source data

	Ci	Cs	Od
# scaffolds	2 501	446	707 767
Avg length(scaf) [nt]	46 674	367 798	759
Max length(scaf) [nt]	972 361	6 019 272	1 371
Min length(scaf) [nt]	3 007	1 797	5
Genome size [nt]	**116 731 843**	**164 037 988**	**537 548 966**
Total genome size	818 318 797		

1. Setting up the database

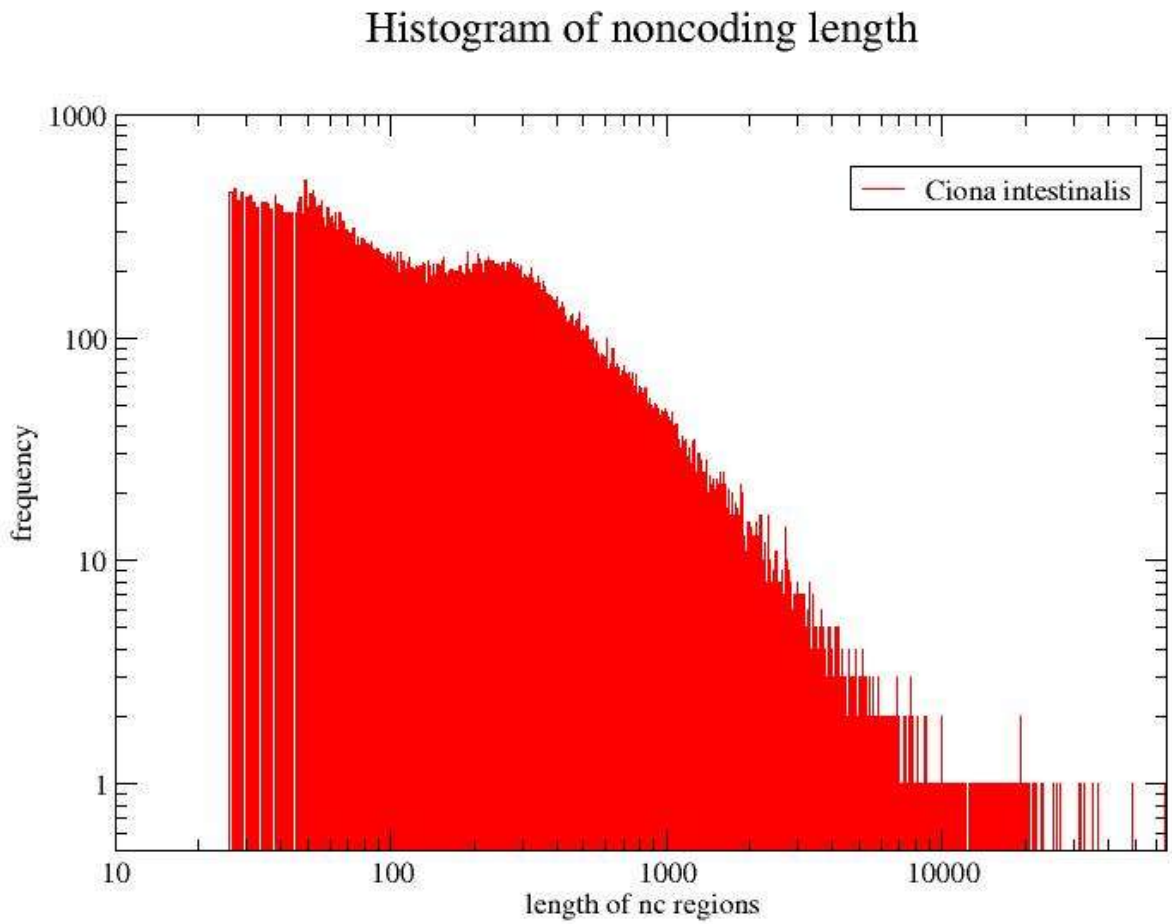
- Calculating nc regions for Ci
  - USCS genome browser provides repeat and gene annotation:

```
Exon   Intron  ____   ____   ____     _____

                ____ Coding region __

Transcription region ____________________________
```

- nc region = all except repeats and <u>coding</u> exons

# repeats	173 030
# genes	15 569
# coding region	15 569
# exon	104 366
# non coding	160 138

1. Setting up the database



nc regions >25 up to ~63 000 [nt]

2. Homology based analyses

- Overview:

Blast	Detect conserved elements
ClustalW	Realignment of blasthits including flanking regions
RNAz	Scan alignments for conserved RNA-substructures
Qrna	Classification: RNA, Coding, Random
Infernal	Which RNA?
Other tools?	

- Result: Ci ncRNA annotation

2. Blast

- Blast searches to find conserved elements...
  1. *Search:* ***nc-Ci*** *against complete* ***Cs genome***
     - nc length > 25
     - eValue 1e-3
  2. *Search:* ***nc-Ci*** *against complete* ***Od genome***
     - Same as above, but
     - only with nc-Ci sequences involved in search 1
- ... producing significant alignments.

Idea: Od is less evolutionary related to Ci than Cs. Hits seem to be important because they are conserved

2. Blast

- Hits with a distance <30 nt are combined and handled as a single hit
- Results:

	CiCs	CiOd
# +strand	281 654	16 991
# -strand	296 207	16 661
# both	577 861	33 652
# combined	2 092	1

2. Realignment

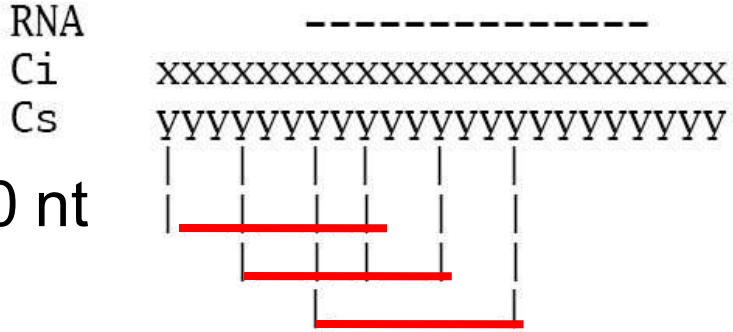
- Realign blast hits (>40nt) and add flanking regions (30nt)
- ClustalW: Global alignments
- Different cases caused by different (unknown) reading directions of alignments:

  (redundancy to recognise every signal)

  – CiCs, 4 cases: PP, PN, NP, NN

  – CiCsOd, 8 cases: PPP, PPN, ..., NNN
- Remove surrounding gaps
- Scan with RNAz
- Stepsize 50 nt, framesize 120 nt

  K. Missal ;)



2. RNAz

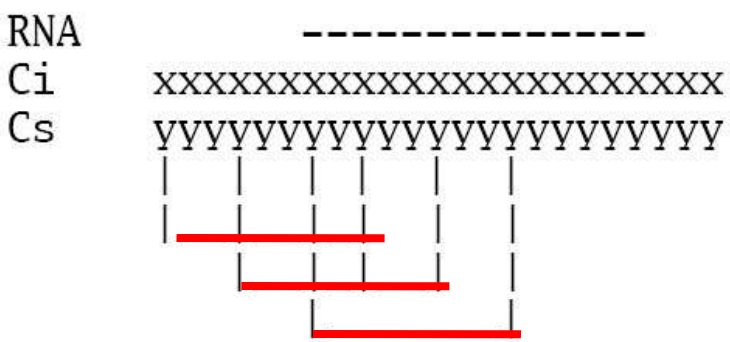
- Example of output
- >RNAz|Data/Al1/al_687.aln.trim|alignLength=127|Slice=1:120|**RNAProb=0.999628**|ConsMFE=-31.60|MeanZScore=-3.18|StrucConsIndex=0.84
- >SEQ|ci_687.1|SeqPos=1:120
- AAGGUACAAUGGACUAAAAGUCUAAAUACAAAAAUUGGGCUCGUCCGGGAUUUGAACCCGGGACCUCUCGCACCCAAAGCGAGAAUCAUACCCCUAGACCAACGAGCCAGACACAACCGC
- >SEQ|cs_687.2|SeqPos=1:119
- AUAAAGCAGAGGACAAGCACUAAAUUUAUCAAAAAUGGGCCCGUCCGGGAUUUGAACCCGGGACCUCUCGCACCCAAAGCGAGAAUCAUGCCCCUAGGACAACGGGCCGCUGUAAAUUC

- >RNAz|Data/Al1/al_687.aln.trim|alignLength=127|Slice=8:127|**RNAProb=0.998288**|ConsMFE=-31.60|MeanZScore=-2.86|StrucConsIndex=0.86
- >SEQ|ci_687.1|SeqPos=8:127
- AAUGGACUAAAAGUCUAAAUACAAAAAUUGGGCUCGUCCGGGAUUUGAACCCGGGACCUCUCGCACCCAAAGCGAGAAUCAUACCCCUAGACCAACGAGCCAGACACAACCGCUUUUCGA
- >SEQ|cs_687.2|SeqPos=8:126
- AGAGGACAAGCACUAAAUUUAUCAAAAAUGGGCCCGUCCGGGAUUUGAACCCGGGACCUCUCGCACCCAAAGCGAGAAUCAUGCCCCUAGGACAACGGGCCGCUGUAAAUUCUUUUCAA

2. RNAz

- Overview of RNAz results

	# RNAz frames			
RNAProb	ALL	>0.5	>0.9	>0.99
CiCs	1 152 951	160 233	102 628	60 386
CiCsOd	108 864	31 734	24 016	15 580



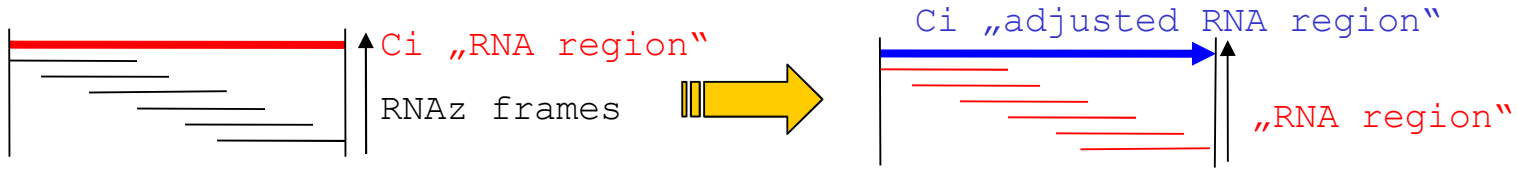
2. RNAz

- Clusters of putative RNA-hits are recognisable

idRnaz	idSigAlign	idScaffolds	NC_ID	scafStart	scafEnd	RNAProb	sliceStart	sliceEnd
5423	287	1	546414	346111	346229	0,853248	1	120
5426	289	1	546414	346111	346230	0,968088	1	120
5420	284	1	546414	346111	346230	0,98312	1	120
5421	284	1	546414	346115	346234	0,954782	5	124
5427	289	1	546414	346117	346236	0,780908	7	126
5429	290	1	546414	346117	346234	0,699971	7	126
5424	287	1	546414	346118	346236	0,754346	8	127
5315	234	1	546788	518798	518917	0,721804	101	220
5278	223	1	546797	521650	521762	0,999968	1	113
5233	213	1	546797	521652	521762	0,999985	1	111
5101	184	1	546797	521652	521762	0,99996	1	111
5250	217	1	546797	521652	521757	0,99974	1	108
5312	232	1	546797	521653	521762	0,999927	1	111
5206	206	1	546797	521653	521762	0,999949	1	111
5194	202	1	546797	521654	521762	0,999978	1	109
5105	185	1	546797	521654	521762	0,99997	1	109
5154	196	1	546797	521654	521762	0,99998	1	109
5051	173	1	546797	521654	521762	0,999991	1	109

Merge the frames to get one “exact” start and one “exact” end

3. Annotation

- 

  n RNAz frames define 1 „RNA region“, look at frames >0.5 RNAProb:
  - Combine RNAz frames from the same alignment directly if they overlap
  - Combine frames from different alignments if they overlap >90%

- Don‘t forget the “redundant” reading directions of the original alignments, last steps were done for all cases!

  (PP,..,NN for CiCs and PPP,..,NNN for CiCsOd)

- Try to get a “exact“ reading direction by defining an “adjusted RNA region“ due to the RNA that represents the cluster optimal

  (n unadjusted „RNA regions“ define 1 „adjusted RNA region“)

➡ The “adjusted region“ only tells you: “There is signal“, statistic analyses were done with each “best RNA“

4. Statistic analyses

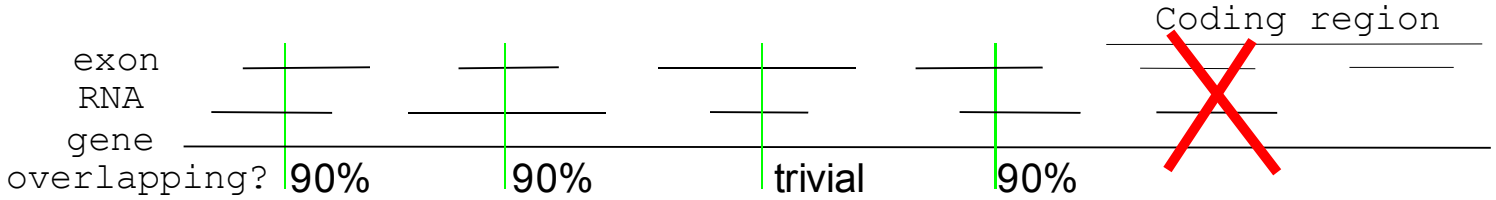
- Results: ncRNAs

RNAProb	>0.5	>0.9	>0.99
# best RNAs			
CiCs	12 861	6 316	2 740
CiCsOd	1 017	726	561
avg length of best RNAs			
CiCs	~125	~130	~130
CiCsOd	~118	~118	~119

4. Statistic analyses

- Do we have ncRNAs on exons?

  (We could if the exon is not a coding one.)



#CiCs: 2 330 (18%) #CiCsOd: 61 (6%)

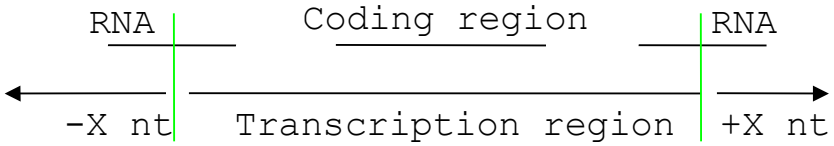
- Do we have ncRNAs on introns?

  (Calculation similar to above)

#CiCs: 700 (5%) #CiCsOd: 67 (7%)

4. Statistic analyses

- Are there RNAs on UTRs (first approach)? Avg length(UTR)=657



	>0.5		>0.9		>0.99	
	±200	±1000	±200	±1000	±200	±1000
CiCs	794 (6%)	515 (12%)	397 (3%)	770 (6%)	152 (1%)	352 (3%)
CiCsOd	39 (4%)	111 (11%)	14 (1%)	73 (7%)	11 (1%)	56 (6%)

second approach: If UTR>1000nt believe its annotation, if it is <1000nt add [200|1000]nt but results differ less.

third approach: Count RNAs at [200|1000]nt up- or downstream from coding region

	>0.5		>0.9		>0.99	
	±200	±1000	±200	±1000	±200	±1000
CiCs	50 (0.4%)	881 (7%)	21 (0.2%)	461 (4%)	11 (0.1%)	244 (2%)
CiCsOd	3 (0.3%)	92 (9%)	1 (0.1%)	63 (6%)	1 (0.1%)	48 (5%)

There seems to be a considerable amount of RNAs on UTRs

4. Statistic analyses

- Additional analyses and validation:
  - miRNA search: miRNA registry

    `http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml`
  - ncRNA search: NONCODE

    `http://www.bioinfo.org.cn/NONCODE/index.htm`
  - rRNA search
- Compare our ncRNAs with known RNAs from additional vertebrate genomes.
- Fold the sequences and hope that there will be exciting structures
- Check existing literature

5. Publicate results

- Results should appear in written form
- Website will be available soon 

Thank you for your attention (patience)

Are there any questions?

;-)