RNA Secondary Structures

1st EVBC WinterSchool

Jörg Fallmann, Kevin Lamkiewicz

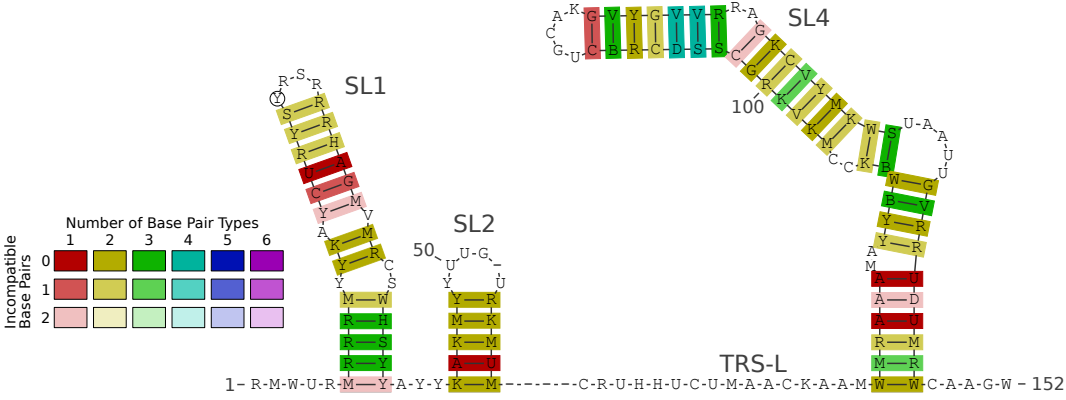
9 – 13.03.2018

European VIRUS BIOINFORMATICS Center

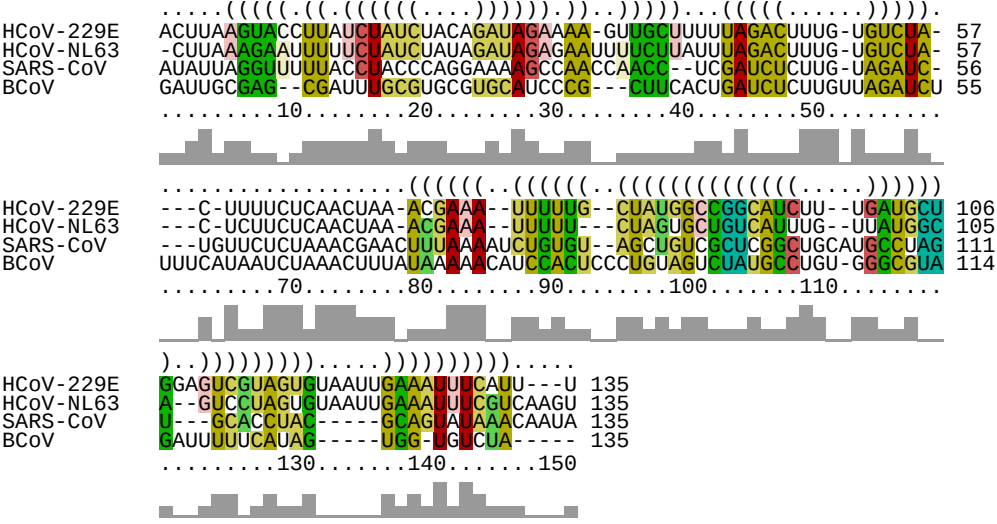
de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Alignments and compensatory mutations

UNCONSERVED SEQUENCE, CONSERVED STRUCTURE



UNCONSERVED SEQUENCE, CONSERVED STRUCTURE



Compensatory Mutations in secondary structures

Importance of such mutations

Compensatory mutations underline the importance of a specific secondary structure.

Compensatory Mutations in secondary structures

Importance of such mutations

Compensatory mutations underline the importance of a specific secondary structure.

But be careful!

If we're assuming a uniform mutation rate, every third pair of mutations is a compensatory mutation.

A	U
A	A
A	C
A	G
C	A
C	C
C	G
C	U

A	U
G	A
G	C
G	G
G	U
U	A
U	C
U	G
U	U

RNA-RNA Long-Range Interactions

WHY LRIS?

- Interaction spans distances between a few hundred and several thousands of nucleotides
- few are described in positive stranded RNA viruses
- often located in loop regions (bulges, hairpins, ...)

WHY LRIS?

- Interaction spans distances between a few hundred and several thousands of nucleotides
- few are described in positive stranded RNA viruses
- often located in loop regions (bulges, hairpins, ...)
  $\Rightarrow$ pseudo-knots!

WHY LRIS?

- Interaction spans distances between a few hundred and several thousands of nucleotides
- few are described in positive stranded RNA viruses
- often located in loop regions (bulges, hairpins, ...)
  $\Rightarrow$ pseudo-knots!
- LRIs may play a very important role in viral replication

How to calculate LRIs

Approach I

- RNAduplex
- RNAplex
- RNAhybrid

How to calculate LRIs

Approach I

- RNAduplex
- RNAplex
- RNAhybrid

Approach II

- RNAcofold

How to calculate LRIs

Approach I

- RNAduplex
- RNAplex
- RNAhybrid

Approach II

- RNAcofold

Approach III

- RNAup
- IntaRNA

How to calculate LRIs

Approach I

- RNAduplex
- RNAplex
- RNAhybrid

Approach II

- RNAcofold

Approach III

- RNAup
- IntaRNA

Approach IV

- inteRNA
- inRNAs

How to calculate LRIs

Approach I

- RNAduplex
- RNAplex
- RNAhybrid

Approach II

- RNAcofold

Approach III

- RNAup
- IntaRNA

Approach IV

- inteRNA
- inRNAs

Approach V

- PETcofold
- RNAaliduplex

LRIscan

Prediction of conserved long-range RNA-RNA interactions in full viral genomes, 2016. M. Fricke, M. Marz

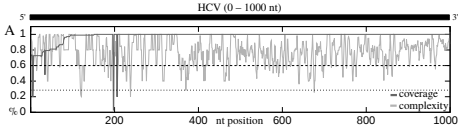
LRIscan

Prediction of conserved long-range RNA-RNA interactions in full viral genomes, 2016. M. Fricke, M. Marz

$\Rightarrow$ `LRIscan`

How does LRIscan work and how do I use it?

WORKFLOW OF LRISCAN



Coverage and Complexity

Coverage of an alignment

Relative number of sequences that do not have a gap on a specific position.

Coverage and Complexity

Coverage of an alignment

Relative number of sequences that do not have a gap on a specific position.

Complexity of the alignment

$$C_i = \frac{1}{m} \sum_{k=1}^{m} \frac{|\delta(a^k_{i \dots i+s-1})|}{|(a^k_{i \dots i+s-1})|}$$

Coverage and Complexity

Coverage of an alignment

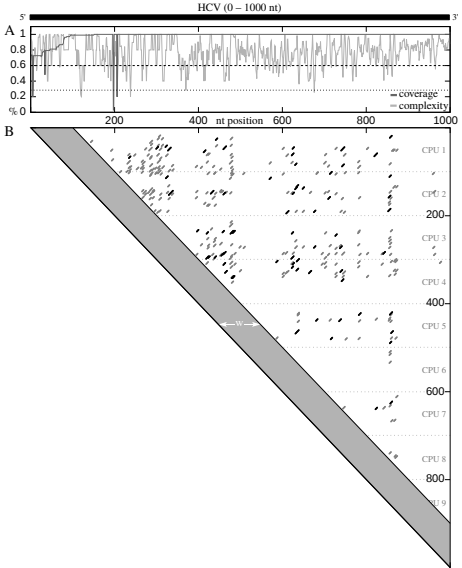
Relative number of sequences that do not have a gap on a specific position.

Complexity of the alignment

$$C_i = \frac{1}{m} \sum_{k=1}^{m} \frac{|\delta(a^k_{i \dots i+s-1})|}{|(a^k_{i \dots i+s-1})|}$$

$$\delta(CCUUUGGAAA) = CUGA$$

WORKFLOW OF LRISCAN – STEP 2

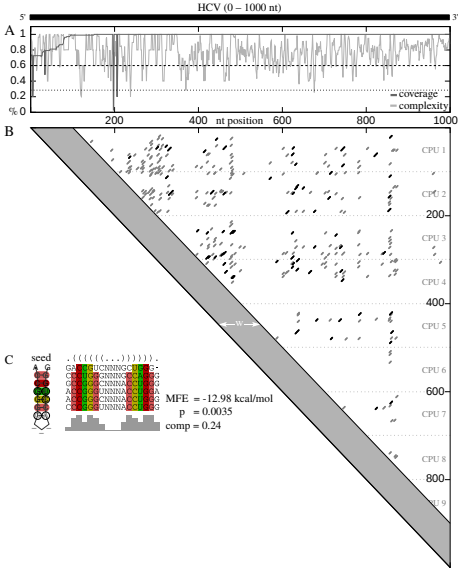


FINDING SEEDS

$$S_{i,j} = (S_{i-1,j+1} + 1) \cdot \Pi_{ij} \cdot \Phi_{ij}$$

- Π_{ij}: do at least t percent of the input sequence form the basepair (i, j)?
- Φ_{ij} : do both alignment columns A_i and A_j meet the coverage threshold?

WORKFLOW OF LRISCAN – STEP 3



Seed Scoring

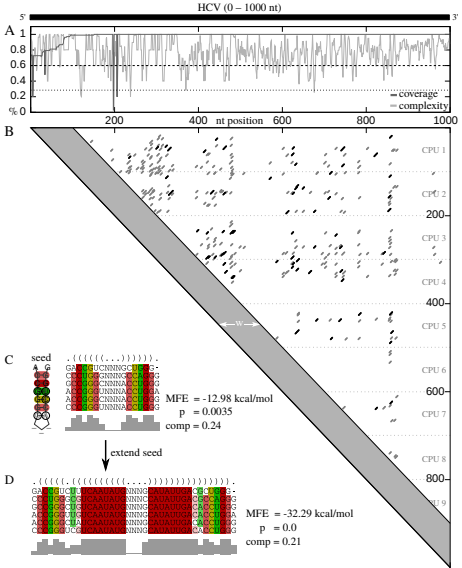
- z-Score analysis for each seed to measure reliability
- compensatory score τ

$$\tau = \frac{\sum_b (u \cdot h)}{6 \cdot |b| \cdot k}$$

with:

- u: number of different base-pair types
- h: number of incompatible base-pairs

WORKFLOW OF LRISCAN – STEP 4



Seed Extension

- each seed is extended 10 nts at the 5' (and 3' respectively)
- calculate MFE with RNAalifold
  - hard constraints for seed region
  - soft constraints for extension, such that intermolecular interactions are formed

LRIscan Results and Output

LRIscan Usage

```
$> ./LRIscan.rb -c 2 -f <ALIGNMENT> -o <OUTPUT>
```

- tabular output in .tsv format
- table and figures in .html
- all figures are also stored in the `ps/` directory

Example Data Set

QUESTIONS

We want to analyze the human coronavirus 229e (HCoV229e).

- Are there conserved structures in the 5' UTR?
- Are there conserved structures in the 3' UTR?
- Can we detect long-range interactions within the Coronaviruses?

- The virus with the sequence of `hcov229e_mutation.fasta` does not replicate as well as the wildtype. Our wet-lab collaboration partner assumes there might be something odd in the 5' UTR.
- Extract the 5'UTR of the `hcov229e_5utr_mutation.fasta`. How does the secondary structure behave?
- There might be a pseudoknot in the 3'UTR. Do you have any ideas how to confirm this hypothesis?
- ...

QUESTIONS

We want to analyze the human coronavirus 229e (HCoV229e).

- Are there conserved structures in the 5' UTR?
- Are there conserved structures in the 3' UTR?
- Can we detect long-range interactions within the Coronaviruses?

- The virus with the sequence of `hcov229e_mutation.fasta` does not replicate as well as the wildtype. Our wet-lab collaboration partner assumes there might be something odd in the 5' UTR.
- Extract the 5'UTR of the `hcov229e_5utr_mutation.fasta`. How does the secondary structure behave?
- There might be a pseudoknot in the 3'UTR. Do you have any ideas how to confirm this hypothesis?
- ...

Note:

If you have your own ideas on what to analyze, feel free to do so. These are just some example questions. Furthermore, if you want to include more related viruses to your analyses, you're welcome to do so as well.