RNA structure analysis and conserved long-range RNA-RNA interaction prediction of full viral RNA genomes

31st TBI Winterseminar
Bled

Dipl. Bioinf. Markus Fricke
RNA Bioinformatics and High Throughput Analysis
Friedrich Schiller University Jena
What is a virus?

- discovered at the end of the 19. century by Louis Pasteur
- virus - Latin *poison* or *slime*
- only 3186 viral species are known today (ICTV 2014)
- at least in mammalian hosts 320,000 viruses are unknown (Anthony 2013)

ICTV classification
- Order (-virales)
- Family (-viridae)
- Subfamily (-virinae)
- Genus (-virus)
- Species

Baltimore classification (1970)
- DNA viruses
  - Group I: dsDNA
  - Group II: ssDNA+
- RNA viruses
  - Group III: dsRNA
  - Group IV: ssRNA+
  - Group V: ssRNA-
- Retro-transcribing viruses
  - Group VI: ssDNA-
d  - Group VII: ssDNA-
ICTV species and level classification

Species:

“a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria”

Criteria:

“criteria may include, but are not limited to, natural and experimental host range, cell and tissue tropism, pathogenicity, vector specificity, antigenicity, and the degree of relatedness of their genomes or genes. The criteria used should be published in the relevant section of the ICTV Report and reviewed periodically by the Study Group”
# Hepatitis C virus and Coronaviruses

<table>
<thead>
<tr>
<th></th>
<th>Hepatitis C virus (HCV)</th>
<th>Coronaviruses (CoV)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td><em>Order</em> Unassigned</td>
<td><em>Order</em> Nidovirales</td>
</tr>
<tr>
<td></td>
<td><em>Family</em> Flaviviridae</td>
<td><em>Family</em> Coronaviridae</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Subfamily</em> Coronavirinae</td>
</tr>
<tr>
<td><strong>Genus</strong></td>
<td>Hepatitis C virus</td>
<td>Alpha-, Beta-, Gamma- and Deltacoronaviruses</td>
</tr>
<tr>
<td><strong>Genome</strong></td>
<td>plus-strand RNA (9.6 kb)</td>
<td>plus-strand RNA (30 kb)</td>
</tr>
<tr>
<td><strong>Transmission</strong></td>
<td>Sexual contact, blood</td>
<td>Respiratory or fecal-oral in humans</td>
</tr>
<tr>
<td><strong>Diseases</strong></td>
<td>acute or chronic liver diseases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>liver cirrhosis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hepatocellular carcinoma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mainly respiratory diseases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(pneumonia) and gastroenteritis</td>
<td></td>
</tr>
</tbody>
</table>
More than 1 MILLION PEOPLE DIE each year from disease caused by hepatitis B & C.

Most of those infected don’t know they have it, INCREASING THE RISK of developing severe liver disease and transmitting the virus to others.

© World Health Organization 2014
Replication of HCV and CoV (Group IV)

HCV

```
5' UTR C E1 E2 NS2 NS3 | NS4B | RdRp 3'
      p7  NS4A
5' UTR

3'```

CoV

```
5' UTR ORF1a ORF1b S E M N RTC 3'
5' Cap

3'```

**Legend:**
- TRS-L
- TRS-B
- mRNA+
- RNA-
Known RNA secondary in HCV

5’ UTR

Coding region

3’ UTR

I

II

IIIa

IIIb

IIIc

IIId

PK

stem II

IRES

V

IV

>8000 nt

5BSL1

5BSL3.1

5BSL3.2

5BSL3.3

CRE

STOP

VR

3' SLIV

3' SLLl

3' SLLII

3' SLLIII

X-Tail

polyU/UC

DLS

3SLI

5BSL2
Conserved RNA secondary in HCV

Conserved RNA secondary structures and long-range interactions in hepatitis C viruses

MARKUS FRICKE, NADIA DÖNNES, MARGARITA ZAYAS, BALT BARTENSCHLAGER, MICHAEL NIEPMANN, and MANIA MARZ.

Alignment of:
- 106 sequences
- 7 genotypes
- 65 subtypes
Conserved RNA secondary in HCV
LRIs

Long-range RNA-RNA interactions

According to their definition a LRI spans distances between a few hundred and several thousands of nucleotides.

100 nt < LRI < ? nt
Conserved LRIs in HCV
Possible HCV circularization

MFE = -93.36 = -44.88 + -48.47 (kcal/mol)

MFE = -91.64 = -39.53 + -52.11 (kcal/mol)
1. Calculate alignment complexity and coverage
2. Find LRI seeds with a sparse dotplot method
3. Filter LRI candidates based on MFE; calculate z-score/p-value and compensatory score.
4. Extend seed interaction

Compression: AAAAGUUUCC = AGUC
- Tested with alignments of HCV, HIV, Flaviviruses, Tombusviruses
- Found 14 out of 16 evolutionarily conserved and experimentally verified LRIs
- Known LRIs are highly ranked
Flavivirus (157 LRIs)

HCV (311 LRIs)

HIV (314 LRIs)

Tombusvirus (529 LRIs)
### LRIsan Output

#### Options
- Show seed extension

#### View
- Types of pairs

#### LRI Table

<table>
<thead>
<tr>
<th>Id</th>
<th>Abit</th>
<th>Ati</th>
<th>Aji</th>
<th>Ai</th>
<th>Bi</th>
<th>Bj</th>
<th>Length</th>
<th>Distance</th>
<th>Complem</th>
<th>Comp</th>
<th>p-value</th>
<th>Mfe</th>
<th>Comp</th>
<th>p-value</th>
<th>Ext</th>
<th>Mfe_ext</th>
<th>Comp_ext</th>
<th>p-value_ext</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>701</td>
<td>796</td>
<td>8838</td>
<td>8843</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.132</td>
<td>0.8</td>
<td>-1.19</td>
<td>0.24</td>
<td>0</td>
<td>12</td>
<td>-17.53</td>
<td>0.22</td>
<td>0.0092548502410789</td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>2322</td>
<td>2326</td>
<td>7173</td>
<td>7177</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>-10.13</td>
<td>0.23</td>
<td>0</td>
<td></td>
<td>0.245041501857275</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>5550</td>
<td>5557</td>
<td>5927</td>
<td>5934</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>-12.95</td>
<td>0.22</td>
<td>0</td>
<td></td>
<td>0.92207332610199528</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>9215</td>
<td>9221</td>
<td>10430</td>
<td>10445</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>-10.95</td>
<td>0.2</td>
<td>0.119</td>
<td>0.22</td>
<td>0.4169584550298515</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>138</td>
<td>1060</td>
<td>10960</td>
<td>10969</td>
<td>10969</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>-18.18</td>
<td>0.18</td>
<td>0</td>
<td></td>
<td>0.651</td>
<td>0.651</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>9999</td>
<td>9795</td>
<td>10171</td>
<td>10177</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>-15.66</td>
<td>0.25</td>
<td>0.57420220681533e-07</td>
<td>19</td>
<td>-23.77</td>
<td>0.18</td>
<td>0.00595087241421653</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>85</td>
<td>92</td>
<td>10690</td>
<td>10997</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>-10.4</td>
<td>0.3</td>
<td>1.9047160548780742e-07</td>
<td>14</td>
<td>-11.29</td>
<td>0.29</td>
<td>0.021589014862537</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>5060</td>
<td>5364</td>
<td>9768</td>
<td>9772</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>-10.71</td>
<td>0.25</td>
<td>3.4375505003442e-07</td>
<td>15</td>
<td>-13.93</td>
<td>0.24</td>
<td>0.03489262643298761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>4814</td>
<td>4818</td>
<td>5357</td>
<td>5361</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>-10.24</td>
<td>0.23</td>
<td>3.476131018567125e-07</td>
<td>14</td>
<td>-16.36</td>
<td>0.23</td>
<td>0.1023014053272357</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The table above shows the alignment results for LRIsan output, including id, position, length, distance, complem, comp, p-value, mfe, comp, and extended values.
RNA secondary structure can be transformed in a undirected graph.

- $b < 3a$
  - closing base pair must be shorter than a hairpin loop

- $b < b + 2a$
  - closing base pair must be shorter than a stacked pair

$\Rightarrow$ polynomial-time algorithms
Dijkstra’s algorithm in combination with a Fibonacci heap in $O(n \log n)$ time

Input: sample of Boltzmann-weighted secondary structures created with `RNAsubopt`
LRI in Tombusviruses

- Tombusvirus eight LRIs are known: SL3-SLB, PRTE-DRTE, UL-DL, sg1-SLB, sg2-SLB, AS1-RS1, AS2-RS2, DE-CE
- graph-distance for the complete genome of Pear latent virus
- starting at the first interaction site of each LRI
- local minimum at second interaction site

Nicholson 2013
Conserved RNA secondary in CoV – 5’ UTR

RNA structure analysis of alphacoronavirus terminal genome regions
Ramakanth Madhugiri, Markus Fricke, Manja Marz, John Ziebuhr

Markus Fricke // RNA Bioinformatics and High Throughput Analysis // Friedrich Schiller University Jena

15.02.2016
Conserved RNA secondary in CoV – 3’ UTR

Bulged stem-loop

PK-SL2

HVR

Octanucleotide

PK-SL1

Octanucleotide

PK-SL2

β

α

15.02.2016

Markus Fricke // RNA Bioinformatics and High Throughput Analysis // Friedrich Schiller University Jena


Thanks for your attention!

The Marz group:
Manja Marz
Martin Hölzer
Emanuel Barth
Konstantin Riege
Franziska Hufsky
Bertram Vogel
Nelly F. Mostajo Berrospi
Michael Gaspar
Akash Srivastava
Holger Bierhoff
Konrad Sachse
Kathrin Schowtka

Cooperation partner:
Michael Niepmann
Ramakanth Madhugiri
John Ziebuhr
Nadia Dünnes
Margarita Zayas
Rolf Backofen
Jing Qin
Peter F. Stadler
Ralf Bartenschlager
Florian Zirkel
Christian Drosten
Sandra Junglen