Fighting SARS-CoV-2 and other viruses with RNA bioinformatics

Michael T. Wolfinger

Research Group Bioinformatics and Computational Biology
& Department of Theoretical Chemistry
University of Vienna
Austria

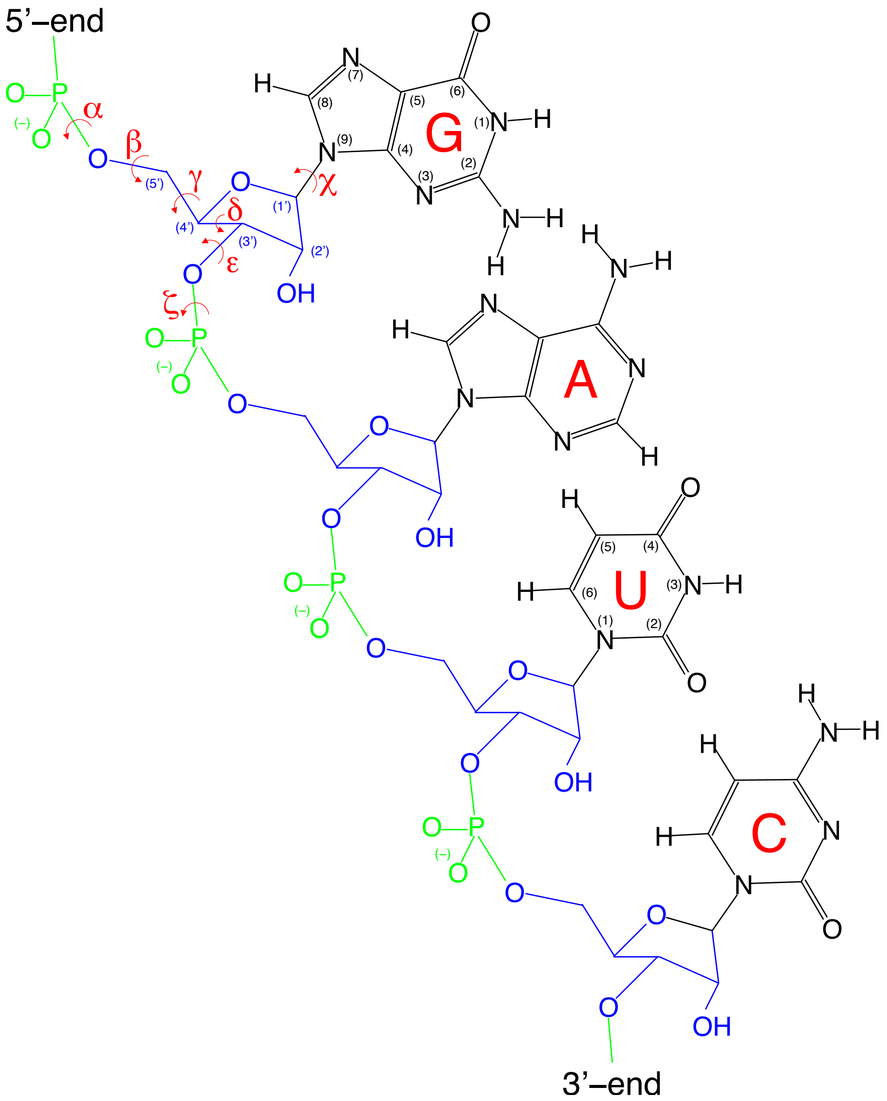
Chulalongkorn University
Bangkok
11 March 2020

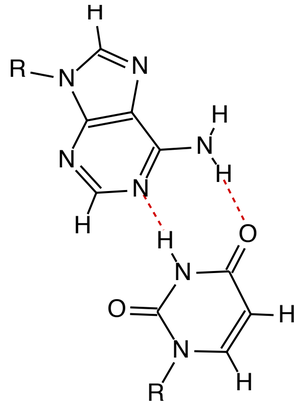
universität wien

Recent Virus Outbreaks

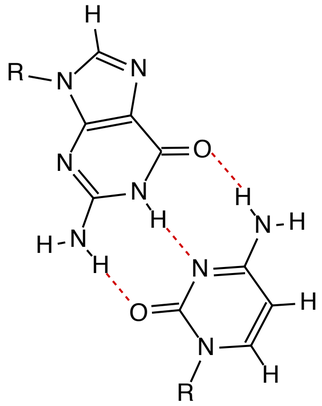
- 2012: MERS CoV in Saudi Arabia
- 2014/2015: Ebola virus (EBOV) in West Africa
- 2015/2016: Zika virus (ZIKV) in Brazil
- 2017: Yellow fever virus (YFV) in Brazil
- 2018: EBOV in DRC
- 2018/2019: Polio virus in Pakistan
- 2019: EEEV in USA
- 2019/2020: SARS-CoV-2 in China

The Chemical Structure of RNA

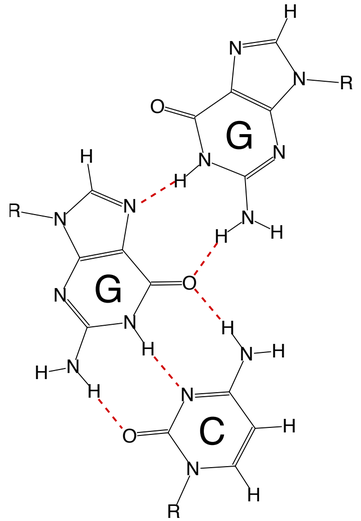




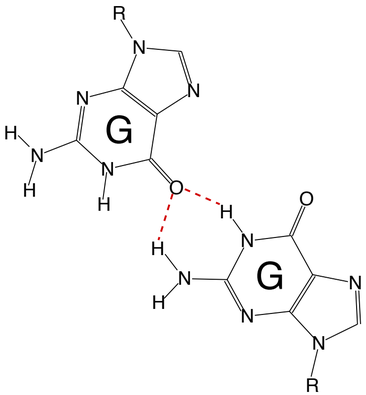
AU
Watson-Crick



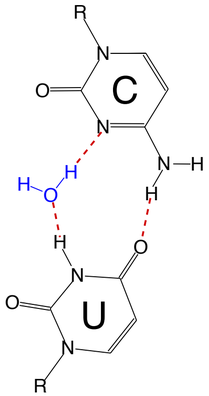
GC
Watson-Crick



(a)



(b)



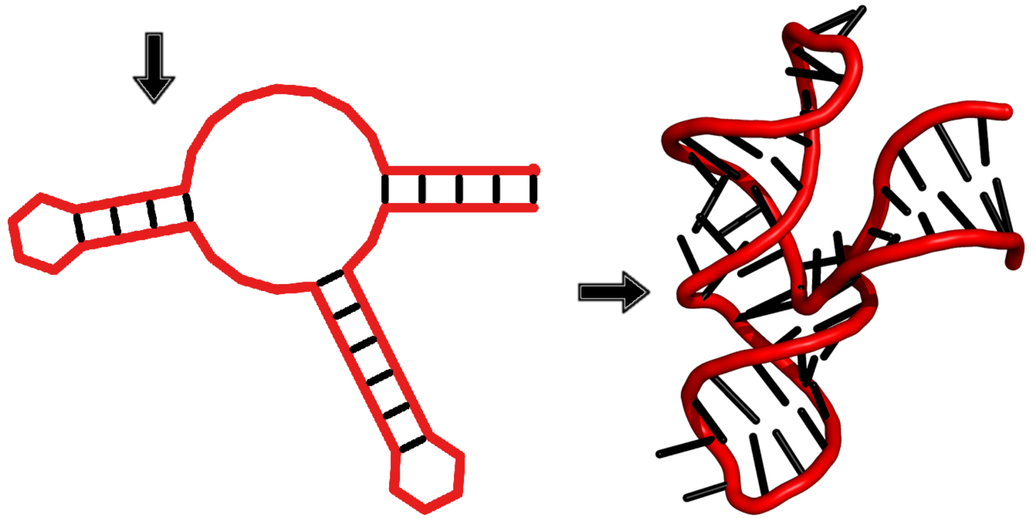
(c)

Part I:

RNA structure prediction

The RNA Folding Problem

GCGCUCUGAUGAGGCCGCAAGGCCGAAACUGCCGCAAGGCAGUCAGCGC

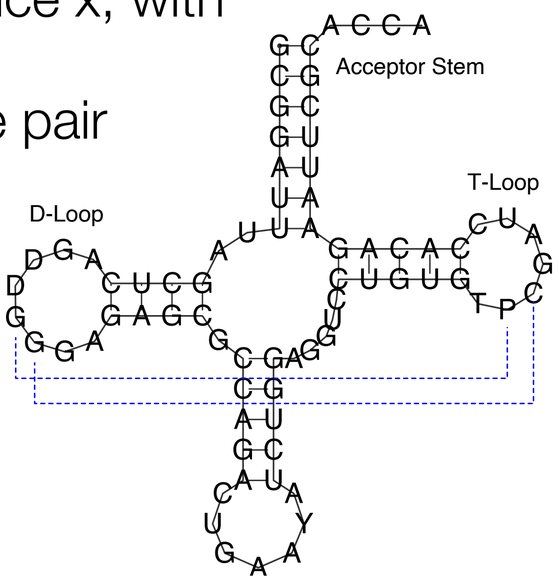


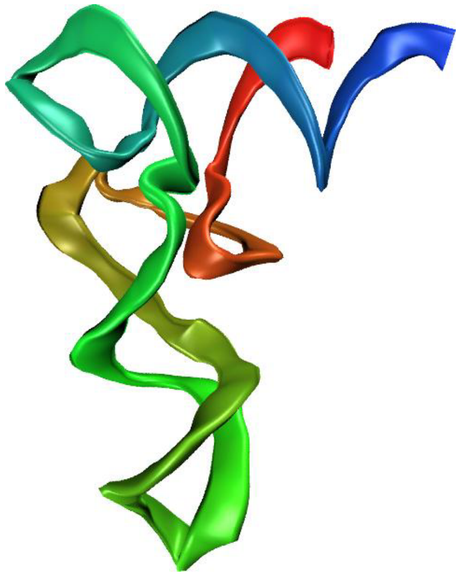
- Hierarchical folding: Secondary structure forms first then helices arrange to form tertiary structure
- Secondary structures cover most most of the folding energy
- Convenient and biologically useful description
- Computationally easy to handle
- Tertiary structure prediction needs knowledge of secondary structure

Secondary Structures

A secondary structure is a list of base pairs (i,j) on a sequence x, with

- Any nucleotide (sequence position) can form at most one pair
- If (i,j) is a pair then $x_i x_j \in \{GC,CG,AU,UA,GU,UG\}$
- If (i,j) is a base pair, then $j - i > 3$
- No pseudo-knots: No pairs (i,j) and (k,l) with $i < k < j < l$

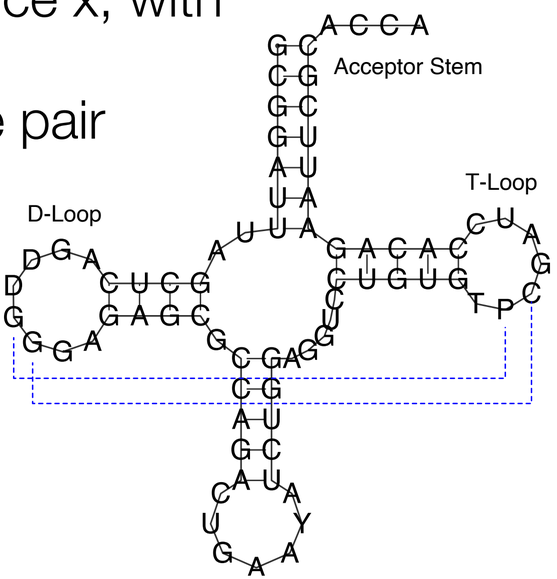


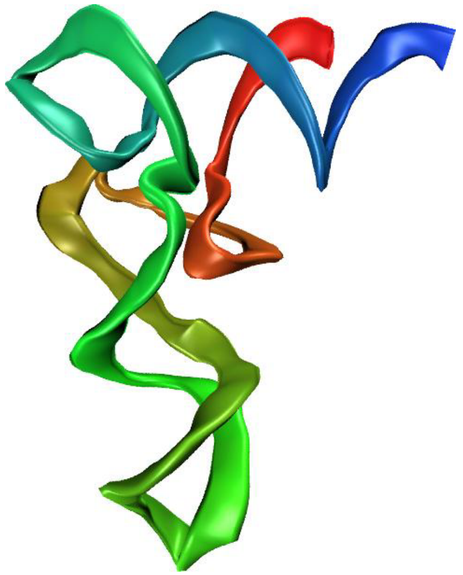


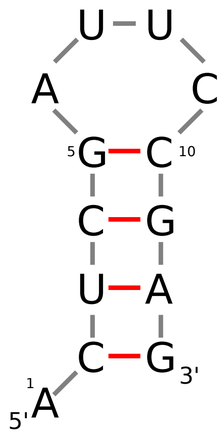
Secondary Structures

A secondary structure is a list of base pairs (i,j) on a sequence x, with

- Any nucleotide (sequence position) can form at most one pair
- If (i,j) is a pair then $x_i x_j \in \{GC,CG,AU,UA,GU,UG\}$
- If (i,j) is a base pair, then $j - i > 3$
- No pseudo-knots: No pairs (i,j) and (k,l) with $i < k < j < l$





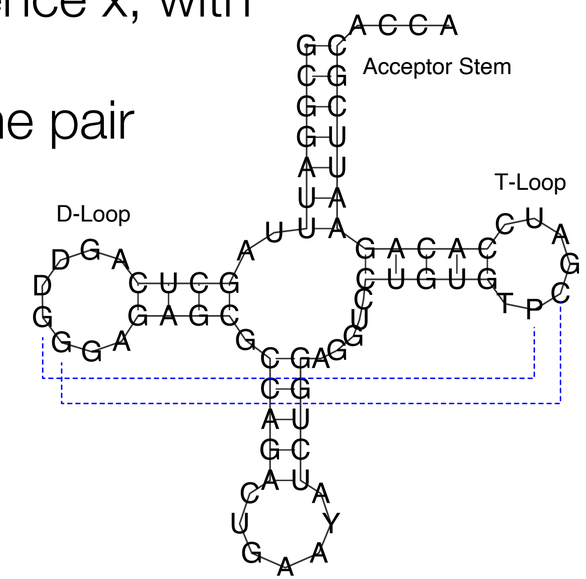


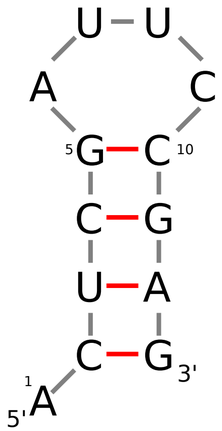

```
A C U C G A U U C C G A G
. ( ( ( ( . . . . ) ) ) )
```

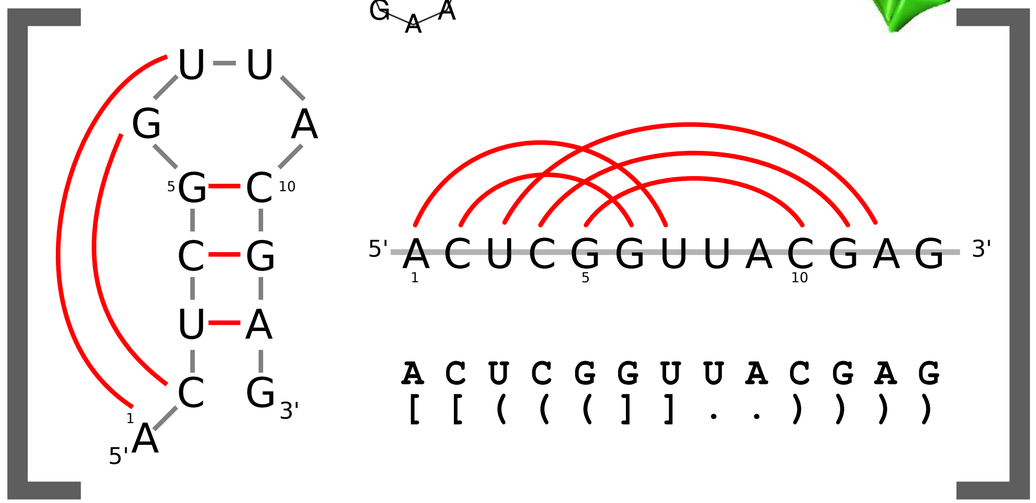
Secondary Structures

A secondary structure is a list of base pairs (i,j) on a sequence x, with

- Any nucleotide (sequence position) can form at most one pair
- If (i,j) is a pair then $x_i x_j \in \{GC,CG,AU,UA,GU,UG\}$
- If (i,j) is a base pair, then $j - i > 3$
- No pseudo-knots: No pairs (i,j) and (k,l) with $i < k < j < l$

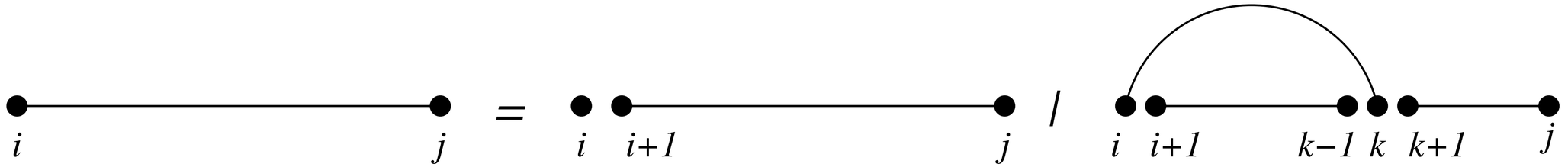






Conformation Space

The number of secondary structures for a sequence $x = x_1 \ldots x_n$ can be computed recursively



$$S_{ij} = S_{i+1,j} + \sum_{k=i+m}^{j} S_{i+1,k-1} S_{k+1,j} \Pi_{kj}$$

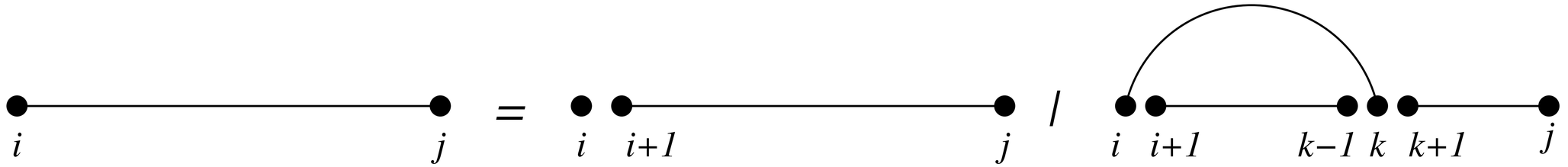
$\Pi_{ik} = 1$ if $x_i x_k \in \{GC, CG, AU, UA, GU, UG\}$, otherwise $\Pi_{ik} = 0$

For sequences with equal {A,U,G,C} content, the number of conformations grows asymptotically with sequence length

$$\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$$

Conformation Space

The number of secondary structures for a sequence $x = x_1 \ldots x_n$ can be computed recursively



$$S_{ij} = S_{i+1,j} + \sum_{k=i+m}^{j} S_{i+1,k-1} S_{k+1,j} \Pi_{kj}$$

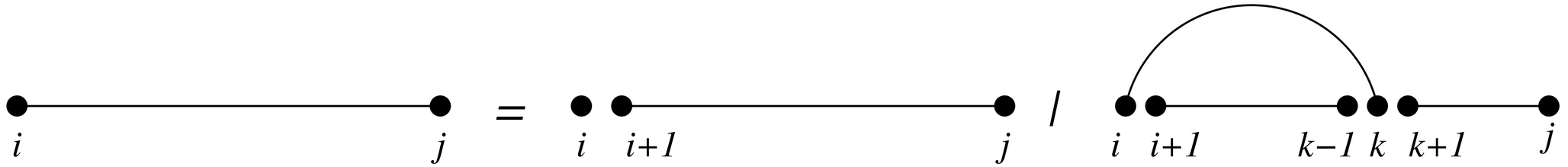
$\Pi_{ik} = 1$ if $x_i x_k \in \{GC,CG,AU,UA,GU,UG\}$, otherwise $\Pi_{ik} = 0$

For sequences with equal {A,U,G,C} content, the number of conformations grows asymptotically with sequence length

$$\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$$

Conformation Space

The number of secondary structures for a sequence $x = x_1 \ldots x_n$ can be computed recursively



$$S_{ij} = S_{i+1,j} + \sum_{k=i+m}^{j} S_{i+1,k-1} S_{k+1,j} \Pi_{kj}$$

$\Pi_{ik} = 1$ if $x_i x_k \in \{GC,CG,AU,UA,GU,UG\}$, otherwise $\Pi_{ik} = 0$

For sequences with equal {A,U,G,C} content, the number of conformations grows asymptotically with sequence length

$$\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$$

Conformation Space

The number of secondary structures for a sequence $x = x_1 \ldots x_n$ can be computed recursively

Many sequences fold into the same structure

$$S_{ij} = S_{i+1,j} + \sum_{k=i+m}^{j} S_{i+1,k-1} S_{k+1,j} \Pi_{kj}$$

$\Pi_{ik} = 1$ if $x_i x_k \in \{GC, CG, AU, UA, GU, UG\}$, otherwise $\Pi_{ik} = 0$

For sequences with equal {A,U,G,C} content, the number of conformations grows asymptotically with sequence length

$$\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$$

Conformation Space

The number of secondary structures for a sequence $x = x_1 \ldots x_n$ can be computed recursively

Many sequences fold into the same structure

Nature ‘exploits’ this property

$\Pi_{ik} = 1$ if $x_i x_k \in \{GC, CG, AU, UA, GU, UG\}$, otherwise $\Pi_{ik} = 0$

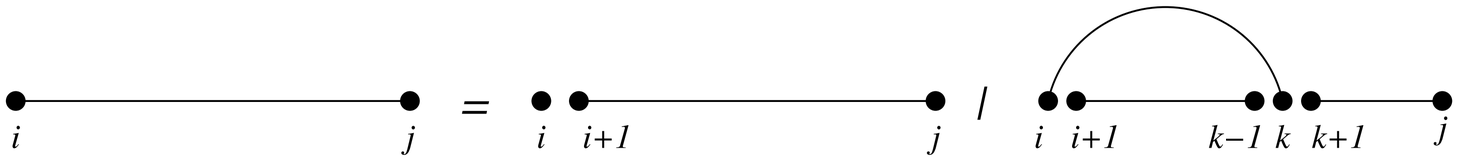
For sequences with equal {A,U,G,C} content, the number of conformations grows asymptotically with sequence length

$$\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$$

Solving the RNA Folding Problem

Toy model for RNA folding: assign energies to base pairs $\varepsilon(x, y)$

Easily solved by Dynamic Programming: recursive computation with tabulation of intermediate results



$$E_{ij} = \min_{i<k\leq j} \left\{ E_{i+1,j} \; ; \quad \left(E_{i+1,k-1} + E_{k+1,j} \; + \varepsilon(x_i, x_k) \right) \right\}$$

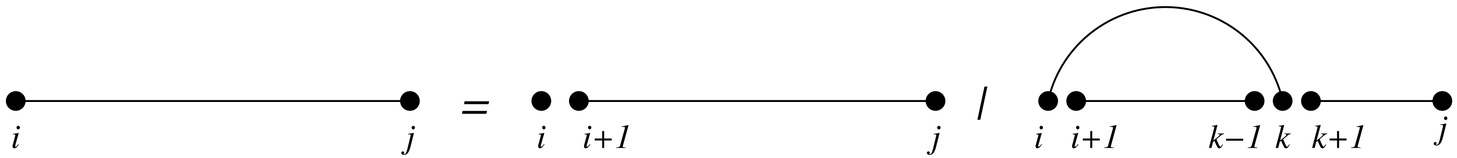
- E_{1n} is the best possible energy for our sequence
- Backtracing through the E table yields the corresponding structure
- The algorithms requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ CPU time

In practice this toy model is not good enough !
We need loop-dependent energies for serious predictions

Solving the RNA Folding Problem

Toy model for RNA folding: assign energies to base pairs $\varepsilon(x, y)$

Easily solved by Dynamic Programming: recursive computation with tabulation of intermediate results

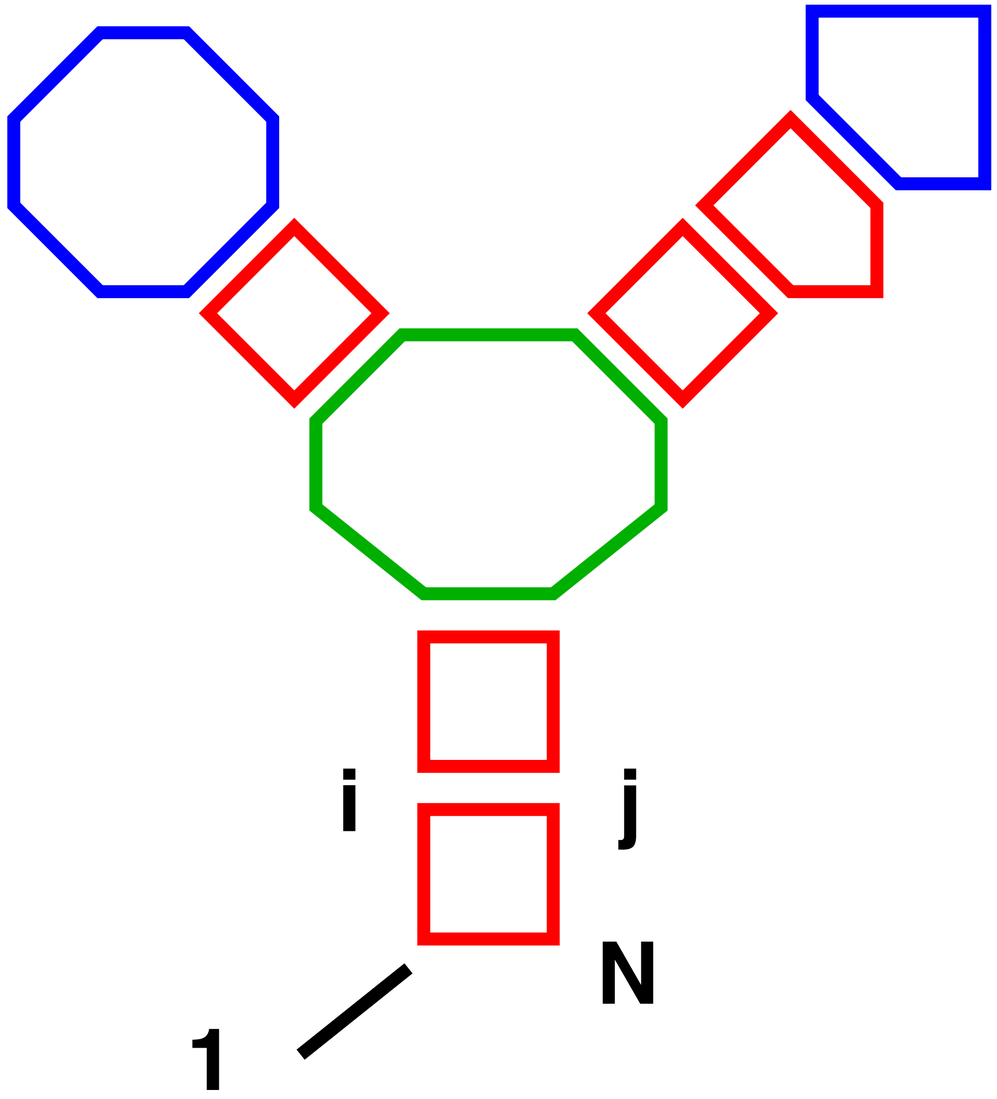


$$E_{ij} = \min_{i<k\leq j} \left\{ E_{i+1,j} \; ; \quad \left(E_{i+1,k-1} + E_{k+1,j} \; + \varepsilon(x_i, x_k) \right) \right\}$$

- E_{1n} is the best possible energy for our sequence
- Backtracing through the E table yields the corresponding structure
- The algorithms requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ CPU time

In practice this toy model is not good enough !
We need loop-dependent energies for serious predictions

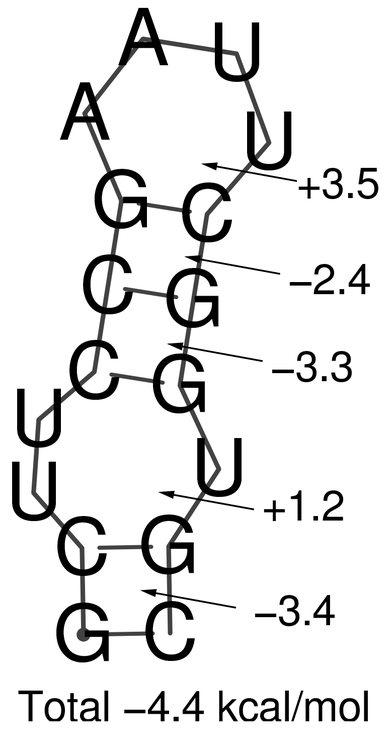
Loop Decomposition



Nearest Neighbour Model

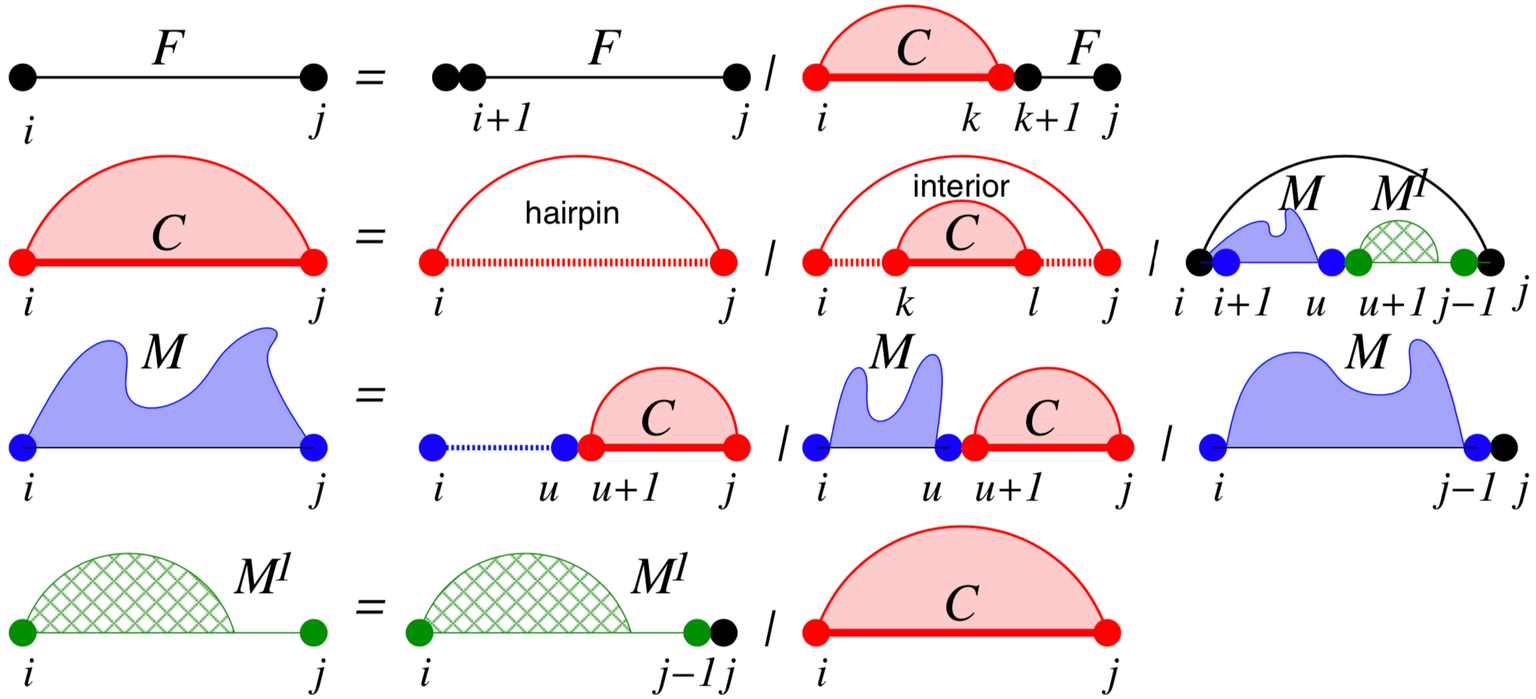
The standard energy model expresses the free energy of a structure as the sum over its loop energies

$$E(S) = \sum_{l \in S} E(l)$$



- Good approximation for most oligonucleotides
- Loop energies depend on loop type/size and some sequence dependence
- Most relevant parameters are experimentally measured; some still guesswork
- Secondary structures are macro-states, hence energies are temperature-dependent free energies
- Training parameters is becoming a viable alternative to experiment

Folding with Loop Based Energies



F_{ij} free energy of the optimal substructure on the subsequence $x[i..j]$.

C_{ij} optimal free energy on $x[i..j]$, where (i, j) pair.

M_{ij} $x[i..j]$ is part of a multiloop and contains at least one pair.

M^1_{ij} same as M_{ij} but contains exactly one component closed by (i, h).

Partition Function

Recall: $\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$

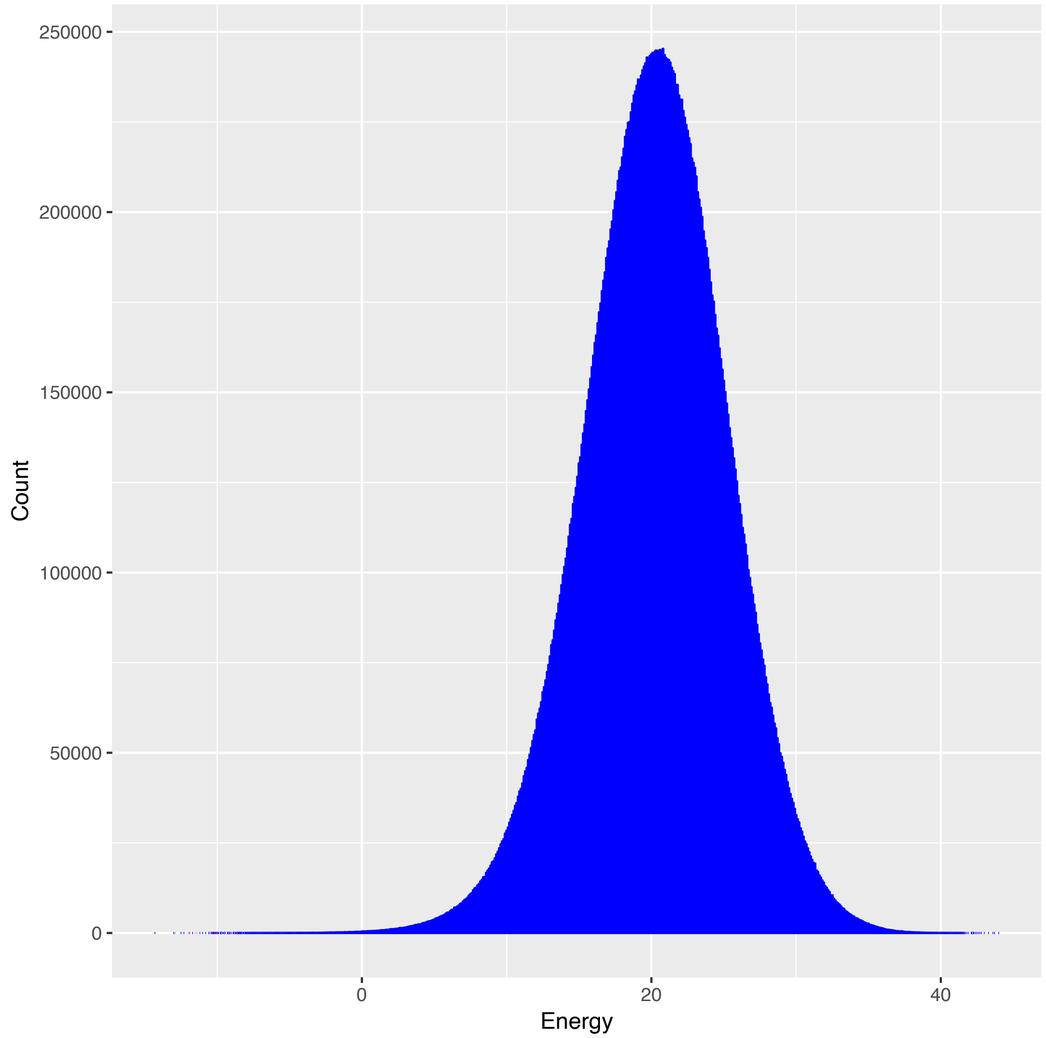
RNA is a biopolymer and ruled by thermodynamics

The partition function is the fundamental quantity of statistical mechanics and all thermodynamic properties can be derived from it

$$Z = \sum_{\Psi} exp\left(-\frac{E(\Psi)}{RT}\right)$$

E.g. the free energy of formation is given by

$$\Delta G = -RT \ln Z$$



Partition Function

Recall: $\bar{S}(n) \sim n^{-\frac{3}{2}} 1.85^n$

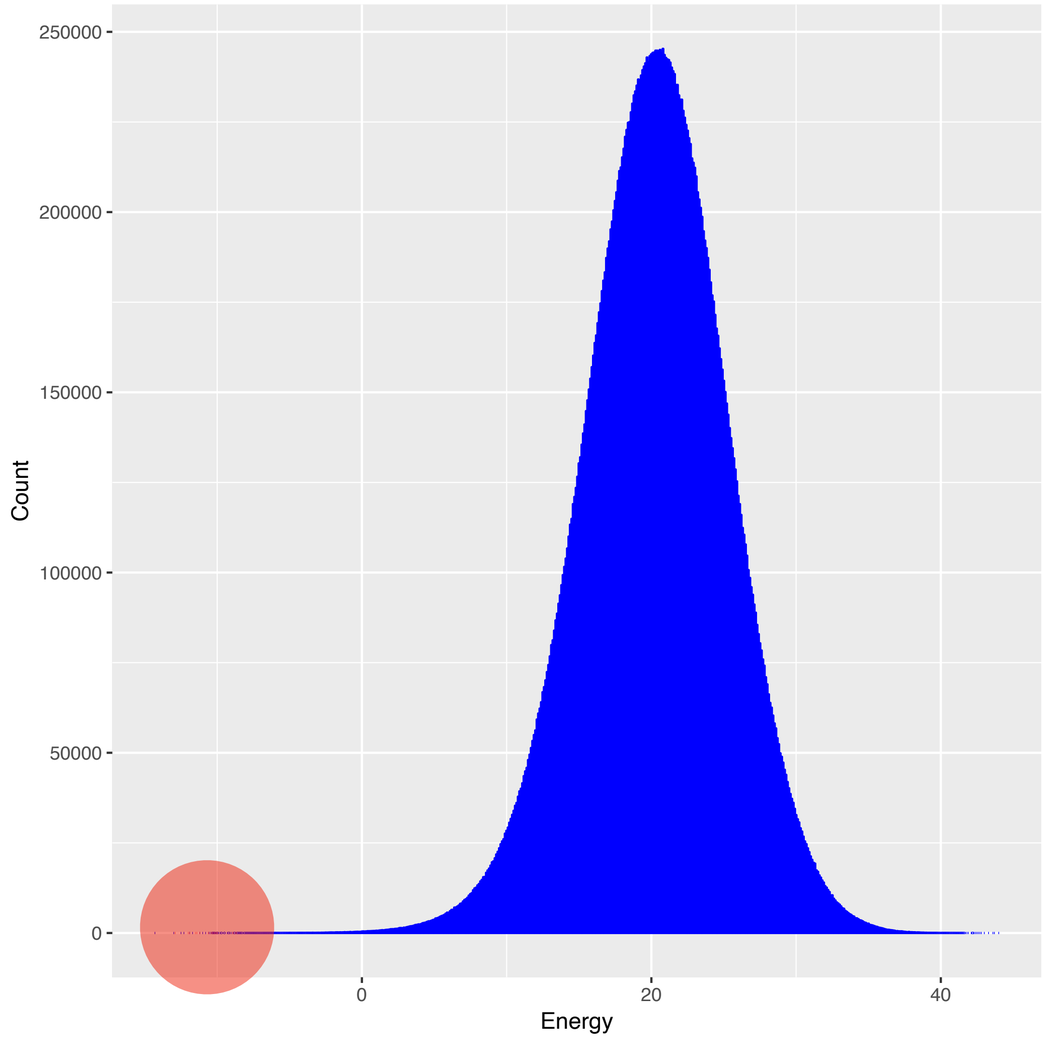
RNA is a biopolymer and ruled by thermodynamics

The partition function is the fundamental quantity of statistical mechanics and all thermodynamic properties can be derived from it

$$Z = \sum_{\Psi} exp\left(-\frac{E(\Psi)}{RT}\right)$$

E.g. the free energy of formation is given by

$$\Delta G = -RT \ln Z$$



Computing the Partition Function

The recursion has the same structure as for energy minimisation, with two differences

- replace minimum operation by sums
- addition of energies by products of partition functions

$$E_{ij} = \min_{i<k\leq j} \left\{ E_{i+1,j}\,;\ \ \left(E_{i+1,k-1} + E_{k+1,j} + \varepsilon(x_i, x_k) \right) \right\}$$

$$Z_{ij} = Z_{i+1,j} + \sum_{k,\,(i,k)\ \text{pairs}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\varepsilon(x_i, x_k)/RT)$$

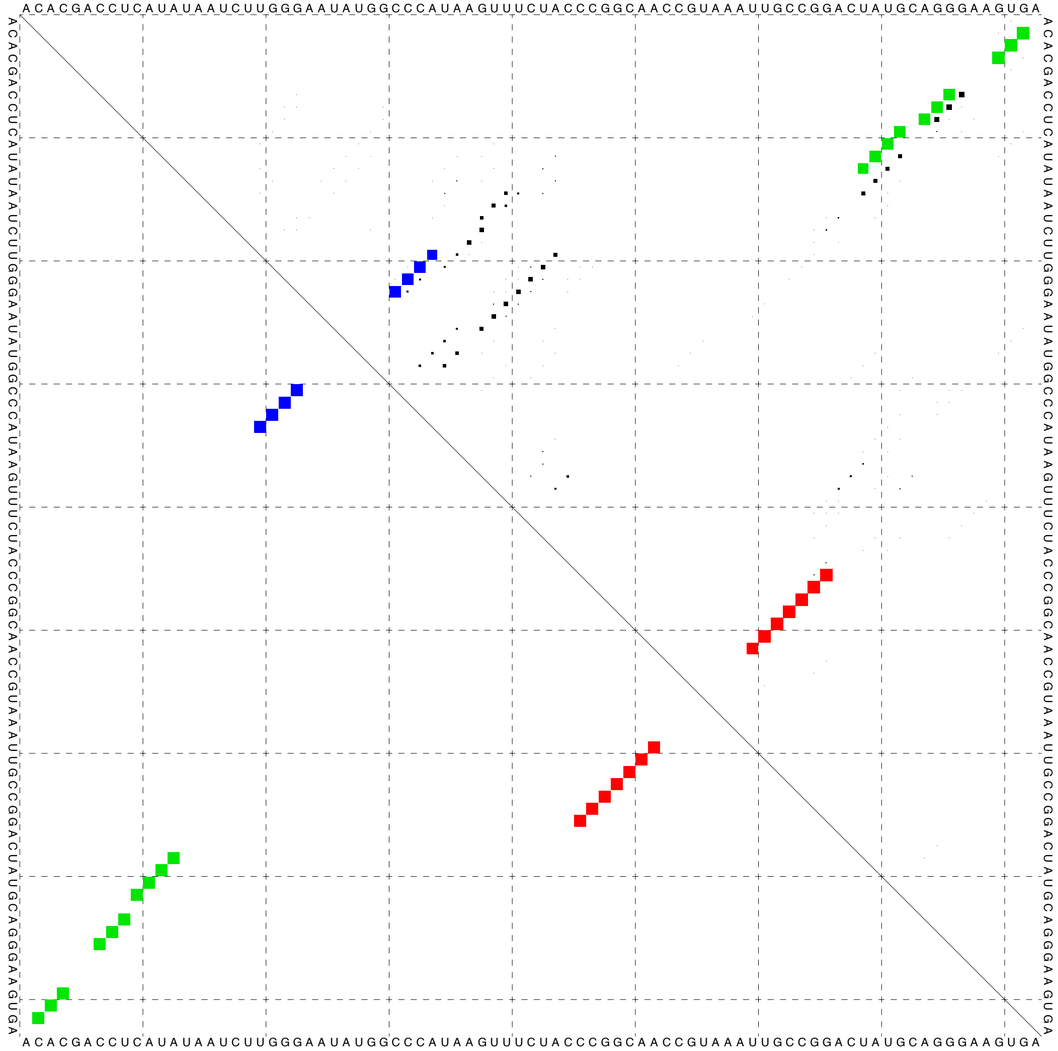
The probability of structure features can be computed from Z, e.g. the probability that a pair is formed

$$p_{ij} = \sum_{\Psi,(i,j)\in\Psi} p(S)$$

For efficient computation define the partition function $\widehat{Z}_{ij}$ for structures outside the subsequence x[i..j]

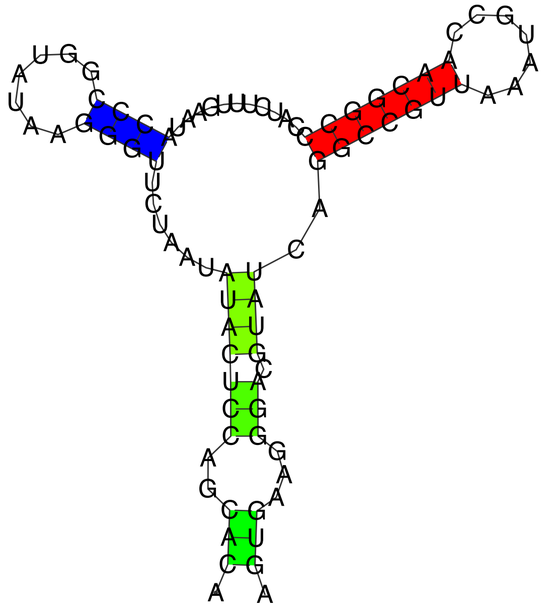
$$p_{ij} = \widehat{Z}_{ij} Z_{i+1,j-1} \exp(-\varepsilon_{ij}/RT)/Z$$

Representing Ensembles of RNA Structures



Ensembles of structures (thermodynamic equilibrium) are best represented by base pair probabilities.

A pair (i , j) with probability p is represented by a square in row i and column j with area p.



The Vienna RNA Package

- Minimum free energy and partition function folding
- Complete suboptimal folding
- Inverse folding / RNA design
- Comparison of secondary structures
- Specific heat curves
- Inclusion of structure probing data
- Analysis of folding kinetics / co-transcriptional folding
- Utilities for plotting and annotation structures
- 2.5D prediction: G-quadruplexes and pseudo-knots
- Prediction of consensus structures

For the programmer:

- A *C* library to link against your programs
- Python/Perl scripting language interface

The Vienna RNA Package

- Minimum free energy and partition function folding
- Complete suboptimal folding
- Inverse folding / RNA design
- Comparison of secondary structures
- Specific heat curves
- Inclusion of structure probing data
- Analysis of folding kinetics / co-transcriptional folding
- Utilities for plotting and annotation structures
- Prediction of consensus structures

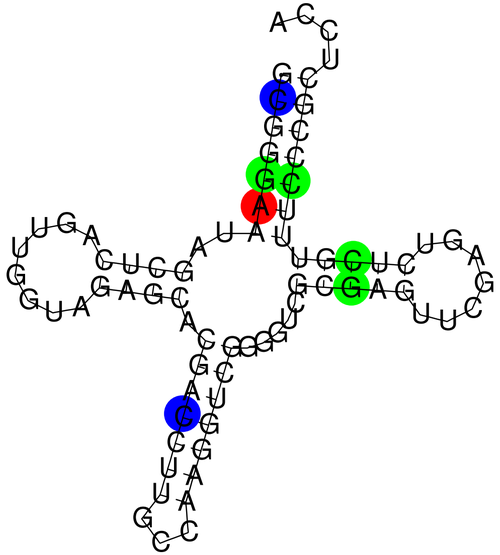
For the programmer:

- A *C* library to link against your programs
- Python/Perl scripting language interface

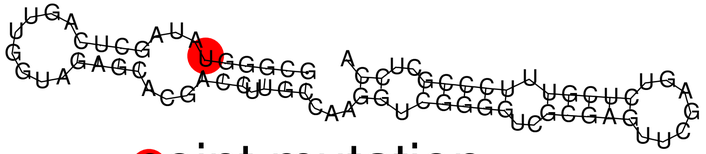
Free software, C source code and fold server available at

http://www.tbi.univie.ac.at/RNA/

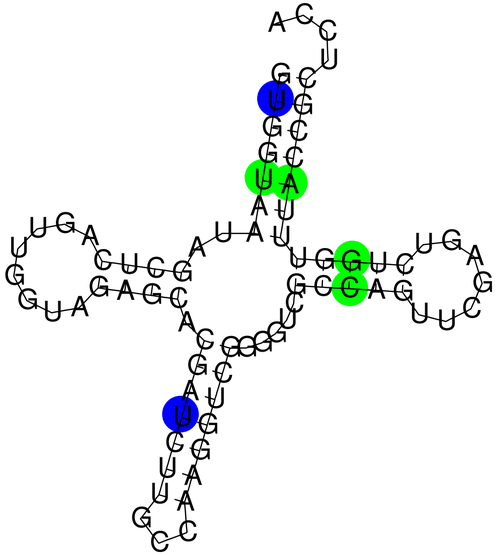
Functional Structures: Point Mutations



original structure

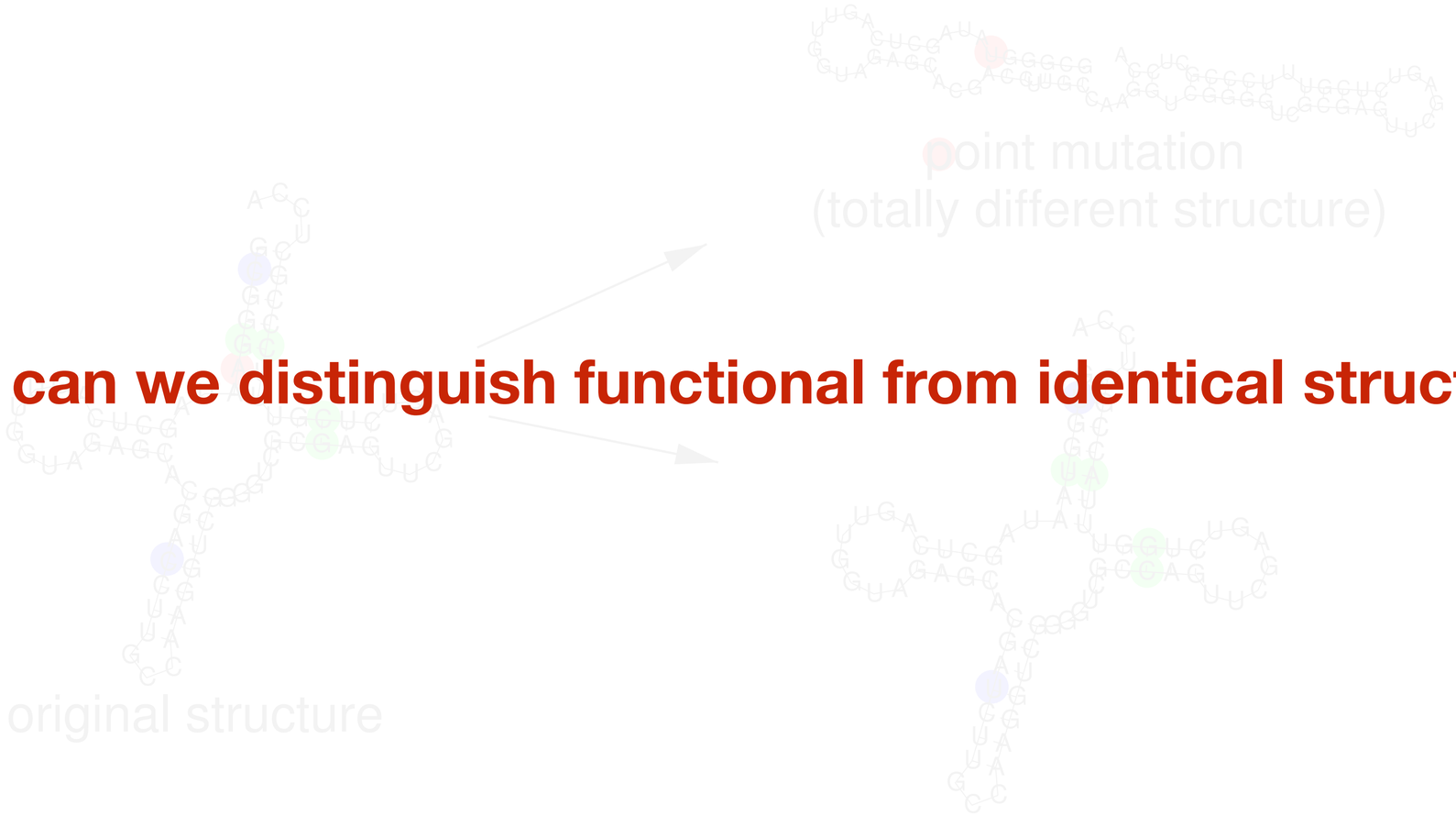


point mutation
(totally different structure)



compensatory and consistent mutations
(no structural change)

Functional Structures: Point Mutations



How can we distinguish functional from identical structures?

Consensus Structures: Alignment Folding

Combine covariance analysis and folding into one DP algorithm

- Apply conventional folding algorithm to alignment
- Use a modified energy function that includes covariance score

$$E_c(A, \Psi) = \sum_k E(A_k, \Psi) + cv \cdot \sum_{(i,j) \in \Psi} B_{ij}$$

- Can be used for all variants: MFE, partition function, …
- Efficient: $\mathcal{O}(N \cdot n^2 + n^3)$ CPU and $\mathcal{O}(n^2)$ memory for alignment length n and N sequences
- Same results as RNAfold for single sequences

Consensus Structures: Alignment Folding

Combine covariance analysis and folding into one DP algorithm

- Apply conventional folding algorithm to alignment
- Use a modified energy function that includes covariance score

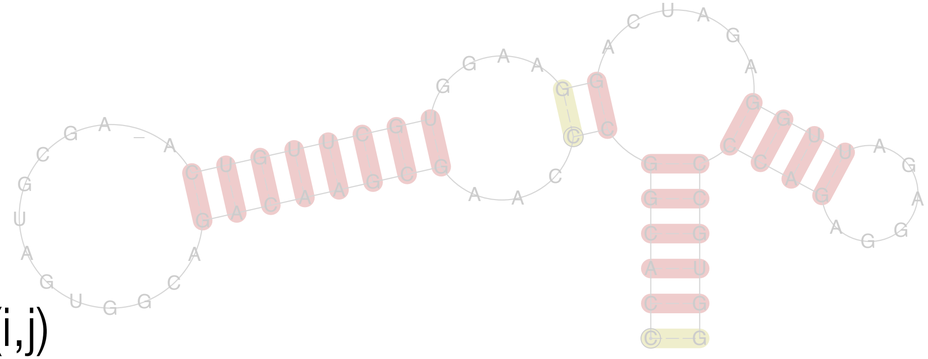
$$E_c(A, \Psi) = \sum_k E(A_k, \Psi) + cv \cdot \sum_{(i,j) \in \Psi} B_{ij}$$

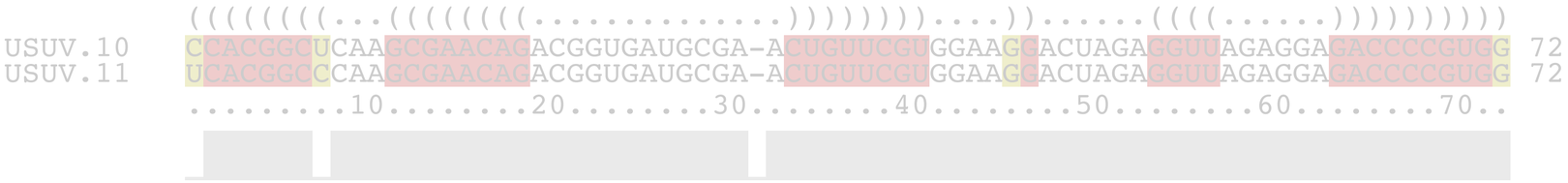
- Can be used for all variants: MFE, partition function, …
- Efficient: $\mathcal{O}(N \cdot n^2 + n^3)$ CPU and $\mathcal{O}(n^2)$ memory for alignment length n and N sequences
- Same results as RNAfold for single sequences

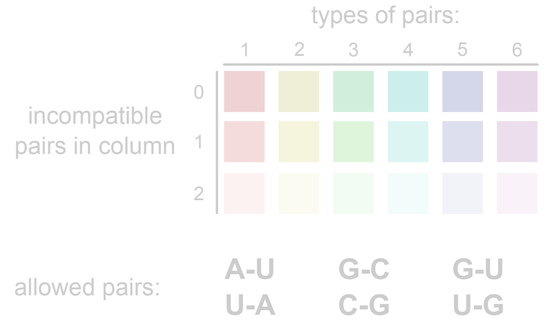
```
GGGCCGGUAGCUCAUUUAGGCAGAGCGUCUGACUCUUAAUCAGACGGUCGCGUGUUCGAAUCGCGUCCGGCCCA
GCCGAAAUAGCUCAGUU-GGGAGAGCGUUAGACUGAAGAUCUAAAGGUCCCCGGUUCAAUCCCGGGUUUCGGCA
GCGGAUUUAGCUCAGUU-GGGAGAGCGCCAGACUGAAGAUUUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCA
CAGAGUGUAGCUUA---ACACAAAGCACCCAACUUACACUUAGGAGAU-UUCAACU-UAACUUGACCGCUCUGA

(((((((..((((.........)))).(((((.......)))))).....(((((.......)))))))))))).
mfe: -29 kcal/mol = (-21.3 +  -7.7) kcal/mol
```

RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

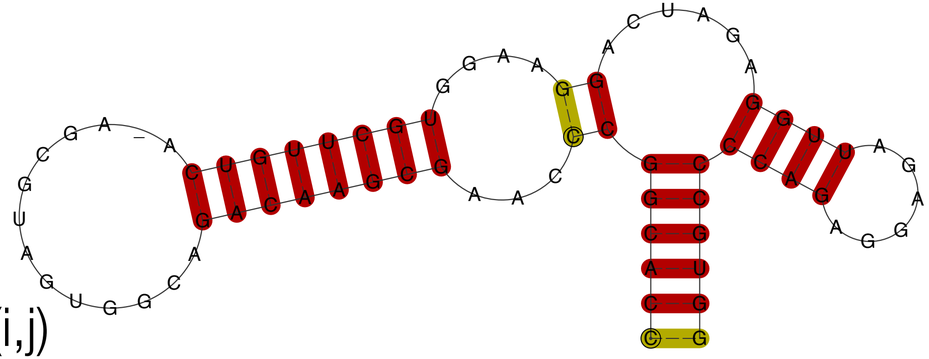




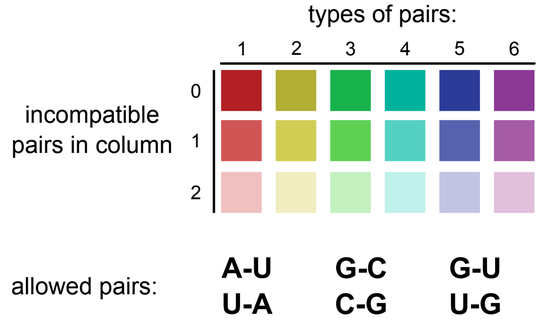


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

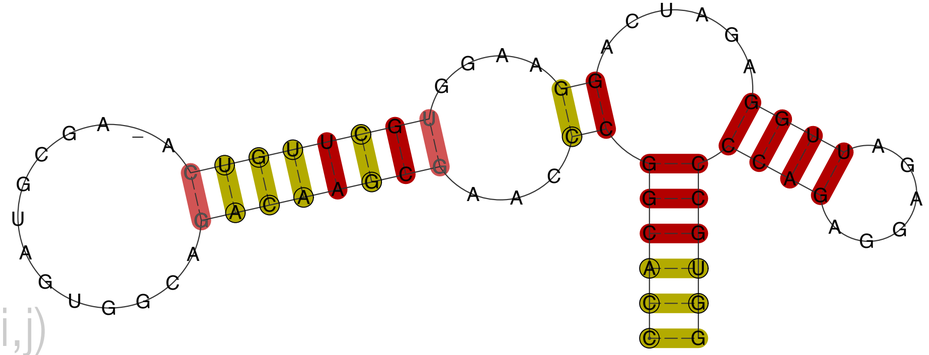


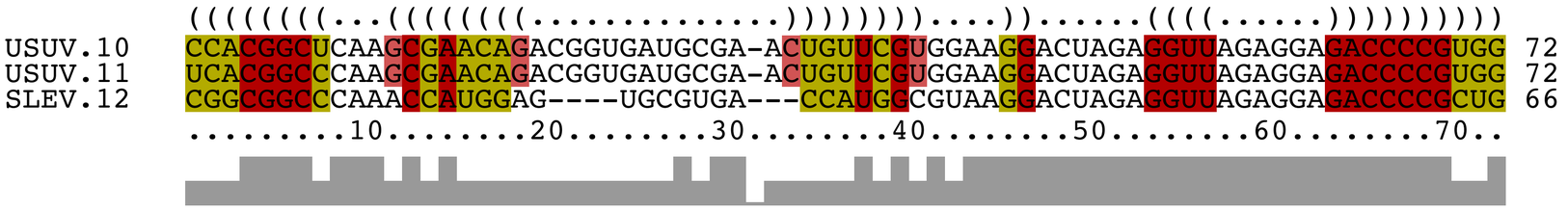
```
                 (((((((((...(((((((((..............)))))))))....)).....(((((.....)))))))))))
USUV.10          CCACGGCUCAAGCGAACAGACGGUGAUGCGA-ACUGUUCGUGGAAGGACUAGAGGUUAGAGGAGACCCCGUGG 72
USUV.11          UCACGGCCCAAGCGAACAGACGGUGAUGCGA-ACUGUUCGUGGAAGGACUAGAGGUUAGAGGAGACCCCGUGG 72
                 .........10........20........30........40........50........60........70..
```

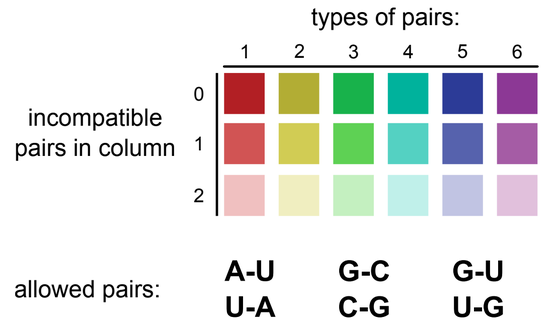


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

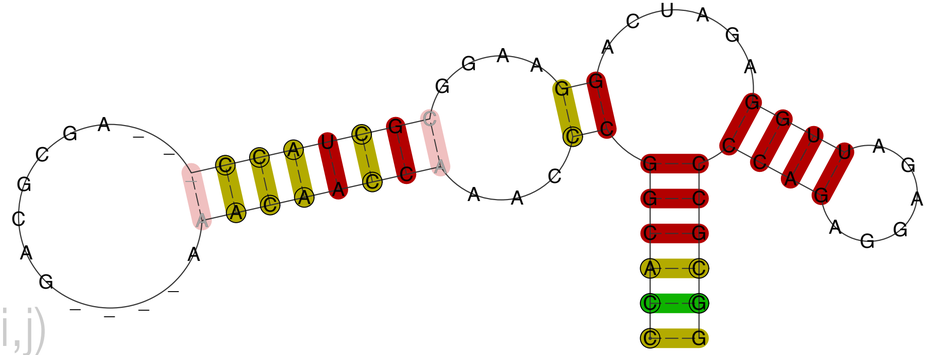


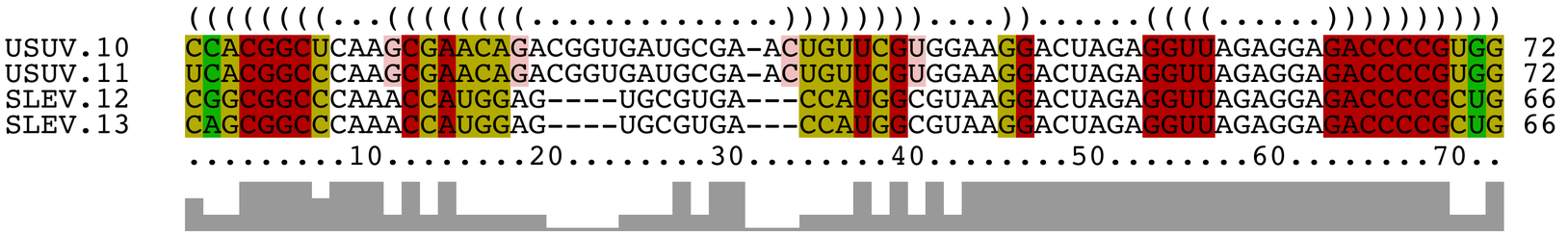


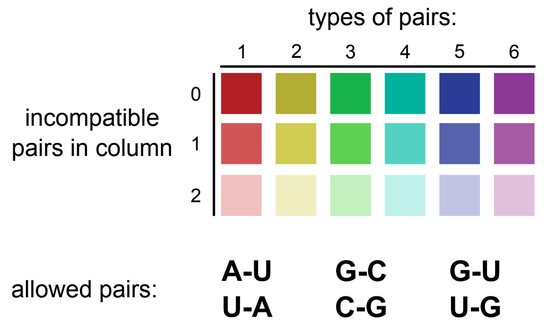


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

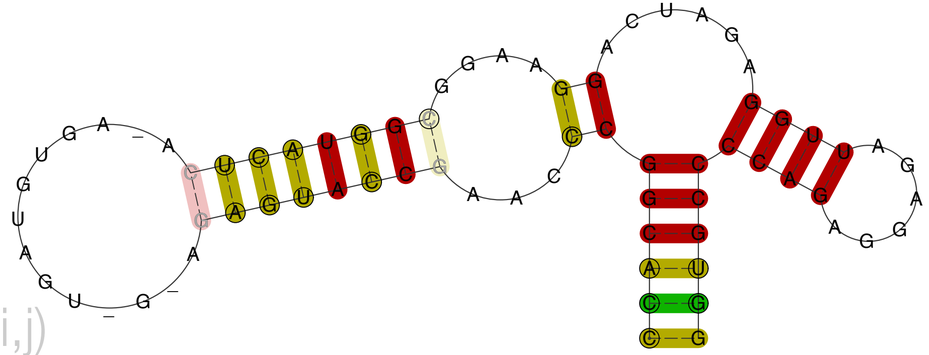


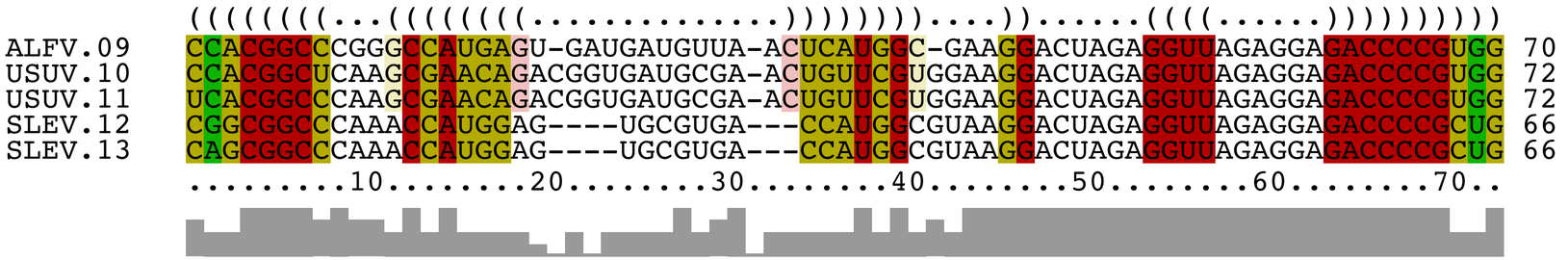


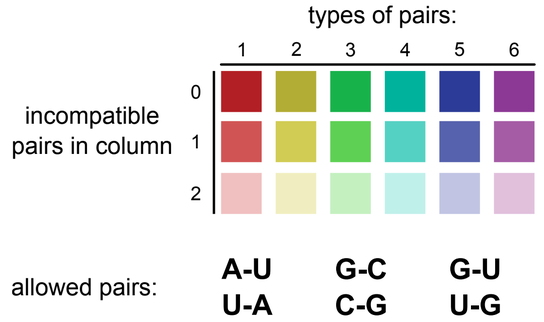


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

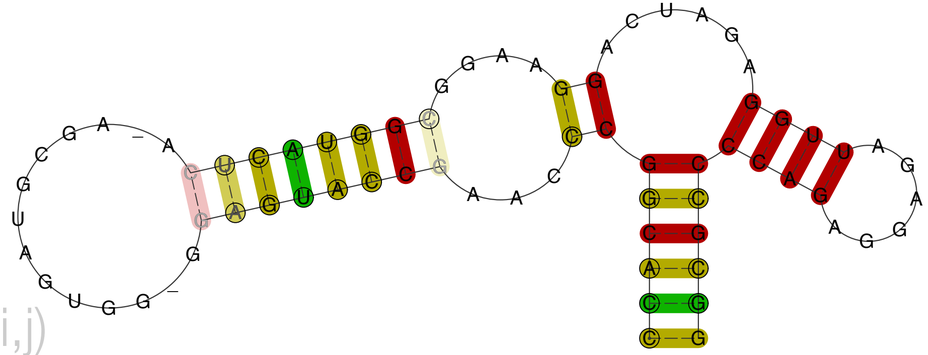


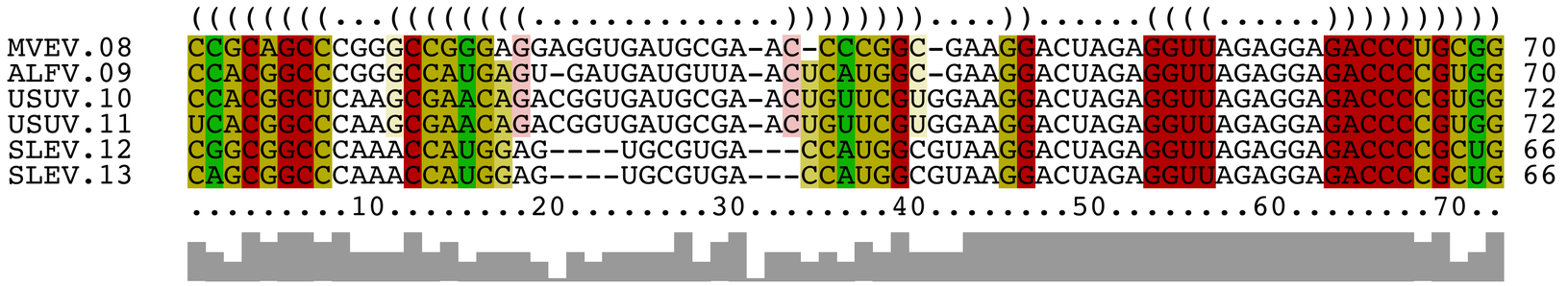


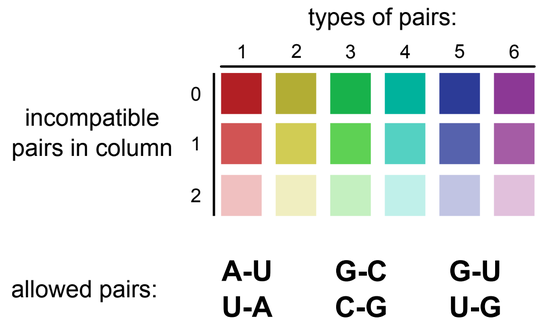


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

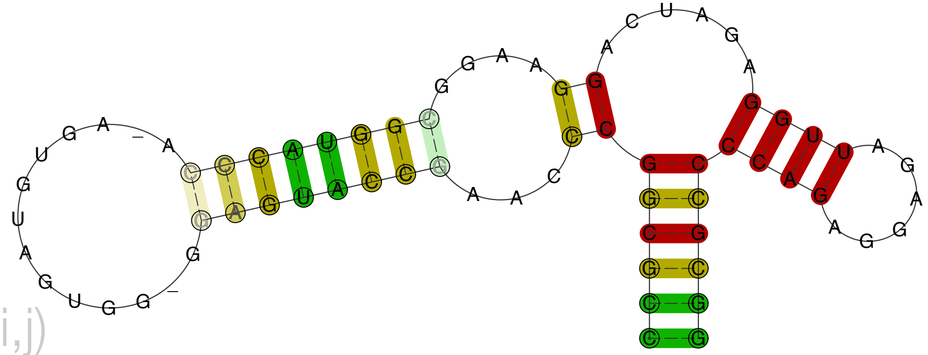


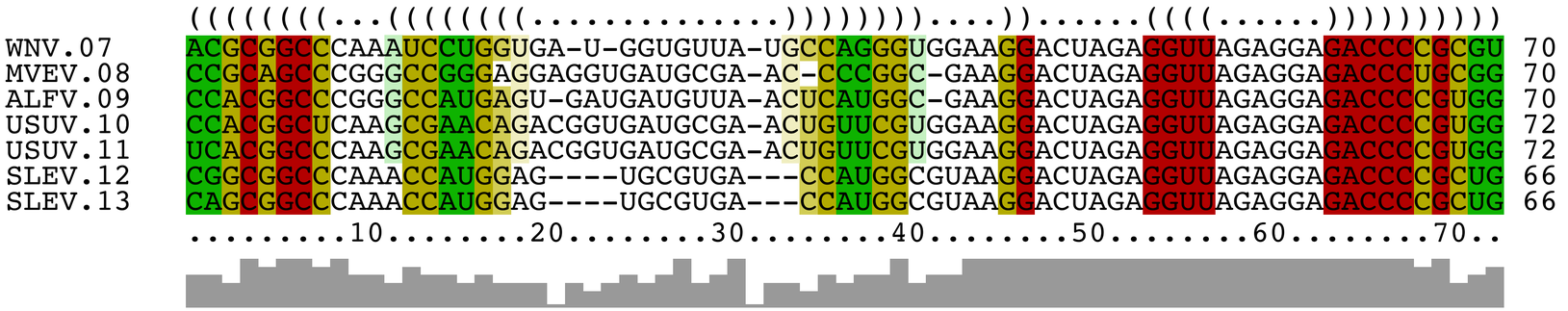


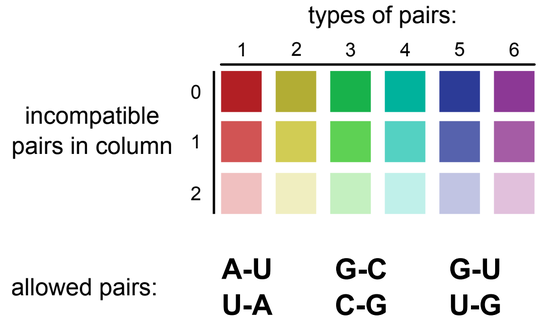


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

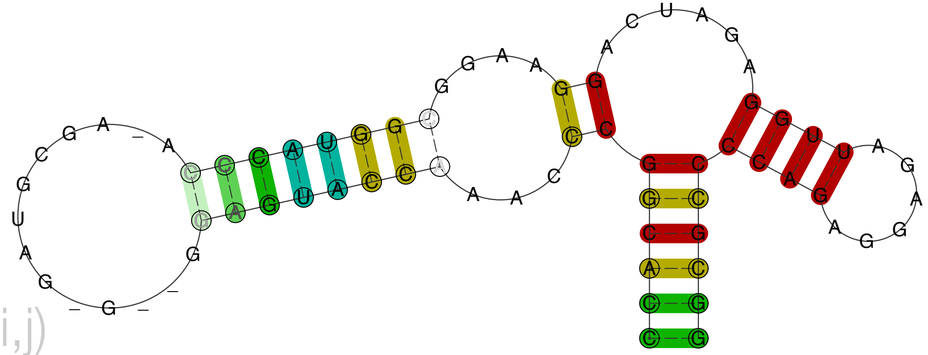


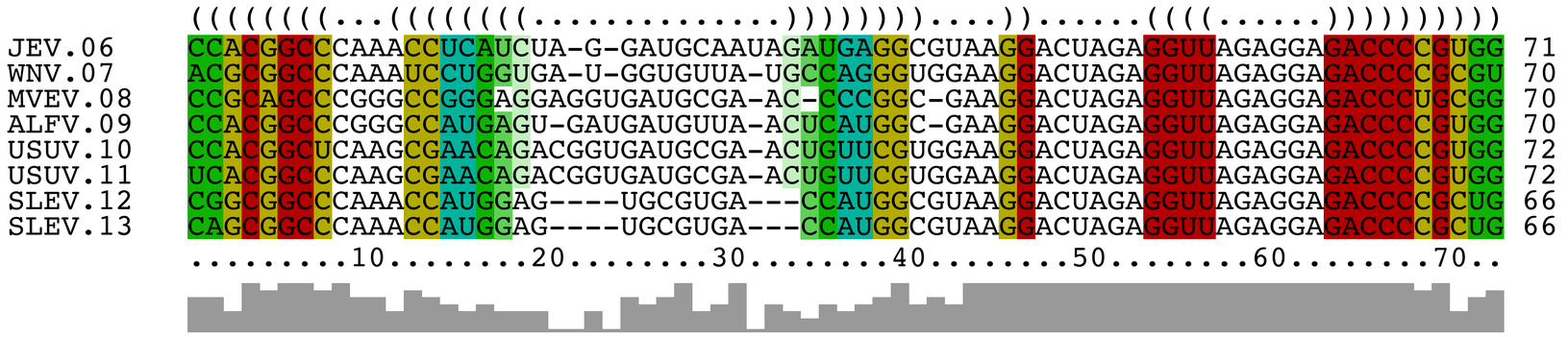


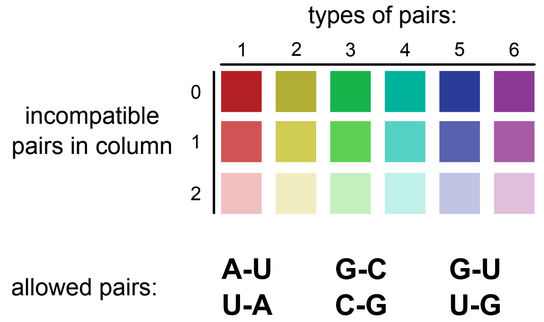


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

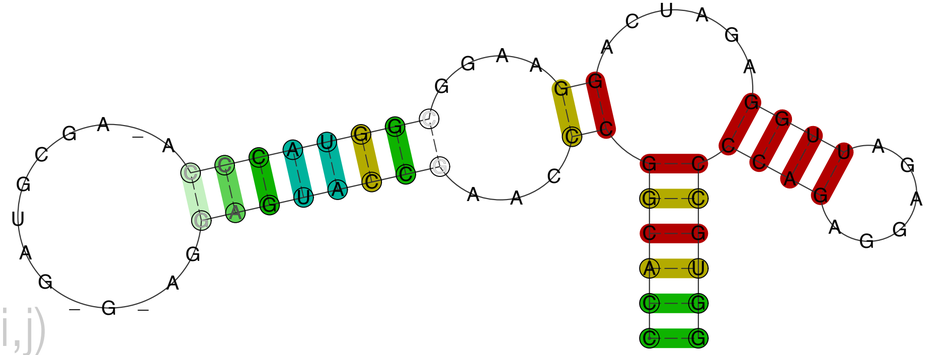


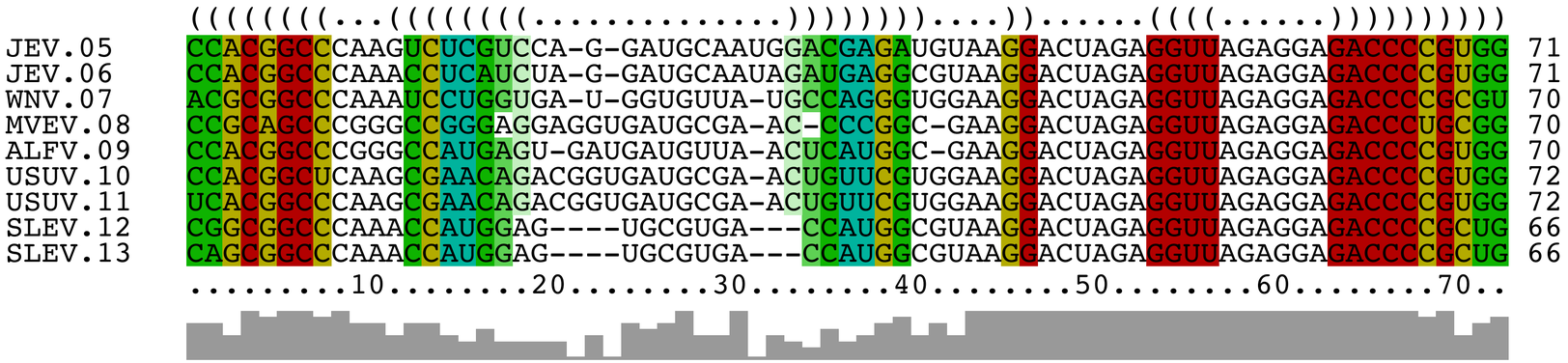


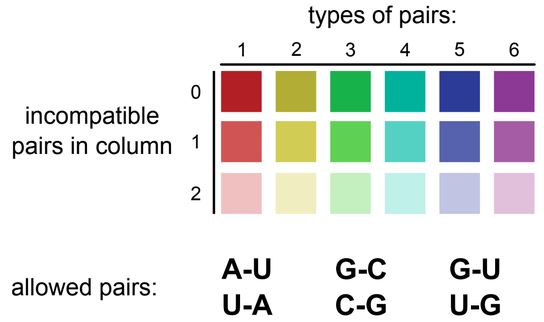


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

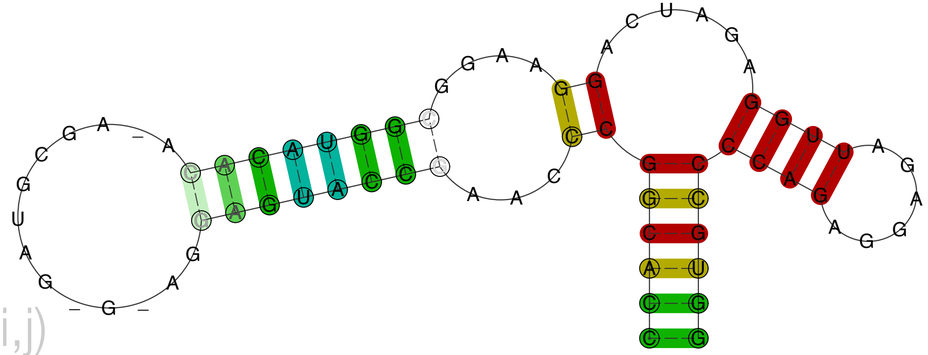


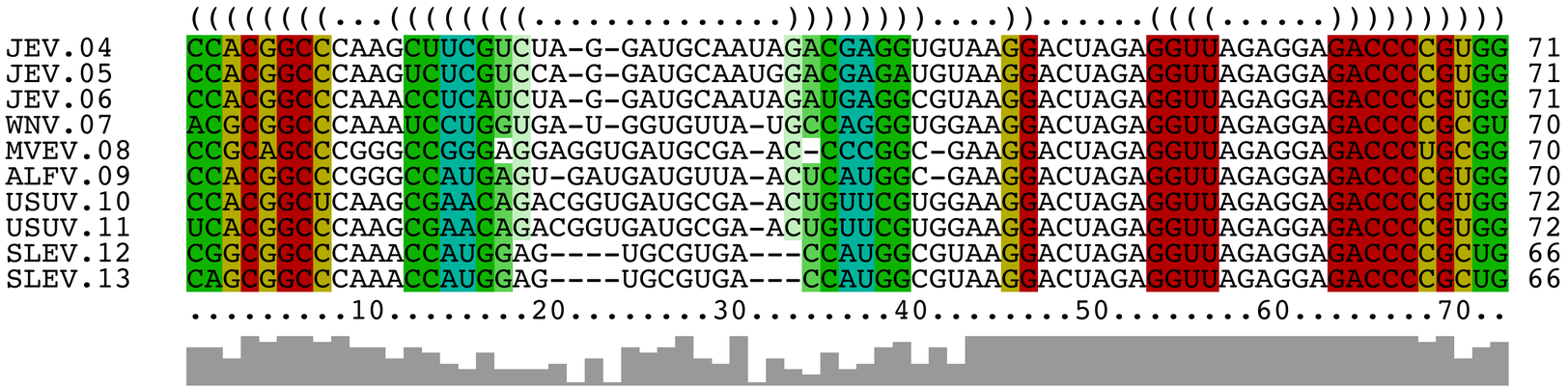


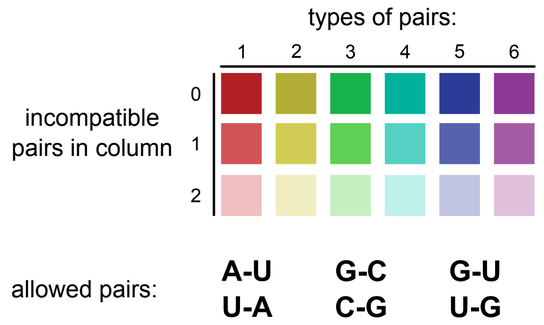


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

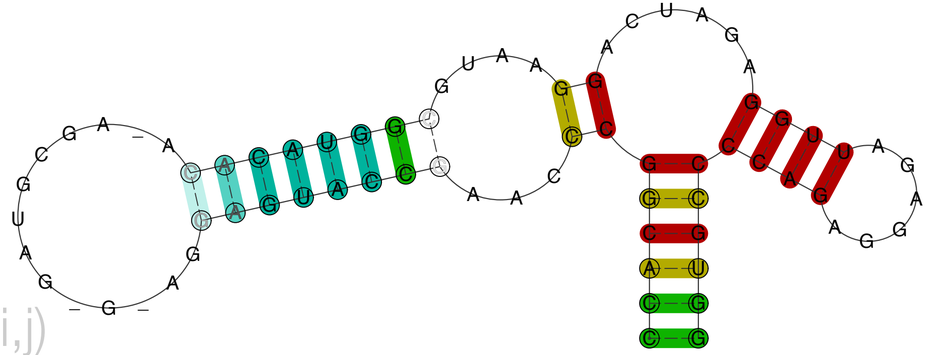


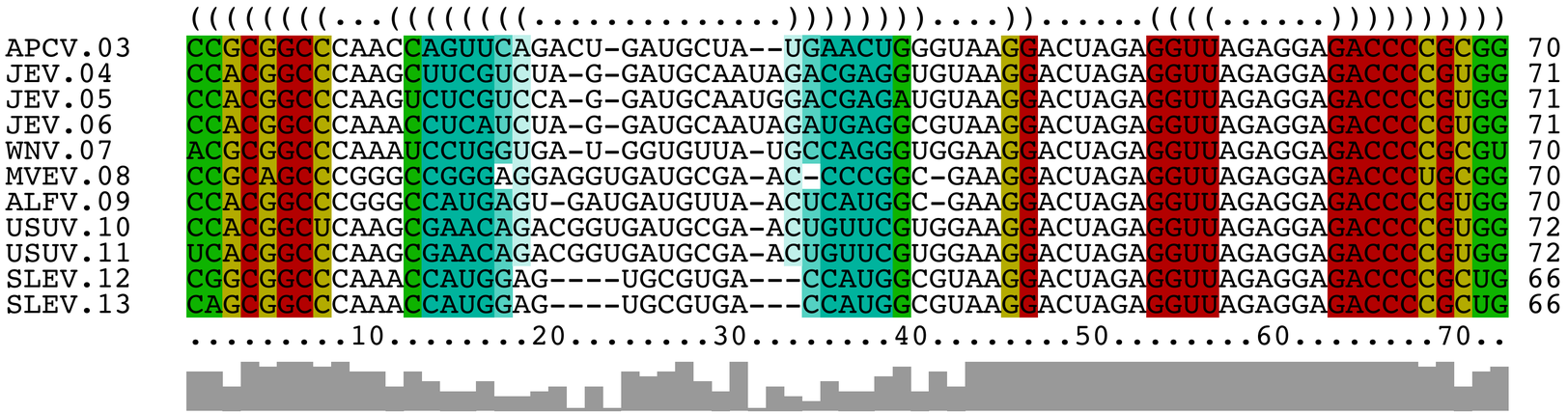


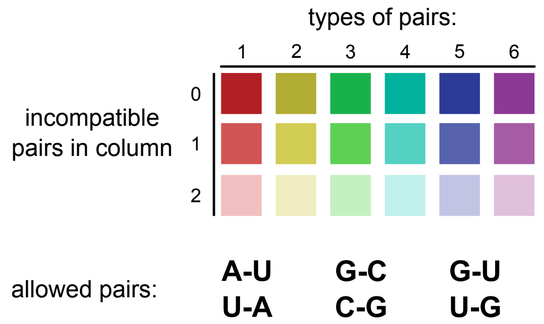


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

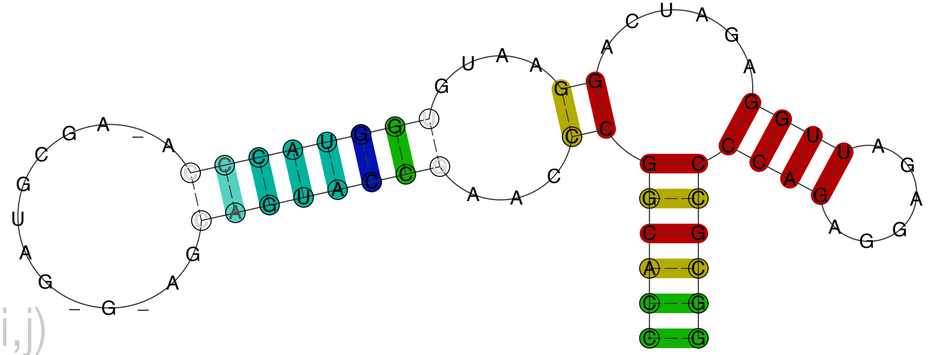


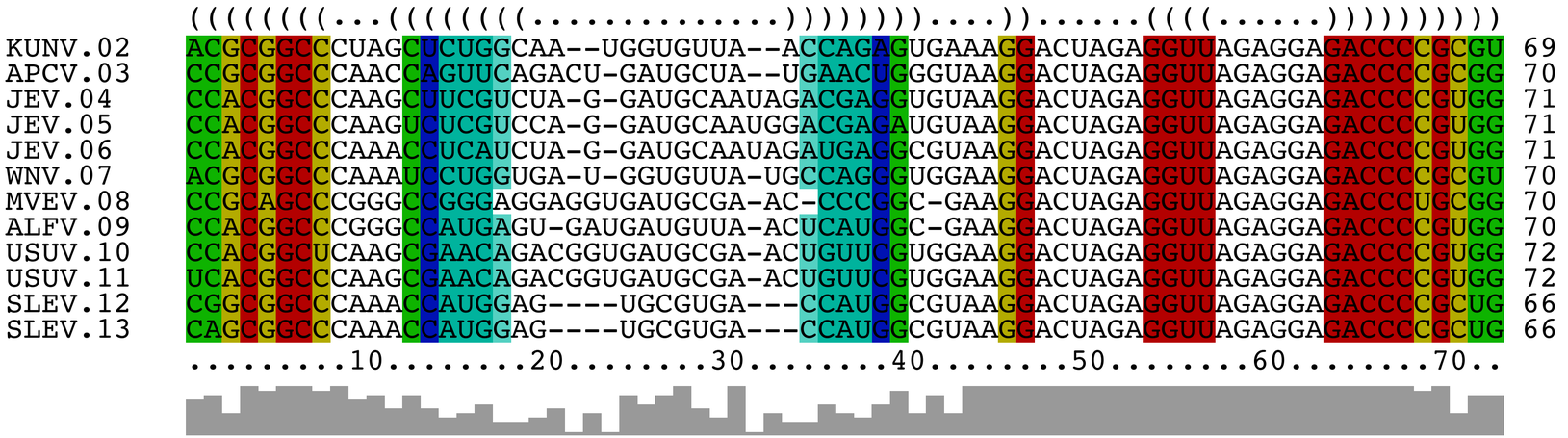


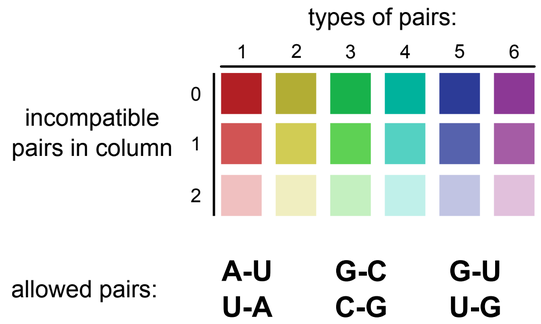


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

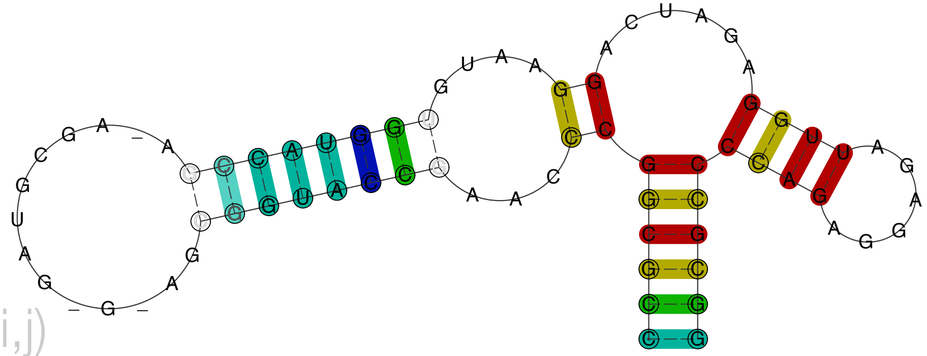


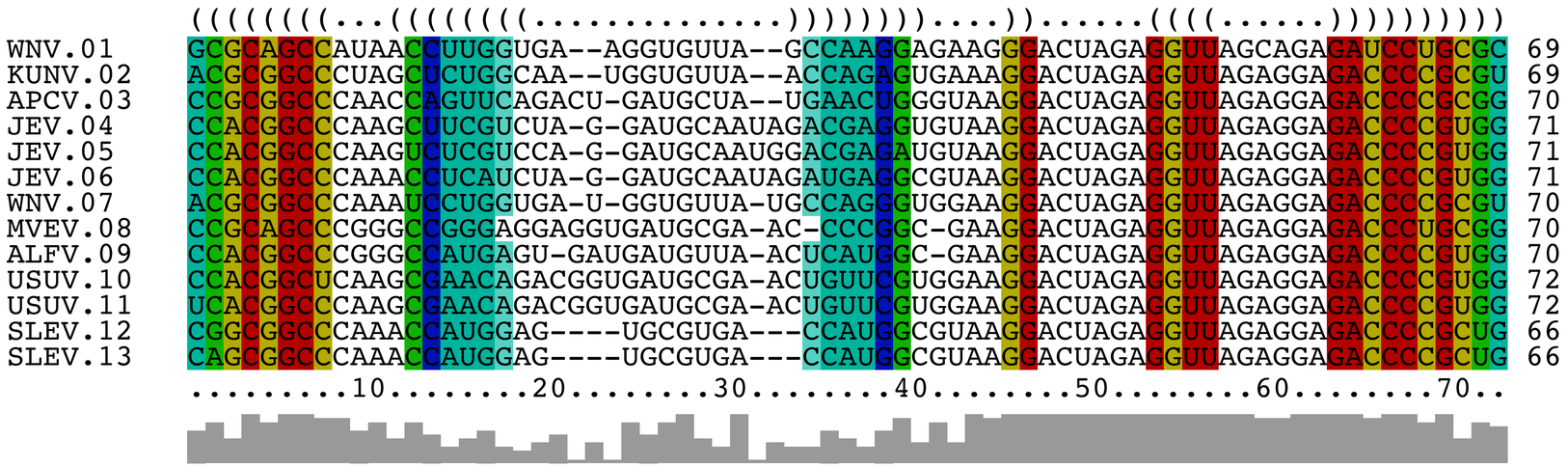


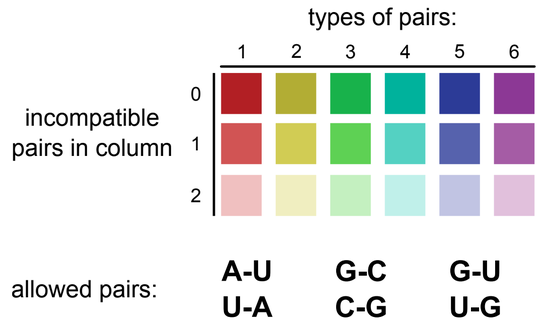


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- For base pair (i,j): GC/CG/AU/UA/GU/UG
- Consistent mutation: different standard combinations
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)



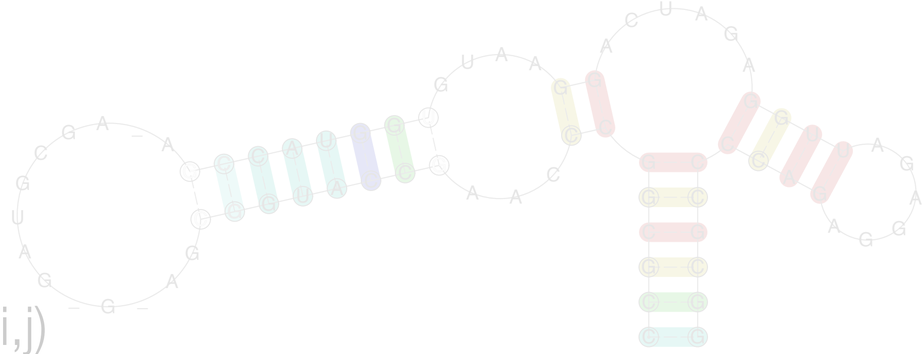


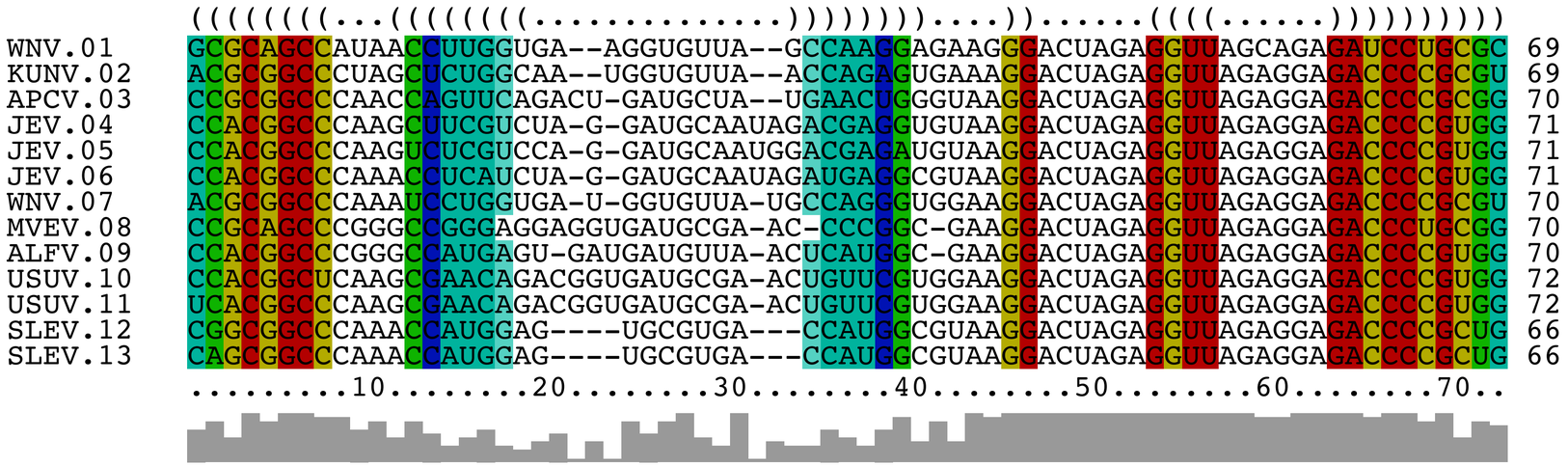


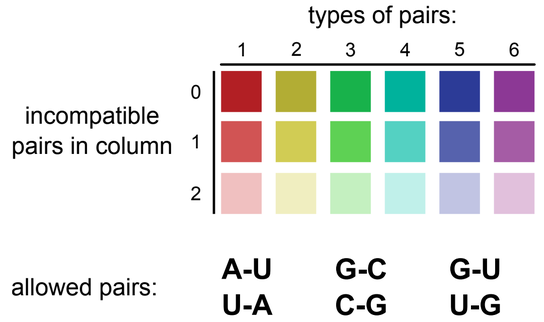
RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- F
- C
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

We use structural RNA alignments and Covariance Models to find conserved elements





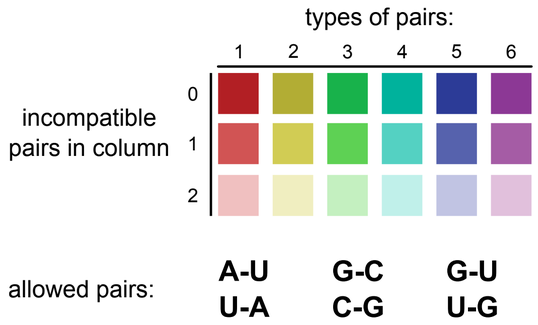


RNA Covariation as Evolutionary Trait

- High mutation rate in RNA viruses due to error-prone RdRP
- R
- C
- Compensatory mutation: both positions are mutated
- Presence of both strongly supports predicted base pair (i,j)

We use structural RNA alignments and Covariance Models to find conserved elements

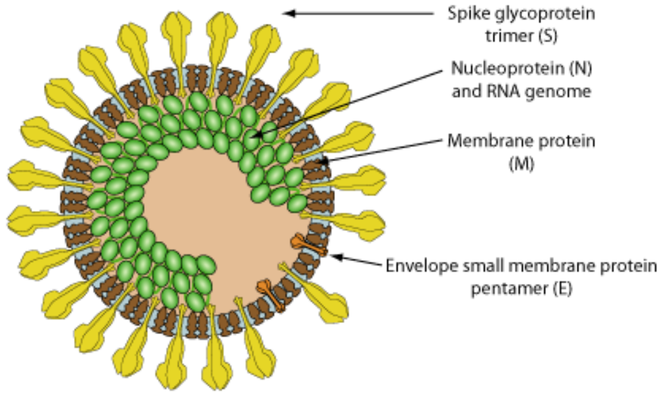

```
                 (((((((((...(((((((.............)))))))....)).....(((.....)))))))))
WNV.01           GCGCAGCCAUAACCUUGGUGA--AGGUGUUA--GCCAAGGAGAAGGGACUAGAGGUUAGCAGAGAUCCUGCGC 69
KUNV.02          ACGCGGCCCUAGCUCUGGCAA--UGGUGUUA--ACCAGAGUGAAAGGACUAGAGGUUAGAGGAGACCCCGCGU 69
APCV.03          CCGCGGCCCAACCAGUUCAGACU-GAUGCUA--UGAACUGGGUAAGGACUAGAGGUUAGAGGAGACCCCGCGG 70
JEV.04           CCACGGCCCAAGCUUCGUCUA-G-GAUGCAAUAGACGAGGUGUAAGGACUAGAGGUUAGAGGAGACCCCGUGG 71
JEV.05           CCACGGCCCAAGUCUCGUCCA-G-GAUGCAAUGGACGAGAUGUAAGGACUAGAGGUUAGAGGAGACCCCGUGG 71
JEV.06           CCACGGCCCAAACCUCAUCUA-G-GAUGCAAUAGAUGAGGCGUAAGGACUAGAGGUUAGAGGAGACCCCGUGG 71
WNV.07           ACGCGGCCCAAAUCCUGGUGA-U-GGUGUUA-UGCCAGGGUGGAAGGACUAGAGGUUAGAGGAGACCCCGCGU 70
MVEV.08          CCGCAGCCCGGGCCGGGAGGAGGUGAUGCGA-AC-CCCGGC-GAAGGACUAGAGGUUAGAGGAGACCCUGCGG 70
ALFV.09          CCACGGCCCGGGCCAUGAGU-GAUGAUGUUA-ACUCAUGGC-GAAGGACUAGAGGUUAGAGGAGACCCCGUGG 70
USUV.10          CCACGGCUCAAGCGAACAGACGGUGAUGCGA-ACUGUUCGUGGAAGGACUAGAGGUUAGAGGAGACCCCGUGG 72
USUV.11          UCACGGCCCAAGCGAACAGACGGUGAUGCGA-ACUGUUCGUGGAAGGACUAGAGGUUAGAGGAGACCCCGUGG 72
SLEV.12          CGGCGGCCCAAACCAUGGAG----UGCGUGA---CCAUGGCGUAAGGACUAGAGGUUAGAGGAGACCCCGCUG 66
SLEV.13          CAGCGGCCCAAACCAUGGAG----UGCGUGA---CCAUGGCGUAAGGACUAGAGGUUAGAGGAGACCCCGCUG 66
                 .........10.........20.........30.........40.........50.........60.........70..
```

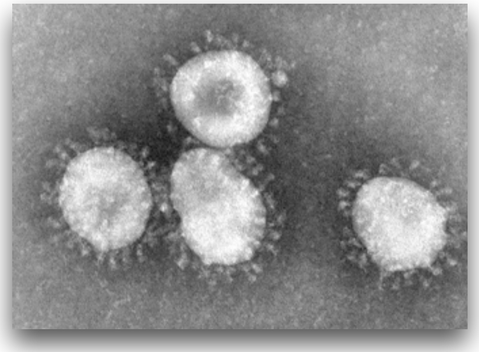


Part II:

Coronaviruses

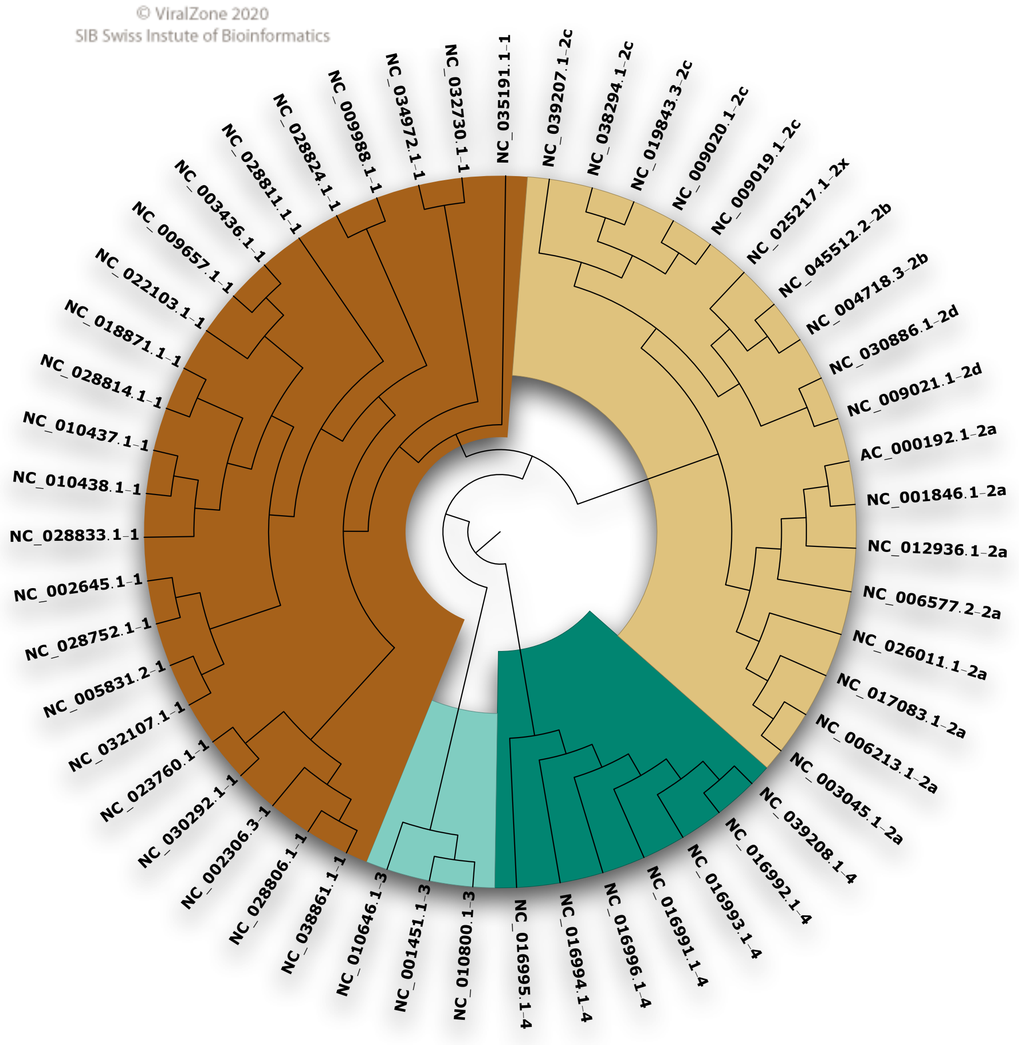
Coronaviruses (CoV)



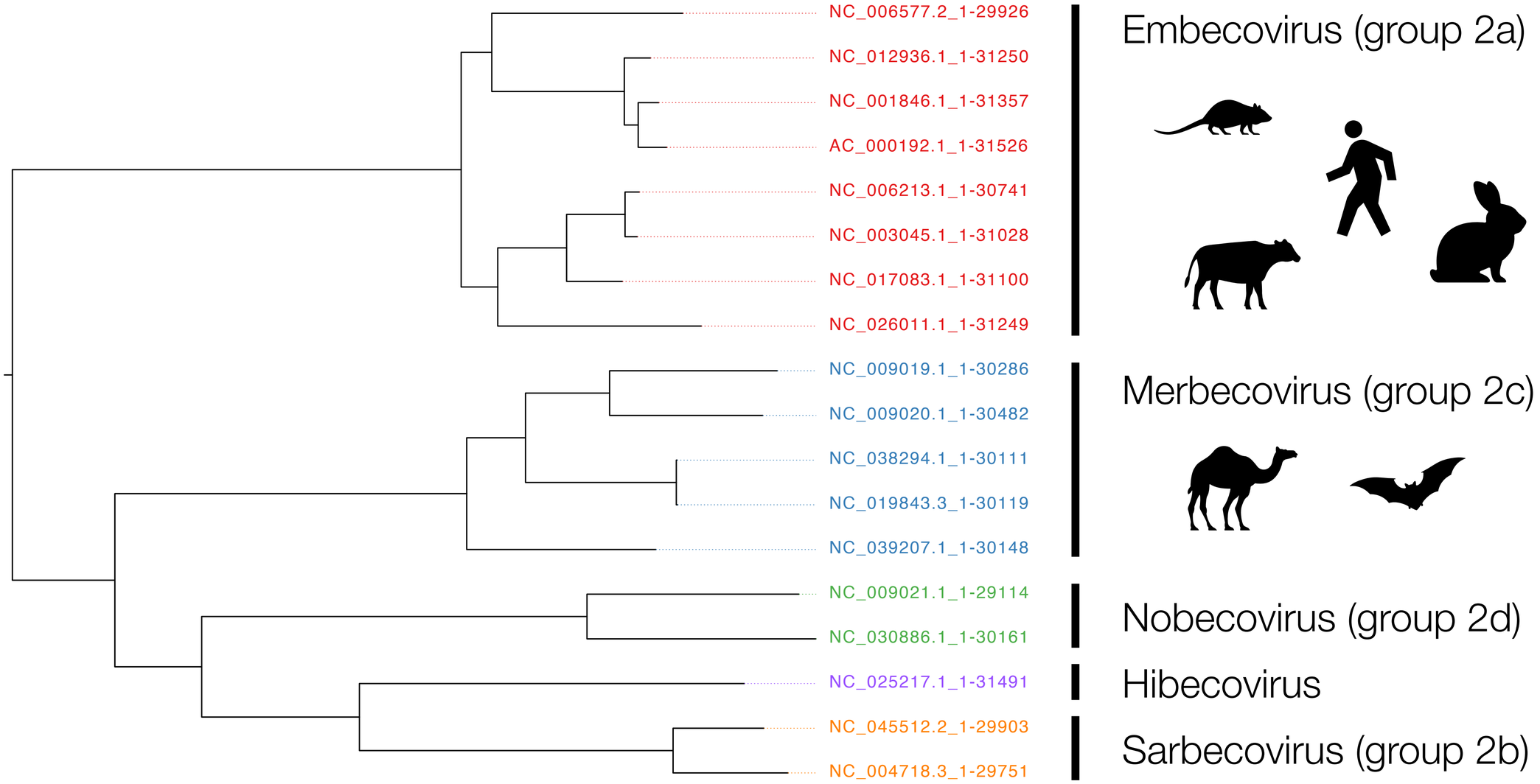


Dr. Fred Murphy, CDC

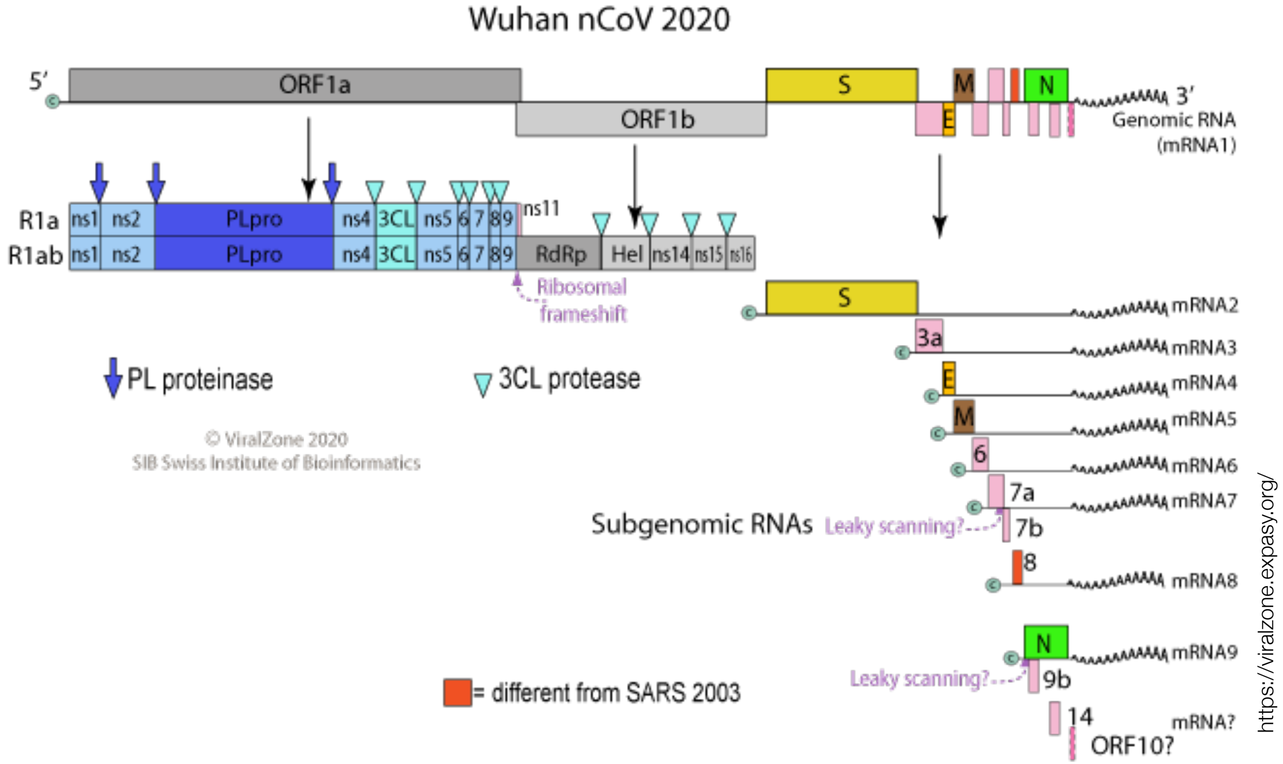
- Order *Nidovirales* / family *Coronaviridae* / genus *Alpha|Beta|Gamma|Deltacoronavirus*
- Cause respiratory infections in humans
- Some CoVs cause outbreaks and are a global health concern (SARS, MERS)
- Others circulate continuously, causing influenza-like illness or common cold
- CoVs infect animals; outbreaks result from spillover events
- Human-to-human transmission
- Only experimental vaccines

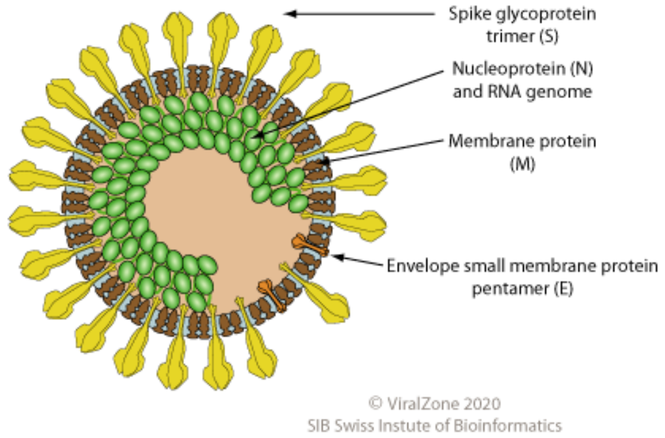


Betacoronavirus phylogeny



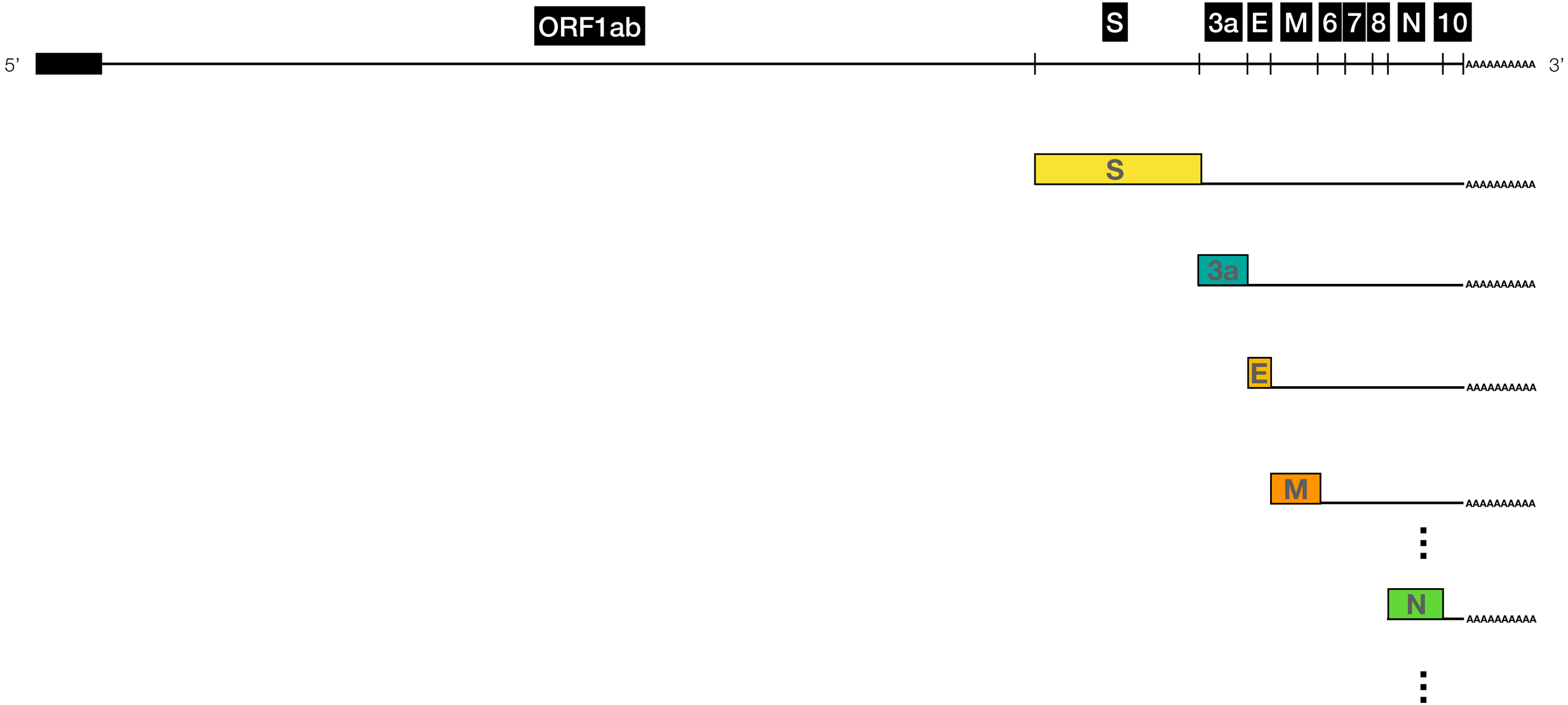
CoV Genome Organization



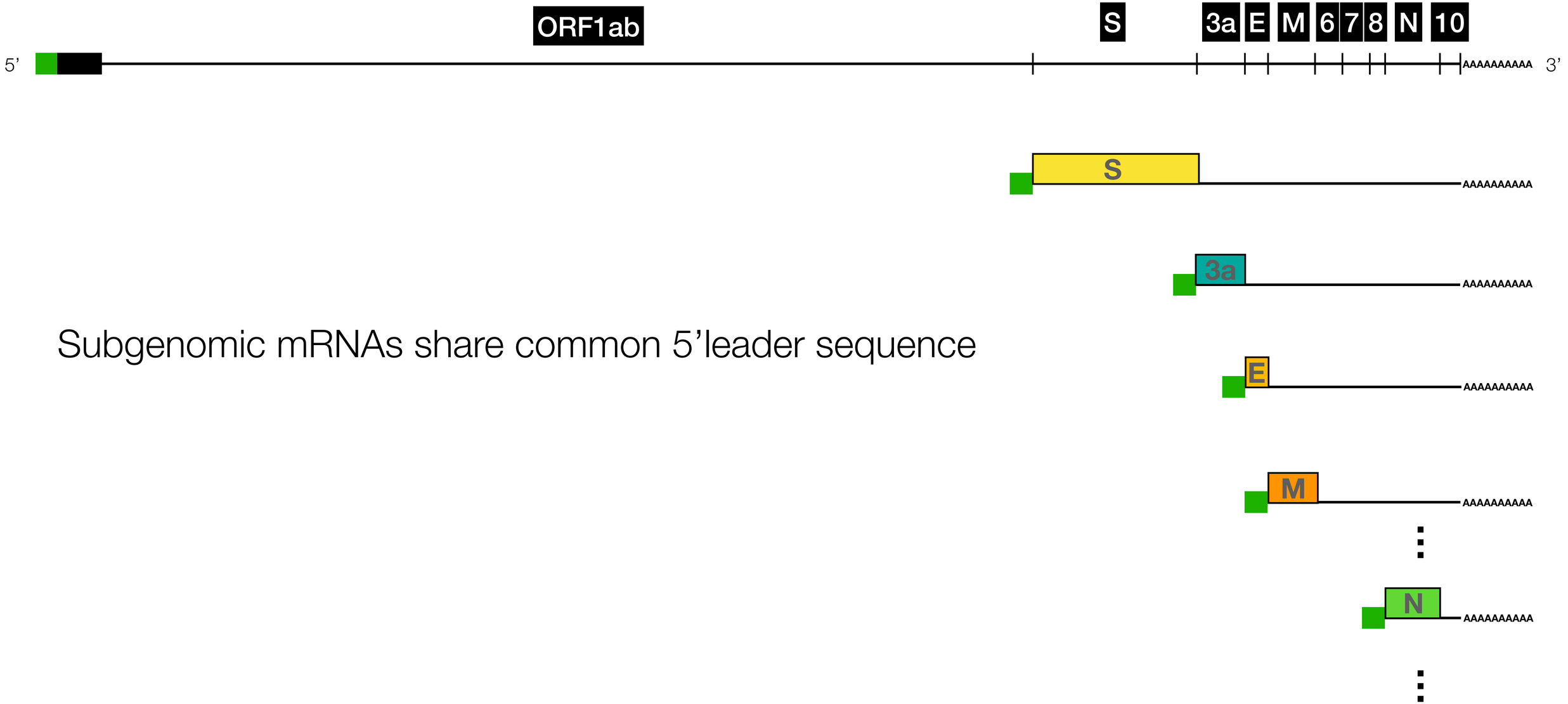


- Non-segmented, linear ssRNA(+) genomes of 27-32kb length (largest RNA viruses)
- Capped and polyadenylated
- Genomic RNA encodes ORF1a & ORF1b, yielding RdRp & non-structural proteins
- Subgenomic RNAs encode structural proteins

CoV RNA species

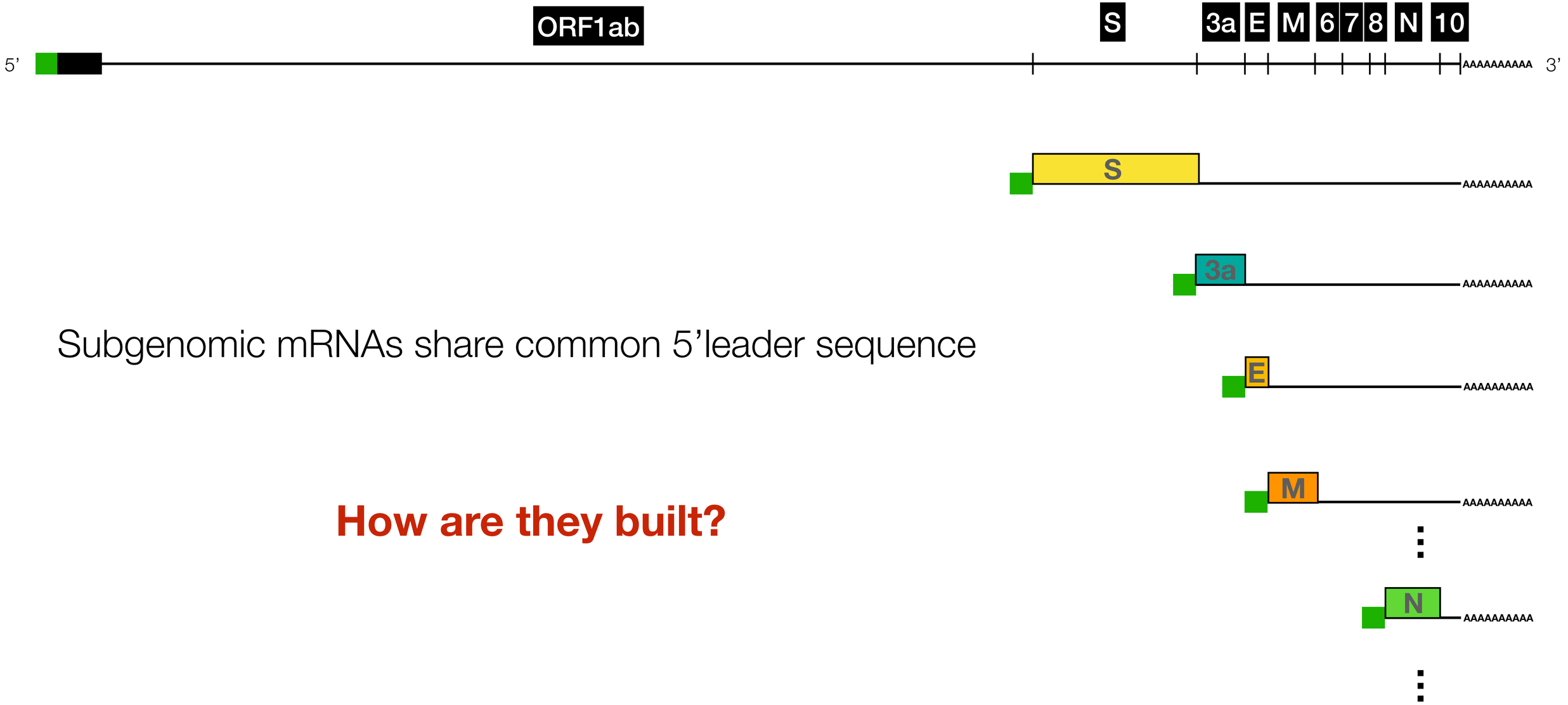


CoV Subgenomic mRNAs



Subgenomic mRNAs share common 5’leader sequence

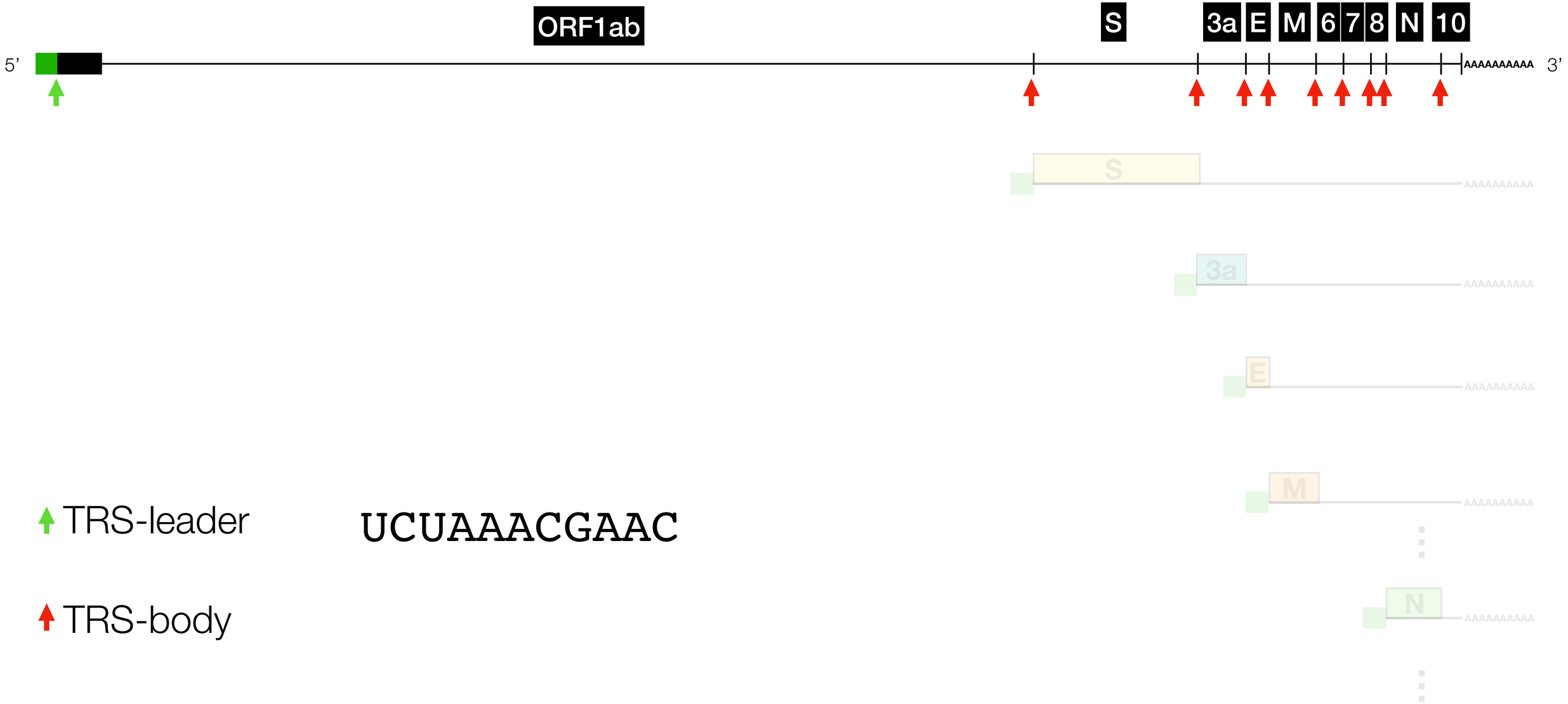
CoV Subgenomic mRNAs



Subgenomic mRNAs share common 5’leader sequence

How are they built?

CoV Discontinuous Transcription

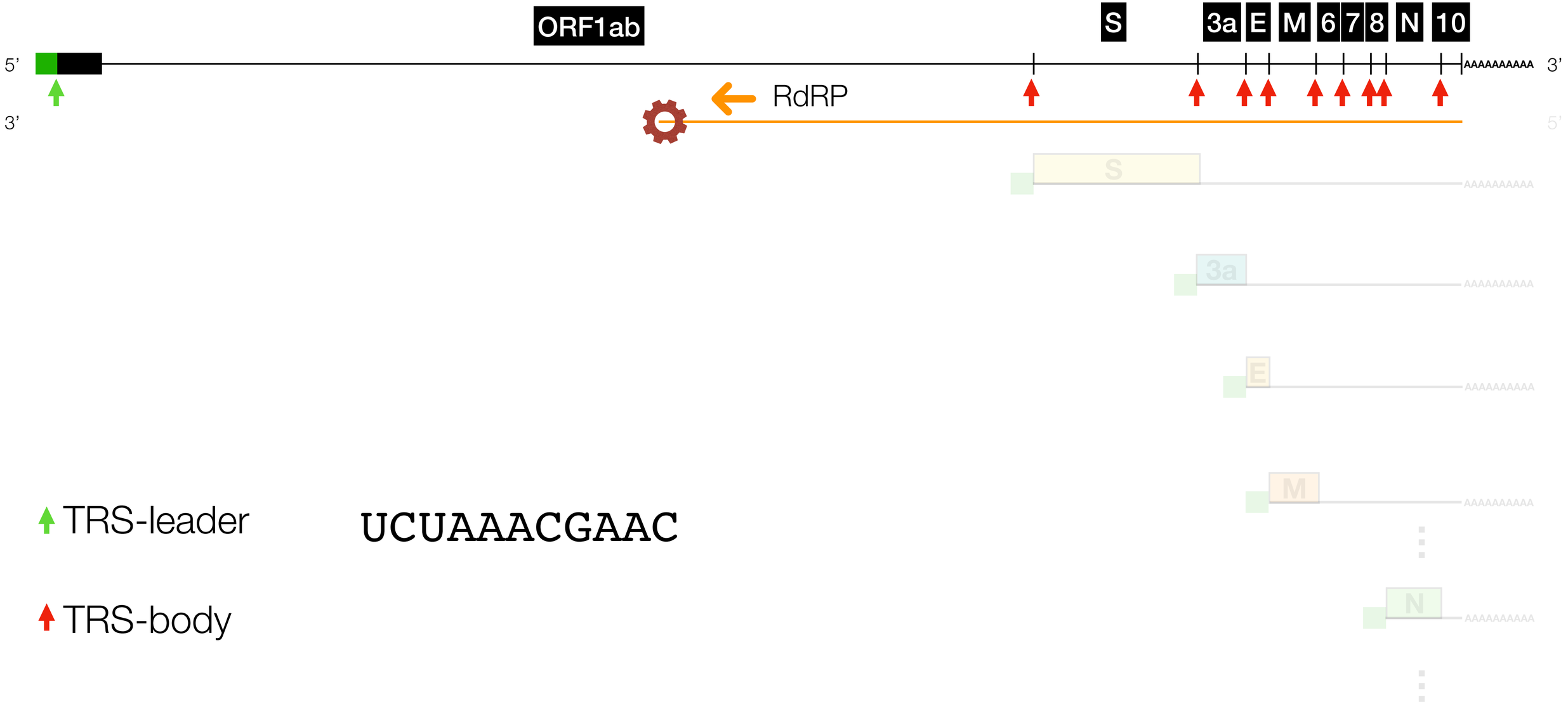


TRS-leader

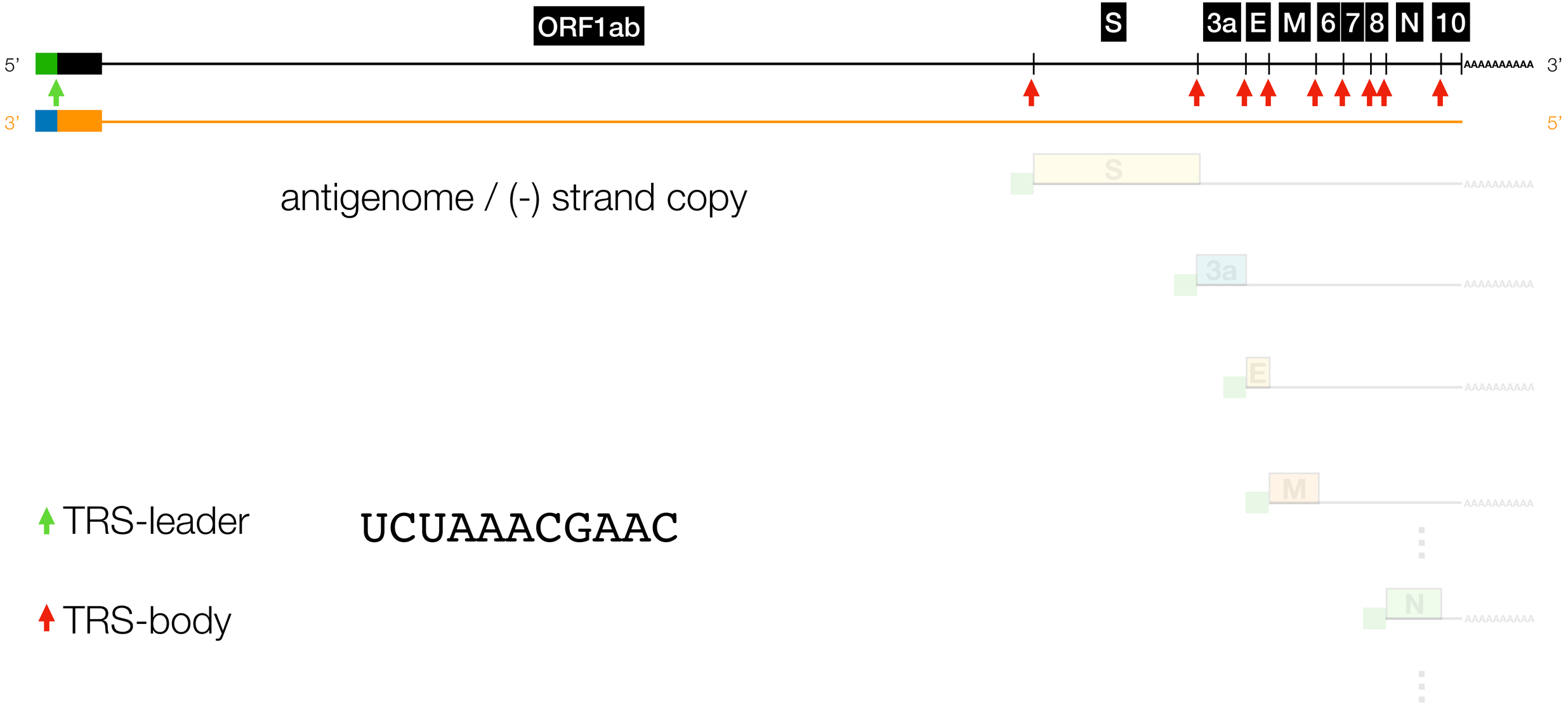
UCUAAACGAAC

TRS-body

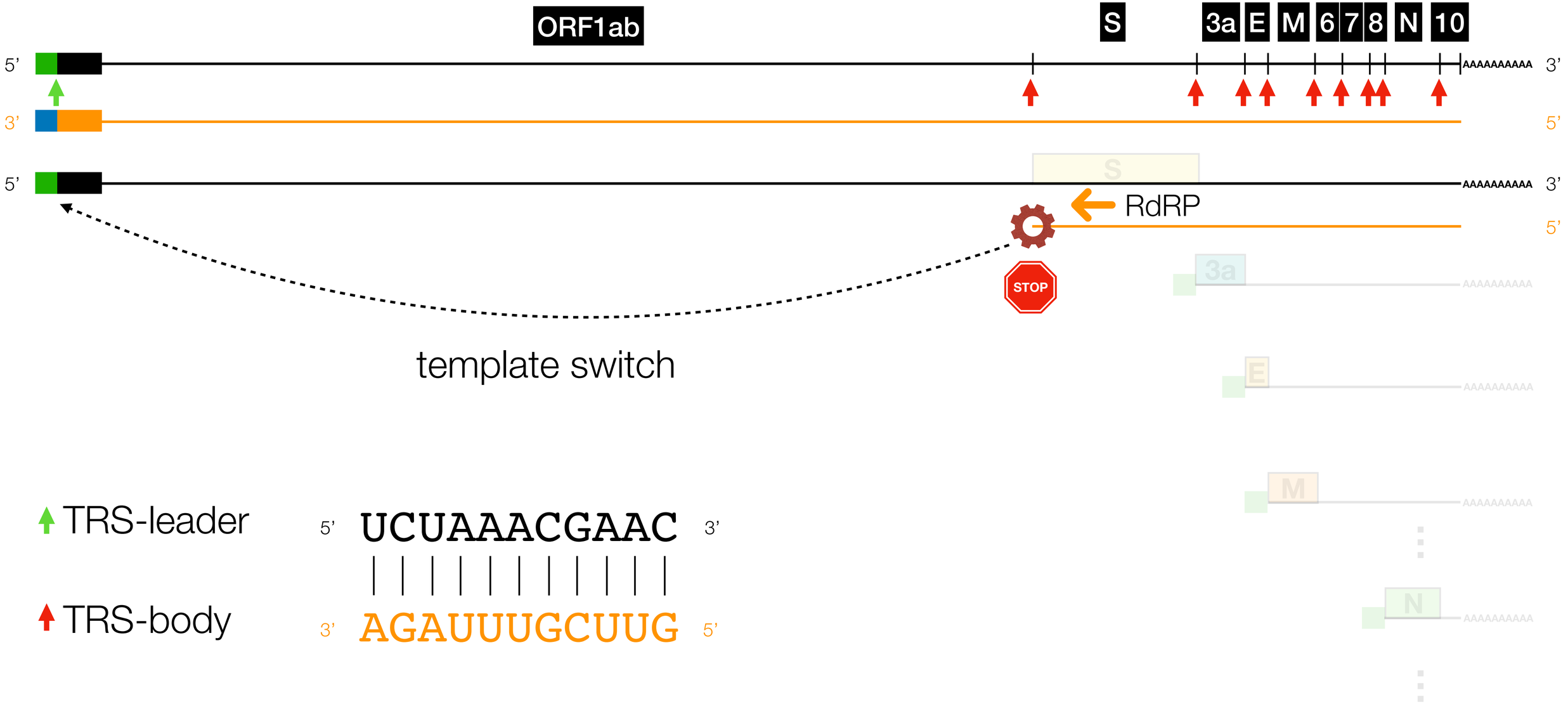
CoV Discontinuous Transcription



CoV Discontinuous Transcription



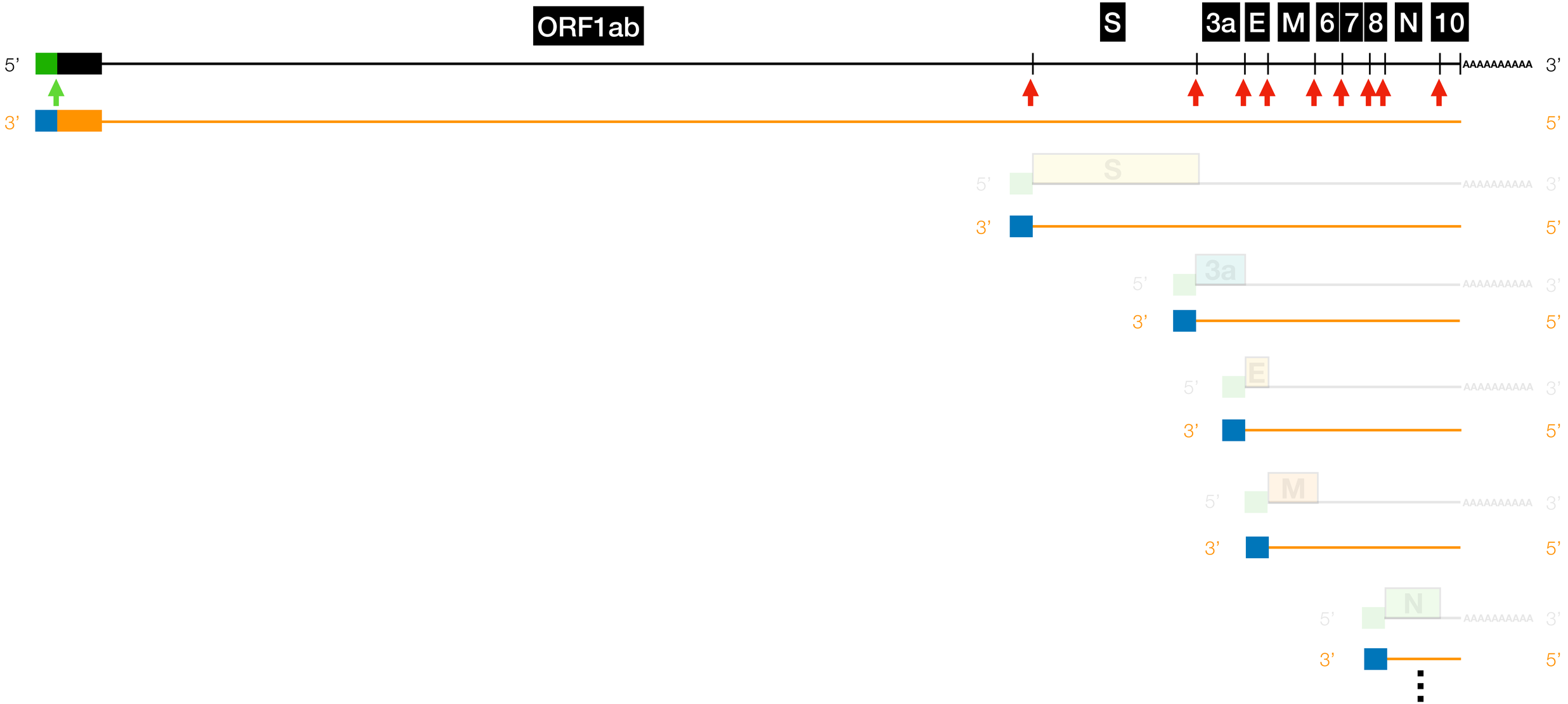
CoV Discontinuous Transcription



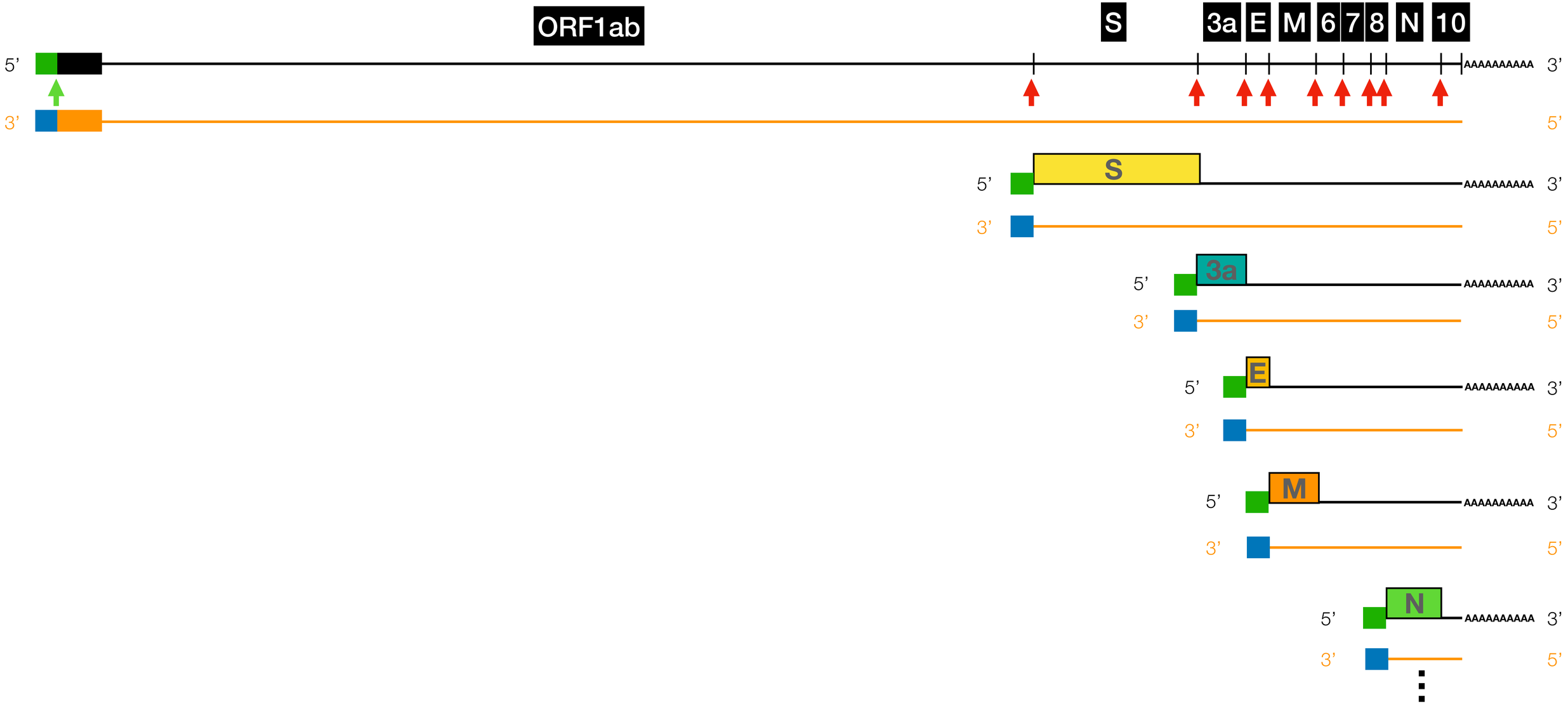
TRS-leader 5’ UCUAAACGAAC 3’

TRS-body 3’ AGAUUUGCUUG 5’

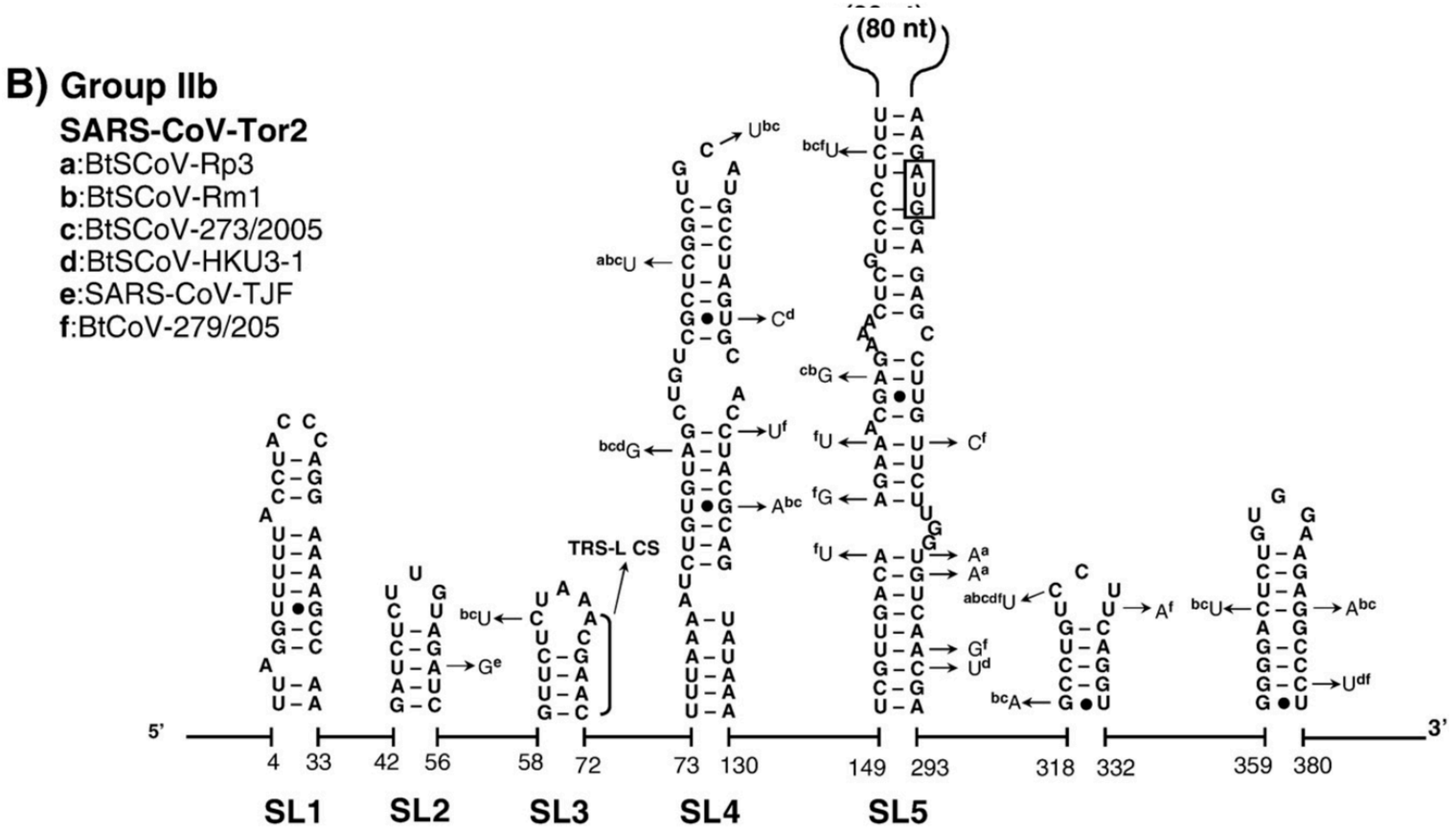
CoV Discontinuous Transcription

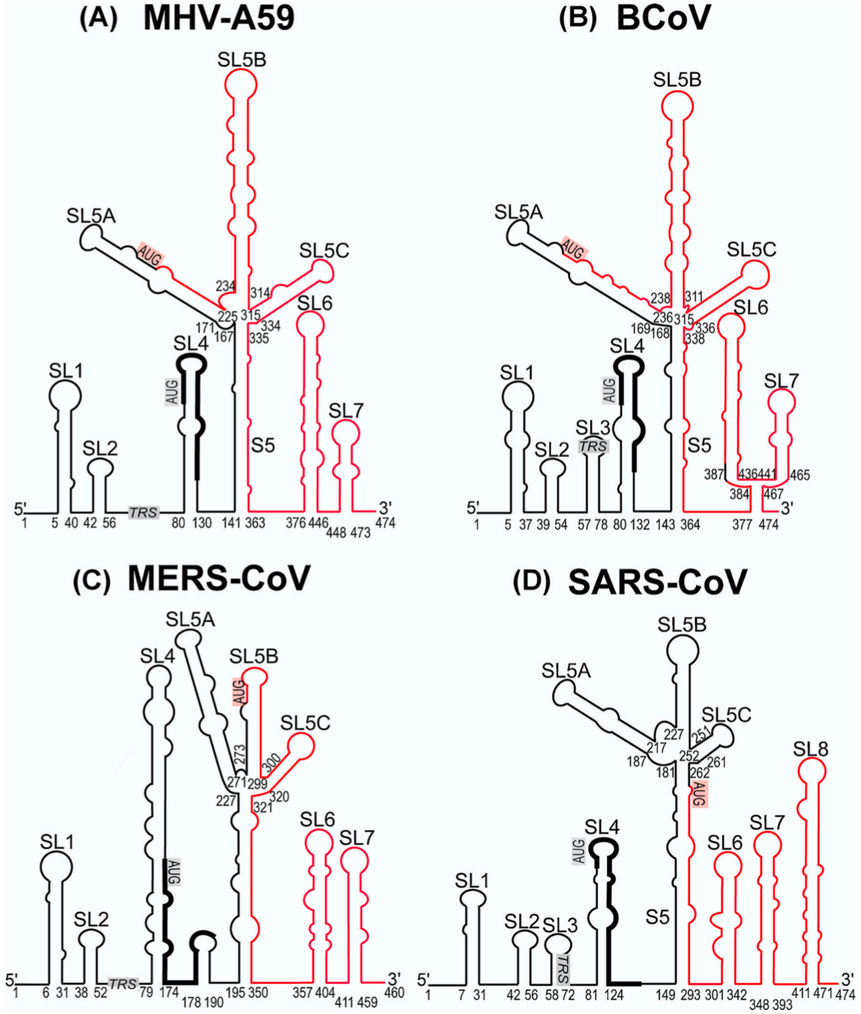


CoV Discontinuous Transcription



CoV 5’UTR & Leader Region



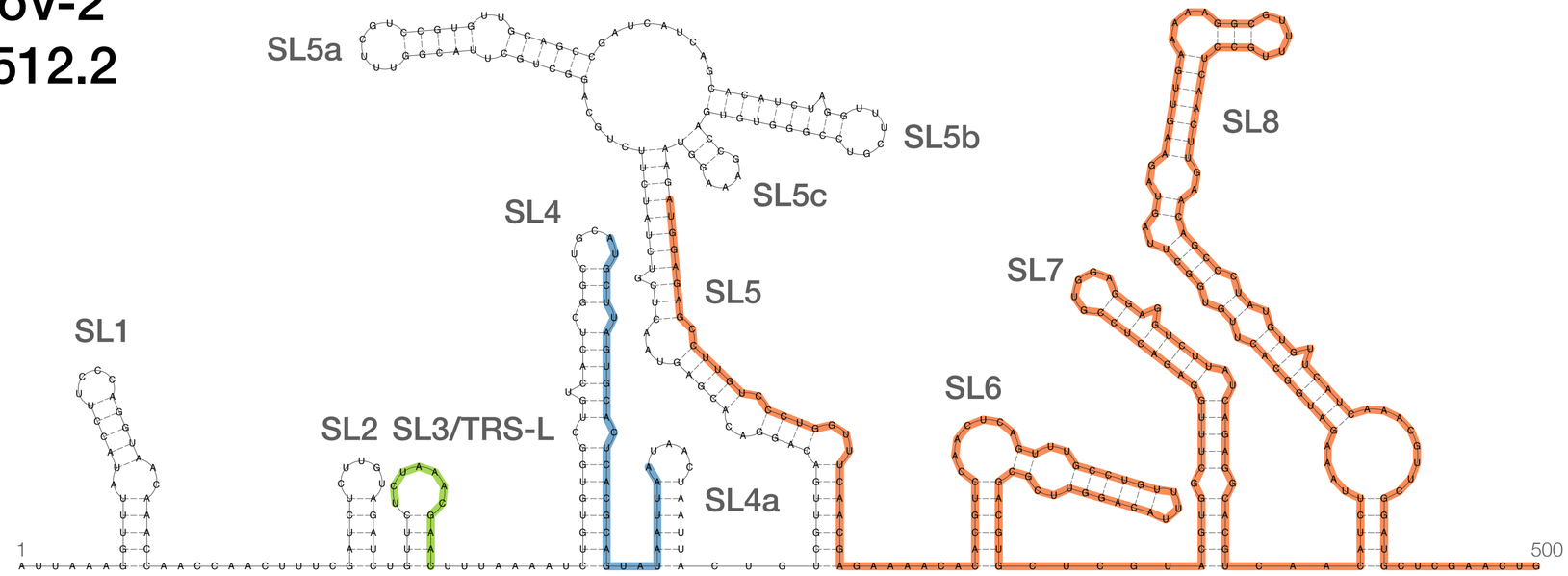


Chen, S. C., & Olsthoorn, R. C. (2010) *Virology*, *401*(1), 29-41.

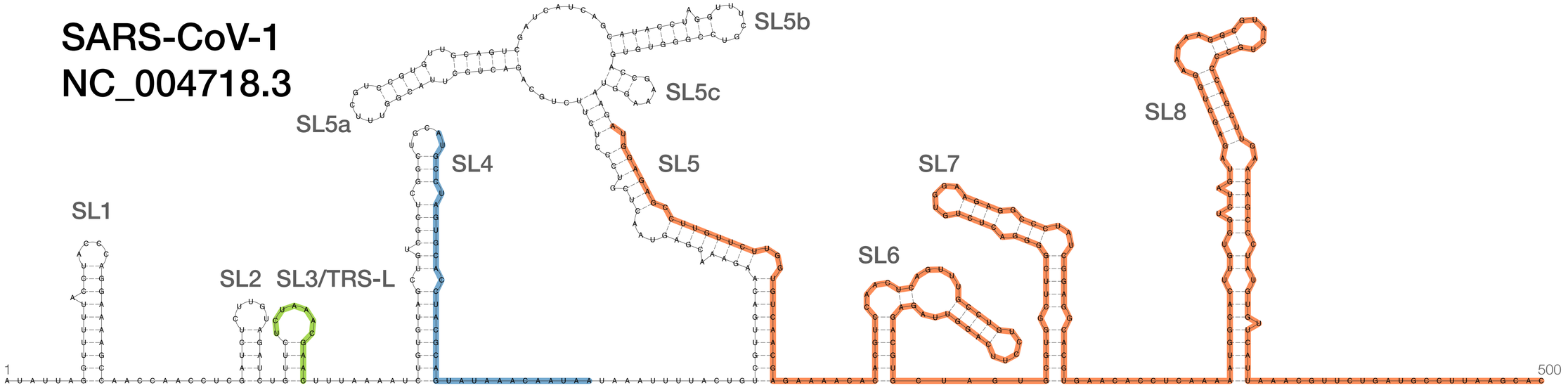
Yang, D., & Leibowitz, J. L. (2015) *Virus research*, *206*, 120-133.

SARS-CoV-1/2 Leader Region

SARS-CoV-2
NC_045512.2



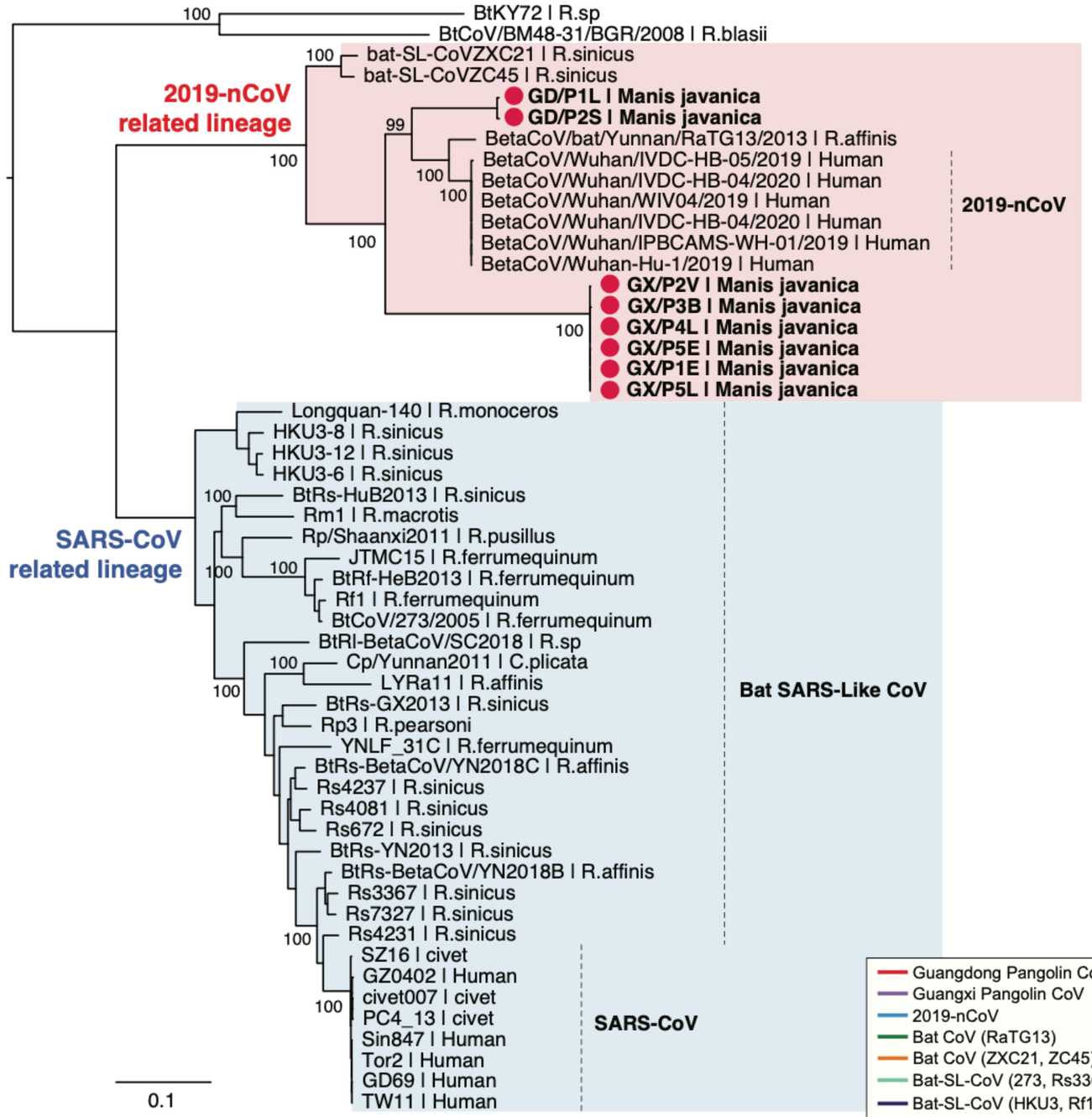
SARS-CoV-1
NC_004718.3



SARS-CoV-1/2 Leader Region

SARS-CoV-2
NC_045512.2

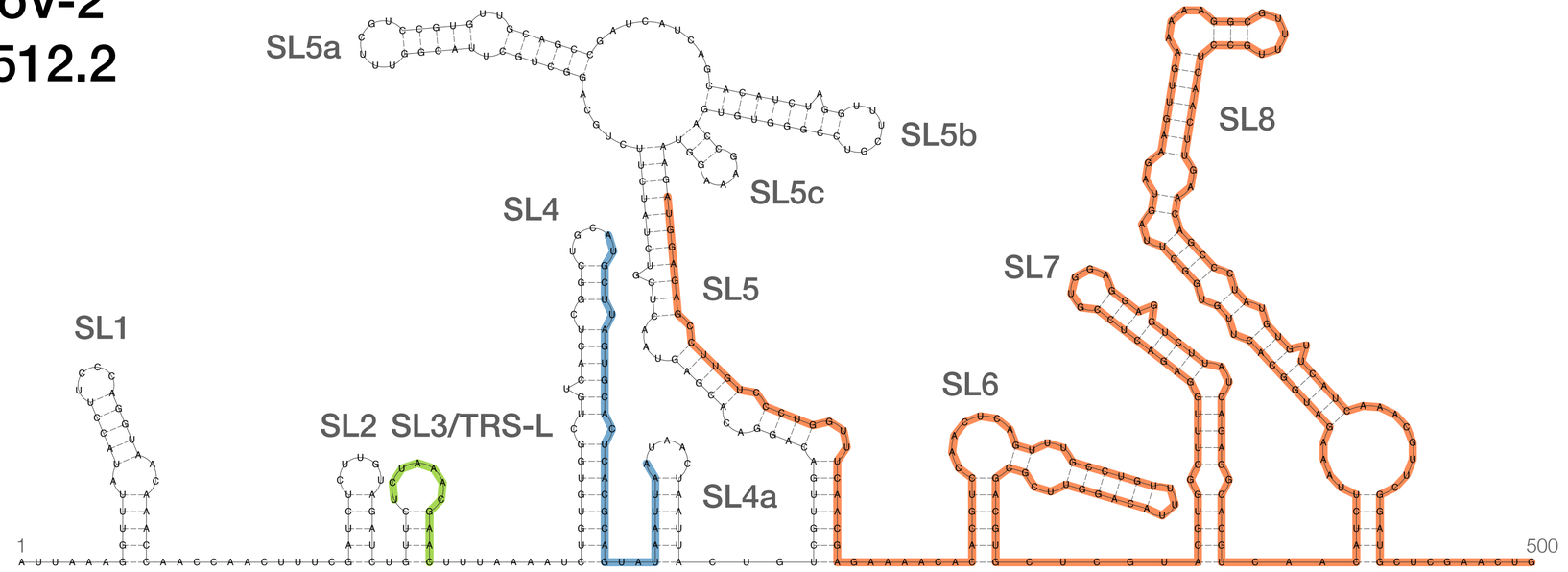
SARS-CoV-1
NC_004718.3



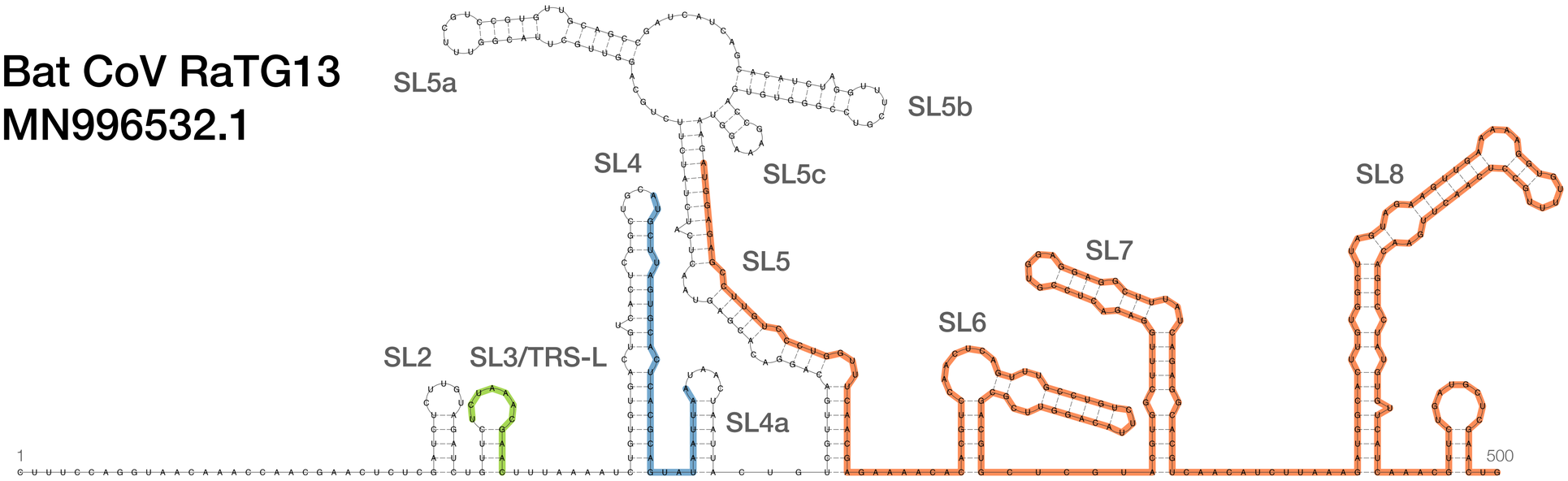
Lam et al. (2020)
bioRxiv 2020.02.13.945485.

Sarbecovirus Leader Region

SARS-CoV-2
NC_045512.2

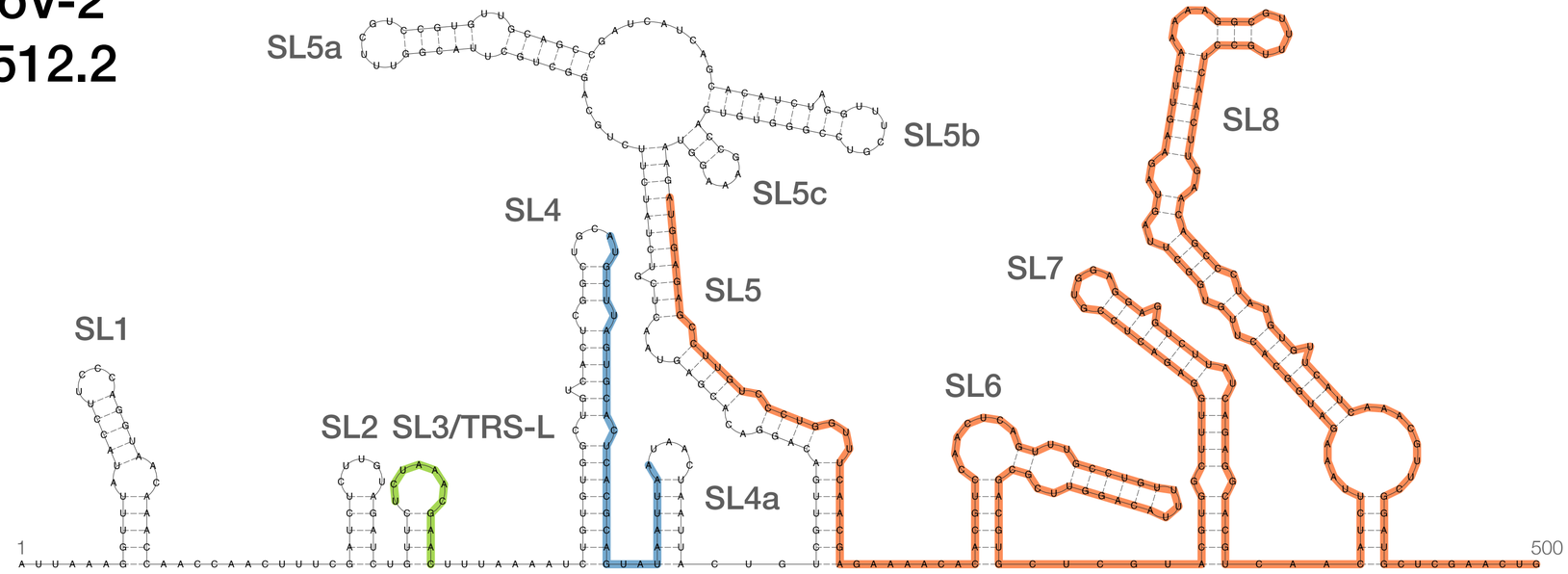


Bat CoV RaTG13
MN996532.1



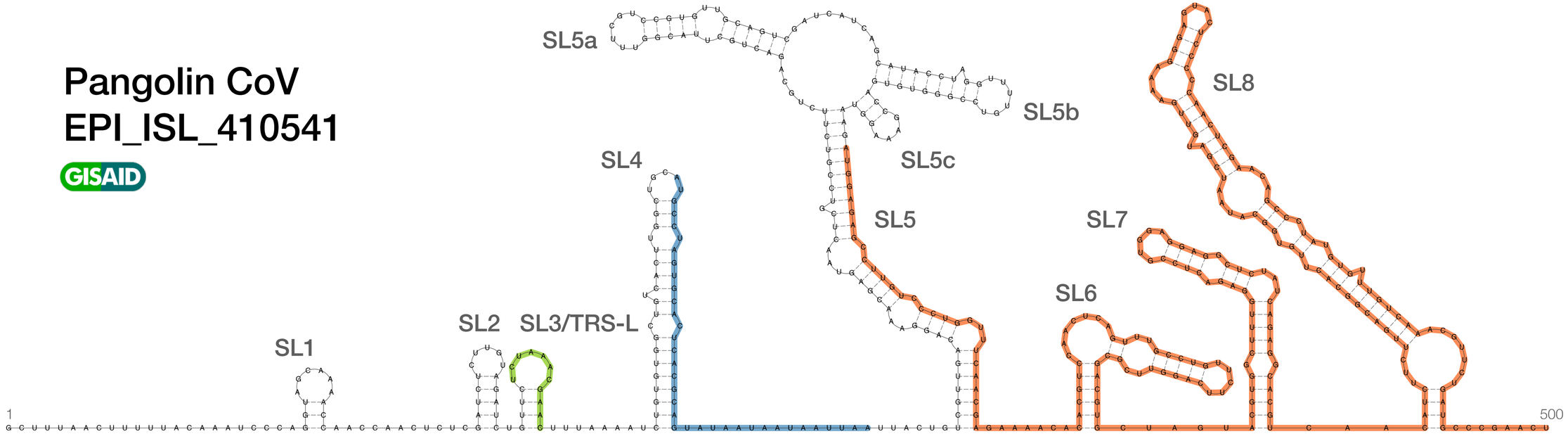
Sarbecovirus Leader Region

SARS-CoV-2
NC_045512.2



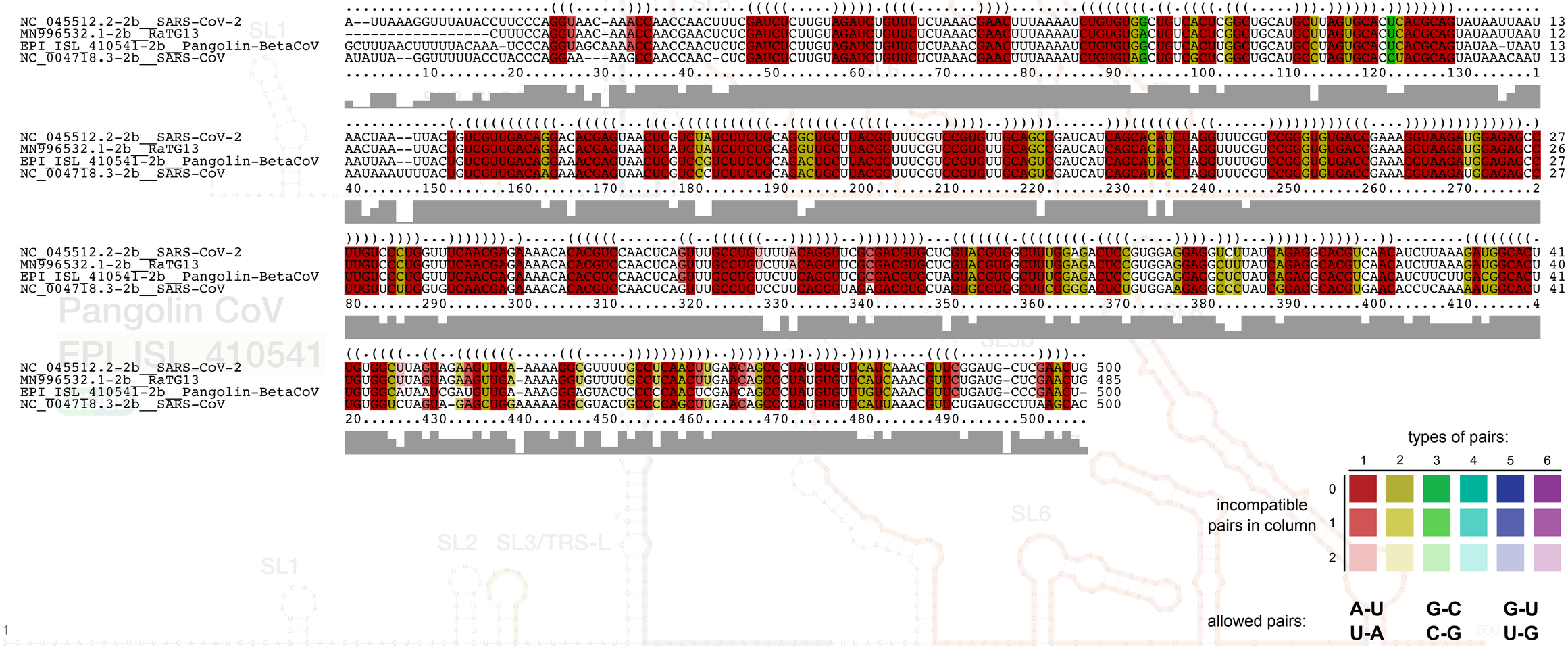
Pangolin CoV
EPI_ISL_410541

GISAID

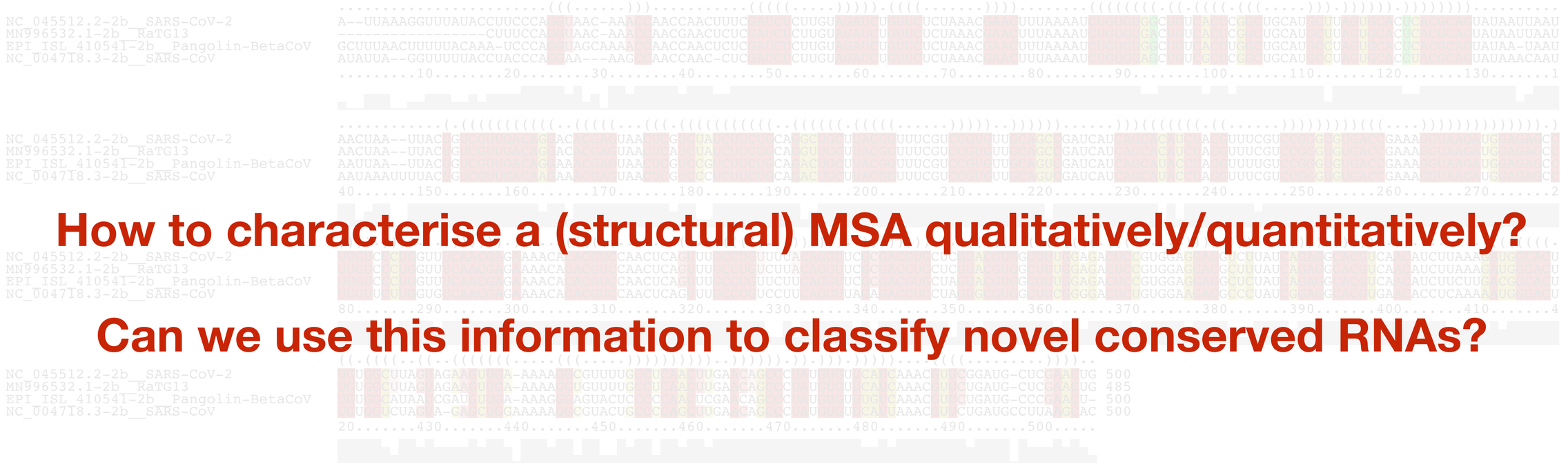


Sarbecovirus Leader Region

Structural multiple sequence alignment (MSA) of four Sarbecovirus leader regions



Sarbecovirus Leader Region



How to characterise a (structural) MSA qualitatively/quantitatively?

Can we use this information to classify novel conserved RNAs?

Using structure prediction for gene finding

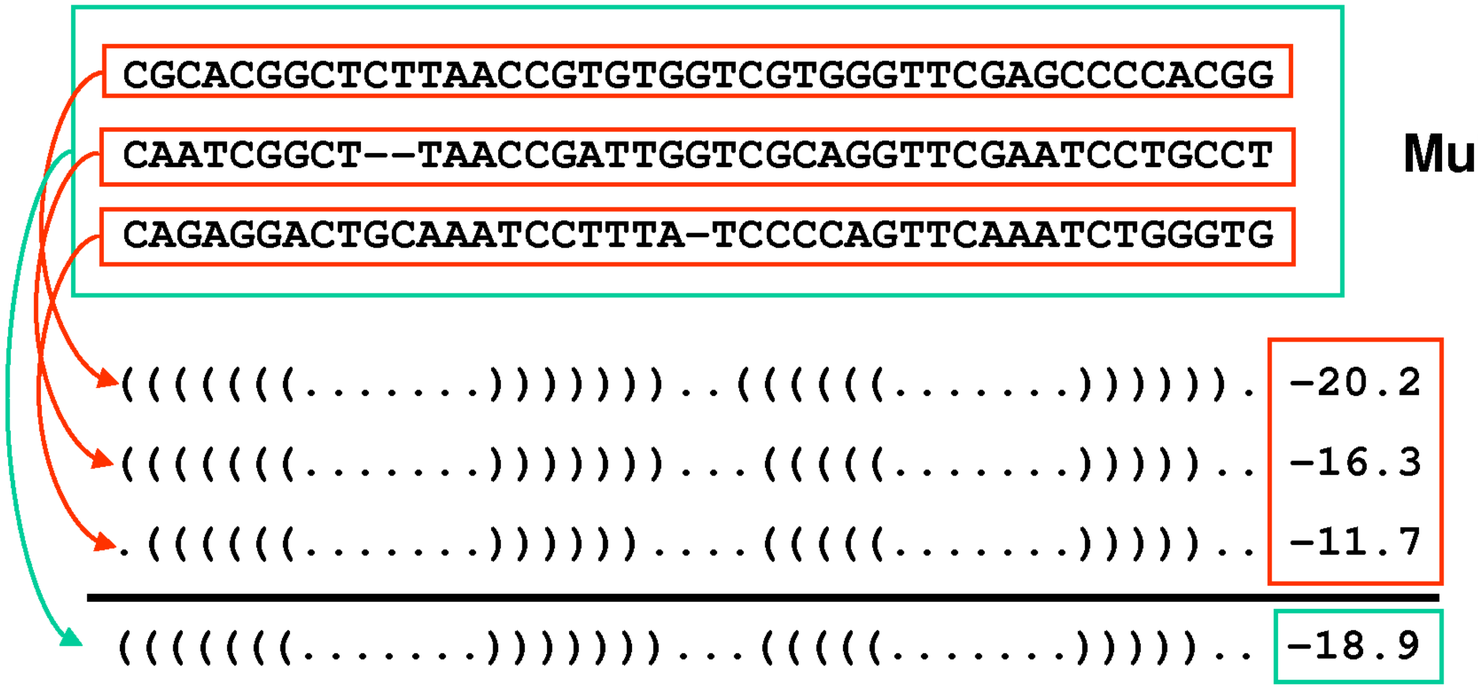
Two traits: Thermodynamic stability and structure conservation

Naturally occurring structured RNAs have a lower folding energy compared to random sequences of the same size and sequence composition.

1. Calculate native MFE m
2. Calculate mean μ and standard deviation σ of MFEs of a large number of shuffled random sequences
3. Express significance in standard deviations from the means as $z = \frac{m - \mu}{\sigma}$

Negative z-scores indicate that the native RNA is more stable than random RNAs

Structure Conservation Index (SCI)



Multiple sequence alignment

RNAfold: single sequence MFEs

RNAalifold: Consensus MFE

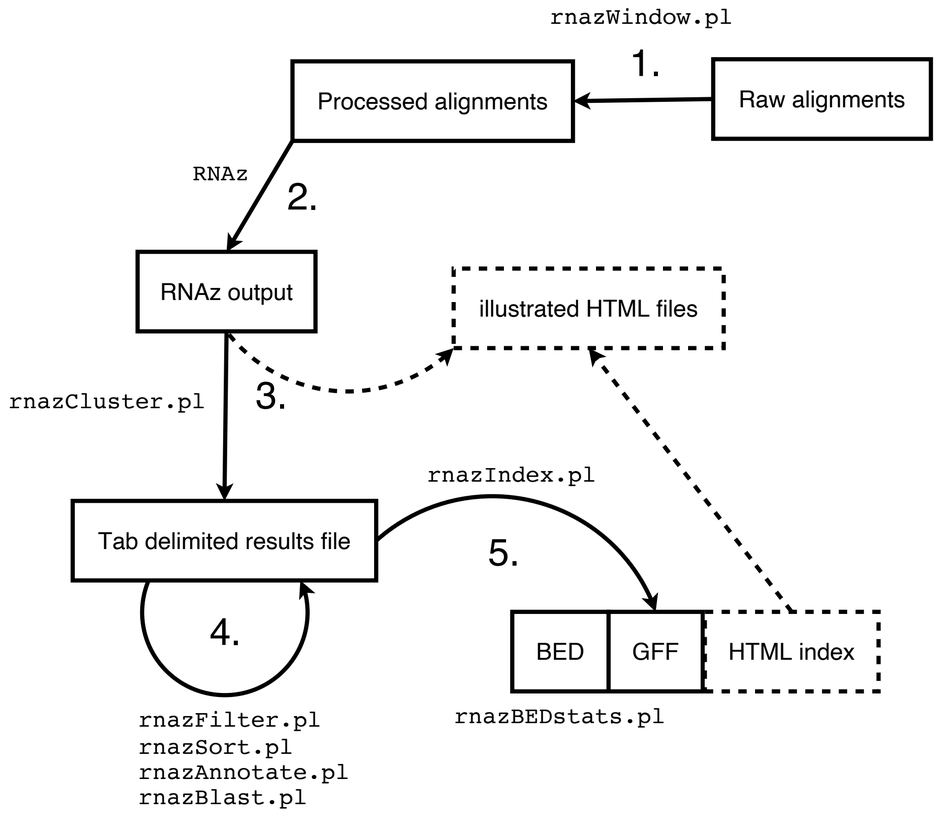
$$\text{SCI} = \frac{\text{Consensus MFE}}{\text{Mean single MFEs}}$$

The SCI is an efficient and convenient measure for secondary structure conservation

RNAz: Classification of Conserved RNAs

- RNAz detects conserved RNAs by screening MSA for secondary structures that are
  1. Thermodynamically stable
  2. Evolutionarily conserved, by SCI
- Uses a Support Vector Machine (SVM) to classify based on these (and a few other) measures

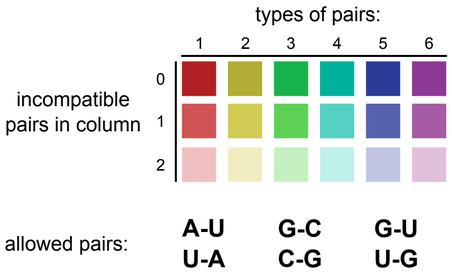
➔ “RNA class probability”

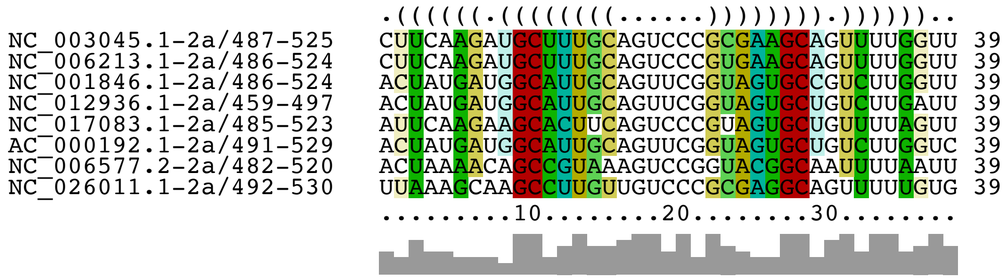


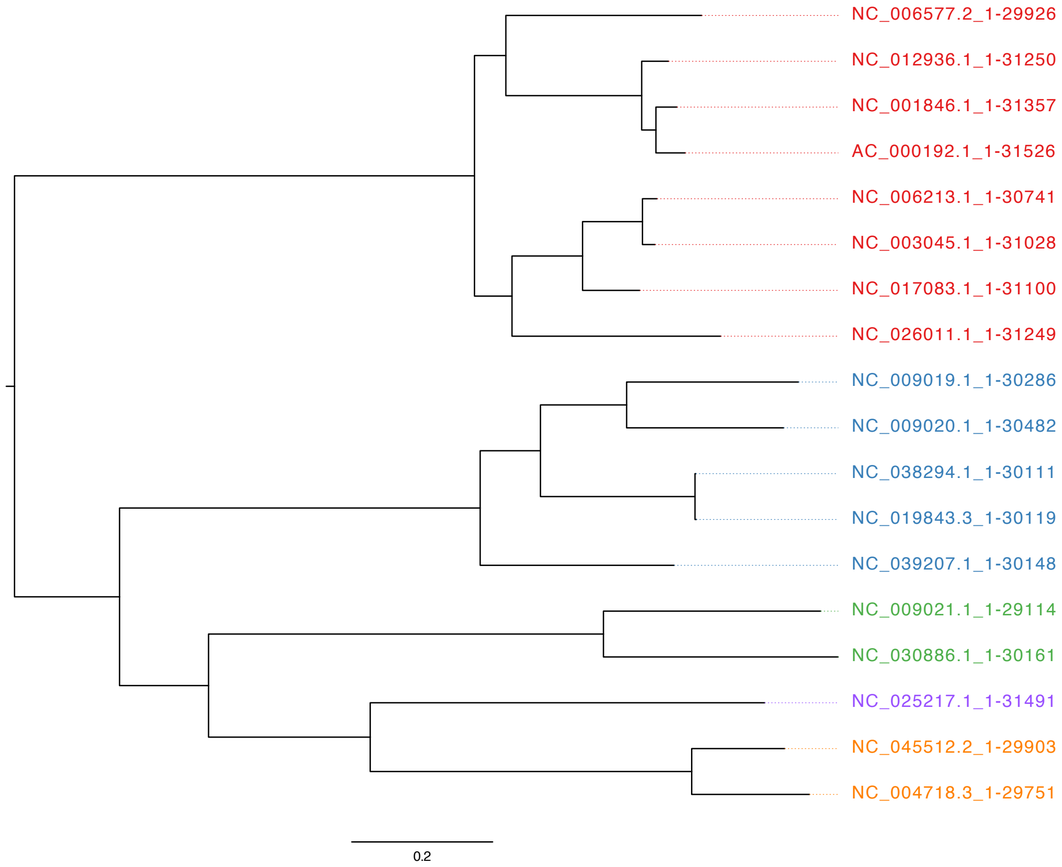
Washietl, S. Hofacker I.L., Stadler, P.F. (2005) *Proc Natlc Acad Sci USA*, 102:2454-2459.

Conserved RNAs in CoV Coding Regions

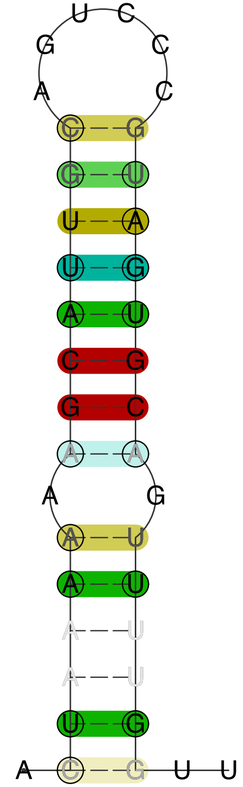
NEW







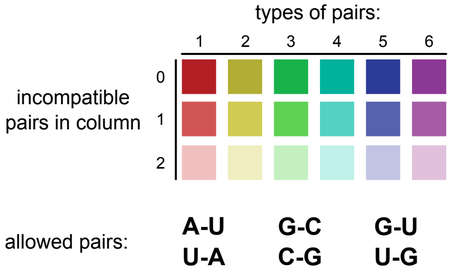
$z = -5.0$



Found in Embecovirus species:

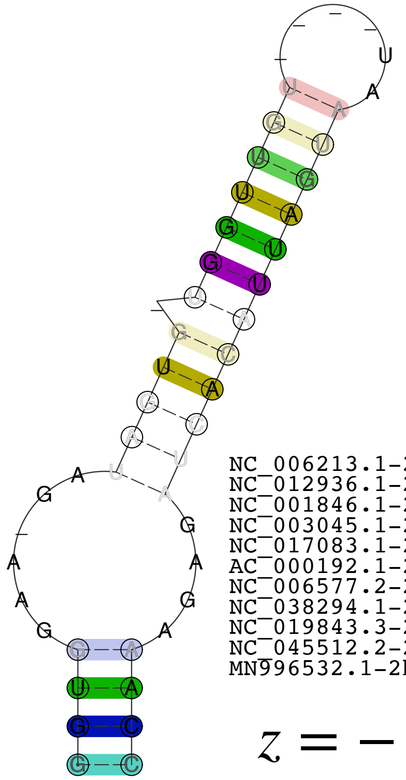
- *Bovine coronavirus*
- *Human coronavirus OC43*
- *Mouse hepatitis virus*
- *Rat coronavirus Parker*
- *Rabbit coronavirus HKU14*
- *Human coronavirus HKU1*
- *Betacoronavirus HKU24*

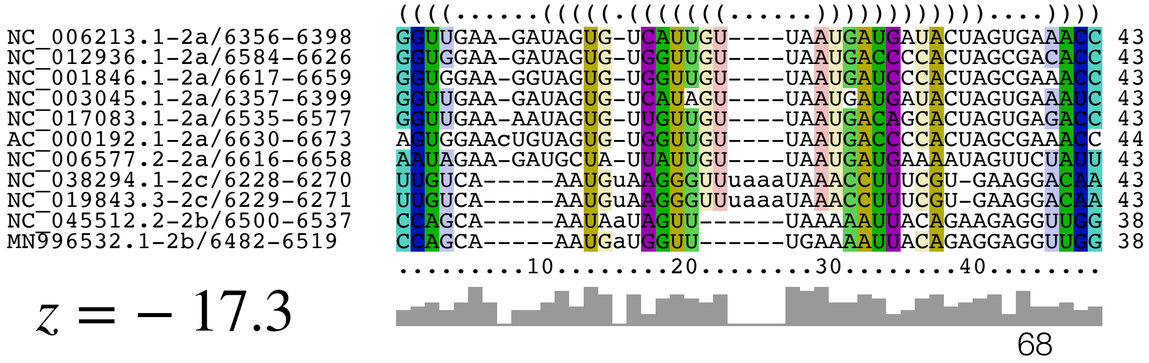
Conserved RNAs in CoV Coding Regions



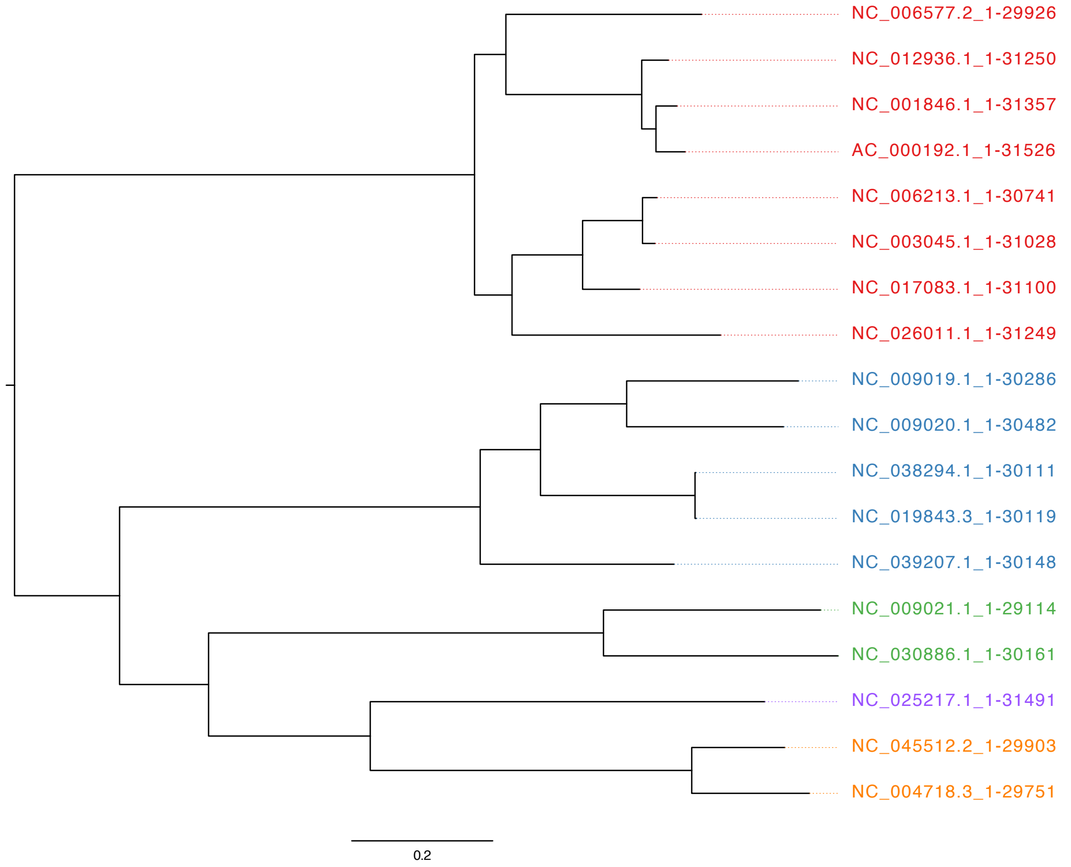
Found in Embecovirus species as well as

- *Betacoronavirus England 1* (Merbecovirus)
- *Middle East respiratory syndrome coronavirus (MERS)*
- *SARS-CoV-2*
- *Bat coronavirus RaTG13*

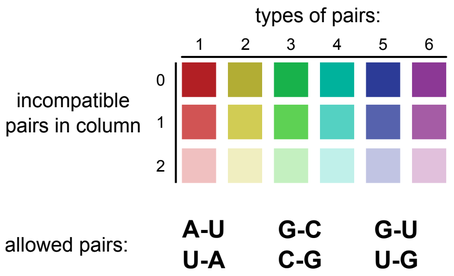


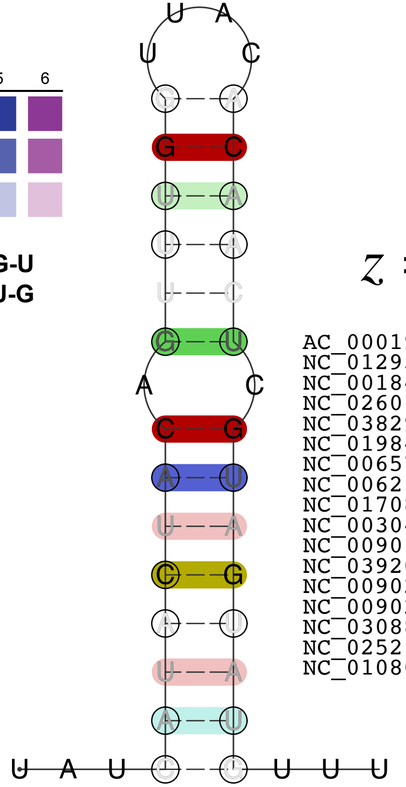


$z = -17.3$



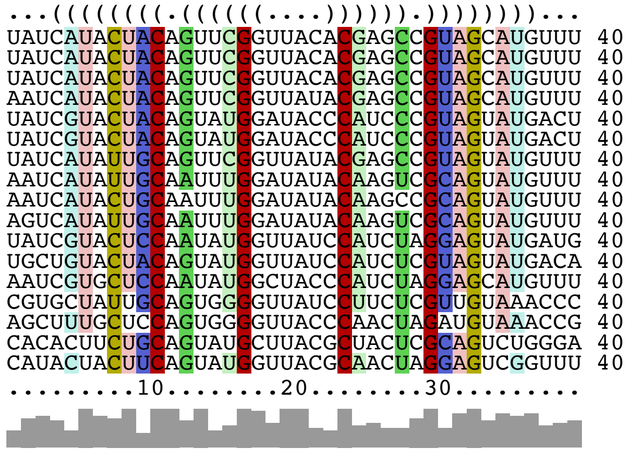
Conserved RNAs in CoV Coding Regions





$z = -9.3$

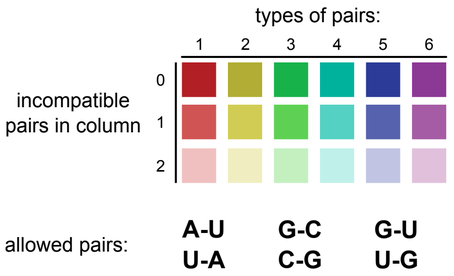
```
AC_000192.1-2a/29255-29294
NC_012936.1-2a/29015-29054
NC_001846.1-2a/29081-29120
NC_026011.1-2a/29021-29060
NC_038294.1-2c/27944-27983
NC_019843.3-2c/27945-27984
NC_006577.2-2a/27731-27770
NC_006213.1-2a/28487-28526
NC_017083.1-2a/28870-28909
NC_003045.1-2a/28801-28840
NC_009019.1-2c/28092-28131
NC_039207.1-2c/27931-27970
NC_009020.1-2c/28261-28300
NC_009021.1-2d/25787-25826
NC_030886.1-2d/26009-26048
NC_025217.1-2x/28708-28747
NC_010800.1-3/24729-24768
```

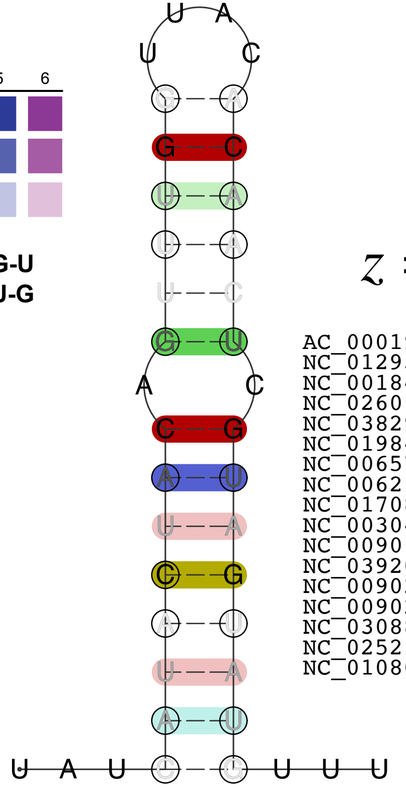


- Found in Embecovirus & Merbecovirus species plus
  - *Bat coronavirus HKU9-1* (Nobecovirus)
  - *Rousettus bat coronavirus* (Nobecovirus)
  - *Bat Hp-betacoronavirus* (Hibecovirus)
  - *Turkey coronavirus* (Gammacoronavirus)

Conserved RNAs in CoV Coding Regions

NEW

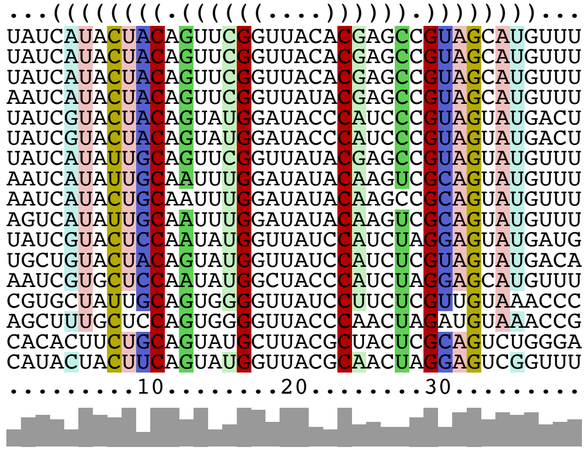


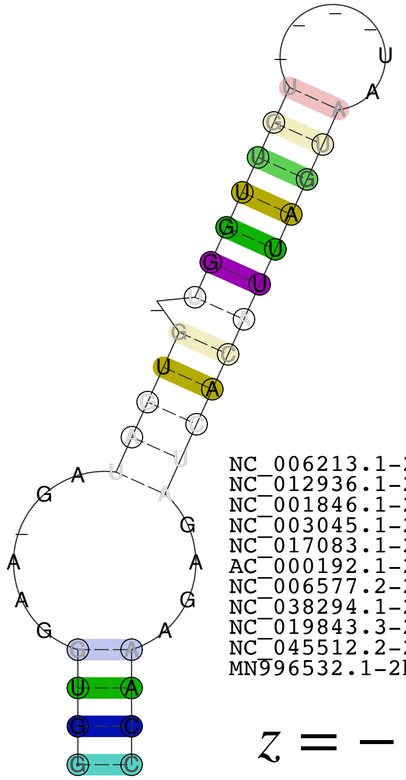


$z = -9.3$

```
AC_000192.1-2a/29255-29294   UAUCAUACUACAGUUCGGUUACACGAGCCGUAGCAUGUUU 40
NC_012936.1-2a/29015-29054   UAUCAUACUACAGUUCGGUUACACGAGCCGUAGCAUGUUU 40
NC_001846.1-2a/29081-29120   UAUCAUACUACAGUUCGGUUACACGAGCCGUAGCAUGUUU 40
NC_026011.1-2a/29021-29060   AAUCAUACUACAGUUCGGUUAUACGAGCCGUAGCAUGUUU 40
NC_038294.1-2c/27944-27983   UAUCGUACUACAGUAUGGAUACCGAUCCCGUAGUAUGACU 40
NC_019843.3-2c/27945-27984   UAUCGUACUACAGUAUGGAUACCGAUCCCGUAGUAUGACU 40
NC_006577.2-2a/27731-27770   UAUCAUAUUGCAGUUCGGUUAUACGAGCCGUAGUAUGUUU 40
NC_006213.1-2a/28487-28526   AAUCAUAUUGCAAUUUGGAUAUACAAGUCGCAGUAUGUUU 40
NC_017083.1-2a/28870-28909   AAUCAUACUGCAAUUUGGAUAUACAAGCCGCAGUAUGUUU 40
NC_003045.1-2a/28801-28840   AGUCAUAUUGCAAUUUGGAUAUACAAGUCGCAGUAUGUUU 40
NC_009019.1-2c/28092-28131   UAUCGUACUCCAAUAUGGUUAUCCAUCUAGGAGUAUGAUG 40
NC_039207.1-2c/27931-27970   UGCUGUACUACAGUAUGGUUAUCGAUCUCGUAGUAUGACA 40
NC_009020.1-2c/28261-28300   AAUCGUGCUCCAAUAUGGCUACCCAUCUAGGAGCAUGUUU 40
NC_009021.1-2d/25787-25826   CGUGCUAUUGCAGUGGGGUUAUCCUUCUCGUUGUAAACCC 40
NC_030886.1-2d/26009-26048   AGCUUUGCUCCAGUGGGGUUACCCAACUAGAUGUAAACCG 40
NC_025217.1-2x/28708-28747   CACACUUCUGCAGUAUGCUUACGCUACUCGCAGUCUGGGA 40
NC_010800.1-3/24729-24768    CAUACUACUUCAGUAUGGUUACGCAACUAGGAGUCGGUUU 40
```

```
...(((((((((.((((((....))))))).)))))))))...
.........10........20........30.........
```

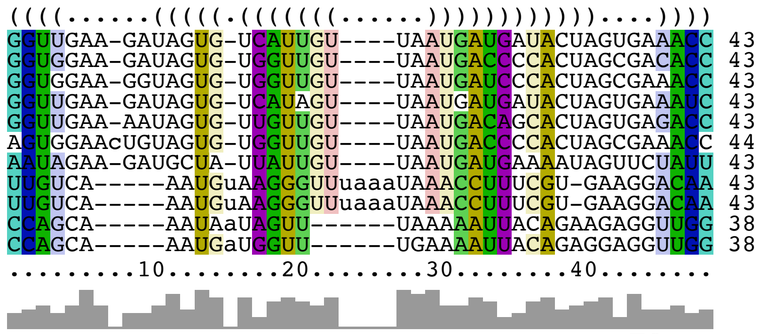




$z = -17.3$

```
NC_006213.1-2a/6356-6398   GCUUGAA-GAUAGUG-UCAUUGU----UAAUGAUGAUACUAGUGAAACC 43
NC_012936.1-2a/6584-6626   GCUGGAA-GAUAGUG-UGGUUGU----UAAUGACCCCACUAGCGACACC 43
NC_001846.1-2a/6617-6659   GCUGGAA-GGUAGUG-UGGUUGU----UAAUGAUCCCACUAGCGAAACC 43
NC_003045.1-2a/6357-6399   GCUUGAA-GAUAGUG-UCAUAGU----UAAUGAUGAUACUAGUGAAAUC 43
NC_017083.1-2a/6535-6577   GCUUGAA-AAUAGUG-UUGUUGU----UAAUGACAGCACUAGUGAGACC 43
AC_000192.1-2a/6630-6673   AGUGGAAcUGUAGUG-UGGUUGU----UAAUGACCCCACUAGCGAAACC 44
NC_006577.2-2a/6616-6658   AAUAGAA-GAUGCUA-UUAUUGU----UAAUGAUGAAAAUAGUUCUAUU 43
NC_038294.1-2c/6228-6270   UUGUCA-----AAUGuAAGGGUUuaaaUAAACCUUUCGU-GAAGGACAA 43
NC_019843.3-2c/6229-6271   UUGUCA-----AAUGuAAGGGUUuaaaUAAACCUUUCGU-GAAGGACAA 43
NC_045512.2-2b/6500-6537   CCAGCA-----AAUAaUAGUU------UAAAAAUUACAGAAGAGGUUGG 38
MN996532.1-2b/6482-6519    CCAGCA-----AAUGaUGGUU------UGAAAAUUACAGAGGAGGUUCG 38
```

```
((((......((((((.(((((((.....)))))))))))))....))))
.........10........20........30........40........
```

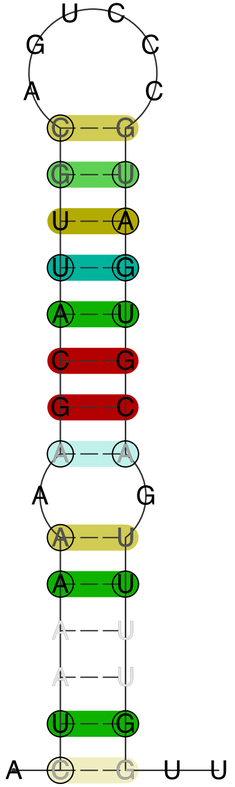


All have RNAz class probabilities > 0.999

```
NC_003045.1-2a/487-525   CUUCAAGAUGCUUUGCAGUCCCGCGAAGCAGUUUUGGUU 39
NC_006213.1-2a/486-524   CUUCAAGAUGCUUUGCAGUCCCGUGAAGCAGUUUUGGUU 39
NC_001846.1-2a/486-524   ACUAUGAUGGCAUUGCAGUUCGGUAGUGCCGUCUUGGUU 39
NC_012936.1-2a/459-497   ACUAUGAUGGCAUUGCAGUUCGGUAGUGCUGUCUUGAUU 39
NC_017083.1-2a/485-523   AUUCAAGAAGCACUUCAGUCCCGUAGUGCUGUUUUAGUU 39
AC_000192.1-2a/491-529   ACUAUGAUGGCAUUGCAGUUCGGUAGUGCUGUCUUGGUC 39
NC_006577.2-2a/482-520   ACUAAAACACGCCGUAAAGUCCGGUACGGCAAUUUUAAUU 39
NC_026011.1-2a/492-530   UUAAAGCAAGCCUUGUUGUCCCGCGAGGCAGUUUUUGUG 39
```

```
.(((((((.(((((((((.....)))))))))))))))).......
.........10........20........30.........
```

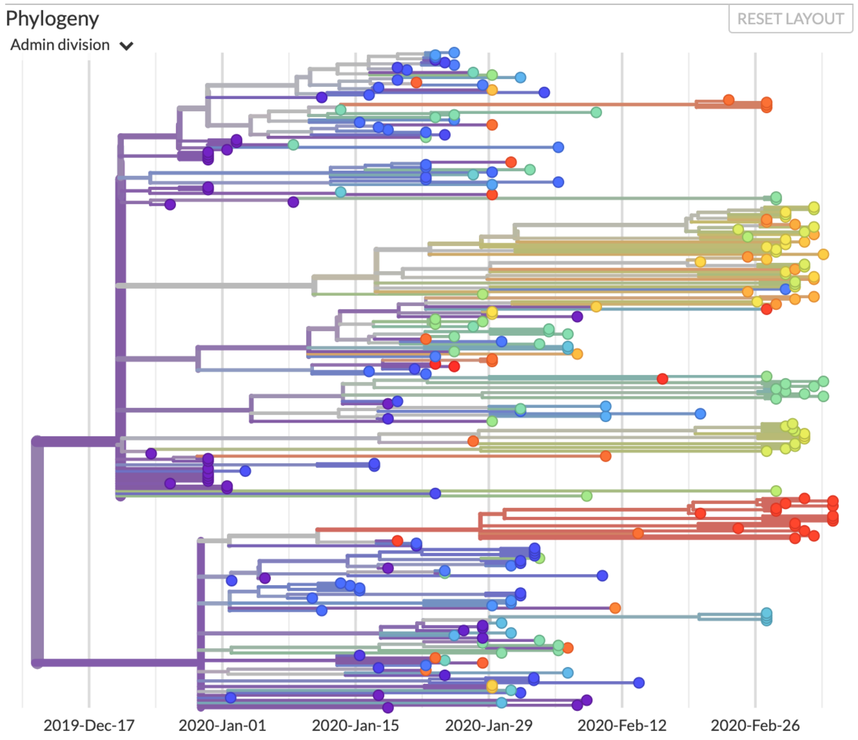

$z = -5.0$

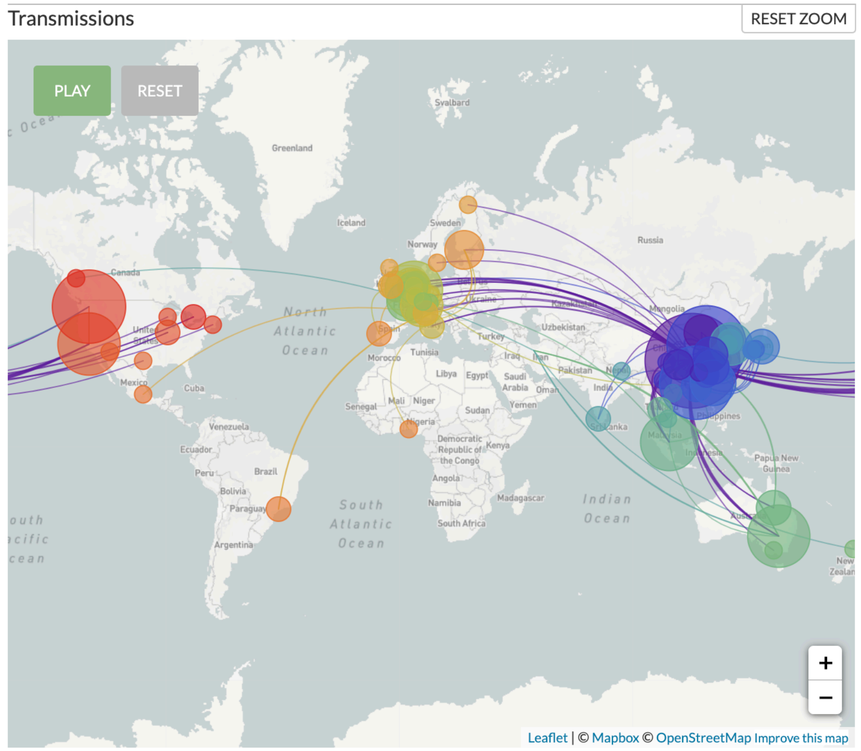


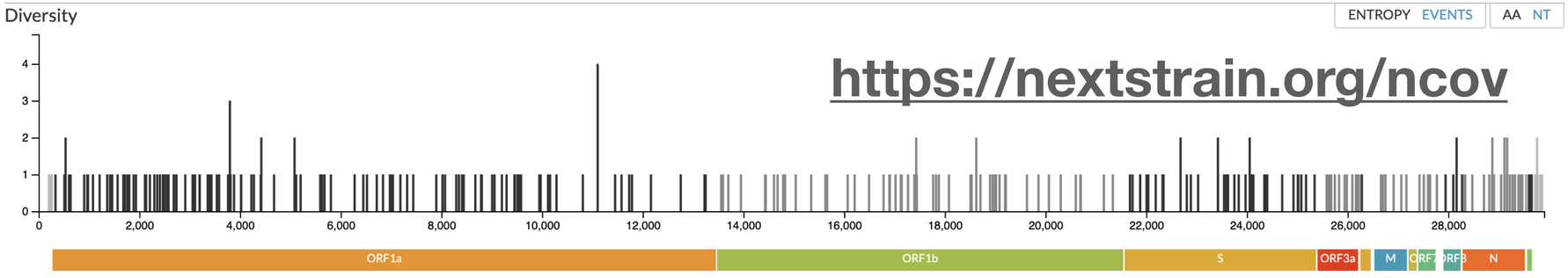
SARS-CoV-2 Molecular Epidemiology

Built with github.com/nextstrain/ncov using data from GISAID.

Showing 264 of 264 genomes sampled between Dec 2019 and Mar 2020.







https://nextstrain.org/ncov

Summary

- RNA secondary structure prediction via dynamic programming algorithms
- Functional RNA structures are evolutionarily conserved in coding and non-coding regions
- Structural alignments and covariance models are used to find conserved RNAs
- SARS-CoV-2 shares ancestral roots with bat and pangolin CoVs
- ViennaRNA Web Services are available at

http://rna.tbi.univie.ac.at/

Acknowledgements

TBI Vienna

Ivo L. Hofacker
Ronny Lorenz
Roman Ochsenreiter
Peter Wolschann

University San Diego

Adriano de Bernardi Schneider

Chulalongkorn University Bangkok

Thanyada Rungrotmongkol
Nitchakan Darai

University of Kent

Martin Michaelis
Mark Wass

Hokkaido University

Hirofumi Sawa
Yasuko Orba
Christida Wastika

@mtwolfinger

tbi

universität wien