Extracting and Evaluating Features from RNA Virus Sequence to Predict Host Species Susceptibility Using Deep Learning

International Conference on Bioinformatics and Biomedical Technology — ICBBT 2021

by

Kevin Sutanto &
Marcel Turcotte

Version May 22, 2021

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]
  - **Wide range** of susceptible **host** species [3, 4, 5, 6, 7]

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]
  - **Wide range** of susceptible **host** species [3, 4, 5, 6, 7]
- Controlling the **spread**

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]
  - **Wide range** of susceptible **host** species [3, 4, 5, 6, 7]
- Controlling the **spread**
  - **Identification and monitoring** of reservoir hosts [8]

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]
  - **Wide range** of susceptible **host** species [3, 4, 5, 6, 7]
- Controlling the **spread**
  - **Identification and monitoring** of reservoir hosts [8]
  - **Manual testing** to identify possible hosts is **demanding**

Motivation

- **COVID-19** pandemic caused by **SARS-CoV-2**
- **RNA** viruses
  - **Jumps between species** are facilitated by **high mutation rates** [1] and **re-assortment** [2]
  - **Wide range** of susceptible **host** species [3, 4, 5, 6, 7]
- Controlling the **spread**
  - **Identification and monitoring** of reservoir hosts [8]
  - **Manual testing** to identify possible hosts is **demanding**
  - **Computational techniques** could be used to **narrow down** possible hosts

Motivation (contd)

Kevin Sutanto and Marcel Turcotte.
Assessing the Use of Secondary Structure Fingerprints and Deep Learning to Classify RNA Sequences.
IEEE International Conference on Bioinformatics and Biomedicine (BIBM),
Seoul, South Korea, December 16-19, 2020.

Motivation (contd)

Kevin Sutanto and Marcel Turcotte.
Assessing the Use of Secondary Structure Fingerprints and Deep Learning to Classify RNA Sequences.
IEEE International Conference on Bioinformatics and Biomedicine (BIBM),
Seoul, South Korea, December 16-19, 2020.

Kevin Sutanto and Marcel Turcotte.
Assessing Global-Local Secondary Structure Fingerprints to Classify RNA Sequences with Deep Learning.
IEEE/ACM Transactions on Computational Biology and Bioinformatics,
Submitted on 2021-02-28.

Motivation (contd)

Kevin Sutanto and Marcel Turcotte.
Assessing the Use of Secondary Structure Fingerprints and Deep Learning to Classify RNA Sequences.
IEEE International Conference on Bioinformatics and Biomedicine (BIBM),
Seoul, South Korea, December 16-19, 2020.

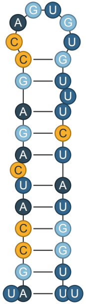
Kevin Sutanto and Marcel Turcotte.
Assessing Global-Local Secondary Structure Fingerprints to Classify RNA Sequences with Deep Learning.
IEEE/ACM Transactions on Computational Biology and Bioinformatics,
Submitted on 2021-02-28.

Kevin Sutanto.
RNA sequence classification using secondary structure fngerprints, sequence-based features, and deep learning.
Master of Computer Science, University of Ottawa, School of Electrical Engineering and Computer Science,
2021.

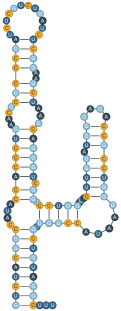
Related Work

- **Deep learning** has been used to identify:
  - **Viruses** from alignment-free metagenomic data [9]
  - **Interactions** between viral and host **proteins** [10]
  - **Hosts** for sequences of “influenza A”, “rabies lyssavirus” and “rotavirus A” [11]
- Data utilized in **prior host identification studies**:
  - **Sequences** of the viruses themselves [11, 12]
  - **Encoded viral proteins** [13]
  - **K-mers** [14, 15]

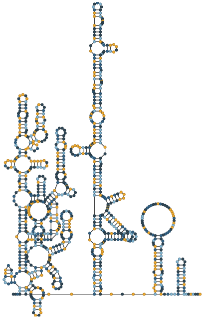
RNA Secondary Structure



piRNA
30 nt, piR-40447



5S Ribosomal RNA
121 nt, CRW V00589



16S Ribosomal RNA
954 nt, CRW J01415

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]
- **Examples:**

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]
- **Examples:**
  - *Secondary structure motif* to evade host viral recognition mechanism in *alphaviruses* [18]

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]
- **Examples:**
  - *Secondary structure motif* to evade host viral recognition mechanism in *alphaviruses* [18]
  - *Conserved structures* "hinting" conserved functions among the coronaviruses [19]

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]
- **Examples:**
  - *Secondary structure motif* to evade host viral recognition mechanism in *alphaviruses* [18]
  - *Conserved structures* “hinting” conserved functions among the coronaviruses [19]
  - *Structural conservation* in addition to nucleotide in SARS-CoV-2 vs. viruses in SARS family [20]

Observations

- **Secondary structure is conserved** despite high nucleotide mutation rate
- Secondary structure often takes part in their **biological processes** [16, 17]
- **Examples:**
  - *Secondary structure motif* to evade host viral recognition mechanism in *alphaviruses* [18]
  - *Conserved structures* "hinting" conserved functions among the coronaviruses [19]
  - *Structural conservation* in addition to nucleotide in SARS-CoV-2 vs. viruses in SARS family [20]
- **Secondary structure** has **not** been used to predict host species susceptibility

Proposed Approach

- Can **features derived from secondary structures** improve virus-host prediction

Proposed Approach

- Can **features derived from secondary structures** improve virus-host prediction
- Separately and combined with **nucleotide-based features**

Proposed Approach

- Can **features derived from secondary structures** improve virus-host prediction
- Separately and combined with **nucleotide-based features**
- **Deep learning**

Methods Overview

- **Features:**
  - **K-mers**
  - **Skip-mers** [21]
  - **Secondary structure fingerprints** [22]
- **Deep learning**
- **Dataset** and filtering

K-mers and Skip-mers

- **K-mers**

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]
  - Unlike k-mers, contain **wild-cards** at certain positions

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]
  - Unlike k-mers, contain **wild-cards** at certain positions
  - Allows to efficiently represent **longer** sequence patterns

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]
  - Unlike k-mers, contain **wild-cards** at certain positions
  - Allows to efficiently represent **longer** sequence patterns
  - **Herein**:

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]
  - Unlike k-mers, contain **wild-cards** at certain positions
  - Allows to efficiently represent **longer** sequence patterns
  - **Herein**:
    - *Match 1 skip 1* (e.g. A*G*A*C) with *length of 7, 9, and 11*,

K-mers and Skip-mers

- **K-mers**
  - $k = 4, 5, 6$
- **Skip-mers** [21]
  - Unlike k-mers, contain **wild-cards** at certain positions
  - Allows to efficiently represent **longer** sequence patterns
  - **Herein**:
    - *Match 1 skip 1* (e.g. A*G*A*C) with *length of 7, 9, and 11*,
    - *Match 2 skip 1* (e.g. AC*GT*) with *length of 6, 7, and 9*.

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**
    1. **Finding structural motif matches** from the sequence

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**
  1. **Finding structural motif matches** from the sequence
  2. Getting **free energy** values of the matches

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**
  1. **Finding structural motif matches** from the sequence
  2. Getting **free energy** values of the matches
  3. **Rescaling and concatenating** the values

Secondary Structure Fingerprints

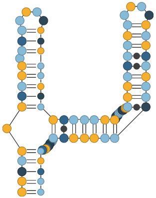
- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**
  1. **Finding structural motif matches** from the sequence
  2. Getting **free energy** values of the matches
  3. **Rescaling and concatenating** the values
- **RNAMotif** [23] was used to **find and match** secondary structures

Secondary Structure Fingerprints

- Curated **common secondary structure motifs** [22]
- **Overview of the approach:**
  1. **Finding structural motif matches** from the sequence
  2. Getting **free energy** values of the matches
  3. **Rescaling and concatenating** the values
- **RNAMotif** [23] was used to **find and match** secondary structures
- **Circumvent issues** associated with the prediction of **RNA secondary structure**

Related Work Using Secondary Structure

⇒

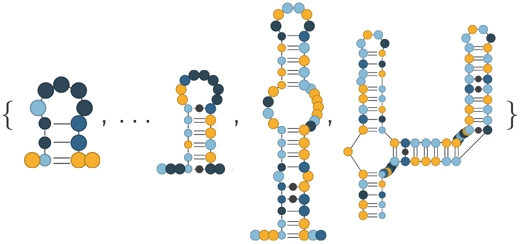


⇒ **MoSS** ⇒ **feature vector**

Fiannaca, A., Rosa, M. L., Paglia, L. L., Rizzo, R. & Urso, A. nRC: non-coding RNA Classifier based on structural features. *BioData Mining* 10, (2017)

.

RNA Secondary Structures Fingerprints



$\Rightarrow$ **RNAMotif** $\Rightarrow$ **feature vector**

Deep Learning

- **For each feature set**, 3 different network architectures:
  - 2 consecutive relu-activated dense layers + a softmax-activated dense layer (total depth = 3);
  - 3 consecutive relu-activated dense layers + a softmax-activated dense layer (total depth = 4); and
  - 4 consecutive relu-activated dense layers + a softmax-activated dense layer (total depth = 5).
- The **best performance** among the 3 = performance of the **feature set**.
- **Width of each layer = number of values in the feature set being used**
  - e.g. 256 for 4-mer

Deep Learning

- **10-fold validation** [24] was used
  - **Each fold:** 90% training, 10% evaluation data
  - Splitting into folds takes **class balance** into account
- **Adam** [25] optimizer, **sparse categorical crossentropy** loss
- **300 epochs** for training
- Starting **learning rate** = 0.001, **decay** by 50% every 100 epochs

Dataset

- **RNA virus sequences** and **their host species**
- From **NCBI Virus** [26] as of September 12, 2020
- **Filtering** – the following were **excluded**:
  - Entries with **partial** sequences only
  - Entries which sequence length **exceeds** 40,000
  - Sequences with **unknown nucleotides and/or host species**
  - Hosts with $<$ 100 entries
- **47,266 entries**

Results

- **1 feature type at a time:** *sequence-based* > *secondary structure based*

Results

- **1 feature type at a time:** *sequence-based* > *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%

Results

- **1 feature type at a time:** *sequence-based* $>$ *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**

Results

- **1 feature type at a time:** *sequence-based* $>$ *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**
  - **Combining** multiple statistics derived from **free energy values** of matches generally **improved** results

Results

- **1 feature type at a time:** *sequence-based* > *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**
  - **Combining** multiple statistics derived from **free energy values** of matches generally **improved** results
  - E.g.: **min free energy** (at 36.75%) < **min, avg, max free energy** (at 59.42%)

Results

- **1 feature type at a time:** *sequence-based* $>$ *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**
  - **Combining** multiple statistics derived from **free energy values** of matches generally **improved** results
  - E.g.: **min free energy** (at 36.75%) $<$ **min, avg, max free energy** (at 59.42%)
  - 6-mer + length 9 match-2-skip-1 skip-mer + min. free energy gives **85.9%** $\pm$ 0.28%

Results

- **1 feature type at a time:** *sequence-based* $>$ *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**
  - **Combining** multiple statistics derived from **free energy values** of matches generally **improved** results
  - E.g.: **min free energy** (at 36.75%) $<$ **min, avg, max free energy** (at 59.42%)
  - 6-mer + length 9 match-2-skip-1 skip-mer + min. free energy gives **85.9%** $\pm$ 0.28%
- Best performing **overall**:

Results

- **1 feature type at a time:** *sequence-based* > *secondary structure based*
  - *Best:* match-2-skip-1 skip-mer of length 9 at **84.92%** $\pm$ 0.25%
- **Secondary structure fingerprints:**
  - **Combining** multiple statistics derived from **free energy values** of matches generally **improved** results
  - E.g.: **min free energy** (at 36.75%) < **min, avg, max free energy** (at 59.42%)
  - 6-mer + length 9 match-2-skip-1 skip-mer + min. free energy gives **85.9%** $\pm$ 0.28%
- Best performing **overall**:
  - 6-mer + length 9 match-2-skip-1 skip-mer at **86.9%** $\pm$ 0.28%

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**
  - We found that **combining different values** to form the fingerprints generally **improved** results

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**
  - We found that **combining different values** to form the fingerprints generally **improved** results
    - e.g. min, avg, max free energy *vs.* min free energy

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**
  - We found that **combining different values** to form the fingerprints generally **improved** results
    - e.g. min, avg, max free energy *vs.* min free energy
  - *Subsequent related study* [27]: derive and use **additional separate scores** based on **locality of matches**

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**
  - We found that **combining different values** to form the fingerprints generally **improved** results
    - e.g. min, avg, max free energy *vs.* min free energy
  - *Subsequent related study* [27]: derive and use **additional separate scores** based on **locality of matches**
    - i.e. whether the secondary structure match is **global or local**; and if local, *which section*

Limitation and Future Work

- Current study **only considered top predictions** by the deep neural network
  - Non-top predictions have not been investigated or used to measure performance
  - *Possible future work:* Take the other predictions (e.g. top 3 hosts instead of just the top) into account, they **may or may not be** susceptible
- Limited performance of **secondary structure fingerprints**
  - We found that **combining different values** to form the fingerprints generally **improved** results
    - e.g. min, avg, max free energy *vs.* min free energy
  - *Subsequent related study* [27]: derive and use **additional separate scores** based on **locality of matches**
    - i.e. whether the secondary structure match is **global or local**; and if local, *which section*
    - Yielded **promising results** per our finding from this study

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**
  - Due to involvement of secondary structures in RNA viruses [18, 19, 20]

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**
  - Due to involvement of secondary structures in RNA viruses [18, 19, 20]
- **Best classification accuracy** at **86.89%** using **6-mer** + **match-2-skip-1 skip-mers of length 9**.

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**
  - Due to involvement of secondary structures in RNA viruses [18, 19, 20]
- **Best classification accuracy** at **86.89%** using **6-mer** + **match-2-skip-1 skip-mers of length 9**.
- **Sequence-based features** performed better overall in this study.

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**
  - Due to involvement of secondary structures in RNA viruses [18, 19, 20]
- **Best classification accuracy** at **86.89%** using **6-mer** + **match-2-skip-1 skip-mers of length 9**.
- **Sequence-based features** performed better overall in this study.
  - However, we found that including **more score variants** to form the fingerprints resulted in **improvements**.

Conclusions

- Proposed and tested a **deep learning pipeline to predict susceptible hosts** from viral sequence
- Unlike previous studies, **secondary structure information** is used and evaluated, in addition to **sequence-based features**
  - Due to involvement of secondary structures in RNA viruses [18, 19, 20]
- **Best classification accuracy** at **86.89%** using **6-mer** + **match-2-skip-1 skip-mers of length 9**.
- **Sequence-based features** performed better overall in this study.
  - However, we found that including **more score variants** to form the fingerprints resulted in **improvements**.
    - Further investigated in a subsequent study [27].

Thank **you**!

Availability

Dataset with the **secondary structure fingerprints** is available at:

- https://www.eecs.uottawa.ca/~turcotte/icbbt2021

Acknowledgments

- This research was enabled in part by funding from the **Natural Sciences and Engineering Research Council of Canada** (NSERC), and support provided by **Compute Ontario** (www.computeontario.ca) and **Compute Canada** (www.computecanada.ca).
- **Indigenous Affirmation** of the Universisity of Ottawa
  - We pay respect to the **Algonquin people**, who are the traditional guardians of this land. We acknowledge their longstanding relationship with this territory, which remains unceded. We pay respect to all **Indigenous people** in this region, from all nations across Canada, who call Ottawa home.
  - We acknowledge the **traditional knowledge keepers**, both young and old.
  - And we honour their **courageous leaders**: past, present, and future.

Kevin **Sutanto**
Kevin.Sutanto@uOttawa.ca

Marcel **Turcotte**
Marcel.Turcotte@uOttawa.ca

School of Electrical Engineering and **Computer Science** (EECS)
University of Ottawa

uOttawa

References I

Martin T. Ferris, Paul Joyce, and Christina L. Burch.
High Frequency of Mutations That Expand the Host Range of an RNA Virus.
Genetics, 176(2):1013–1022, June 2007.

Dhanasekaran Vijaykrishna, Reshmi Mukerji, and Gavin J. D. Smith.
RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion.
PLOS Pathogens, 11(7):e1004902, July 2015.

Ben Longdon, Michael A. Brockhurst, Colin A. Russell, John J. Welch, and Francis M. Jiggins.
The Evolution and Genetics of Virus Host Shifts.
PLoS Pathogens, 10(11):e1004395, November 2014.

S. Cleaveland, M.K. Laurenson, and L.H. Taylor.
Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence.
Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 356(1411):991–999, July 2001.

References II

T. Jonathan Davies and Amy B Pedersen.
Phylogeny and geography predict pathogen community similarity in wild primates and humans.
Proceedings of the Royal Society B: Biological Sciences, 275(1643):1695–1701, July 2008.

Louise H. Taylor, Sophia M. Latham, and Mark E.J. Woolhouse.
Risk factors for human disease emergence.
Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 356(1411):983–989, July 2001.

Mark E.J. Woolhouse, Daniel T. Haydon, and Rustom Antia.
Emerging pathogens: the epidemiology and evolution of species jumps.
Trends in Ecology & Evolution, 20(5):238–244, May 2005.

John S Mackenzie and Martyn Jeggo.
Reservoirs and vectors of emerging viruses.
Current Opinion in Virology, 3(2):170–179, April 2013.

References III

Jie Ren, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun.
Identifying viruses from metagenomic data using deep learning.
Quantitative Biology, 8(1):64–77, March 2020.

Nantao Zheng, Kairou Wang, Weihua Zhan, and Lei Deng.
Targeting Virus-host Protein Interactions: Feature Extraction and Machine Learning Approaches.
Current Drug Metabolism, 20(3):177–184, May 2019.

Florian Mock, Adrian Viehweger, Emanuel Barth, and Manja Marz.
VIDHOP, viral host prediction with Deep Learning.
Bioinformatics, 08 2020.
btaa705.

Clovis Galiez, Matthias Siebert, François Enault, Jonathan Vincent, and Johannes Söding.
WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs.
Bioinformatics, 33(19):3113–3114, 07 2017.

References IV

Christine LP Eng, Joo Chuan Tong, and Tin Wee Tan.
Predicting host tropism of influenza A virus proteins using random forest.
BMC Medical Genomics, 7(3):S1, December 2014.

Mengge Zhang, Lianping Yang, Jie Ren, Nathan A. Ahlgren, Jed A. Fuhrman, and Fengzhu Sun.
Prediction of virus-host infectious association by supervised learning methods.
BMC Bioinformatics, 18(3):60, March 2017.

Nathan A Ahlgren, Jie Ren, Yang Young Lu, Jed A Fuhrman, and Fengzhu Sun.
Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences.
Nucleic Acids Research, 45(1):39–53, 11 2016.

P. Simmonds and D. B. Smith.
Structural Constraints on RNA Virus Evolution.
Journal of Virology, 73(7):5787–5794, July 1999.

References V

Ivo L Hofacker and Peter F Stadler.
Automatic detection of conserved base pairing patterns in RNA virus genomes.
Computers & Chemistry, 23(3-4):401–414, June 1999.

J. L. Hyde, C. L. Gardner, T. Kimura, J. P. White, G. Liu, D. W. Trobaugh, C. Huang, M. Tonelli, S. Paessler, K. Takeda, W. B. Klimstra, G. K. Amarasinghe, and M. S. Diamond.
A Viral RNA Structural Element Alters Host Recognition of Nonself RNA.
Science, 343(6172):783–787, February 2014.

Ilaria Manfredonia, Chandran Nithin, Almudena Ponce-Salvatierra, Pritha Ghosh, Tomasz K Wirecki, Tycho Marinus, Natacha S Ogando, Eric J Snijder, Martijn J van Hemert, Janusz M Bujnicki, et al.
Genome-wide mapping of sars-cov-2 rna structures identifies therapeutically-relevant elements.
Nucleic acids research, 2020.

References VI

Ramya Rangan, Ivan N. Zheludev, Rachel J. Hagey, Edward A. Pham, Hannah K. Wayment-Steele, Jeffrey S. Glenn, and Rhiju Das.
RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look.
RNA, 26(8):937–959, August 2020.

Bernardo J. Clavijo, Gonzalo Garcia Accinelli, Luis Yanes, Katie Barr, and Jonathan Wright.
Skip-mers: increasing entropy and sensitivity to detect conserved genic regions with simple cyclic q-grams.
bioRxiv, 2017.

Kevin Sutanto and Marcel Turcotte.
Assessing the use of secondary structure fingerprints and deep learning to classify RNA sequences.
In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 42–49, Seoul, Korea (South), December 2020. IEEE.

References VII

T. J. Macke.
RNAMotif, an RNA secondary structure definition and search algorithm.
Nucleic Acids Research, 29(22):4724–4735, November 2001.

Tadayoshi Fushiki.
Estimation of prediction error by using K-fold cross-validation.
Statistics and Computing, 21(2):137–146, April 2011.

Diederik P. Kingma and Jimmy Ba.
Adam: A method for stochastic optimization, 2017.

J. Rodney Brister, Danso Ako-adjei, Yiming Bao, and Olga Blinkova.
NCBI Viral Genomes Resource.
Nucleic Acids Research, 43(D1):D571–D577, January 2015.

Kevin Sutanto and Marcel Turcotte.
Assessing global-local secondary structure fingerprints to classify RNA sequences with deep learning.
Submitted 2021-02-28.

Appendix: All the Results (1/4)

K-mer	"Skip-mer" [21]			Secondary Structure Fingerprints	10-Fold Cross Validation Averaged Accuracy		
	Length	Match	Skip		3-Layers Model	4-Layers Model	5-Layers Model
4-mer		-		-	62.48% $\pm$ 0.51%	**64.86%** $\pm$ 0.76%	62.09% $\pm$ 0.77%
5-mer		-		-	**77.29%** $\pm$ 0.22%	75.24% $\pm$ 0.53%	74.31% $\pm$ 0.46%
6-mer		-		-	**84.56%** $\pm$ 0.28%	83.55% $\pm$ 0.48%	83.55% $\pm$ 0.57%
-	6	2	1	-	61.74% $\pm$ 0.31%	**61.85%** $\pm$ 0.94%	59.45% $\pm$ 1.0%
-	7	1	1	-	**55.89%** $\pm$ 0.34%	54.38% $\pm$ 0.99%	48.39% $\pm$ 1.86%
-	7	2	1	-	**77.32%** $\pm$ 0.5%	75.76% $\pm$ 0.8%	71.74% $\pm$ 1.83%
-	9	1	1	-	**75.16%** $\pm$ 0.41%	73.23% $\pm$ 0.46%	65.53% $\pm$ 4.57%
-	9	2	1	-	**84.92%** $\pm$ 0.25%	84.0% $\pm$ 0.36%	82.2% $\pm$ 1.13%
-	11	1	1	-	**84.08%** $\pm$ 0.21%	81.78% $\pm$ 0.98%	81.15% $\pm$ 0.88%
-		-		min. free energy	35.91% $\pm$ 0.42%	**36.75%** $\pm$ 0.76%	35.94% $\pm$ 0.51%
-		-		min., avg. free energy	50.65% $\pm$ 0.57%	52.04% $\pm$ 0.86%	**52.6%** $\pm$ 0.65%
-		-		min., avg., max. free energy	57.37% $\pm$ 0.52%	59.39% $\pm$ 0.58%	**59.42%** $\pm$ 0.76%
4-mer	6	2	1	-	71.57% $\pm$ 0.4%	**71.69%** $\pm$ 0.41%	71.15% $\pm$ 0.49%
4-mer	7	1	1	-	70.52% $\pm$ 0.39%	**71.91%** $\pm$ 0.38%	69.63% $\pm$ 1.01%
5-mer	7	2	1	-	**82.14%** $\pm$ 0.29%	82.08% $\pm$ 0.46%	80.1% $\pm$ 0.65%
5-mer	9	1	1	-	**81.77%** $\pm$ 0.47%	81.13% $\pm$ 0.39%	80.39% $\pm$ 0.69%
6-mer	9	2	1	-	**86.89%** $\pm$ 0.28%	86.09% $\pm$ 0.21%	84.68% $\pm$ 0.83%
6-mer	11	1	1	-	**86.7%** $\pm$ 0.38%	86.17% $\pm$ 0.61%	84.73% $\pm$ 1.58%

Appendix: All the Results (2/4)

K-mer	"Skip-mer" [21]			Secondary Structure Fingerprints	10-Fold Cross Validation Averaged Accuracy		
	Length	Match	Skip		3-Layers Model	4-Layers Model	5-Layers Model
4-mer		-		min. free energy	$67.96\% \pm 0.56\%$	$70.97\% \pm 0.54\%$	**72.6%** $\pm 0.63\%$
5-mer		-		min. free energy	$78.93\% \pm 0.24\%$	$80.49\% \pm 0.62\%$	**81.05%** $\pm 0.46\%$
6-mer		-		min. free energy	**84.33%** $\pm 0.51\%$	$84.05\% \pm 0.71\%$	$77.7\% \pm 5.36\%$
4-mer		-		min., avg. free energy	$69.92\% \pm 0.56\%$	$72.28\% \pm 0.52\%$	**75.38%** $\pm 0.6\%$
5-mer		-		min., avg. free energy	$74.93\% \pm 1.66\%$	**81.28%** $\pm 0.43\%$	$81.02\% \pm 0.33\%$
6-mer		-		min., avg. free energy	$83.42\% \pm 0.39\%$	**83.73%** $\pm 0.32\%$	$82.23\% \pm 0.32\%$
4-mer		-		min., avg., max. free energy	$71.14\% \pm 0.49\%$	$74.63\% \pm 0.54\%$	**75.85%** $\pm 0.54\%$
5-mer		-		min., avg., max. free energy	$79.28\% \pm 0.75\%$	$80.74\% \pm 0.52\%$	**81.23%** $\pm 0.75\%$
6-mer		-		min., avg., max. free energy	$83.21\% \pm 0.37\%$	**83.53%** $\pm 0.13\%$	$81.87\% \pm 0.44\%$
-	6	2	1	min. free energy	$66.94\% \pm 0.58\%$	$69.98\% \pm 0.65\%$	**71.02%** $\pm 0.83\%$
-	7	1	1	min. free energy	$66.83\% \pm 0.22\%$	$69.69\% \pm 0.48\%$	**71.23%** $\pm 0.34\%$
-	7	2	1	min. free energy	$78.66\% \pm 0.55\%$	$80.35\% \pm 0.41\%$	**80.72%** $\pm 0.59\%$
-	9	1	1	min. free energy	$77.78\% \pm 0.27\%$	**80.05%** $\pm 0.29\%$	$79.43\% \pm 1.73\%$
-	9	2	1	min. free energy	**84.58%** $\pm 0.34\%$	$79.17\% \pm 3.18\%$	$80.65\% \pm 1.28\%$
-	11	1	1	min. free energy	$83.61\% \pm 0.52\%$	**83.88%** $\pm 0.32\%$	$77.77\% \pm 5.17\%$

Appendix: All the Results (3/4)

K-mer	"Skip-mer" [21]			Secondary Structure Fingerprints	10-Fold Cross Validation Averaged Accuracy		
	Length	Match	Skip		3-Layers Model	4-Layers Model	5-Layers Model
-	6	2	1	min., avg. free energy	69.62% $\pm$ 0.49%	71.83% $\pm$ 0.74%	**74.16%** $\pm$ 0.6%
-	7	1	1	min., avg. free energy	68.11% $\pm$ 0.99%	71.45% $\pm$ 0.73%	**73.93%** $\pm$ 0.59%
-	7	2	1	min., avg. free energy	78.64% $\pm$ 0.35%	79.75% $\pm$ 0.85%	**80.9%** $\pm$ 0.32%
-	9	1	1	min., avg. free energy	78.48% $\pm$ 0.59%	79.58% $\pm$ 0.55%	**81.29%** $\pm$ 0.32%
-	9	2	1	min., avg. free energy	83.2% $\pm$ 0.54%	**83.37%** $\pm$ 0.43%	82.45% $\pm$ 0.57%
-	11	1	1	min., avg. free energy	**82.83%** $\pm$ 0.41%	82.75% $\pm$ 0.44%	82.3% $\pm$ 0.45%
-	6	2	1	min., avg., max. free energy	70.77% $\pm$ 0.42%	74.04% $\pm$ 0.75%	**75.38%** $\pm$ 0.36%
-	7	1	1	min., avg., max. free energy	69.28% $\pm$ 0.75%	74.32% $\pm$ 0.52%	**74.74%** $\pm$ 0.75%
-	7	2	1	min., avg., max. free energy	79.02% $\pm$ 0.58%	80.43% $\pm$ 0.52%	**81.16%** $\pm$ 0.52%
-	9	1	1	min., avg., max. free energy	78.73% $\pm$ 0.45%	80.42% $\pm$ 0.7%	**81.3%** $\pm$ 0.33%
-	9	2	1	min., avg., max. free energy	**83.93%** $\pm$ 0.2%	83.38% $\pm$ 0.42%	82.34% $\pm$ 0.63%
-	11	1	1	min., avg., max. free energy	**83.04%** $\pm$ 0.4%	83.0% $\pm$ 0.18%	82.47% $\pm$ 0.35%

Appendix: All the Results (4/4)

K-mer	"Skip-mer" [21]			Secondary Structure Fingerprints	10-Fold Cross Validation Averaged Accuracy		
	Length	Match	Skip		3-Layers Model	4-Layers Model	5-Layers Model
4-mer	6	2	1	min. free energy	74.26% $\pm$ 0.46%	76.11% $\pm$ 0.66%	**78.78%** $\pm$ 0.3%
4-mer	7	1	1	min. free energy	74.22% $\pm$ 0.29%	77.44% $\pm$ 0.83%	**78.54%** $\pm$ 0.47%
5-mer	7	2	1	min. free energy	**83.74%** $\pm$ 0.39%	83.54% $\pm$ 0.39%	83.65% $\pm$ 0.18%
5-mer	9	1	1	min. free energy	82.17% $\pm$ 0.47%	**83.21%** $\pm$ 0.42%	83.07% $\pm$ 0.45%
6-mer	9	2	1	min. free energy	**85.9%** $\pm$ 0.28%	84.37% $\pm$ 0.71%	83.1% $\pm$ 0.73%
6-mer	11	1	1	min. free energy	**85.86%** $\pm$ 0.35%	84.94% $\pm$ 0.34%	82.39% $\pm$ 0.57%
4-mer	6	2	1	min., avg. free energy	75.06% $\pm$ 0.53%	77.33% $\pm$ 0.66%	**78.89%** $\pm$ 0.48%
4-mer	7	1	1	min., avg. free energy	74.78% $\pm$ 0.46%	77.11% $\pm$ 0.39%	**78.47%** $\pm$ 0.48%
5-mer	7	2	1	min., avg. free energy	82.52% $\pm$ 0.38%	**82.77%** $\pm$ 0.41%	82.68% $\pm$ 0.27%
5-mer	9	1	1	min., avg. free energy	81.26% $\pm$ 0.38%	**82.59%** $\pm$ 0.6%	82.37% $\pm$ 0.41%
6-mer	9	2	1	min., avg. free energy	**84.39%** $\pm$ 0.52%	84.2% $\pm$ 0.3%	82.54% $\pm$ 1.01%
6-mer	11	1	1	min., avg. free energy	**84.33%** $\pm$ 0.53%	84.19% $\pm$ 0.65%	82.06% $\pm$ 0.7%
4-mer	6	2	1	min., avg., max. free energy	75.73% $\pm$ 0.57%	79.56% $\pm$ 0.2%	**79.65%** $\pm$ 0.67%
4-mer	7	1	1	min., avg., max. free energy	75.77% $\pm$ 0.61%	77.99% $\pm$ 0.38%	**79.23%** $\pm$ 0.44%
5-mer	7	2	1	min., avg., max. free energy	82.54% $\pm$ 0.39%	**82.92%** $\pm$ 0.28%	82.16% $\pm$ 0.61%
5-mer	9	1	1	min., avg., max. free energy	81.41% $\pm$ 0.34%	**83.13%** $\pm$ 0.17%	81.55% $\pm$ 1.31%
6-mer	9	2	1	min., avg., max. free energy	**84.73%** $\pm$ 0.32%	83.71% $\pm$ 0.16%	82.33% $\pm$ 0.64%
6-mer	11	1	1	min., avg., max. free energy	**84.61%** $\pm$ 0.26%	83.33% $\pm$ 0.68%	80.4% $\pm$ 2.18%

Appendix: Included Hosts

- **47 different host species:**
  - *Allium sativum*, *Anas carolinensis*, *Anas clypeata*, *Anas platyrhynchos*, *Anatidae*, *Apodemus agrarius*, *Aves*, *Bos taurus*, *Canis lupus familiaris*, *Capra hircus*, *Capsicum annuum*, *Columbidae*, *Corvus brachyrhynchos*, *Cricetulus griseus*, *Culex*, *Culex pipiens*, *Culex quinquefasciatus*, *Culicidae*, *Culiseta melanura*, *Cyanocitta cristata*, *Equus caballus*, *Felis catus*, *Gallus gallus*, *Glycine max*, *Homo sapiens*, *Macaca mulatta*, *Malus domestica*, *Meleagris gallopavo*, *Melogale*, *Mus musculus*, *Oryza sativa*, *Ovis aries*, *Procyon lotor*, *Prunus*, *Prunus avium*, *Prunus persica*, *Pyrus communis*, *Rattus norvegicus*, *Rosa sp.*, *Solanum lycopersicum*, *Solanum tuberosum*, *Sus scrofa*, *Sus scrofa domesticus*, *Triticum aestivum*, *Vitis vinifera*, *Vulpes vulpes*, and *Zea mays*