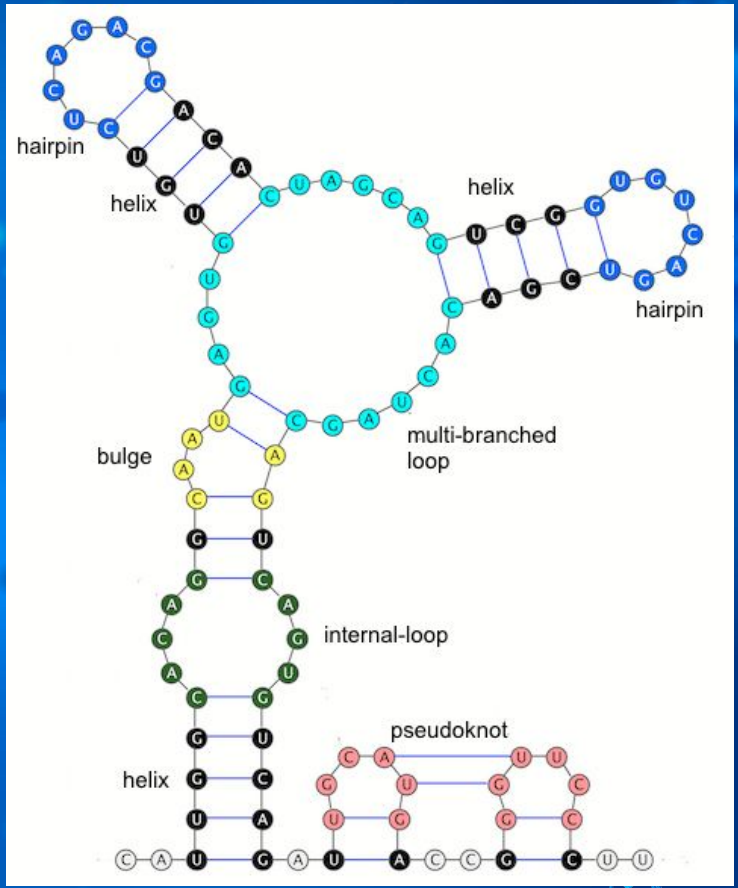
Using RNA Secondary Structure to Analyze Variations in the S (spike) Protein of SARS-CoV-2 and related Coronaviruses

By Lawrence Chillrud

1. Introduction / Background

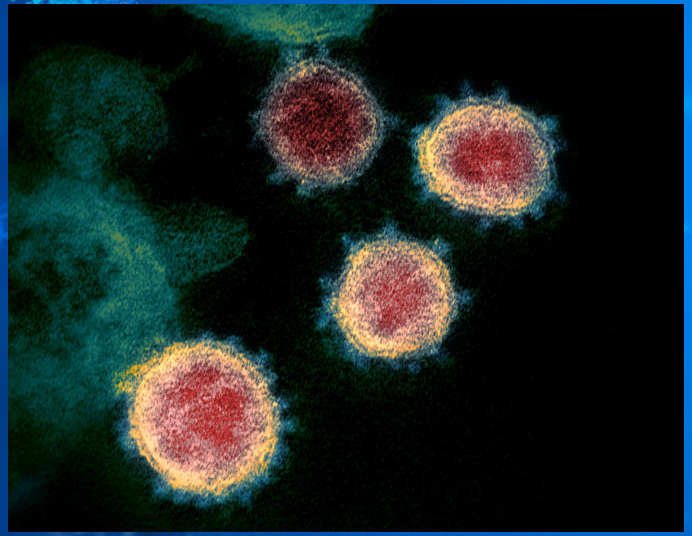
RNA Secondary Structure, Coronaviruses, S (spike) Protein

RNA Secondary Structure

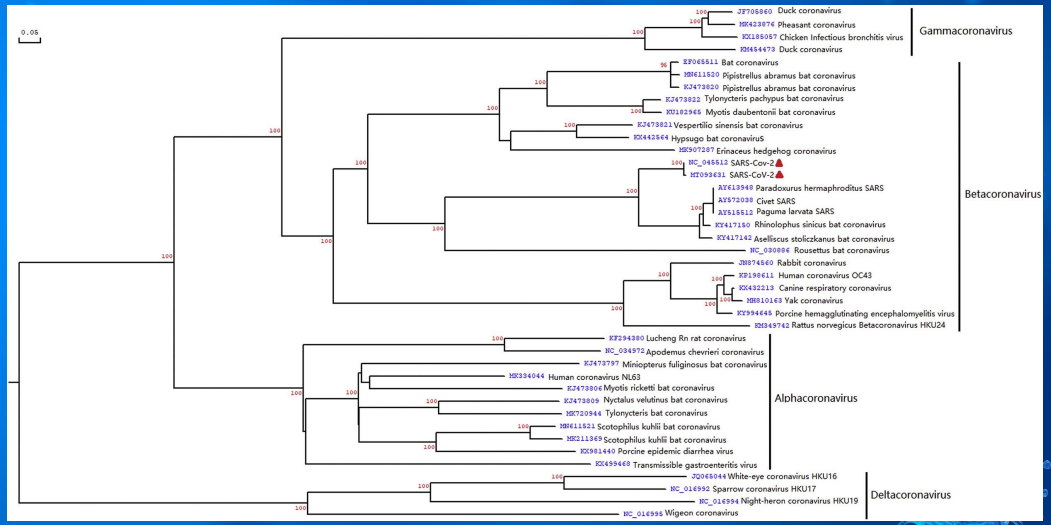


Reference: Mamuye, Adane & Merelli, Emanuela & Tesei, Luca. (2016). A Graph Grammar for Modelling RNA Folding. Electronic Proceedings in Theoretical Computer Science. 231. 31-41. 10.4204/EPTCS.231.3.

Coronaviruses and SARS-CoV-2

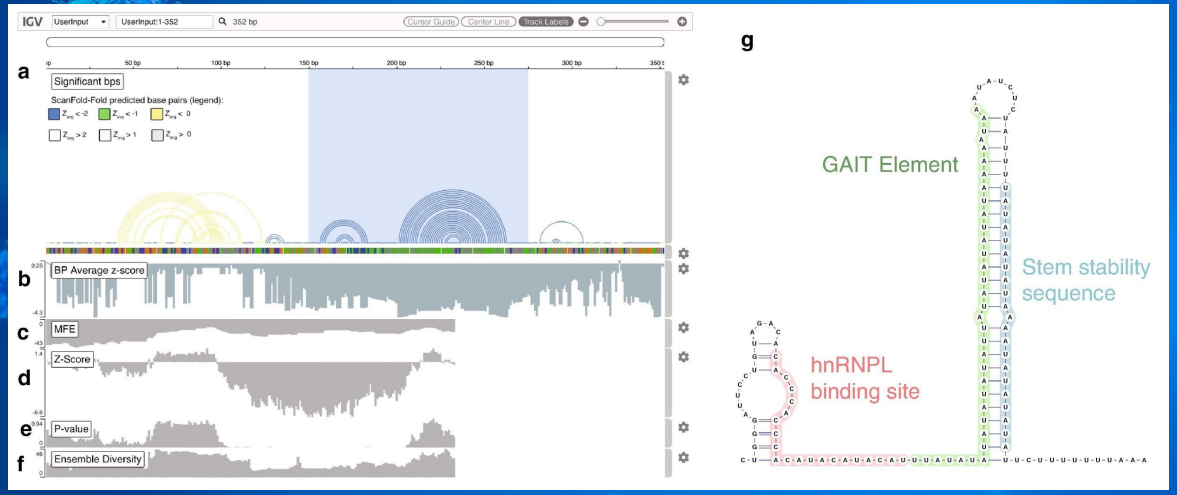


Reference: NIAID-RML.
Link: https://www.flickr.com/photos/niaid/49534865371/



Reference: Li, Chun, et al. "Genetic Evolution Analysis of 2019 Novel Coronavirus and Coronavirus from Other Species." *Infection, Genetics and Evolution*, vol. 82, 10 Mar. 2020, p. 104285., doi:10.1016/j.meegid.2020.104285.

Secondary Structure Prediction Algorithms



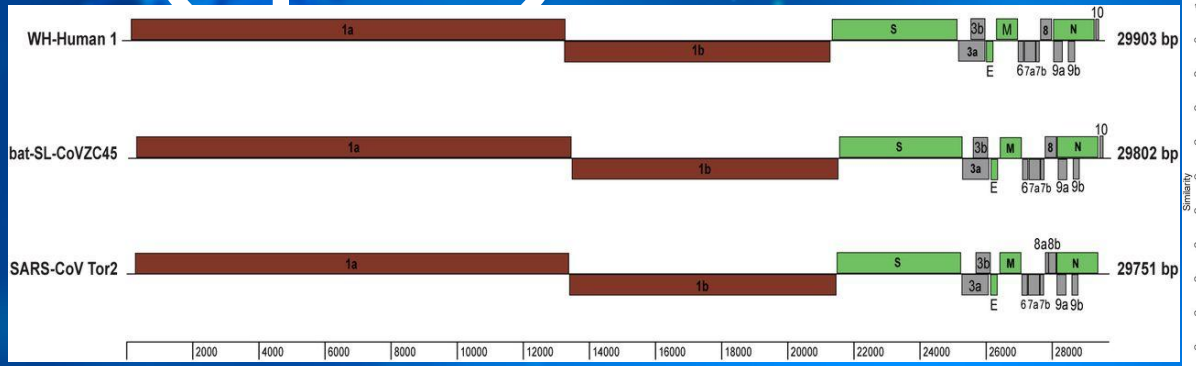
Reference: Andrews, Ryan J., et al. "Mapping the RNA Structural Landscape of Viral Genomes." *Methods*, 8 Nov. 2019, doi:10.1016/j.ymeth.2019.11.001.

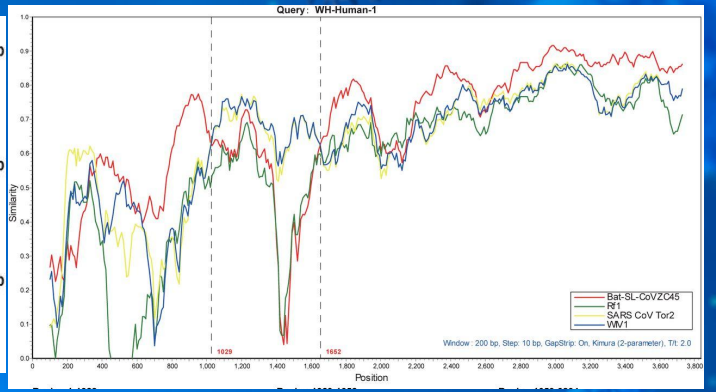
Name	Description	Knots [Note 1]	Links	References
CentroidFold	Secondary structure prediction based on generalized centroid estimator	No	sourcecode webserver	[1]
CentroidHomfold	Secondary structure prediction by using homologous sequence information	No	sourcecode webserver	[2]
Context Fold	An RNA secondary structure prediction software based on feature-rich trained scoring models.	No	sourcecode webserver	[3]
CONTRAfold	Secondary structure prediction method based on conditional log-linear models (CLLMs), a flexible class of probabilistic models which generalize upon SCFGs by using discriminative training and feature-rich scoring.	No	sourcecode webserver	[4]
Crumple	Simple, cleanly written software to produce the full set of possible secondary structures for one sequence, given optional constraints.	No	sourcecode	[5]
CyloFold	Secondary structure prediction method based on placement of helices allowing complex pseudoknots.	Yes	webserver	[6]
E2Efold	A deep learning based method for efficiently predicting secondary structure by differentiating through a constrained optimization solver, without using dynamic programming.	Yes	sourcecode	[7][8]
GTFold	Fast and scalable multicore code for predicting RNA secondary structure.	No	link sourcecode	[9]
IPknot	Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming.	Yes	sourcecode webserver	[10]
KineFold	Folding kinetics of RNA sequences including pseudoknots by including an implementation of the partition function for knots.	Yes	linuxbinary, webserver	[11][12]
Mfold	MFE (Minimum Free Energy) RNA structure prediction algorithm.	No	sourcecode, webserver	[13]
pKiss	A dynamic programming algorithm for the prediction of a restricted class (H-type and kissing hairpins) of RNA pseudoknots.	Yes	sourcecode, webserver	[14]
Pknots	A dynamic programming algorithm for optimal RNA pseudoknot prediction using the nearest neighbour energy model.	Yes	sourcecode	[15]
			sourcecode,	

Reference: https://en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software

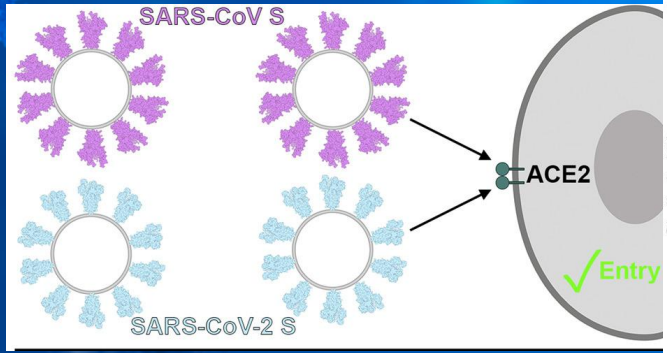
Dynamic Programming Algorithms, Nearest Neighbor Classifiers, NLP approaches involving CFGs, Other approaches

S (spike) Protein

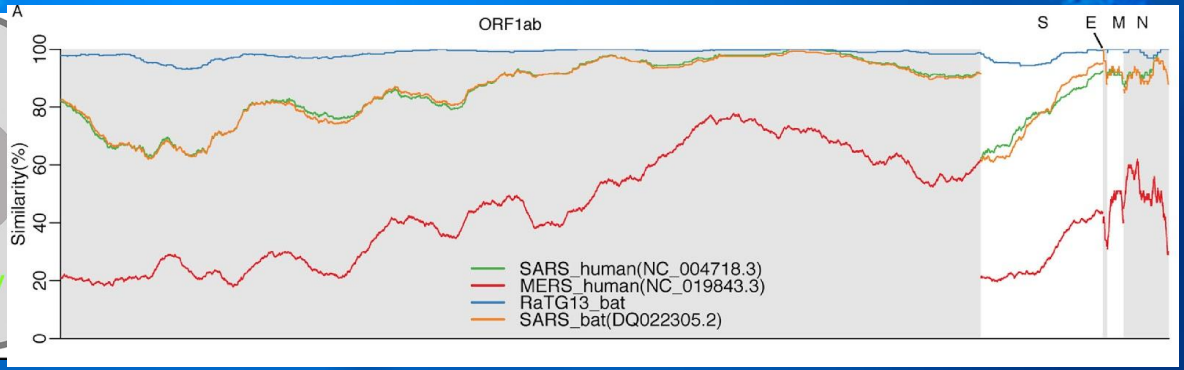




Reference: Wu, Fan, et al. "Complete Genome Characterisation of a Novel Coronavirus Associated with Severe Human Respiratory Disease in Wuhan, China." *BioRxiv*, 2020, doi:10.1101/2020.01.24.919183.



Reference: Walls, Alexandra C., et al. "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein." *Cell*, 9 Mar. 2020, doi:10.1016/j.cell.2020.02.058.

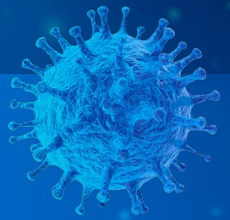


Reference: Wen, Feng, et al. "Identification of the Hyper-Variable Genomic Hotspot for the Novel Coronavirus SARS-CoV-2." *Journal of Infect*
doi:10.1016/j.jinf.2020.02.027.

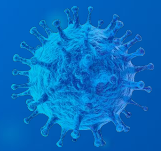
2. Task / Questions

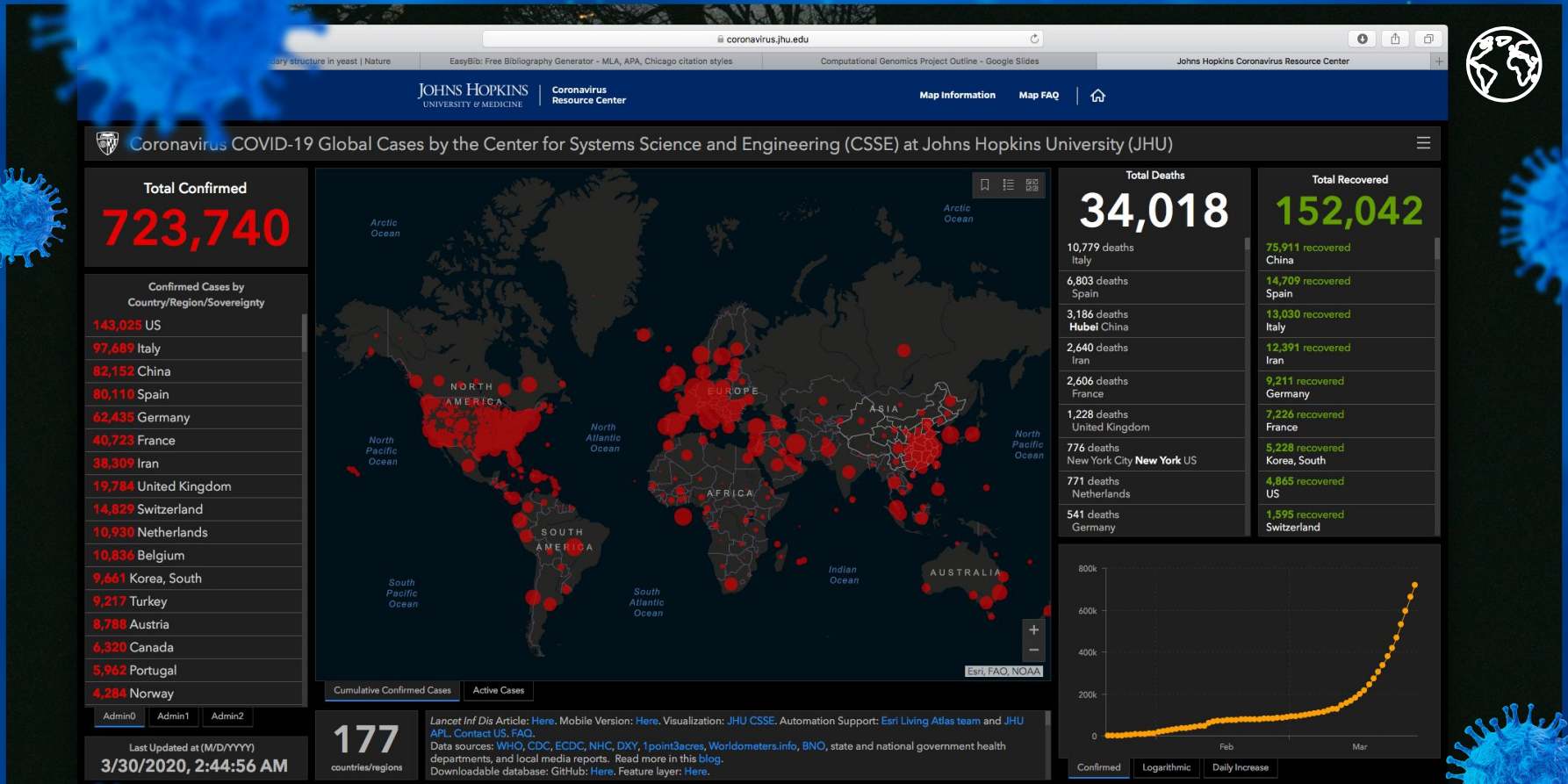
Can we link secondary structure to variation and even selection?

2. Task / Questions



- Can we link mutation rates / the mutations we see in the spike protein regions across the coronavirus genome with what we see in the corresponding secondary structures?
- What about across the SARS-CoV-2 genome?
- From any links we can establish between variation in the primary and secondary structures, what do they tell us about selection processes?
- Do structural insights help at all with developing a vaccine to exploit potential structural weaknesses?

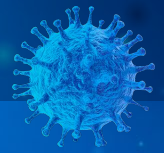



coronavirus.jhu.edu
EasyBib: Free Bibliography Generator - MLA, APA, Chicago citation styles
Computational Genomics Project Outline - Google Slides
Johns Hopkins Coronavirus Resource Center
JOHNS HOPKINS UNIVERSITY & MEDICINE | Coronavirus Resource Center
Map Information
Map FAQ
Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)
Total Confirmed
723,740
Confirmed Cases by Country/Region/Sovereignty
143,025 US
97,689 Italy
82,152 China
80,110 Spain
62,435 Germany
40,723 France
38,309 Iran
19,784 United Kingdom
14,829 Switzerland
10,930 Netherlands
10,836 Belgium
9,661 Korea, South
9,217 Turkey
8,788 Austria
6,320 Canada
5,962 Portugal
4,284 Norway
Admin0
Admin1
Admin2
Last Updated at (M/D/YYYY)
3/30/2020, 2:44:56 AM
Arctic Ocean
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Esri, FAO, NOAA
Cumulative Confirmed Cases
Active Cases
177
countries/regions
Lancet Inf Dis Article: Here. Mobile Version: Here. Visualization: JHU CSSE. Automation Support: Esri Living Atlas team and JHU APL. Contact US. FAQ.
Data sources: WHO, CDC, ECDC, NHC, DXY, 1point3acres, Worldometers.info, BNO, state and national government health departments, and local media reports. Read more in this blog.
Downloadable database: GitHub: Here. Feature layer: Here.
Total Deaths
34,018
10,779 deaths Italy
6,803 deaths Spain
3,186 deaths Hubei China
2,640 deaths Iran
2,606 deaths France
1,228 deaths United Kingdom
776 deaths New York City New York US
771 deaths Netherlands
541 deaths Germany
Total Recovered
152,042
75,911 recovered China
14,709 recovered Spain
13,030 recovered Italy
12,391 recovered Iran
9,211 recovered Germany
7,226 recovered France
5,228 recovered Korea, South
4,865 recovered US
1,595 recovered Switzerland
800k
600k
400k
200k
0
Feb
Mar
Confirmed
Logarithmic
Daily Increase

3. Project Plan

Determine + Gather Data, Determine + Run Software, Analyse Findings

Project Plan



Step 1

Figure out what the relevant RNA sequences I want to analyze are, then gather them. Both within SARS-CoV-2 genome as well as outside of it, i.e. in the broader coronavirus genome.

Step 2

Determine appropriate secondary structure prediction software + mutation detection / seq. alignment software and apply it to my collection of sequences gathered in Step 1.

Step 3

Write code to organize and analyse the output from Step 2 in the context of my task / questions.

Rough Timeline

Determine and Gather all Relevant Sequences

By 4/12

By 4/15

Determine Relevant Software

Run Software on Seqs.

By 4/18

By 4/26

Write + run code to organize + analyse output

Synthesize it all + write up findings

By 4/30

Questions?