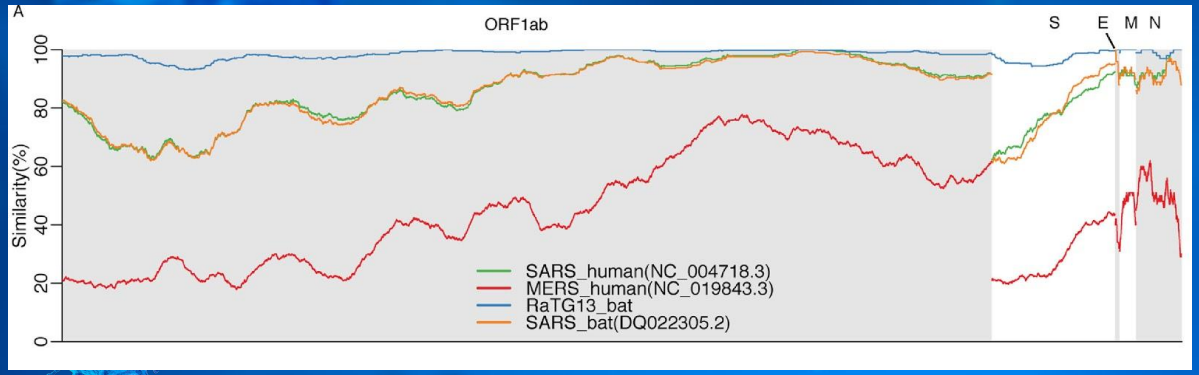
Using RNA Secondary Structure to Analyze Variations in the S (spike) Protein of SARS-CoV-2 and related Coronaviruses

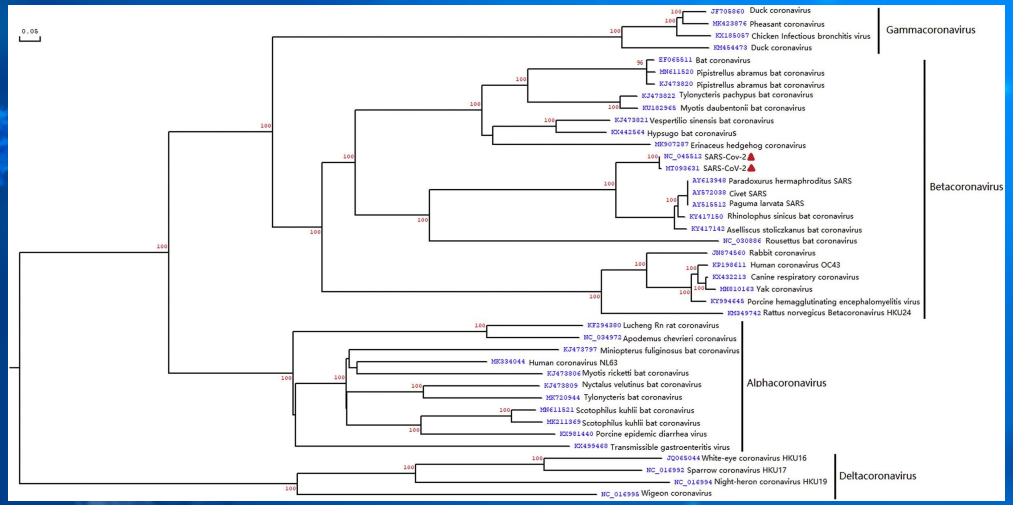
By Lawrence Chillrud

1. Introduction

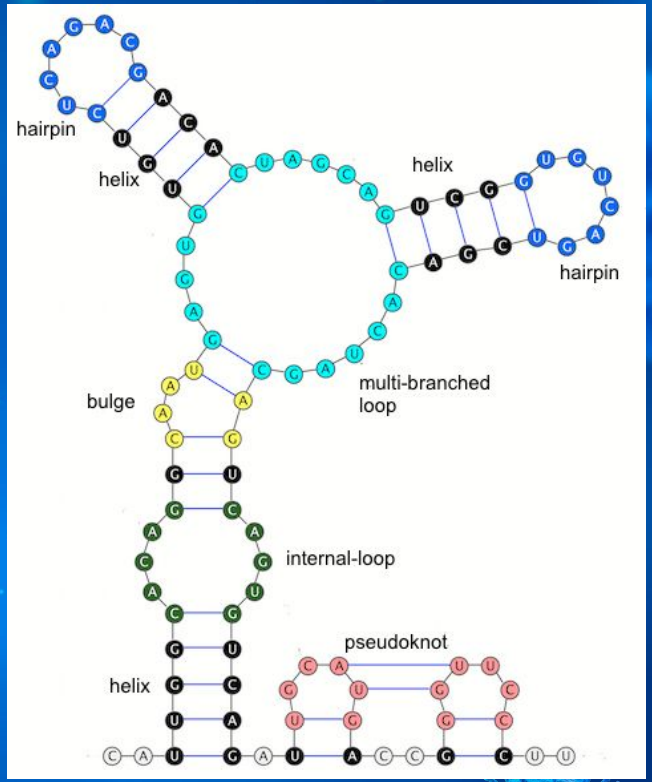
RNA Secondary Structure, Coronaviruses, S (spike) Protein



Reference: Wen, Feng, et al. "Identification of the Hyper-Variable Genomic Hotspot for the Novel Coronavirus SARS-CoV-2." *Journal of Infection*, 4 Mar. 2020, doi:10.1016/j.jinf.2020.02.027.



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.

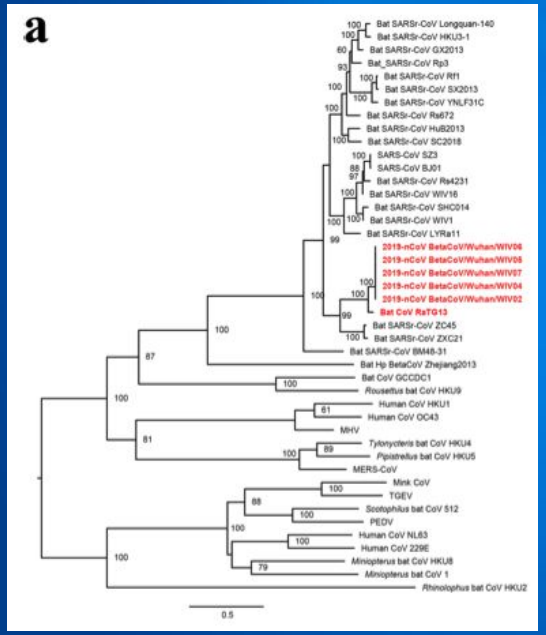


Reference: Mamuye, Adane & Merelli, Emanuela & Tesei, Luca. (2016). A Graph Grammar for Modelling RNA Folding. Electronic Proceedings in Theoretical Computer Science. 231. 31-41. 10.4204/EPTCS.231.3.

2. Methods

Sequence Selection, Alignment, Folding, Analysis

Interspecies Sequence Selection

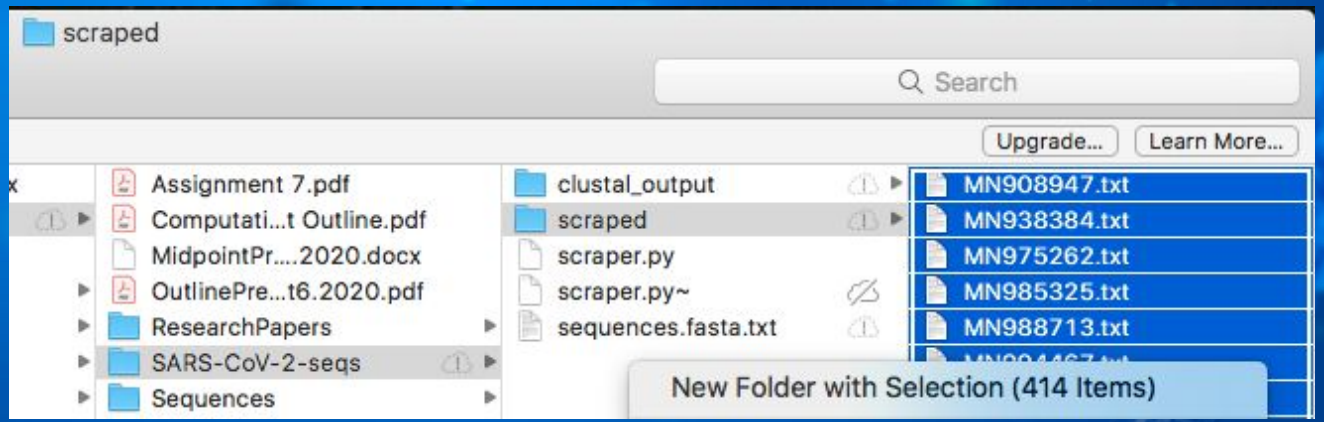


Reference: Zhou, Peng, et al. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." *Nature News*, Nature Publishing Group, 3 Feb. 2020, www.nature.com/articles/s41586-020-2012-7.

- 5 beta coronaviruses I am comparing across the beta coronavirus genus.
  - MN996532.1 (RaTG13)
  - NC_004718.3 (SARS)
  - NC_014470.1 (BM48-31 Bat CoV)
  - NC_019843.3 (MERS)
  - NC_045512.2 (SARS-CoV-2)
- This makes up the interspecies data set.

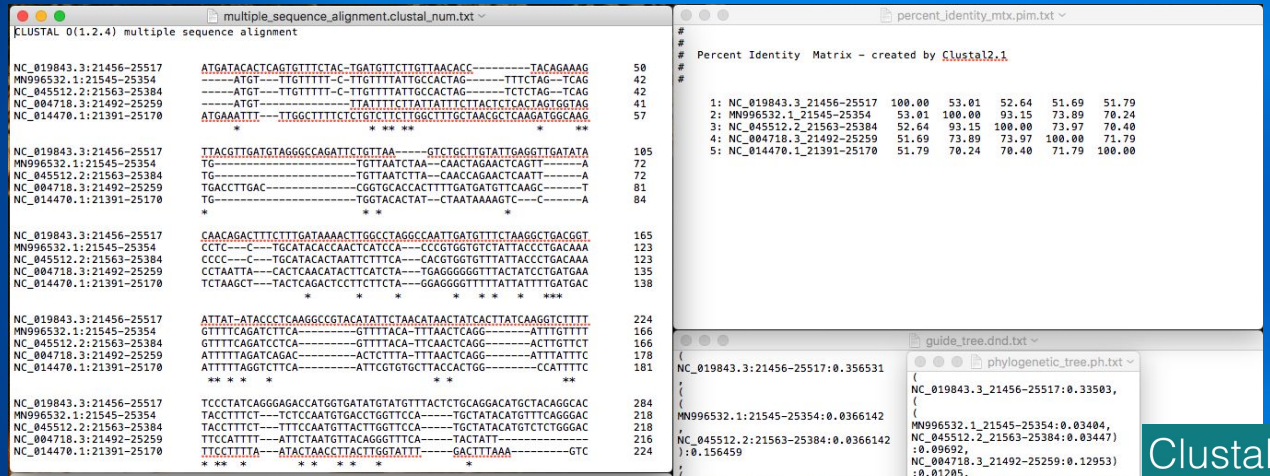
Intraspecies Sequence Selection

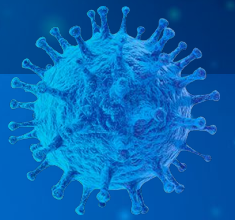
- 414 intraspecies sequences that capture the SARS-CoV-2 S gene.



Left: Screenshot of my python code to scrape genomic information from NCBI. Right: Screenshot of my folder containing the 414 scraped sequences encoding the S protein.

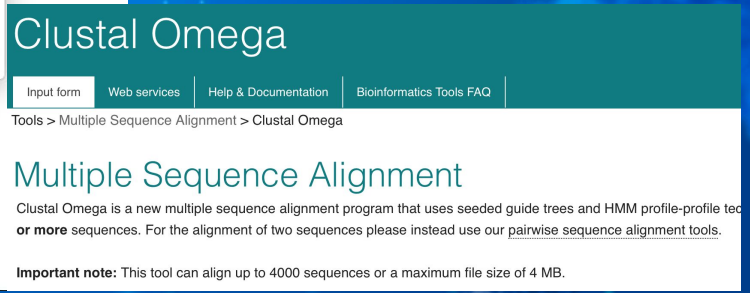
Sequence Alignment



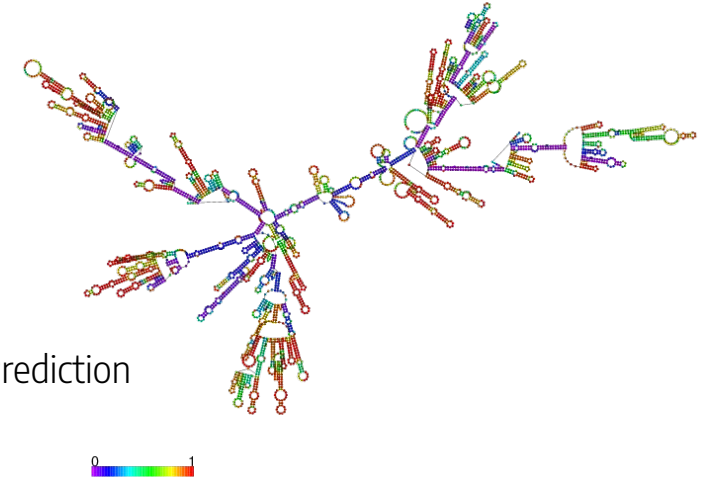


Left: Screenshot of the files output by the Clustal Omega algorithm for multiple sequence alignment.

Bottom: Screenshot of Clustal Omega algorithm's webserver:
https://www.ebi.ac.uk/Tools/msa/clustalo/

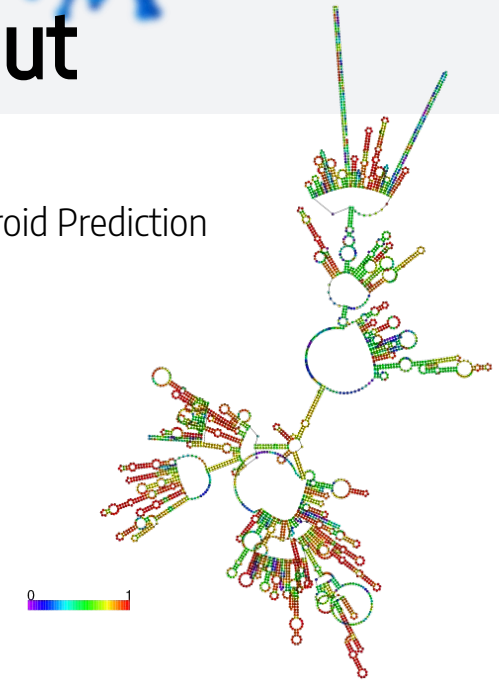


Vienna RNAfold: Example Output



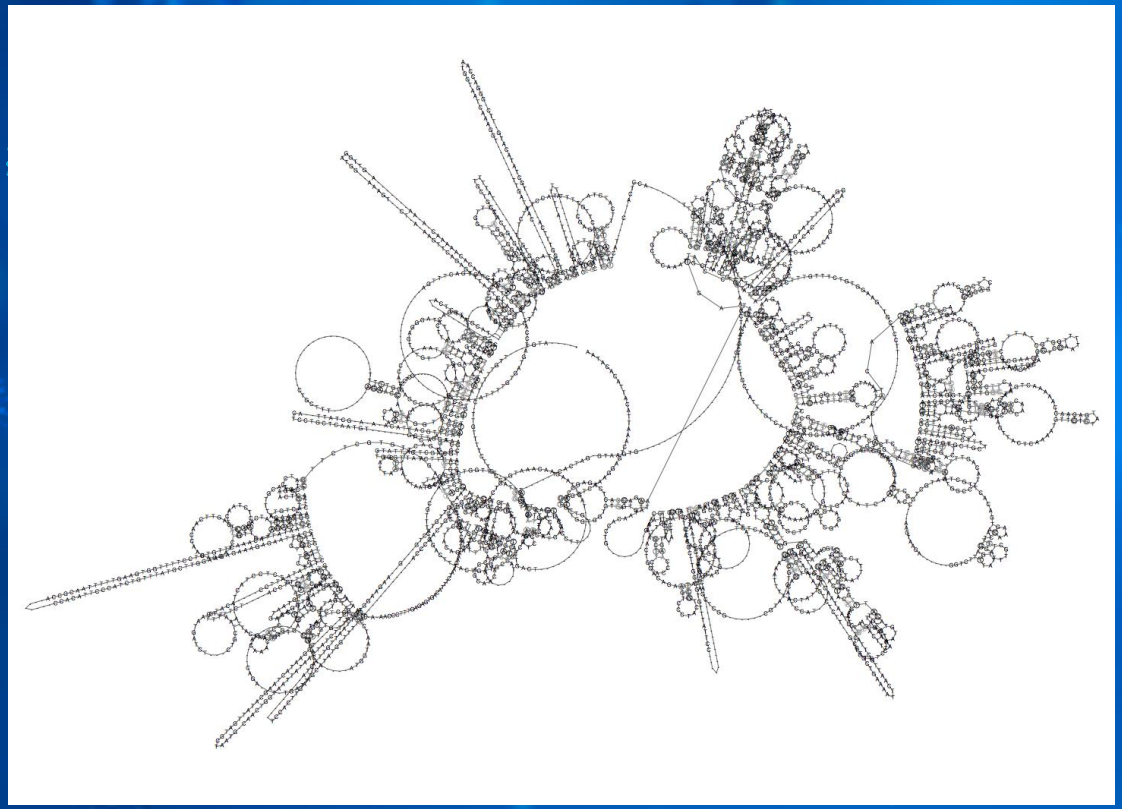
MFE Prediction

Centroid Prediction



Downloaded output from Vienna RNAfold's webserver:
http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi

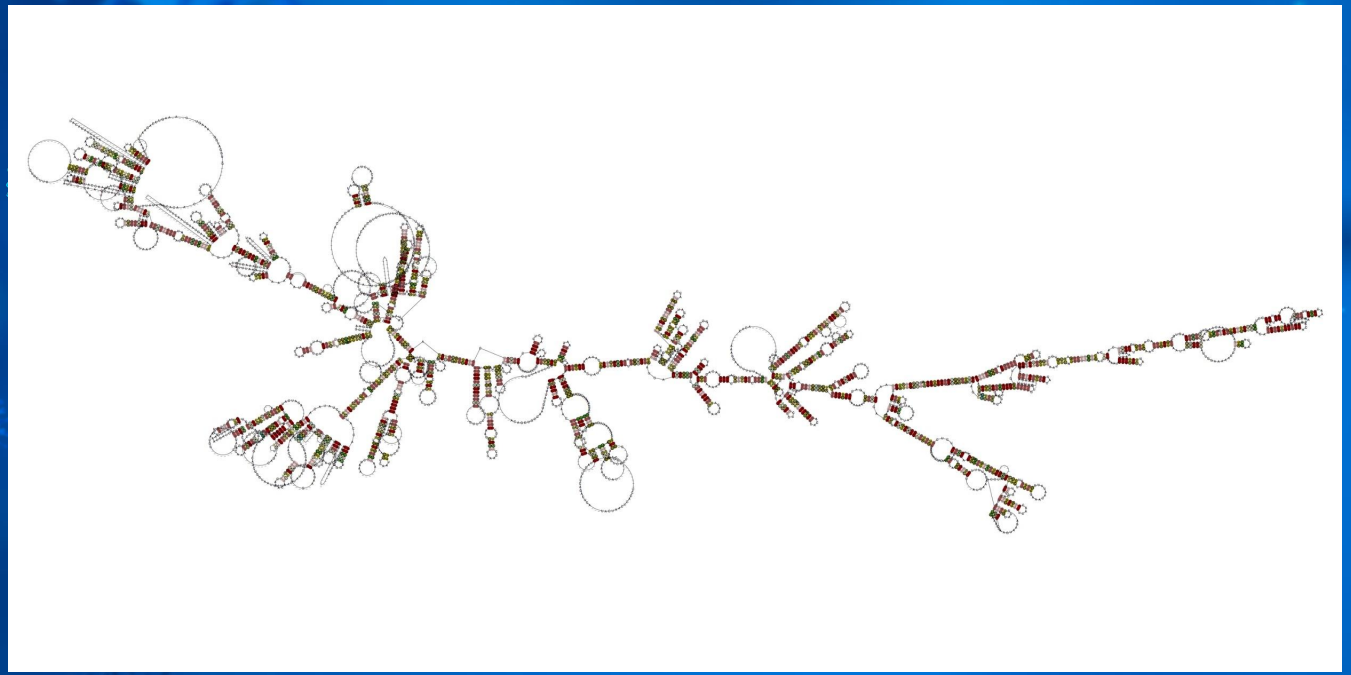
ViennaRNA alidot



"The program `alidot` detects conserved secondary structure elements in relatively small sets of RNAs by combining multiple sequence alignments and secondary structure predictions. Both a (good) sequence alignment and predicted secondary structure predictions for each sequence in the alignment must be provided as inputs. `alidot` works either with predicted mfe structures, or with base pairing probability matrices. The basic idea behind `alidot` is to sort the individual base pairs by their *credibility* and to reduce the number of entries in the list by subsequent filtering steps until only those secondary structure elements are left that are consistetly predicted."

Screenshot of the conserved secondary structures predicted by ViennaRNA's `alidot` program when given the interspecies dataset

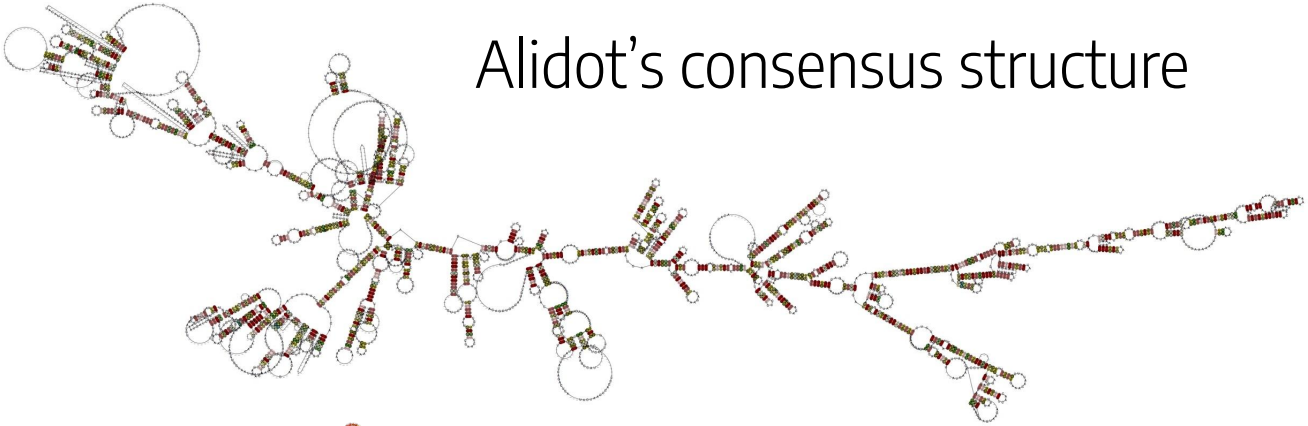
ViennaRNA alifold



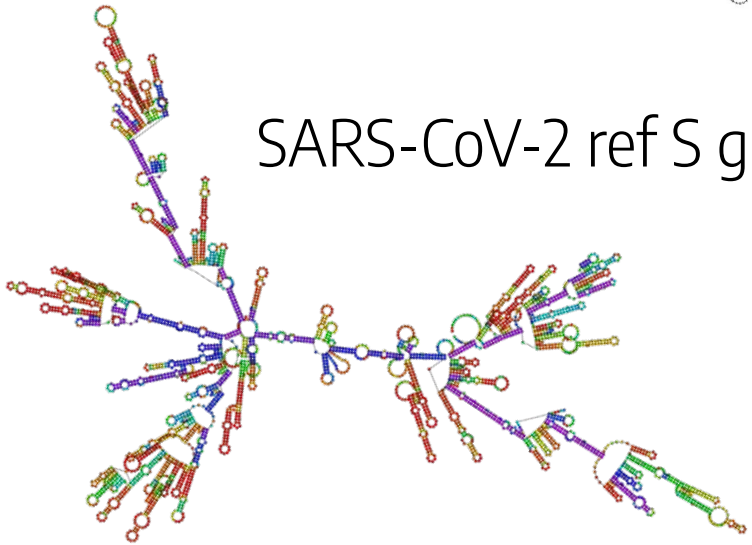
"The program `alifold` predict a consensus secondary structure of a set of aligned sequences.."

Screenshot of the consensus secondary structures predicted by ViennaRNA's `alifold` program when given the interspecies dataset

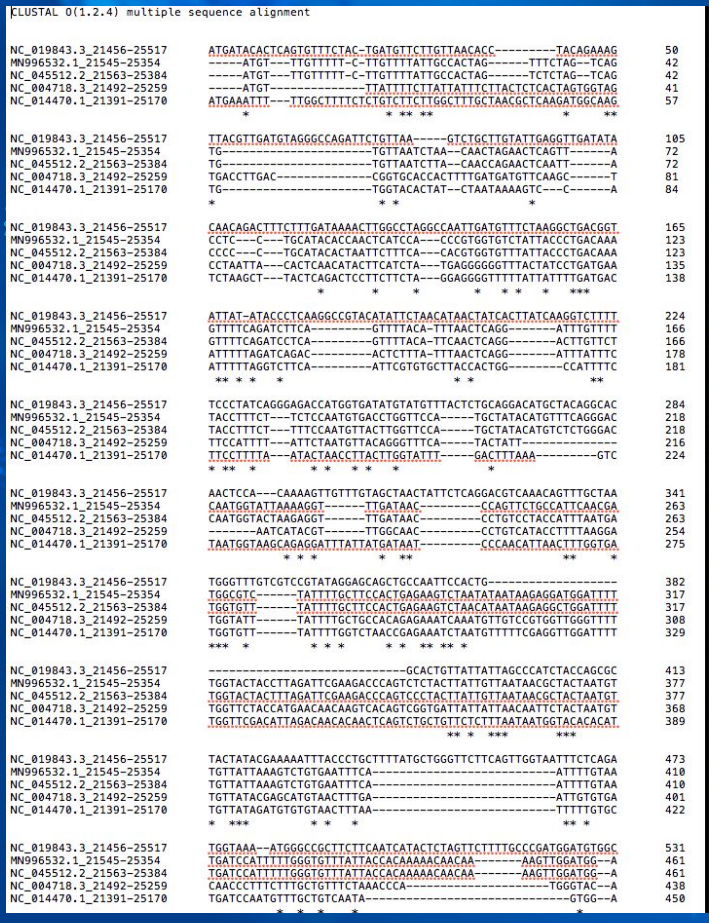
Alidot's consensus structure

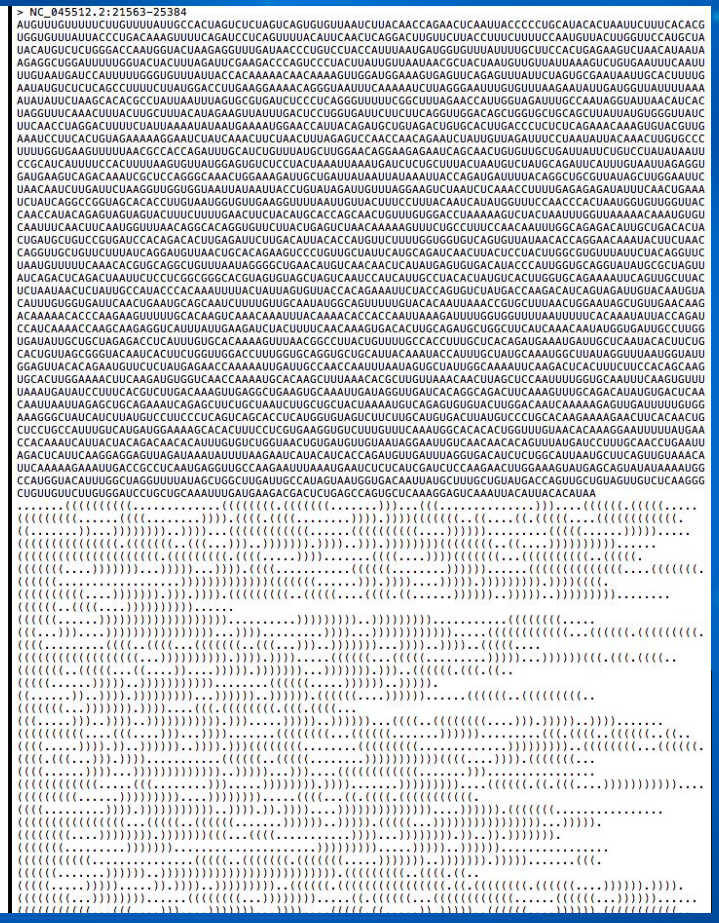


SARS-CoV-2 ref S gene



My Own Algorithm



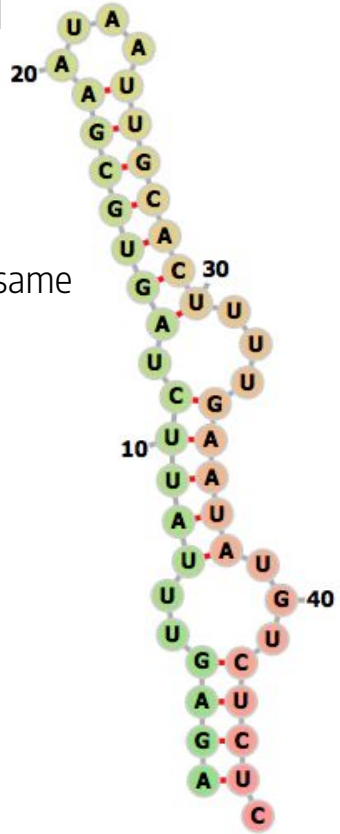


My Own Algorithm

	1	2	3	4	5	6	7
SC2	A	U	G	-	C	A	U
Bat	A	U	G	C	C	A	U
SARS	A	C	A	U	U	C	U
BM48-31	A	C	A	G	U	C	U
MERS	A	C	U	A	G	C	U
	(	(	(	.	)	)	)
	1	2	3	4	5	6	7
SC2	X	X	X	X	X	X	X
Bat	X	X	X	Y	X	X	X
SARS	X	Y	Y	Z	Y	Y	X
BM48-31	X	Y	Y	P	Y	Y	X
MERS	X	Y	Z	Q	Z	Y	X
	(	(	(	.	)	)	)

```
> 274-292 log likelihood: -37.62211326680391random seq of same size log likelihood: -52.7956961917466
UUGCUUCCACUGAGAAGUC
..(((((......))))).

> 471-516 log likelihood: -119.79646528234582random seq of same size log likelihood: -112.13864671127861
AGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAUGUCUCUC
(((((..(((((.(((((((.....)))))))...))))))...))))).

> 1267-1286 log likelihood: -36.05579777198838random seq of same size log likelihood: -45.16881916213479
AUAAAUUACCAGAUGAUUUU
..((((((.....)))))).

> 1294-1304 log likelihood: -26.6124989581944random seq of same size log likelihood: -28.890622866778113
GCGUUAUAGCU
((......)).

> 1571-1592 log likelihood: -45.9933061656108random seq of same size log likelihood: -57.10043842636637
UUGUGGACCUAAAAAGUCUACU
..((((((.......)))))).

> 2182-2203 log likelihood: -49.70515948468804random seq of same size log likelihood: -44.66402909690834
CAGUGUCUAUGACCAAGACAUC
..((((((.......)))))).

> 2244-2264 log likelihood: -55.533570155296665random seq of same size log likelihood: -33.862962813153295
UGCAGCAAUCUUUUGUUGCAA
(((((((((...)))))))))

> 2276-2297 log likelihood: -49.93569809155589random seq of same size log likelihood: -48.72767996143369
UUGUACACAAUUAAACCGUGCU
..((((...........)))).

> 2452-2471 log likelihood: -47.222710506602894random seq of same size log likelihood: -56.07092404684964
UUGAAGAUCUACUUUUCAAC
(((((((.....)))))))

> 2740-2758 log likelihood: -25.877639661851013random seq of same size log likelihood: -43.556302151640764
AUGUUCUCUAUGAGAACCA
..(((((...)))))..

> 3045-3065 log likelihood: -39.01146005367711random seq of same size log likelihood: -52.74933696093571
GCAGAAAUCAGAGCUUCUGCU
((((((........))))))

> 3165-3179 log likelihood: -25.621606268582298random seq of same size log likelihood: -41.41538476623717
GCACCUCAUGGUGUA
((((((....))))))

> 3435-3455 log likelihood: -44.41372625555621random seq of same size log likelihood: -39.82257756574254
GACUCAUUCAAGGAGGAGUUA
(((((.(((...)))))))).

> 3578-3593 log likelihood: -23.686253578019425random seq of same size log likelihood: -45.7479335967179
AAAUGAAUCUCUCAUC
..((((.....)))).

> 3676-3688 log likelihood: -25.95255005869852random seq of same size log likelihood: -31.210463338852193
CCAUAGUAAUGGU
((((....)))).
```

Largest potentially conserved secondary structure across interspecies dataset found at positions 471-516.
46 nt long.
Less “likely” than a random, uniformly distributed seq. of same size.



```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#


    1: NC_019843.3_21456-25517  100.00   53.01   52.64   51.69   51.79
    2: MN996532.1_21545-25354    53.01  100.00   93.15   73.89   70.24
    3: NC_045512.2_21563-25384   52.64   93.15  100.00   73.97   70.40
    4: NC_004718.3_21492-25259   51.69   73.89   73.97  100.00   71.79
    5: NC_014470.1_21391-25170   51.79   70.24   70.40   71.79  100.00
```

> 2805–3019 log likelihood: −44.147656945127004random seq of same size log likelihood: −59.62788118593763
GACUCACUUUCUUCCACAGCAAGUGCACUUGGAAAACUUCAAGAUGUGGUCAACCAAAAUGCACAAGCUUUAAACACGCUUGUUAAACAACUUAGCUCCAAUUUUGGUGCAAUUUCAAGUGUUUUAAAUGA
UAUCCUUUCACGUCUUGACAAAGUUGAGGCUGAAGUGCAAAUUGAUAGGUUGAUCACAGGCAGACUUCAAAGUUUGCAGACAUA
....(((((.((.....)).)))))..((((((....)))))).((((((((((..(((((((....((((((.(((((((.............((.
(((....))).))....)))))))))))).........((((.(((((((((...))))))))))))))))..)).....)))))))))).(((((((....)))))))).......

> 3045–3065 log likelihood: −4.121216772689696random seq of same size log likelihood: −3.2054185692852672
GCAGAAAUCAGAGCUUCUGCU
((((((........)))))).

> 3111–3187 log likelihood: −10.836212570759677random seq of same size log likelihood: −10.634812993118459
AAAAGAGUUGAUUUUUGUGGAAAGGGCUAUCAUCUUAUGUCCUUCCCUCAGUCAGCACCUCAUGGUGUAGUCUUCUU
........((((...((.((.((((((..........)))))).)).)))))))((((....)))))).........

> 3191–3205 log likelihood: −3.676568939767363random seq of same size log likelihood: −4.807758015843309
UGUGACUUAUGUCCC
...(((....)))..

> 3298–3340 log likelihood: −5.751592160071572random seq of same size log likelihood: −14.162117007367469
CACACUGGUUUGUAACACAAAGGAAUUUUUAUGAACCACAAAU
.....((((((((((.............))))))))))).....

> 3347–3413 log likelihood: −12.660082206683999random seq of same size log likelihood: −15.305094237511675
UACAGACAACACAUUUGUGUCUGGUAACUGUGAUGUUGUAAUAGGAAUUGUCAACAACACAGUUUAU
..((((((.((....))))))))..(((((((.(((((((((....)))).)))))..)))))))...

> 3429–3463 log likelihood: −12.85357645522643random seq of same size log likelihood: −6.951105167652991
GAAUUAGACUCAUUCAAGGAGGAGUUAGAUAAAUA
..(((.((((((.(((...)))))))))..)))).....

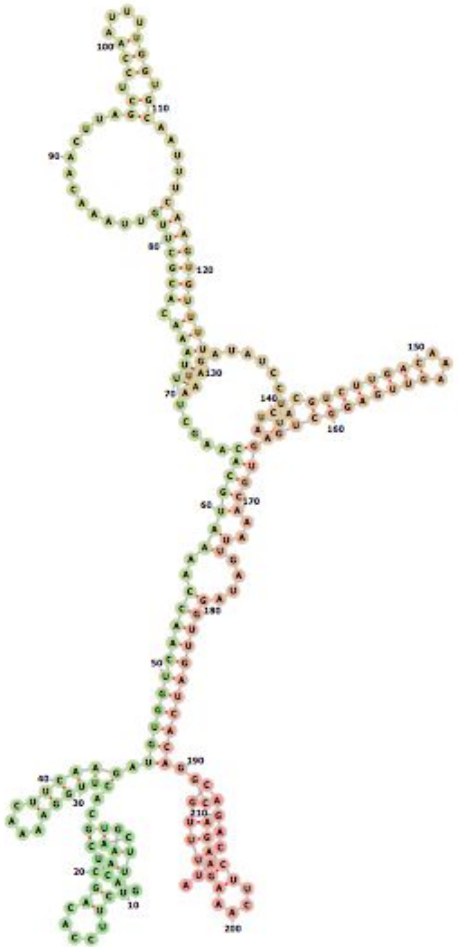
> 3508–3523 log likelihood: −1.1857275544595498random seq of same size log likelihood: −7.5064467303370765
CUGGCAUUAAUGCUUC
..((((....))))..

> 3533–3551 log likelihood: −6.537950185843786random seq of same size log likelihood: −5.359652399086683
CAUUCAAAAAGAAAUUGAC
...((((.......)))).

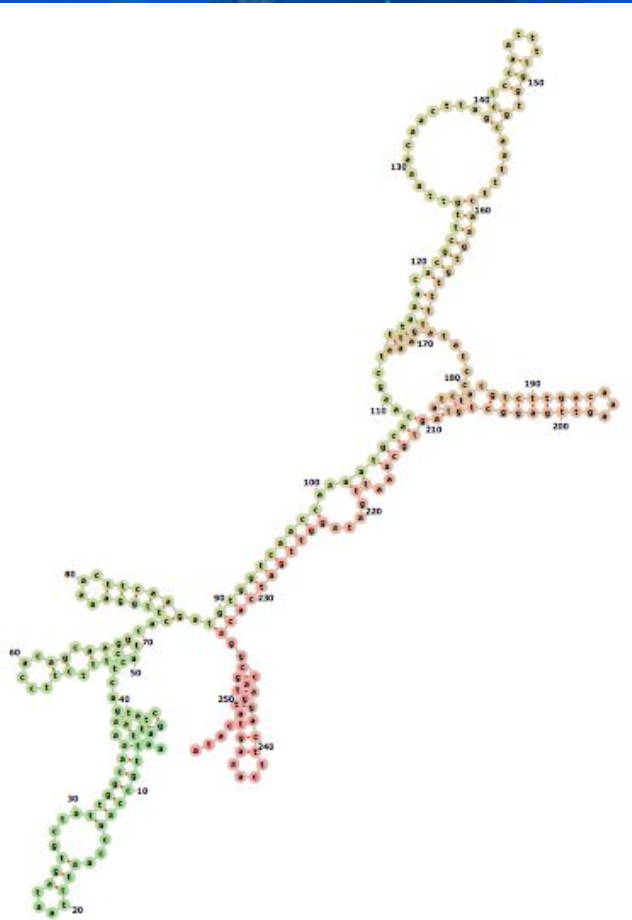
> 3576–3593 log likelihood: −3.898892856228527random seq of same size log likelihood: −5.068194018884908
UUAAAUGAAUCUCUCAUC
....((((.....)))).

> 3598–3617 log likelihood: −1.4821594430744374random seq of same size log likelihood: −7.714708669323717
UCCAAGAACUUGGAAAGUAU
(((((....)))))......

> 3625–3766 log likelihood: −25.91302819059777random seq of same size log likelihood: −32.79113113416216
AUAUAAAAUGGCCAUGGUACAUUUGGCUAGGUUUUAUAGCUGGCUUGAUUGCCAUAGUAAUGGUGACAAUUAUGCUUUGCUGUAUGACCAGUUGCUGUAGUUGUCUCAAGGGCUGUUGUUCUUGUGGAUCC
UGCUGCAAAUU
.........((((((.........))))))(((((.((((((.(((.((((((((((....))))...)))))))..)))..)))))).))))..((((.((((..
((((((((((....)))))))).))))).))))...



```
> 1865-1949 log likelihood: -16.5963577363989 64random seq of same size log likelihood: -35.24422772784434
tgctattcatgcagatcaacttactcctacttggcgtgtttattctacaggttctaatgtttttcaaacacgtgcaggctgttta
(((.......)))..................(((((((((((.....(((........))).....)))))))))).))))).......

> 1959-2022 log likelihood: -13.9306486802511 32random seq of same size log likelihood: -25.481418801349353
gaacatgtcaacaactcatatgagtgtgacatacccattggtgcaggtatatgcgctagttatc
....((((((.((.(((....)))))))))))....((((((((((......))))))))))....

> 2031-2109 log likelihood: -8.50282247119687random seq of same size log likelihood: -46.357351269672876
actaattctcctcggcgggcacgtagtgtagctagtcaatccatcattgcctacactatgtcacttggtgcagaaaatt
.....((((...((.((((.((((((((((.((((.........)))).)))))))))))..)))))..)).))))....

> 2142-2264 log likelihood: -13.23857169566095random seq of same size log likelihood: -65.7495641670686
cccacaaattttactattagtgttaccacagaaattctaccagtgtctatgaccaagacatcagtagattgtacaatgtacatttgtggtgattcaactgaatgcagcaatcttttgttgcaa
.................(((((((((((((((...(((((..((((((.......))))))..)))))).(((((...)))))))))))))))...)))))).(((((((((....)))))))).

> 2275-2299 log likelihood: -2.690766604809136random seq of same size log likelihood: -10.729816900266393
tttgtacacaattaaacccgtgcttt
...((((..........))))...

> 2324-2367 log likelihood: -51.934008672339765random seq of same size log likelihood: -17.020280061126137
caaaaacacccaagaagtttttgcacaagtcaaacaaatttaca
(((((((.........)))))))..................

> 2426-2492 log likelihood: -45.836305262247855random seq of same size log likelihood: -31.092225146443187
atcaaaaccaagcaagaggtcatttattgaagatctacttttcaacaaagtgacacttgcagatgct
..........(((((..(((((((.(((((((.....)))))))..)))))))..)))))).....

> 2497-2517 log likelihood: -2.2602439480396743random seq of same size log likelihood: 4.597933888818781
tcatcaaacaatatggtgatt
((((((.......))))))..

> 2545-2604 log likelihood: -6.457839851541927random seq of same size log likelihood: -36.61917484631782
tcatttgtgcacaaaagtttaacggccttactgttttgccacctttgctcacagatgaaa
((((((((..((((.((..(((((.....)))))).....)).)))))..))))))..

> 2636-2711 log likelihood: -8.179930478619774random seq of same size log likelihood: -22.792889427825823
gggtacaatcacttctggttggacctttggtgcaggtgctgcattacaaataccatttgctatgcaaatggcttat
((((.((((((....)))))).)))).((((((((...))))))))......((((((((...))))))))).....

> 2736-2757 log likelihood: -2.3678746122320398random seq of same size log likelihood: -6.253946944069332
cagaatgttctctatgagaacc
......(((((((...)))))).

> 2761-3019 log likelihood: -95.20654599662484random seq of same size log likelihood: -132.28422885280503
aattgattgccaaccaatttaatagtgctattggcaaaattcaagactcactttcttccacagcaagtgcacttggaaaacttcaagatgtggtcaaccaaaatgcacaagctttaaacacgcttgttaaacaacttag
ctccaatttggtgcaatttcaagtgttttaaatgatatcctttcacgtcttgacaaagttgaggctgaagtgcaaattgataggttgatcacaggcagacttcaaagtttgcagacata
..((((((((((...(((....)))....))))))...))))....(((((.((.....)).)))))..((((((....)))))).((((((((((..(((((((....((((((.(((((((.............
((.(((....))).))....)))))))))))))).........((((.(((((((((...))))))))))))))))..)).....)))))))))).(((((((....)))))))).......

> 3045-3074 log likelihood: -3.2289199257709633random seq of same size log likelihood: -6.901323212637057
gcagaaatcagagcttctgctaatcttgct
((((((........))))))..........

> 3111-3187 log likelihood: -69.92152455286477random seq of same size log likelihood: -3.774327179254454
aaaagagttgatttttgtggaaagggctatcatcttatgtccttccctcagtcagcacctcatggtgtagtcttctt
........((((...((.((.((((((...........)))))).)).)))))))((((....))))).........
```



3. Results

Difficulty getting past the bugs...

Predictions...

When only looking at SARS-CoV-2 and RaTG13, found 54 potential conserved structures, the longest one (that is more likely than expected) is 214 nucleotides long... (32/54 “likely”)

As we add in species to look at, we get less potential conserved structures found, they’re usually shorter, and usually they aren’t as likely as in the lower dimensional cases...

When looking across all 5 species, found 15 potential conserved structures, the longest one is 45 nucleotides long, although it’s not as likely as we would expect. (10/15 “likely”)

4. Conclusions

Summary, Limitations, Prospective Future Work

Summary, Limitations, Future Work

Picked out potential conserved secondary structures but with limited confidence – bugs really dragged this project down along with limited resources.

Loads of assumptions were made that decrease the accuracy of the project – e.g., collapsation of the quintets, only utilizing the SARS-CoV-2 secondary structure, abbreviated application of the MSA. Notion of “likely” is fuzzy as well; statistical significance analysis needed.

Future work would involve getting rid of the limitations discussed, making the methods more valid! Also a need to take into account thermodynamic data as other analyses do...

Questions?