
Reverse engineering the coronavirus (SARS-CoV-2)

Start here: [corona.py](#)

:thought_balloon: Background

This project applies techniques from reverse engineering to understand the SARS-CoV-2 virus. The goal here is simply to build an understanding of the virus from first principles.

Biology vs. software

Biological systems are fundamentally information processing systems. While not a perfect analogy, software provides a useful framework for thinking about biology. The table below provides a rough outline of this analogy.

:microscope: Biology | :computer: Software | Notes --- | --- | --- nucleotide | byte | genome | bytecode | translation | disassembly | 3 byte wide instruction set with arbitrary “reading frames” protein | function | a polypeptide is a function with multiple pieces protein secondary structure | basic blocks | 80% accuracy in prediction protein tertiary structure | | This seems like the hard one to predict: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205819> quaternary structure | compiled function with inlining | https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction_prediction gene | library | bacteria are statically linked, viruses are dynamically linked transcription | loading protein structure prediction | library identification genome analysis | static analysis | molecular dynamics simulations of protein folding | dynamic analysis | Simulation doesn’t seem to work yet. Constrained by tooling and compute. no equivalent | execution | We are reverse engineering a CAD format. Runs more like FPGA code, all at once. No serial execution. (What are the FPGA reverse engineering tools?)

:wrench: Progress

Downloading the SARS-CoV-2 genome

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA and RNA sequences. The SARS-CoV-2 sequences available in GenBank have been downloaded in [download_sequences.py](#).

Translating RNA to proteins

`lib.py` contains a function `translate` that converts an RNA sequence to a chain of amino acids. This function is used in `corona.py`.

Annotating functions

The `translate` function is used in `corona.py` to identify and annotate functions for all proteins encoded by the genome.

Folding proteins

The OpenMM toolkit is used for molecular simulation of protein folding in `fold.py`.

:bulb: Work to be done

- Automatic extraction of genes from different coronaviruses
- Good multisequence compare tool
- Molecular dynamics?
- Secondary Structure prediction on orf1a?

:question: Open questions

- How is orf1ab cleaved into polypeptides? Can we predict this from the sequence?
- How do the researchers know (guess?) where orf1ab cleaves?
 - nsp3 and nsp5 do it – <https://www.pnas.org/content/pnas/103/15/5717.full.pdf>
- Which protein is the immune system responding to?
 - “spike” and “nucleocapsid” – <http://www.cmi.ustc.edu.cn/1/3/193.pdf>
 - Are some people already immune from exposure to other coronavirus?
- Find the “furin cleavage site” in the “spike glycoprotein”
 - It might be at the “PRRA” – <https://www.sciencedirect.com/science/article/pii/S0166354220300528>
 - Use ProP or PiTou to predict? – <https://en.wikipedia.org/wiki/Furin>
- How similar are the other coronaviruses? (causes colds, not either SARS or MERS)
 - alpha

-
- ★ https://en.wikipedia.org/wiki/Human_coronavirus_229E (simpler, though targets APN)
 - ★ https://en.wikipedia.org/wiki/Human_coronavirus_NL63 (targets ACE2!)
 - <https://www.ncbi.nlm.nih.gov/nuccore/MG772808>
 - beta
 - ★ https://en.wikipedia.org/wiki/Human_coronavirus_OC43 (targets Neu5Ac)
 - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2095096/pdf/JIDMM17330.pdf>
 - Specifically, how similar is the N protein OC43, SARS v1, and SARS v2?
 - ★ https://en.wikipedia.org/wiki/Human_coronavirus_HKU1 (targets Neu5Ac)
 - ★ MERS-CoV
 - ★ SARS-CoV
 - ★ SARS-CoV-2
 - What adds the phosphate group to the N protein? Kinase?

:droplet: Testing

How tests work

- All based on https://en.wikipedia.org/wiki/Reverse_transcription_polymerase_chain_reaction
- USA – <https://www.fda.gov/media/134922/download>
 - selected from regions of the virus nucleocapsid (N) gene
 - 28286–28308–28332–28358
 - 29163–29187–29210–29230
 - https://biosearchtech.a.bigcontent.io/v1/static/coa_KIT-NCOV-PP1-1000_Lot-No-143503
- South Korea – http://www.kogene.co.kr/eng/about_us/news/listbody.php?h_gcode=board&h_code=7&po_no
 - E gene detection (same for all coronavirus)
 - specific RdRp detection

Homemade test?

- Isolation of viral RNA (no matter what)
 - <https://www.qiagen.com/us/products/diagnostics-and-clinical-research/sample-processing/qiaamp-viral-rna-mini-kit/#orderinginformation>
- Primers and probes (to detect SARS-CoV-2)

-
- <https://www.biosearchtech.com/products/pcr-kits-and-reagents/pathogen-detection/2019-ncov-cdc-probe-and-primer-kit-for-sars-cov-2>
 - Wouldn't need if using a nanopore sequencer (nanopore MinION)
 - RT-qPCR Master Mix (to PCR)
 - <https://www.thermofisher.com/order/catalog/product/A15300#/A15300>
 - Probably wouldn't need if using a nanopore sequencer
 - All in one?
 - <https://www.chaibio.com/coronavirus>
 - Open qPCR, understand <https://www.chaibio.com/openqpcr>
 - FAM and HEX fluorophores?

:pill: Possible treatments and prophylactics

:warning: **Disclaimer:** The information in this repository is for informational purposes only. It is *not* medical advice.

Hydroxychloroquine + zinc

- Zinc blocks RdRp
- <https://jvi.asm.org/content/91/21/e00754-17> – how similar is Hep E RdRp?
- <https://www.ncbi.nlm.nih.gov/pubmed/21079686>
- Chloroquine is a Zinc Ionophore (allows zinc into the cell)
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4182877/>
- Without zinc (seems ineffective)
 - <https://www.recoverytrial.net/files/hcq-recovery-statement-050620-final-002.pdf> (where is study?)
- With zinc
 - https://www.infezmed.it/media/journal/Vol_28_2_2020_9.pdf
 - <https://www.sciencedirect.com/science/article/pii/S0306987720306435>
 - No results yet: <https://www.clinicaltrials.gov/ct2/show/NCT04377646>

RdRP inhibitors

- Favipiravir (prodrug for favipiravir-RTP)

-
- Adenosine Analog
 - Remdesivir (prodrug for GS-441524)
 - ★ <https://www.nejm.org/doi/pdf/10.1056/NEJMoa2007764?articleTools=true>
 - Galidesivir

Dexamethasone

- corticosteroid = anti-inflammatory + immunosuppressant
- <https://www.medrxiv.org/content/10.1101/2020.06.22.20137273v1.full.pdf>
- Effects in late stage (blocking immune response?)

Lopinavir-Ritonavir (AIDS cocktail)

- Both drugs are protease inhibitors
- Wouldn't expect it to work, don't think covid has protease
- No effect: https://www.recoverytrial.net/files/lopinavir-ritonavir-recovery-statement-29062020_final.pdf

:books: Resources

Coronavirus-related publications

- Chapter 4 - Coronavirus Pathogenesis – <https://www.sciencedirect.com/science/article/pii/B9780123858856000000>
- <https://www.futuremedicine.com/doi/pdf/10.2217/fvl-2018-0008>
- <https://www.sciencedirect.com/science/article/pii/S2095177920302045>
- <http://korkinlab.org/wuhan>
- https://github.com/mattroconnor/deep_learning_coronavirus_cure

Biology

- textbooks
 - Molecular Biology of the Cell
- classes
 - better tests - <https://ocw.mit.edu/courses/biology/7-012-introduction-to-biology-fall-2004/index.htm>

-
- suspected better lectures - <https://ocw.mit.edu/courses/biology/7-014-introductory-biology-spring-2005/index.htm>
 - alternative lectures - <https://youtube.com/playlist?list=PLGhmZX2NKiNldpyRUBBEzNoWL0Cso1jip>
 - basics - <https://www.khanacademy.org/science/biology>

Bioinformatics

- Nextstrain
 - <https://nextstrain.org/ncov>
 - <https://github.com/nextstrain/ncov>
 - <https://github.com/nextstrain/augur>
- <https://biopython.org/>
- <https://github.com/cogent3/cogent3>
- https://www.reddit.com/r/bioinformatics/comments/191ykr/resources_for_learning_bioinformatics/
- <https://en.wikipedia.org/wiki/Bioinformatics>

Epidemic modeling

- <https://gabgoh.github.io/COVID/index.html>
- <http://epidemicforecasting.org>

Antibodies

- <https://www.medrxiv.org/content/10.1101/2020.07.09.20148429v1.full.pdf>
- <https://www.medrxiv.org/content/10.1101/2020.05.11.20086439v2.full.pdf>

Masks

- Looking for large controlled studies measuring infection rate with/without wearing mask
- <https://bmjopen.bmj.com/content/bmjopen/5/4/e006577.full.pdf> – cloth masks bad surgical masks nothing
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2662657/pdf/08-1167_finalRCME.pdf – unclear, low adherence
- <https://sci-hub.tw/10.1001/archpedi.1987.04460060111049> – masks + goggles help, unclear on mask type

-
- <https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1750-2659.2011.00198.X> – N95 ~2x more effective than medical masks
 - <https://jamanetwork.com/journals/jama/article-abstract/184819> – N95 and surgical similar in effectiveness

Vaccines

- https://en.wikipedia.org/wiki/COVID-19_vaccine
- <https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html>
- generic vaccine types – <https://www.niaid.nih.gov/research/vaccine-types>
- Nucleic Acid Vaccines (none are currently used)
 - Use mRNA to produce viral proteins
 - ★ “It’s a very unique way of making a vaccine and, so far, no (such) vaccine has been licenced for infectious disease”
 - ★ Moderna, mRNA-1273 – <https://clinicaltrials.gov/ct2/show/NCT04283461>
 - ★ <https://www.biorxiv.org/content/10.1101/2020.04.22.055608v1.full.pdf>
 - ★ <https://www.medrxiv.org/content/10.1101/2020.06.30.20142570v1.full.pdf>
 - DNA vaccine
 - ★ “No DNA vaccines have been approved for human use in the United States”
 - ★ Based on injecting DNA (plasmid) that expresses the spike protein
 - ★ <https://siasky.net/bACLKGMcmX4NCp47WwOOJf0lU666VLeT5HRWpWVtqZPJEA>
 - Viral Vector Vaccines
 - ★ “There are no viral vector vaccines currently on market for use in humans.”
 - ★ adenovirus vector, AZD1222
 - ★ recombinant adenovirus type 5 vector, Ad5-nCoV
- Protein-Based Vaccines (subunit vaccines?)
- Whole-Virus Vaccines (inactivated vs live-attenuated)
 - Like old vaccines (live attenuated?)
 - In phase 3 already, Sinopharm = 15,000 people
- Nanoparticle vaccine?
 - <https://www.ipd.uw.edu/2019/03/ipds-first-nanoparticle-vaccine/>
- Reverse Engineering
 - <https://berthub.eu/articles/posts/reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/>

-
- <https://berthub.eu/articles/posts/part-2-reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/>

Genome studies (what genes = bad covid)

- <https://www.nejm.org/doi/full/10.1056/NEJMoa2020283?source=nejmtwitter&medium=organic-social>

DNA Synthesis

- Produce “oligos” of around 170-200 bp
 - “Phosphoramidite chemistry” – it’s a chemical process
 - Size limited by $0.99^n \rightarrow 13\%$ yield on 200bp if the process is 99% good
 - Column based / Array based
 - <https://www.nature.com/articles/nmeth.2918>
- Then assemble them
 - https://en.wikipedia.org/wiki/Gibson_assembly
 - 20-40 bp overlap
 - 5-15 oligos can be combined at once (~2 kb produced)
- New technologies
 - <https://static1.squarespace.com/static/5c981af7ebfc7fc8528d6564/t/5f4048cda662a571341b9a60/15980+MDD+DNA+Writing.pdf>