Data Analytics and Computational Statistics
Prof. Dr. Bernhard Renard

Master Project 2021

Continuously tracking and filtering SARS-CoV-2 mutations for molecular surveillance

The ongoing pandemic caused by SARS-CoV-2 emphasizes the importance of molecular surveillance to understand the evolution of the virus and to monitor and plan the epidemiological responses.

In this project you will improve and extend CovRadar, a tool used at the Robert Koch Institute (RKI) for molecular surveillance of the Coronavirus, which means that we are closely monitoring the temporal and geographical course of the virus at the epidemiological and, in particular, genetic level, in order to allow deriving actions to control the pandemic. CovRadar shows upcoming changes at the Corona Spike protein, an important target for vaccines.

CovRadar consists of an analysis pipeline written in Python and Snakemake and a web application written in Flask that enable the analysis and visualization of over 360,000 genetic sequences. The app is hosted on a cloud and available at `https://covradar.net/`.

The analysis is based on Next Generation Sequencing (NGS) data obtained from different databases for genomic data and internal RKI resources. First, CovRadar extracts the regions of interest using local alignment, then builds a multiple sequence alignment, infers variants, consensus sequences and phylogenetic trees and finally presents the results in an interactive PDF-like app.

Biological background

The coronavirus has its genetic material stored in single-stranded RNA. This RNA is extracted, reverse-transcribed, amplified and identified in PCR tests to confirm infection with the coronavirus. We use Next Generation Sequencing (NGS) to determine the SARS-CoV-2 viral RNA sequences of many different infected patients. We are interested in where in the genome of the virus changes evolve, what effects they have on the protein structure of the virus (e.g., can it bind more strongly to the human host as a result of the mutation?), and where in the world these variants occur. For now, we focus on the Spike protein - the connection to the human host cell. I.e., we first make a local alignment to the reference genome (the first case in Wuhan) and separate the Spike protein region. All these extracted sequences are aligned (multiple sequence alignment) to detect differences between them. From this, mutations (variants) can be inferred. In order to be able to make a temporal and local statement, we form a consensus sequence for each calendar week and country from the corresponding sequences (i.e. we take the most frequent nucleotide per position). From the variants we can calculate a phylogenetic relationship, which is visualized in phylogenetic trees.

Possible projects Depending on your interests and skills, the project offers many possibilities that can be explored:

1. Algorithmic development: We want new features to be developed, such as automatic updates, analysis of all genes in SARS-CoV-2 (currently only the spike protein is analyzed), reporting of protein positions and mutation tracking.

2. Scaling: Right now CovRadar accesses over 360,000 SARS-CoV-2 sequences. It is expected that this number will increase drastically in the next months. We also want to extend our analysis to the full viral genome. To keep up with this development the performance in both backend and frontend needs to be improved.

3. Software quality: Writing automated tests and implementing CI/CD pipelines to complete our test suite would allow an easier integration of new features, as well as facilitate future improvements in the source code.

4. Data visualization: New visualizations are required to improve the effectiveness of RKI's analysis and their predictions for the evolution of the pandemic in Germany.

5. Subproject benchmarking: There are many similar tools available in the literature for each step of our analysis. It is possible within the project to benchmark the most popular and novel software for each step in the analysis to find the best combination for our research question.

For those interested in biology, we can offer the following topics (no prior knowledge of biology required):

6. Subproject genomic epidemiology: CovRadar's analysis pipeline generates phylogenetic trees that have yet to be incorporated into the web app. Furthermore, with the upcoming German sequences, new analyses will be possible that could impact future lockdown politics and help to monitor the effectiveness of current vaccines.

7. Subproject positive selection: An algorithmic and frontend framework for long-range interactions between variants or sites that have a positive selection signal needs to be developed.

8. Subproject compensatory mutations: An algorithmic and frontend framework for the investigation of compensatory mutations also need to be developed. It is likely that when the sequence on one site changes (mutates), another site mutates as well to keep a certain protein structure or functionality. This is also known as "compensatory mutations", so mutations that occur in close or even far proximity and are evolutionarily connected.

A combination of different topics is also possible.

Why should you join this project? You have the unique opportunity to work on an urgent, important topic, together with the Robert Koch Institute. The outcome is a tool which will be used to help control the pandemic in Germany.

What we expect from you We are looking for up to 4 students enrolled in Data Engineering, IT Systems Engineering, Cyber Security or Digital Health with good programming skills (beyond introductory level courses for DH students), either in Python with focus in algorithmic design or in web development and systems administration. It is preferred that you have experience in some of the following technologies: Snakemake, MySQL, NGINX, Flask or Python Dash. You also need to sign a visiting student contract at the Robert Koch Institute in order to be allowed to work with the data.

Advisors

Prof. Dr. Bernhard Renard - bernhard.renard@hpi.de
Dr. Katharina Baum - katharina.baum@hpi.de
Fábio Miranda - fabio.malchermiranda@hpi.de
Alice Wittig - alice.wittig@hpi.de

Please don't hesitate to email us with any question or to set up a zoom meeting.

Project partners

Prof. Dr. Tobias Friedrich
Dr. Andreas Goebel
Karen Seidel
Marcus Pappik
Dr. Sarel Cohen

Dr. Stephan Fuchs
Dr. Martin Hölzer