Doctoral thesis project

Probabilistic approach for monitoring the proportions of SARS-CoV-2 variants in wastewater

February 14, 2022

1 Context

The emergence of SARS-CoV-2 variants complicates the management of the health crisis. The rapid local availability of the proportions of each variant could facilitate this management. Early detection of the emergence of new variants is also crucial.

Obépine is currently analysing wastewater from 200 wastewater treatment plants on a bi-weekly basis. This involves measuring the amount of total SARS-CoV-2 genome in these samples by Reverse-Transcriptase quantitative Polymerase Chain Reaction (RT-qPCR) [1].

RT-qPCR measurements specific to some of the variants also make it possible to monitor the proportions of these variants in the epidemic, but this method requires the conception of an RT-qPCR specific to each variant of interest, does not make it possible to detect the emergence of new variants and does not exploit transient mutations (non-specific to a variant but more present in some than in others).

In addition, some of the samples collected by *Obépine* have undergone full genetic sequencing for SARS-CoV-2. The aim here is to exploit these data.

Master 2 internship This project is a continuation of a Master 2 internship carried out between July and December 2021 at SUMMIT (*Sorbonne University Maison des Modélisations, Ingénieries et Technologies*) as part of a research collaboration with the *Obépine* Scientific Interest Group (GIS).

The internship focused on estimating the proportions of variants already identified in wastewater samples from the raw sequencing data (BAM files).

During the internship, a statistical model was proposed to solve the problem by making the best use of all the available data (co-occurrence of mutations on some of the reads, lack of reads for some parts of the genome, etc.). It is based on an unsupervised classification of the reads obtained from the PCR and takes into account the occurrence of sequencing errors at a fixed rate. An algorithm called Expectation-Maximization was derived and validated on simulated data and on real data from four *Obépine* stations in *Ile-de-France*.

The goal now is to extend and better adapt the proposed model and the associated inference algorithm.

2 Objectives

The Master 2 internship and the Ph.D. thesis are part of a research project that aims at developing a statistical tool that:

- accurately quantifies the presence of known variants,
- detects the emergence of a new variant,
- predicts the evolution of the mix of variants in time and space.

Detecting the emergence of a new variant requires the addition of an unknown element in the characteristics of the variants.

Predicting the evolution of the mix of variants over time requires the addition of a temporal dimension to the model, which can be done by adding an auto-regressive component affecting the monitored quantities. The addition of such a component will make the quantification of the mix of variants more robust and refined by exploiting the temporal dependence between samples. The same applies to the addition of a spatial component in the model. Approaches exploiting the phylogenetic tree of SARS-CoV-2 could also be explored.

Taking into account the variability of the error rate, which depends on the type of mutation and its location in the genome, will also be considered.

3 Profile of the Candidate

The PhD candidate should have a Master 2 in statistics or a related field. Knowledge in genetics and skills in bioinformatics or biostatistics would be an asset.

4 Involved researchers and entities

Grégory Nuel, CNRS Research Director at LPSM (Laboratory of Probability, Statistics and Modeling, UMR 8001) and director of the previous Master 2 internship, will be director of the thesis [1, 3].

Researchers from the ***Obépine* GIS** will participate in the follow-up of the project. In particular, **Vincent Maréchal**, Professor of Virology and Director of the UFR of Life Sciences at Sorbonne University and Director of the GIS Obépine, will be co-director of the thesis [7, 1, 4, 5, 8, 9, 6].

Research engineers from **SUMMIT**, specialists in applied statistics and artificial intelligence and who participated in the supervision of the previous Master 2 internship, will provide scientific and technical support to the PhD student [1, 3, 2]. In particular, **Marie Courbariaux**, PhD, will co-supervise the thesis.

References

[1] Nicolas Cluzel, Marie Courbariaux, Siyun Wang, Laurent Moulin, Sébastien Wurtzer, Isabelle Bertrand, Karine Laurent, Patrick Monfort, Christophe Gantzer, Soizick Le Guyader, Mickaël Boni, Jean-Marie Mouchel, Vincent Maréchal, Grégory Nuel, and Yvon Maday. A nationwide indicator to smooth and normalize heterogeneous sars-cov-2 rna data in wastewater. *Environment International*, 158:106998, 2022.

[2] Nicolas Cluzel, Amaury Lambert, Yvon Maday, Gabriel Turinici, and Antoine Danchin. Leçons biochimiques et statistiques de l'évolution du virus SARS-CoV-2 : nouveaux chemins pour combattre les virus. *Comptes Rendus. Biologies*, 343(2):177–209, 2020.

[3] Marie Courbariaux, Nicolas Cluzel, Siyun Wang, Vincent Maréchal, Laurent Moulin, Sébastien Wurtzer, Obépine Consortium, Jean-Marie Mouchel, Yvon Maday, and Grégory Nuel. A flexible smoother adapted to censored data with outliers and its application to sars-cov-2 monitoring in wastewater. *Frontiers in Applied Mathematics and Statistics*, 8, 2022.

[4] Grigoris T Gerotziafas, Mariella Catalano, Yiannis Theodorou, Patrick Van Dreden, Vincent Marechal, Alex C Spyropoulos, Charles Carter, Nusrat Jabeen, Job Harenberg, Ismail Elalamy, et al. The covid-19 pandemic and the need for an integrated and equitable approach: an international expert consensus paper. *Thrombosis and haemostasis*, 121(08):992–1007, 2021.

[5] Olivier Terrier, Sébastien Dilly, Andrés Pizzorno, Dominika Chalupska, Jana Humpolickova, Evžen Bouřa, Francis Berenbaum, Stéphane Quideau, Bruno Lina,

Bruno Fève, et al. Antiviral properties of the nsaid drug naproxen targeting the nucleoprotein of sars-cov-2 coronavirus. *Molecules*, 26(9):2593, 2021.

[6] S Wurtzer, P Waldman, M Levert, N Cluzel, JL Almayrac, C Charpentier, S Masnada, M Gillon-Ritz, JM Mouchel, Y Maday, et al. Sars-cov-2 genome quantification in wastewaters at regional and city scale allows precise monitoring of the whole outbreaks dynamics and variants spreading in the population. *Science of The Total Environment*, 810:152213, 2022.

[7] Sébastien Wurtzer, V Marechal, JM Mouchel, Yvon Maday, Remy Teyssou, E Richard, JL Almayrac, and L Moulin. Evaluation of lockdown effect on sars-cov-2 dynamics through viral genome quantification in waste water, greater paris, france, 5 march to 23 april 2020. *Eurosurveillance*, 25(50):2000776, 2020.

[8] Sébastien Wurtzer, Vincent Maréchal, Isabelle Bertrand, Mickaël Boni, Soizick Le Guyader, Laurent Moulin, Yvon Maday, Christophe Gantzer, and Jean-Marie Mouchel. Maladies infectieuses virales vues au travers des eaux usées. *Virologie*, 25(1):8–11, 2021.

[9] Sebastien Wurtzer, Prunelle Waldman, Audrey Ferrier-Rembert, Gaelle Frenois-Veyrat, Jean-Marie Mouchel, Mickael Boni, Yvon Maday, Vincent Marechal, Laurent Moulin, et al. Several forms of sars-cov-2 rna can be detected in wastewaters: implication for wastewater-based epidemiology and risk assessment. *Water Research*, 198:117183, 2021.