Unsupervised detection and fitness estimation of emerging SARS-CoV-2 variants: Application to wastewater samples (ANRS0160)

Alexandra Lefebvre*†[1,2], Vincent Maréchal[3], Arnaud Cloagen[4], Amaury Lambert[2,5], and Yvon Maday[1]

[1]Laboratoire Jacques-Louis Lions – Sorbonne Universite, Centre National de la Recherche Scientifique, Université Paris Cité – France

[2]Centre interdisciplinaire de recherche en biologie – Labex MemoLife, Collège de France, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique – France

[3]Biologie et thérapeutiques du cancer [CRSA] – Centre de Recherche Saint-Antoine – France

[4]Centre National de Recherche en Génomique Humaine – Institut de Biologie François JACOB – France

[5]Institut de biologie de l'ENS Paris – Département de Biologie - ENS-PSL, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique – France

Abstract

Repeated waves of emerging variants during the SARS-CoV-2 pandemics have highlighted the urge of collecting longitudinal genomic data and developing statistical methods based on time series analyses for detecting new threatening lineages and estimating their fitness early in time. Most models study the evolution of the prevalence of particular lineages over time and require a prior classification of sequences into lineages. Such process is prone to induce delays and biases. More recently, few authors studied the evolution of the prevalence of mutations over time with alternative clustering approaches, avoiding specific lineage classification. Most of the aforementioned methods are however either non parametric or unsuited to pooled data characterizing, for instance, wastewater (WW) samples. The pooled nature of WW data, with a mixture of fragmented and incomplete sequences associated with potentially several lineages and secreted by multiple infected individuals, involves specific statistical challenges. However the analysis of WW samples has recently been pointed out as a valuable complementary approach to clinical sample analysis (where one sample is associated to one viral sequence), as it is representative of the viral circulation at a population level. All infected individuals indeed participate to the sampling. In this context, we propose an alternative unsupervised method for clustering mutations according to their frequency trajectory over time and estimating group fitness from time series of pooled mutation prevalence data. Our model is a mixture of observed count data and latent group assignment and we use the expectation-maximization algorithm for model selection and parameter estimation. We apply our method to time series of SARS-CoV-2 sequencing data collected from wastewater treatment plants in France from October 2020 to April 2021 and we compare our results to supervised methods (that track specific mutations over time) and retrospective analyses. We show that our model agnostically group mutations in a consistent way with lineages B.1.160,

*Speaker

†Corresponding author: alexandra.lefebvre@math.cnrs.fr

Alpha, B.1.177 and Beta, with selection coefficient estimates per group in coherence with the viral dynamics in France reported by Nextstrain. Moreover, our method detected the Alpha variant as threatening as early as supervised methods with the noticeable difference that, since unsupervised, it does not require any prior information on the set of mutations.

Keywords: Time series analysis, Mixture model, EM algorithm, Clustering trajectories, Wastewater surveillance, Variant fitness.