Projet de thèse de doctorat

Approche probabiliste pour le suivi des proportions de variants de SARS-CoV-2 dans les eaux usées

14 février 2022

1 Contexte

L'apparition de variants de SARS-CoV-2 complique la gestion de la crise sanitaire. Disposer rapidement et localement des proportions de chaque variant pourrait faciliter cette gestion. Il est de plus crucial de détecter précocement l'émergence de nouveaux variants. Le suivi de SARS-CoV-2 dans les eaux usées constitue une opportunité pour obtenir ces informations.

Obépine analyse actuellement les eaux usées de 200 stations d'épuration de façon bi-hebdomadaire. Il s'agit de mesures de la quantité de génome de SARS-CoV-2 totale dans ces échantillons par *Reverse-Transcriptase quantitative Polymerase Chain Reaction* (RT-qPCR) [1].

Des mesures par RT-qPCR spécifiques à certains variants permettent de plus un suivi de la part de ces variants dans l'épidémie [4], mais cette méthode nécessite la mise en place d'une RT-qPCR spécifique à chaque variant d'intérêt, ne permet pas la détection de l'apparition de nouveaux variants et n'exploite pas les mutations passagères (non spécifiques à un variant mais plus présentes chez certains que chez d'autres).

Certains des échantillons prélevés par Obépine ont de plus fait l'objet d'un séquençage génétique complet (c'est-à-dire comportant plusieurs millions de *reads*) pour SARS-CoV-2, il s'agit ici d'exploiter ces données.

Stage de Master 2 Ce projet fait suite à un stage de Master 2 effectué entre juillet et décembre 2021 à SUMMIT (Sorbonne Université Maison des Modélisations, Ingénieries et Technologies) dans le cadre d'une collaboration de recherche avec le Groupement d'Intérêt Scientifique (GIS) Obépine.

Le stage s'est concentré sur l'estimation des proportions des variants déjà identifiés dans des échantillons d'eaux usées à partir des données de séquençage brutes (fichiers BAM).

Au cours du stage, un modèle statistique a été proposé pour résoudre le problème en exploitant au mieux la totalité des données disponibles (co-occurrence de mutations sur certains *reads*, absence de lecture pour certaines parties du génome,...). Il repose sur une classification non supervisée des fragments lus à l'issue de la PCR et prend en compte l'existence d'erreurs de séquençage à un taux fixe. Un algorithme *Expectation-Maximization* a pour cela été dérivé et validé sur des données simulées et sur des données réelles issues de quatre stations Franciliennes d'Obépine.

Il s'agirait à présent d'étendre et d'adapter le modèle proposé et l'algorithme d'inférence associé.

2 Objectifs

Le projet de recherche dans lequel s'inscrivent le stage de Master 2 et la thèse de doctorat a pour but de développer un outil statistique permettant :

— de quantifier de façon précise la présence des variants déjà connus,
— de détecter l'apparition d'un nouveau variant,
— de prédire l'évolution du mix des variants dans le temps et l'espace.

Détecter l'apparition d'un nouveau variant nécessite l'ajout d'une part d'inconnue dans les caractéristiques des variants.

Prédire l'évolution du mix des variants dans le temps nécessite l'ajout d'une dimension temporelle au modèle, ce qui peut se faire par l'ajout d'une composante auto-régressive affectant les quantités suivies. L'ajout d'une telle composante permettra de rendre plus robuste et d'affiner la quantification du mix de variants par l'exploitation de la dépendance temporelle entre les échantillons. Il en va de même de l'ajout d'une composante spatiale dans le modèle. Des approches exploitant l'arbre phylogénétique de SARS-CoV-2 pourront aussi être explorées.

D'autres extensions sont aussi à prévoir, telles que la prise en compte de la variabilité du taux d'erreur, selon le type de mutation et son emplacement dans le génome par exemple.

3 Profil du candidat

Le ou la candidat.e devra être titulaire d'un Master 2 en statistiques ou d'un domaine proche. Des connaissances en génétique et des compétences en bioinformatique ou en biostatistiques seraient un atout.

4 Entités et chercheurs impliqués

Les données à traiter (directement issues des séquençages) sont très massives et en très grande dimension, ce qui rend nécessaire le recours à des méthodes de calcul particulièrement performantes et inscrit bien ce projet dans les thèmes de recherche de l'**Institut des Sciences du Calcul et des Données (ISCD)**. Il s'inscrirait de plus dans un de ses domaines d'application, celui de la médecine.

Le GIS Obépine a pour objectif de poursuivre le développement des activités de recherche appliquées dans le domaine de l'utilisation des eaux usées pour le suivi du SARS-CoV-2 et de ses variants, ainsi que d'autres pathogènes ou agents chimiques.
Des chercheurs du GIS Obépine participeront au suivi de la thèse.

SUMMIT est une unité de service de Sorbonne Université. SUMMIT a pour mission de développer l'activité partenariale entre Sorbonne Université et le monde des entreprise et de contribuer à la promotion des savoir-faire des unités et laboratoires de Sorbonne Université.
Des ingénieurs de recherche de SUMMIT, spécialistes en statistiques appliquées et en intelligence artificielle et ayant participé à l'encadrement du précédent stage de master 2, fourniront un appui scientifique et technique au doctorant [1, 3, 2]. En particulier, **Marie Courbariaux**, PhD, co-encadrera la la thèse.

Le Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001) est une unité de recherche de Sorbonne Université, de l'Université de Paris et du CNRS. **Grégory Nuel**, Directeur de Recherche CNRS au LPSM et directeur du précédent stage de master 2, assurera la direction de la thèse [1, 3].

Références

[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, Grégory Nuel, Isabelle Bertrand, Mickaēl Boni, Christophe Gantzer, Soizick F. Le Guyader, Yvon Maday, Vincent Maréchal, Jean-Marie Mouchel, Laurent Moulin, Rémy

Teyssou, and Sébastien Wurtzer. 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] S Wurtzer, P Waldman, M Levert, Jm Mouchel, O Gorgé, M Boni, Y Maday, V Marechal, L Moulin, et al. Monitoring the propagation of sars cov2 variants by tracking identified mutation in wastewater using specific rt-qpcr. *medRxiv*, 2021.