Proposition de stage de Master en Bioinformatique

VirHostX : *caractérisation large échelle du virome des animaux, vecteurs et environnements réservoirs de pathogènes humains à partir des données de séquençage de l'archive internationale SRA.*

Lieu : Université Claude Bernard Lyon1, PRABI-AMSB (DOUA, Villeurbanne) – HCL GENEPII (Lyon, Croix rousse)

Durée : 6 mois, dès février 2025

Co-Encadrement : **Vincent Navratil** (PRABI, Univ Lyon 1) vincent.navratil@univ-lyon1.fr ; **Laurence Josset** (CIRI, HCL) laurence.josset@chu-lyon.fr ; **Jocelyn Turpin** (IVPC, INRAE) jocelyn.turpin@univ-lyon1.fr ; **Oldrich Navratil** (EVS, Univ Lyon 2) oldrich.navratil@univ-lyon2.fr

Mots clés : *bioinformatique, données massives, métagénomique, virome, relations virus/hôtes*, OneHealth

1. Contexte scientifique et local du projet

La grande majorité des virus responsables de maladies infectieuses humaines sont zoonotiques, c'est à dire présentant une origine animale [1,2]. Comprendre les facteurs moléculaires, écologiques et biogéographiques complexes responsables du passage de virus animaux à des individus humains (*« spillover »*) - ou inversement le retour de virus humains à des animaux ou de manière plus générale dans l'environnement (« *spillback* ») - est une tâche qui se révèle extrêmement complexe [3]. Elle nécessite tout d'abord de dresser un catalogue complet des associations virus/hôtes présentes dans les populations humaines et animales, ainsi que d'une annotation approfondie de leur distribution géospatiale et temporelle dans les différents environnements. A l'heure actuelle, seul un nombre limité de virus et d'associations virus/hôte a été recensé dans les bases de données et la majorité des relations connues restent biaisées autour de virus humains et/ou de virus causant des maladies chez ce dernier et les animaux domestiqués d'intérêts. Une grande partie des virus et de leur spectre d'hôtes reste donc inconnue par les scientifiques, limitant notre compréhension de l'émergence de nouveaux pathogènes à l'échelle planétaire. L'essor récent de la métagénomique ouvre de nombreuses perspectives dans la caractérisation systématique du virome – c'est à dire l'ensemble des génome viraux - des organismes animaux domestiques ou sauvages, des insectes vecteurs mais également de leurs écosystèmes en lien avec les activités humaines [3]. Dans ce contexte, l'archive internationale SRA (*Short Read Archive*) donne accès ouvertement à plusieurs millions d'échantillons de séquençage métagénomique (ADN) et métatranscriptomique (ARN) associé à l'état de l'art des publications dans le domaine. Cette ressource peut être exploitée à large échelle pour identifier plus finement la présence de virus à l'aide d'algorithmes bioinformatiques et d'infrastructures hautement performantes [4].

Le stage de Master s'inscrit dans le cadre plus large du projet Virome@tlas, lauréat de l'AAP Structurant ShapeMed@Lyon. Le projet Virome@tlas vise à la structuration à Lyon d'un consortium inter-disciplinaire autour d'une plateforme numérique pour l'exploration et la surveillance géographique à large échelle de la virosphère à partir de l'archive SRA.

2. Objectifs scientifiques du projet

L'objectif du projet de Master est de caractériser de manière exhaustive par une approche bioinformatique l'ensemble des relations virus/hôtes, mieux comprendre leur structuration à large échelle en lien avec des variables biogéographiques et potentiellement de mettre en évidence de nouveaux événements de *spillover/spillback* à partir de l'exploitation de millions d'échantillons biologiques SRA et leurs assignations taxonomiques SRA-STAT [5] associées.

Après une étude de l'état de l'art des bases de données et méthodes bioinformatiques disponibles, la première partie du projet consistera à construire un réseau de co-occurence des virus avec leurs hôtes animaux à partir des données d'assignation taxonomique fournies par l'outil SRA-STAT. Cette étude sera

menée sur l'ensemble des organismes animaux domestiques et d'élevage connus mais également sur les animaux sauvages mais également des insectes vecteurs associés à l'homme suspectés de jouer un rôle dans l'émergence de nouveaux pathogènes. Une analyse exploratoire sera menée pour quantifier la présence d'une sélection de virus dans les différents environnements échantillonnés disponibles dans SRA (hôpitaux, eaux usées, lacs, rivières, estuaires). La deuxième partie visera 1) à une analyse descriptive de la structure du réseau de co-occurence afin d'identifier les principales tendances (largeur du spectre d'hôte des virus, distance entre les hôtes basée sur la composition de leur virome, associations positives ou négatives de virus pouvant signer des co-infections ou compétitions, facteurs biogéographiques influençant, etc...) 2) à une évaluation de la qualité des données en comparant le réseau généré à l'état actuel des connaissances (*VirHostRange*[1] - communication personnelle) et aux dernières publications utilisant des méthodologie bioinformatiques différentes [6,7]**.**

Le·la candidat·e travaillera en collaboration étroite avec les responsables scientifiques du stage et les collaborateurs du projet Virome@tlas pour développer ses compétences à l'interface de la virologie, de la bioinformatique et de la géographie. Il.elle pourra bénéficier d'un support technique de la plateforme PRABI-amsb (http://amsb.prabi.fr), GENEPII et OMEAA pour être guidé au mieux dans les choix méthodologiques et leur mise en œuvre (bonnes pratiques de développement, infrastructure de calcul). Les bases de données, les jeux de données, l'infrastructure informatique et les outils bioinformatiques nécessaires à la bonne réalisation des tâches listées ci-dessus seront disponibles dès le début du projet. L'ensemble des développements seront réalisés sous des Notebook Jupyter hébergés sur des machines virtuelles du cloud PRABI/LBBE et versionnés sous le dépôt git du CCIN2P3 afin de faciliter la reproductibilité des analyses.

Ce master ouvre la possibilité de continuer en Thèse dans le cadre de l'école doctorale E2M2 ou l'école doctorale EID@Lyon.

3. Compétences recherchées

Nous recherchons un·e candidat·e de niveau bac **+3 ou + 4 en bioinformatique**. Il·elle devra être capable d'interagir avec différents chercheurs et ingénieurs du projet Virome@tlas pour développer et mettre en place des méthodes innovantes en sciences de la vie.

Compétences requises :

- Python, R, algorithmes bioinformatiques, théorie des graphes, *machine learning*

- Versioning sous git

- Fortes capacités relationnelles et d'organisation, autonomie

- Capacités rédactionnelles, esprit d'analyse et de synthèse

4. Informations pratiques

Le stagiaire sera hébergé au PRABI-AMSB et sur la plateforme GENEPII afin de favoriser les échanges interdisciplinaires. Ce stage bénéficiera d'une gratification de 6 mois à partir de février 2025 (552 euros/mois). Pour tout renseignement ou candidature, merci d'adresser un CV et une lettre de motivation le plus rapidement possible pour démarrage le 01/02/2025 à l'attention de : Vincent Navratil (PRABI-AMSB, Univ Lyon 1) vincent.navratil@univ-lyon1.fr ; Laurence Josset (CIRI, HCL) laurence.josset@chu-lyon.fr ; Jocelyn Turpin (IVPC, INRAE) jocelyn.turpin@univ-lyon1.fr ; Oldrich Navratil (EVS, Univ Lyon 2) oldrich.navratil@univ-lyon2.fr

5. Bibliographie

1. Woolhouse, M., Scott, F., Hudson, Z., Howey, R. & Chase-Topping, M. Human viruses: discovery and emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 2864–2871 (2012).
2. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).
3. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
4. Edgar, R. C. *et al.* Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147 (2022).
5. Katz, K. S. *et al.* STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biology* **22**, 270 (2021).
6. Chen, Y.-M. *et al.* Host traits shape virome composition and virus transmission in wild small mammals. *Cell* **186**, 4662-4675.e12 (2023).
7. He, W.-T. *et al.* Virome characterization of game animals in China reveals a spectrum of emerging pathogens. *Cell* **185**, 1117-1129.e8 (2022).

[1] https://virhostnet.prabi.fr/d3-virhostnet-layout/example