université de BORDEAUX

LaBRI

Étude du virus H5N1
Partie génétique

Paul GERVAUD Jade GOUPIL Ceny KETANI
Djemilatou OUANDAOGO Théo MARCHAL
Noé MATHIEUX

Projets de Conception et Architecture Logiciel

Encadrant :

Mme. Beurton-Aimar

Table des matières

Introduction

La grippe aviaire est une maladie virale, touchant en particulier les oiseaux, notamment les oiseaux d'élevages. Cette maladie est causée par des virus appartenant à la famille des virus de la grippe de type A. Dans certains cas, ils existent une infection possible entre oiseaux et hommes. [1]

Notre étude se portera plus spécifiquement sur un sous-type spécifique du virus de la grippe aviaire, H5N1. Ce virus a été détecté en premier lieu à Hong Kong, en 1997, causant notamment la mort de 6 personnes. On le prénomme ainsi puisqu'il présente à sa surface deux protéines spécifiques, une Hémagglutinine de type 5, permettant au virus de se fixer et de pénétrer la cellule, et une Neuraminidase de type 1, qui libère le virus après réplication. [2]

Il est primordiale pour les biologistes d'étudier ce virus, puisqu'il a une capacité de mutation très importante, et est également capable d'échanger ses gènes avec des virus d'espèces différentes. À terme et sans surveillance de l'Organisation Mondiale de la Santé (OMS), nous pourrions nous retrouver face à une pandémie grippale. Actuellement, ce sont les oiseaux des régions Asiatiques, Européennes et Africaines qui sont le plus touchés. [3]

Notre objectif ici sera de centraliser un ensemble d'outils sur une page web, permettant à un scientifique d'effectuer une analyse génétique rapide et de qualité.

Dans un premier temps, nous allons revenir sur les différentes connaissances du virus, sur ce que nous pouvons apporter et mettre en place, et avec quels outils. Nous verrons ensuite les besoins de ce projet, et l'organisation allouée pour mener à bien sa réalisation.

Chapitre 1

Analyse du sujet

1.1 Contexte

Actuellement, l'analyse génétique des virus et plus particulièrement H5N1 s'effectue de cette manière : Une première partie de manipulation, avec l'extraction de l'ARN virale, la conversion en ADN complémentaire via une enzyme nommée transcriptase inverse, le séquençage de cette dernière (notamment avec PacBio ou Illumina), et l'assemblage des reads pour obtenir une séquence du génome viral. Ensuite, viens la partie de l'analyse, avec l'annotation de la séquence, nécessaire à l'identification des gènes et protéines, la détection de mutants et de variants comparés à une séquence de référence, et enfin l'analyse phylogénétique, comparant notre séquence à des séquences d'une base de donnée, nous permettant de comprendre son origine et son évolution.

1.2 État de l'art

1.2.1 Connaissance actuelle

Les premières observations du virus H5N1 débutent en 1997, dans la région de Guangdong, il s'agit alors du premier génotype de la grippe aviaire qui est en mesure d'infecter les humains. Malgré sa très faible transmission, il n'en reste pas moins très dangereux, avec une létalité très elevée, notamment dû au fait que sa réplication se développe au niveau des cellules du poumon. De plus, ce virus est un réel risque pour les élevages de volailles, il possède en effet une virulence plus grande que les autres virus, mais aussi en dehors des périodes de circulation/migration. [4]

Actuellement, on retrouve près de 240 séquences génétiques de H5N1, provenant des oiseaux sauvages, du bétail ou des vaches laitières. Le partage de toutes ces données permet de mieux comprendre les mécanismes de propagation du virus, notamment l'identification des mutations qui pourraient permettre le passage du virus entre les espèces animales, et potentiellement aux humains.

Pour faire face à l'ensemble de ces virus grippaux, la plateforme internationale GISAID (Global Initiative on Sharing All Influenza Data) a été créee. Elle a été lancée en 2008 pour permettre aux chercheurs, gouvernements et organisations de santé publique de partager rapidement les séquences génétiques des virus de la grippe, facilitant ainsi la surveillance mondiale des épidémies et des pandémies potentielles. Les chercheurs peuvent télécharger et comparer les séquences de nouvelles souches avec celles déjà connues pour identifier des mutations importantes, telles que

celles qui influencent la virulence, la transmissibilité, ou la résistance aux vaccins. C'est une plateforme importante pour suivre les relations évolutives entre les différentes souches du virus, notamment à l'aide d'arbres phylogénétiques.

1.2.2 Développement de la page web interactive

RShiny est un framework open-source basé sur le langage R, conçu pour le développement d'applications web interactives. Très apprécié dans le domaine de la bioinformatique, il permet notamment de créer des interfaces conviviales où des utilisateurs non spécialistes, tels que les biologistes possédant les séquences du virus, peuvent interagir avec des algorithmes complexes ou des bases de données. Il intègre pleinement les capacités de traitement des données de R, facilitant ainsi la visualisation et l'analyse exploratoire de données biologiques, telles que les séquences génomiques du virus H5N1. Grâce à son extensibilité, il permet l'intégration d'analyses bioinformatiques, tout en permettant l'importation et l'exportation de fichiers au format standardisé (FASTA ou autre). [5]

Une seconde possibilité est le langage HTML (HyperText Markup Language). C'est un langage de balisage standard utilisé pour structurer et afficher du contenu sur le web. il détermine l'organisation des éléments sur une page (texte, images) et est souvent combiné avec JavaScript (pour les fonctionnalités interactives) et/ou CSS (pour le style). Pour créer des applications web interactives dans des projets de bioinformatique, telles que l'analyse de séquences du virus H5N1, HTML doit être couplé avec des scripts de back-end en utilisant des technologies telles que Python, R, ou des frameworks (Django notamment) pour traiter et analyser les données. L'ajout d'interactions dynamiques nécessite également des bibliothèques JavaScript. Bien que HTML offre une grande flexibilité pour concevoir des interfaces web sur mesure, il exige ainsi un travail appronfondi en développement web, notamment en matière d'architecture de type front-end et back-end. [6]

Ainsi, RShiny et HTML offrent deux approches distinctes pour le développement d'applications web interactives. RShiny est conçu pour les scientifiques et les analystes de données qui souhaitent créer des applications rapidement, en intégrant directement les capacités analytiques du langage R dans une interface utilisateur simple. En revanche, HTML fournit une base beaucoup plus flexible pour créer des interfaces web sur mesure, mais cette flexibilité vient au prix d'une plus grande complexité. Les développeurs doivent combiner HTML avec des technologies comme CSS, JavaScript, et des frameworks back-end. Nous avons donc fait le choix d'utiliser RShiny lors ce projet.

1.2.3 Algorithme global de comparaison

Un algorithme global de comparaison est conçu pour la recherche de séquences homologues, le but étant d'obtenir un score significatif sur une longueur proche de la longueur des deux séquences. Nous comparons donc notre séquence en entrée, et recherchons dans le jeu de données celle qui s'en rapproche le plus, pour tenter une identification globale de la séquence. Puisque nous travaillons ici sur le virus H5N1, nous allons seulement construire un petit jeu de données, voire

une seule séquence entière pour vérifier que la séquence rentrée correspond bien au virus. Pour cela, nous allons utiliser un algorithme de Needleman Wunsch, qui utilise une matrice de score pour comparer les séquences.

Concrètement, nous utiliserons le logiciel open source Emboss. Il existe des packages R permettant de réaliser un alignement global entre deux séquences, comme Biostrings, mais leurs utilisations seront limités par la taille des séquences. Il existe en effet un lien entre la longueur des séquences, et la perte de vitesse et d'efficacité, pouvant poser problème lors de l'utilisation de séquences entières. L'alternative pour cela serait d'utiliser des outils, comme le module créé à partir du logiciel Emboss, ce que nous allons utiliser dans ce projet.

1.2.4 Algorithme local de comparaison

Un algorithme d'alignement local est conçu pour trouver des zones de similitude entre des séquences de protéines ou d'ADN, qu'elles soient homologues ou non. Il cherche le meilleur chevauchement entre deux séquences afin d'identifier des domaines conservés ou des segments partagés entre différentes séquences, ce qui permet d'identifier des gènes ou des régions fonctionnelles. L'algorithme BLAST (Basic Local Alignment Search Tool) est l'un des plus utilisés dans ce domaine. Il compare des séquences à des bases de données en utilisant une matrice de similarité (basée sur l'algorithme de Smith-Waterman), et évalue la signification statistique des alignements. BLAST produit des résultats sous forme d'alignements par paires (requête vs séquences de la base de données), et fournit des informations comme l'identifiant de la séquence, le nom du gène, un score d'alignement, la e-value (probabilité que l'alignement soit dû au hasard), le pourcentage d'identité et la longueur de l'alignement.

Il existe deux moyens d'utiliser l'algorithme BLAST sur sa propre plateforme : en l'exécutant en local à l'aide de Blast+, ou en l'exécutant en ligne à l'aide d'une API. La première méthode présente l'avantage d'être rapide dans ses requêtes, mais nécessite une grande capacité de stockage puisque l'utilisateur doit créer/stocker son propre jeu de données. Il existe des packages tel que rBLAST, implémentant directement l'utilisation de BLAST+ depuis R. La seconde méthode présente l'avantage d'utiliser la banque de données en ligne de BLAST, mais en contre-partie limitera les requêtes (pas plus d'une requête toutes les 10 secondes). Les API utilisées sont de type REST (Representational State Transfer), la plus complète étant l'API Entrez. C'est cette API que nous allons implémenter, afin de pouvoir travailler avec une base de donnée assez conséquente.

1.2.5 Phylogénie

La phylogénie moléculaire a pour objectif de reconstruire les relations de parenté entre des séquences de nucléotides ou d'acides aminés. Généralement, les arbres phylogénétiques sont construits pour différentes espèces, permettant ainsi de retracer l'histoire évolutive. Cependant, dans le cadre de notre plateforme dédiée à l'étude du virus H5N1, un arbre des espèces pour une séquence donnée donnerait toujours le même résultat. Ainsi, il est donc plus pertinent de créer un arbre phylogénétique des variants du virus H5N1. Cela permettrait de situer la séquence d'entrée parmi les variants existants et d'identifier celui auquel elle est le plus proche. Notre jeu

de données devra ainsi contenir différentes séquences connues de variants du H5N1, disponibles dans des bases de données, présentant divers degrés de divergence.

Pour construire un arbre phylogénétique, on part du principe que les séquences choisies sont homologues, ce qui sera le cas ici. L'étape suivante consiste alors à effectuer un alignement multiple global. Il existe divers outils pour réaliser ce travail, tel que MUSCLE, ou T-COFFEE, chacun ayant ses spécificités en termes de rapidité, qualité de l'alignement, et proximité des séquences. Dans notre cas, le nombre de séquences à aligner ne sera pas trop élevé, et la qualité d'alignement sera importante, car les séquences sont relativement proches (contrairement à des séquences inter-espèces).

Parmi les outils les plus adaptés, Clustal O et MUSCLE se distinguent, étant à la fois libres de droits et bien adaptés à ce type d'analyse. Pour les implémenter, nous utiliserons des packages R tels que le package muscle ou le package msa, qui intègre à la fois MUSCLE et Clustal.

L'étape finale est la construction de l'arbre phylogénétique. Différentes méthodes existent, les principales étant : la méthode de distance (calcul des distances évolutives entre toutes les paires de séquences de l'alignement), maximum de vraisemblance (probabilise entièrement le processus évolutif, et trouve quel scénario a la plus forte probabilité de donner les séquences analysées), et la méthode de parcimonie (création d'un arbre évolutif en choisissant la théorie contenant le moins de phénomènes évolutifs possibles). Chacune présente des avantages en fonction des caractéristiques des séquences analysées et du type d'arbre à construire. Nous avons donc deux types de méthodes différentes : de distance (une valeur par séquence) ou de caractère (probabilité de chaque mutation par séquence). Pour avoir une approche plus précise, nous choisirons une méthode de caractère. Les deux principales, la méthode de parcimonie et le maximum de vraisemblance, ont des temps de calcul équivalent, à la différence que la méthode de parcimonie ne produit pas de longueurs de branches précises. Nous utiliserons donc la méthode de vraisemblance, à l'aide notamment du package R phangorn. Il implémente de plus une fonction de bootstrap (technique de rééchantillonnage), qui permet d'estimer la fiabilité d'un arbre.

Enfin, pour obtenir une représentation de l'arbre, le serveur de l'EMBL propose l'outil Itol, ainsi que itol.tookit, mettant à disposition différentes API permettant d'inclure cet outil dans une pipeline. Une autre possibilité serait de construire directement l'arbre avec des packages déjà fournis comme ggtree ou phytools, afin d'avoir notamment plus de contrôle sur l'affichage.

1.2.6 Intégration d'un génome de référence

Pour comparer la séquence d'entrée ou explorer les variations génétiques du virus H5N1, nous envisageons d'utiliser D3 Genome Browser (D3GB). Cet outil interactif de navigation génomique est particulièrement adapté à notre projet, car il est libre de droit et compatible avec des formats standards tels que GenBank et FASTA, ce qui facilitera l'intégration des données. D3GB permet de visualiser les génomes de manière intuitive et offre des fonctionnalités d'analyse interactives, ce qui le rend idéal pour l'exploration des mutations et variations génétiques du virus. De plus, son intégration avec des fonctions en R s'aligne parfaitement avec les technologies de notre projet, notamment avec l'utilisation de RShiny pour développer l'application web. D3GB représente ainsi une solution robuste et bien documentée, capable de répondre aux exigences de notre

projet collaboratif.

1.3 Présentation des données

Comme vu auparavant, nous cherchons à analyser et à comparer des séquences du virus H5N1. Les fichiers d'entrée sont donc des fichiers au format FASTA. Ce type de format est largement utilisé en bioinformatique, notamment pour stocker des séquences génomiques, que ce soit de l'ADN, de l'ARN ou des protéines. De plus, il est supporté par de nombreux outils d'analyse bioinformatique (comme BLAST ou Clustal Omega), et est compatible avec de nombreux algorithmes et pipelines. Enfin, il sera facilement intégrable dans notre plateforme RShiny via des packages comme Biostrings ou seqinr.

```
>CC.M2_9_ID317| Inga_chartacea
AAACTGCATGCATTTGCCATGACTAGCATTTG
```

Figure 1.1: Exemple de format Fasta.
En jaune : Symbole d'initiation, en orange : Descripteur de la séquence et en gris : Séquence d'ADN.

1.4 Sortie des données

Suivant la demande de l'utilisateur, plusieurs formats de sortie sont possibles. Pour l'alignement global, nous récupérerons un fichier texte contenant l'ensemble de statistiques sur l'alignement réalisé, tel que le score global d'alignement, le pourcentage d'identité (pourcentage de correspondances exactes entre les deux séquences), le nombre et la position des gaps (insertions/délétions) et la longueur des régions alignées. Pour l'alignement local, le fichier de sortie pourra être de différents formats, comme par exemple XML, JSON ou TXT. Enfin, pour la phylogénie, la sortie des fichiers produit se fera au format PHYLIP ou FASTA.

1.5 Analyse des besoins

1.5.1 Fonctionnels

L'objectif de notre projet est de créer une plateforme web interactive dédiée à l'analyse génétique du virus H5N1. Cette interface se devra d'être intuitive pour des biologistes, afin qu'il soit simple de charger et de traiter des séquences virales rapidement et efficacement. Cette plateforme devra offrir une variété d'outils bioinformatiques, chacun accessible via une interface simple, pour faciliter l'exploration et la comparaison des données génomiques complexes. À travers cette plateforme, les utilisateurs pourront réaliser des analyses d'alignement global, local et phylogénétique.

L'interface intégrera notamment des fonctionnalités de visualisation avancées pour la navigation génomique, avec la possibilité de comparer directement les séquences importées aux séquences de référence disponibles en base de données. Ces séquences pourront être analysées à l'aide d'algorithmes de comparaison, tels que Needleman-Wunsch pour l'alignement global ou BLAST pour l'alignement local, permettant aux utilisateurs de rechercher des séquences homologues ou de détecter des mutations. De plus, le système proposera des fonctionnalités pour l'analyse phylogénétique des séquences virales, en construisant des arbres phylogénétiques pour étudier l'évolution et la propagation du virus. L'ensemble des résultats obtenus pourra être exporté dans des formats standards, tels que FASTA, PHYLIP ou encore JSON, afin de faciliter leur exploitation et visualisation.

1.5.2 Non Fonctionnels

Du point de vue des exigences non fonctionnelles, il est impératif que l'application soit robuste, afin de traiter des volumes de données potentiellement conséquents sans compromettre la rapidité et/ou la fiabilité des résultats. nous utiliserons une architecture modulaire à l'aide de Rshiny, avec des composants indépendants pour les différentes fonctionnalités (alignements, analyses phylogénétiques, visualisations). Cela permettra d'intégrer facilement de nouvelles fonctionnalités ou d'adapter les algorithmes utilisés selon les demandes et besoins.

1.6 Organisation

Pour ce qui est de l'organisation du projet, nous le diviserons en plusieurs parties : La création de la page web, l'implémentation de l'entrée, la visualisation et la récupération d'un fichier de sortie. Cette structure nous permettra de nous concentrer sur chaque aspect du développement, assurant ainsi une intégration fluide et efficace des différentes fonctionnalités.

La première étape sera celle de la création de notre plateforme web avec Rshiny. Nous allons devoir trouver un ensemble de pluging ou de scripts nécessaire à l'implémentation de chaque outils présentés auparavant. Cela inclura la conception de l'interface utilisateur, la structuration des pages et l'optimisation de l'expérience utilisateur afin que l'ensemble soit accessible et intuitif

La seconde étape sera la possibilité d'entrer une séquence au format FASTA. Nous veillerons à ce que le système puisse effectuer toutes les analyses requises à partir de cette séquence.

Ensuite, la troisième étape consistera à mettre en place un système de visualisation clair et interactif. L'utilisateur devra pouvoir observer les résultats des analyses rapidement et facilement.

Enfin, nous nous concentrerons sur les types de données que l'utilisateur pourra obtenir en sortie après les différentes analyses. Cela inclura la possibilité d'exporter les résultats sous forme de fichiers, tels que FASTA ou TXT, afin que les utilisateurs puissent les consulter ultérieurement ou les intégrer dans d'autres analyses.

Références

[1] «Grippe aviaire». Institut Pasteur, 6 octobre 2015, https://www.pasteur.fr/fr/centre-medical/fiches-maladies/grippe-aviaire.

[2] «Grippe aviaire : symptômes, traitement et prévention». Institut Pasteur de Lille, https://pasteur-lille.fr/centre-prevention-sante-longevite/vaccins-et-voyages/grippe-aviaire/.

[3] Grippe aviaire | MesVaccins. https://www.mesvaccins.net/web/diseases/50-grippe-aviaire.

[4] Höfle, Ursula. «A Brief History of H5N1, an Avian Influenza Virus Devastating to Birds and Low Risk to Humans». SMC España, https://sciencemediacentre.es/en/brief-history-h5n1-avian-influenza-virus-devastating-birds-and-low-risk-humans.

[5] «Formation Shiny». R-atique, http://perso.ens-lyon.fr/lise.vaudor/tuto-shiny/.

[6] Notions de Base En HTML - Apprendre Le Développement Web | MDN. 27 juillet 2024, https://developer.mozilla.org/fr/docs/Learn/Getting_started_with_the_web/HTML_basics.