UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES POUR L'IDENTIFICATION ET L'ANALYSE DES SIGNATURES GÉNOMIQUES DANS LE CONTEXTE DE LA CLASSIFICATION DES SÉQUENCES BIOLOGIQUES VIRALES

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE

PAR

DYLAN LEBATTEUX

AOÛT 2025

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma profonde reconnaissance au Professeur Abdoulaye Baniré Diallo pour sa direction subtile et perspicace. Ses conseils avisés et nos échanges enrichissants ont été déterminants dans la réussite de ce projet. C'est une personne d'une grande valeur avec laquelle j'espère poursuivre de futures collaborations.

Ma gratitude va également au Dr Isabelle Boucoiran et au Professeur Soren Gantt, qui ont accepté de conseiller et de financer une partie de ma recherche. Leur expertise en biologie virale et leurs précieuses connaissances cliniques ont apporté une dimension essentielle à mon projet de doctorat.

Je suis particulièrement reconnaissant envers le Professeur Hugo Soudeyns pour le temps consacré à la relecture minutieuse de mes travaux et la pertinence de ses retours, qui ont significativement enrichi la qualité de cette thèse. Ma reconnaissance va également à Florian Corso pour nos collaborations fructueuses qui ont grandement contribué à approfondir l'interprétation biologique de mes résultats.

Je tiens à exprimer ma reconnaissance à l'ensemble des membres du laboratoire de l'UQAM, avec qui j'ai eu le privilège de partager ce parcours. Je remercie tout particulièrement Mohamed Amine Remita pour ses conseils pertinents et la rigueur scientifique qu'il a su me transmettre.

Sur un plan plus personnel, je souhaite exprimer toute ma gratitude à ma femme pour sa patience et son soutien inébranlable, particulièrement durant les moments difficiles de ce parcours doctoral. Je remercie également mes parents pour leurs encouragements constants. Une mention spéciale à mon père, qui m'a transmis sa persévérance et son sens de l'accomplissement, des qualités qui se sont révélées précieuses dans la réalisation de cette thèse.

Enfin, je remercie sincèrement tous ceux qui ont contribué, de près ou de loin, à la concrétisation de ce projet doctoral.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

ADN Acide désoxyribonucléique (Deoxyribonucleic Acid).

ARN Acide ribonucléique (Ribonucleic Acid).

DENV Virus de la dengue (Dengue Virus).

GA Algorithme génétique (Genetic Algorithm).

GPCR Récepteurs couplés aux protéines G (G Protein-Coupled Receptors).

HBV Virus de l'hépatite B (Hepatitis B Virus).

HCV Virus de l'hépatite C (Hepatitis C Virus).

HCMV Cytomégalovirus humain (Human Cytomegalovirus).

HDV Virus de l'hépatite D (Hepatitis D Virus).

HIV-1 Virus de l'immunodéficience humaine de type 1 (Human Immunodeficiency Virus-1).

HTS Séquençage à haut débit (High-Throughput Sequencing).

ICTV Comité international de taxonomie des virus (International Committee on Taxonomy of Viruses).

MSA Alignement multiple de séquences (Multiple Sequence Alignment).

NCBI Centre national pour l'information biotechnologique (National Center for Biotechnology Information).

SARS-CoV-2 Coronavirus 2 du syndrome respiratoire aigu sévère (Severe Acute Respiratory Syndrome Coronavirus 2).

SVM Machine à vecteurs de support (Support Vector Machine).

RÉSUMÉ

L'avènement du séquençage à haut débit a révolutionné l'analyse des génomes viraux, entraînant une croissance exponentielle des données génomiques disponibles. Cette abondance de données exige un système de classification taxonomique robuste des organismes viraux, indispensable pour approfondir notre compréhension de leur biologie, notamment face à l'émergence de pandémies (SARS-CoV-2) et d'infections persistantes (HIV-1, HCMV). Bien que l'ICTV actualise régulièrement la taxonomie officielle, la classification des nouvelles souches virales demeure complexe en raison de leur diversité génétique exceptionnelle, de leur évolution rapide et de leur hétérogénéité structurale. Les approches traditionnelles de classification, fondées sur des propriétés phénotypiques ou des alignements de séquences, montrent leurs limites face à la forte divergence des séquences, aux réarrangements génomiques fréquents, et au volume croissant de données à analyser. Dans ce contexte, le développement d'approches taxonomiques basées sur des signatures génomiques intrinsèques, comme les k-mers, se présente comme une alternative prometteuse. L'analyse des variations de ces signatures peut fournir des informations essentielles sur la différenciation des sous-espèces virales et leurs caractéristiques pathogéniques. Cette thèse présente trois approches méthodologiques complémentaires pour l'identification, l'analyse et la hiérarchisation des signatures génomiques dans un contexte de classification des séquences virales. KEVOLVE combine un algorithme génétique avec des méthodes d'apprentissage automatique pour identifier des sous-ensembles minimaux de k-mers discriminants et construire des modèles prédictifs. KANALYZER exploite les alignements par paires pour extraire et caractériser les variations de ces k-mers discriminants, en fournissant des informations sur leur localisation génomique et leur impact sur les séquences protéiques. Le *KmerSignificance Score* (KSS) propose un système de notation qui évalue la pertinence des k-mers/variations. Il intègre leur pouvoir discriminant via un cadre d'évaluation supervisé et leur signification biologique, à travers l'importance fonctionnelle des protéines affectées et l'impact des mutations sur leurs propriétés physicochimiques. Ces méthodes ont été validées sur des données simulées et appliquées à des virus à forte variabilité nucléotidique et cliniquement pertinents (SARS-CoV-2, HIV-1, HCMV). Pour le HIV-1, les modèles prédictifs de KEVOLVE surpassent les outils de référence pour la classification des sous-types, tout en restant robustes face à des taux de mutation élevés. Sur le SARS-CoV-2, KEVOLVE a démontré son efficacité en surpassant les méthodes statistiques de référence dans l'extraction de motifs hautement discriminants, permettant la construction de modèles de prédiction précis pour des centaines de milliers de séquences. L'analyse par KANALYZER des motifs extraits a révélé leur association avec des mutations connues influençant l'infectiosité et la pathogénicité virales. L'application du KSS a confirmé la pertinence des critères mutationnels définis pour les génotypes des gènes UL55, UL73 et US28 du HCMV, tout en permettant l'annotation de jeux de séquences à plus grande échelle. Couplé à une analyse par clustering hiérarchique, le KSS a permis d'identifier cinq génotypes distincts du gène UL33, chacun caractérisé par un profil mutationnel unique. Les mutations significatives, localisées dans les domaines N-terminal et extracellulaires, suggèrent un rôle dans la modulation de la glycosylation du récepteur, son activité constitutive et l'échappement immunitaire. Ces avancées favorisent une meilleure compréhension de la diversité virale et de ses implications fonctionnelles. Les travaux futurs viseront à étendre ces approches à la classification globale des espèces virales et à les diffuser via une plateforme bioinformatique dédiée.

Mots clés : Classification des séquences virales, apprentissage automatique, algorithme génétique, signatures génomiques, k-mers, mutations.

INTRODUCTION

0.1 Mise en contexte et motivation

0.1.1 Diversité des génomes viraux à l'ère du séquençage haut débit

L'avènement des technologies de Séquençage à haut débit (High-Throughput Sequencing) (HTS) a profondément transformé la collecte, l'assemblage et l'analyse des génomes viraux, conduisant à un accroissement considérable des bases de données issues d'échantillons cliniques, environnementaux, vétérinaires et végétaux (Simmonds et Aiewsakun, 2018). À titre d'illustration, la plateforme GenBank du Centre national pour l'information biotechnologique (National Center for Biotechnology Information) (NCBI) (Benson *et al.*, 2018) recense aujourd'hui plus de trois millions de génomes viraux complets, un nombre en croissance constante.

Cette abondance de données génomiques exige une compréhension approfondie de la biologie des virus, d'autant plus cruciale face à l'émergence de pandémies et d'infections persistantes. Par exemple, le Coronavirus 2 du syndrome respiratoire aigu sévère (Severe Acute Respiratory Syndrome Coronavirus 2) (SARS-CoV-2), responsable de la pandémie de COVID-19, a démontré une capacité d'évolution rapide, conduisant à l'émergence de variants influençant sa transmissibilité et sa virulence (Wu *et al.*, 2020). De même, le Virus de l'immunodéficience humaine de type 1 (Human Immunodeficiency Virus-1) (HIV-1) se caractérise par une variabilité génétique importante, complexifiant la mise au point de vaccins et nécessitant une surveillance continue pour l'adaptation des traitements antirétroviraux (Hemelaar, 2013). Par ailleurs, le Cytomégalovirus humain (Human Cytomegalovirus) (HCMV) constitue une cause majeure d'infections congénitales et opportunistes chez les individus immunodéprimés ; sa complexité génomique entrave la caractérisation fine de ses souches et complique le développement de traitements ciblés (Griffiths et Reeves, 2021). Dans ce contexte, l'étude minutieuse des génomes viraux, leur classification ainsi que la mise en évidence de signatures génomiques s'avèrent indispensables pour le diagnostic, la surveillance épidémiologique, le développement de thérapies et l'élaboration de stratégies préventives (Slezak *et al.*, 2020).

Les virus se distinguent par leur diversité remarquable, qui surpasse celle de la plupart des autres organismes. Cette diversité se manifeste à plusieurs niveaux : la nature du matériel génétique (ADN ou ARN), le type de brin (simple ou double), l'organisation et la structure des gènes, la composition protéique, ainsi que la morphologie et la taille de la capside (King *et al.*, 2012). Les modes de réplication virale, les interactions avec l'hôte (de l'antagonisme au mutualisme) et l'éventuelle segmentation du génome (réparti en

plusieurs fragments indispensables à l'infection) varient également de manière significative d'un virus à l'autre (Simmonds *et al.*, 2017; Simmonds et Aiewsakun, 2018). Cette hétérogénéité se reflète notamment dans la grande variabilité de la taille des génomes viraux : certains, comme le Virus de l'hépatite D (Hepatitis D Virus) (HDV), ne comptent que 1 700 nucléotides et deux gènes (Pascarella et Negro, 2011), tandis que d'autres, qualifiés de « virus géants », comme *Pandoravirus salinus*, atteignent plus de 2,5 millions de paires de bases et contiennent plus de 2 500 gènes (Abergel *et al.*, 2015). Dans cette mosaïque d'organismes, l'élaboration et l'application de méthodes de classification et d'analyse robustes se révèlent essentielles pour appréhender la complexité et l'hétérogénéité du paysage génomique viral.

0.1.2 Défis et perspectives dans la classification taxonomique des virus

La classification hiérarchique des virus est actuellement établie par l'Comité international de taxonomie des virus (International Committee on Taxonomy of Viruses) (ICTV) (Siddell *et al.*, 2023), qui repose sur le travail de plus d'une centaine de comités d'experts. Dans sa dernière mise à jour, l'ICTV recense 314 familles, 3 522 genres et 14 690 espèces virales (`https://ictv.global/taxonomy`). Bien que cette classification soit rigoureuse et régulièrement actualisée, les analyses métagénomiques mettent constamment en évidence de nouvelles séquences virales non répertoriées dans la taxonomie officielle (Matthews, 2018). Par ailleurs, malgré ces mises à jour fréquentes, plusieurs difficultés persistent dans le système taxonomique actuel (Caetano-Anollés *et al.*, 2023; Simmonds, 2024) :

— **Absence de standards universels et complexité de la nomenclature** : L'espèce constitue le rang taxonomique le plus bas actuellement reconnu par l'ICTV, et aucun niveau de sous-espèce n'est encore formellement établi. De plus, les seuils génétiques distinguant les différents rangs (famille, genre, espèce) varient notablement selon les groupes viraux et demeurent souvent arbitraires. L'absence de normes harmonisées complexifie la nomenclature, d'autant plus que la taxonomie virale fait intervenir des expertises variées (virologie, bioinformatique, génomique), rendant le consensus difficile à atteindre.

— **Dynamique évolutive accélérée** : Les virus, particulièrement ceux à ARN, évoluent à un rythme soutenu (Duffy, 2018). Les substitutions, recombinaisons et réassortiments fréquents compromettent la stabilité des classifications et compliquent la représentation fidèle d'une réalité évolutive en perpétuel changement.

— **Complexité biologique et pluralité des formes virales** : Les virus présentent une grande variété de structures, de stratégies de réplication et de modes d'interaction avec l'hôte. Parallèlement, le nombre de virus découverts croît sans cesse, rendant la classification virale d'autant plus exigeante et sujette à des ajustements continus.

— **Variabilité de la gamme d'hôtes et connaissances partielles** : Certains virus infectent un spectre d'hôtes extrêmement restreint, tandis que d'autres peuvent affecter de multiples organismes, complexifiant toute tentative de classification fondée sur l'hôte. De plus, nos connaissances demeurent incomplètes concernant les modalités de transmission, les interactions hôte-virus et les mécanismes évolutifs, ce qui limite la précision et la portée du système taxonomique actuel.

Face à ces contraintes, les méthodes de classification et d'analyse traditionnelles, basées essentiellement sur des propriétés phénotypiques (manifestations cliniques, gamme d'hôtes, distribution géographique) ou des alignements de séquences, ne suffisent plus pour appréhender la diversité virale contemporaine (Simmonds et Aiewsakun, 2018). Compte tenu de l'accumulation exponentielle de données, de la complexité génétique, des réarrangements génomiques fréquents et de l'évolution rapide des virus, il devient essentiel de proposer des méthodes de classification s'appuyant sur des caractéristiques intrinsèques (signatures génomiques) directement dérivées des séquences nucléotidiques, tout en maintenant une cohérence avec les critères historiques (De la Fuente *et al.*, 2023; Adriaenssens *et al.*, 2023; Mayne *et al.*, 2024; Simmonds, 2024).

Dans ce cadre, les approches indépendantes de l'alignement de séquences et les signatures génomiques telles que les fréquences de k-mers offrent une alternative prometteuse. Ces signatures couplées à diverses analyses statistiques, calculs de distances et approches d'apprentissage automatique, permettent d'évaluer les similarités et différences entre génomes, d'estimer les relations évolutives et d'inférer les parentés taxonomiques. Deux principes fondamentaux étayent cette démarche (Gusfield, 1997) :

1. **La structure est liée à la fonction** : Une forte similarité de séquences implique souvent une conservation fonctionnelle et structurale. Toutefois, l'existence de convergences évolutives, où des fonctions analogues émergent de séquences initialement divergentes, illustre la complexité de cette relation.

2. **L'origine commune des séquences biologiques** : Toutes les séquences biologiques dérivent d'un ancêtre commun. Leurs trajectoires évolutives, marquées par des mutations ponctuelles, des délétions ou des insertions, peuvent être retracées par analyses comparatives, permettant ainsi de reconstituer l'histoire évolutive des virus et d'établir leurs relations phylogénétiques.

0.2 Structure de la thèse

Cette thèse s'articule autour de la conception et de la validation de nouvelles approches bio-informatiques pour l'identification et l'analyse des signatures génomiques dans le contexte de la classification des séquences biologiques virales. Le manuscrit comprend ce chapitre introductif, suivi de sept chapitres principaux et d'une conclusion. Les chapitres progressent logiquement des concepts fondamentaux aux développements méthodologiques, en passant par des applications concrètes sur différents systèmes viraux d'importance clinique :

— **Chapitre 1** (*Concepts fondamentaux*) : expose les notions essentielles de biologie, virologie, bio-informatique et d'apprentissage automatique nécessaires à la compréhension des contributions. Le niveau de détail est adapté à leur importance dans nos travaux.

— **Chapitre 2** (*Objectif de recherche et revue de littérature*) : présente en détail les objectifs et hypothèses de recherche, ainsi qu'une analyse critique de la littérature scientifique portant sur les méthodes de classification et d'analyse des séquences virales.

— **Chapitre 3** (*Méthode développée*) : présente **KEVOLVE**, une approche combinant algorithmes génétiques et apprentissage automatique pour l'identification de sous-ensembles minimaux de *k*-mers discriminants. Ces signatures génomiques sont intégrées dans des modèles prédictifs basés sur des ensembles de SVMs. L'évaluation des performances des modèles inclut des tests avec insertion de bruit dans les séquences et l'utilisation de différents segments génomiques à travers une étude comparative avec les outils de référence pour la prédiction des sous-types du HIV-1.

— **Chapitre 4** (*Méthode développée*) : introduit **KANALYZER**, un algorithme combinant alignements par paires et calcul parallèle pour identifier et caractériser les variations des *k*-mers discriminants et leurs informations associées (localisation, fréquence, mutations d'acides aminés dans les régions

codantes). L'évaluation de **KANALYZER** a été réalisée sur des données simulées (en faisant varier la longueur des séquences, la taille des *k*-mers discriminants et le pourcentage de variations synthétiques) ainsi que sur des jeux de données virales réelles (HBV, HIV-1, SARS-CoV-2, DENV et HCV).

— **Chapitre 5** (*Cas d'application*) : applique **KEVOLVE** et **KANALYZER** à un large ensemble de séquences du SARS-CoV-2. Les performances de ces approches sont comparées à des méthodes statistiques de référence pour l'identification de motifs discriminants et la construction de modèles prédictifs. Les résultats démontrent que les modèles prédictifs basés sur les *k*-mers identifiés par **KEVOLVE** offrent de meilleures performances de prédiction. Les variations et informations identifiées par **KANALYZER** quant à elles montrent que les *k*-mers identifiés par nos approches correspondent majoritairement à des régions fonctionnelles ou des mutations d'intérêt validées par la littérature.

— **Chapitre 6** (*Méthode développée*) : présente le développement du ***KmerSignificance Score* (KSS)**, un système de notation heuristique évaluant la pertinence des *k*-mers discriminants et de leurs variations au sein des sous-espèces virales. Ce système intègre à la fois la capacité discriminative des *k*-mers, évaluée par un cadre d'apprentissage automatique, et leur significativité biologique, basée sur (i) l'importance fonctionnelle de leur localisation génomique dans les régions codantes et (ii) l'impact potentiel des substitutions d'acides aminés sur la dynamique protéique. La validation du **KSS** a été effectuée en confrontant ses résultats aux données expérimentales et à la littérature scientifique sur trois systèmes viraux d'importance clinique distincte : SARS-CoV-2, HIV-1 et HCMV. Notre analyse démontre une forte corrélation avec les mutations fonctionnelles décrites dans la littérature tout en identifiant de nouvelles variations potentiellement significatives pour le gène *UL55* du HCMV.

— **Chapitre 7** (*Cas d'application*) : se concentre sur l'analyse du gène *UL33* (un récepteur couplé aux protéines G) du HCMV. L'étude inclut une analyse comparative des GPCRs viraux, confirmant la forte variabilité d'UL33. La reconstruction de variants dans une cohorte mère-enfant kényane révèle, par analyse phylogénétique, une transmission de la souche maternelle majoritaire vers l'enfant, associée à certains génotypes d'*UL33*. Cette analyse a été étendue à l'ensemble des séquences *UL33* disponibles sur GenBank et à une seconde cohorte en Ouganda, auxquelles nous avons appliqué le **KSS**. Les résultats mettent en évidence des mutations discriminantes significatives pour la définition de génotypes d'*UL33*.

— **Conclusion** : synthétise les contributions majeures de la thèse, discute des limitations actuelles et propose des perspectives pour les recherches futures.

0.3 Protocole de validation

Les méthodes développées dans cette thèse font l'objet d'évaluations rigoureuses visant à garantir leur robustesse et leur fiabilité. Ces évaluations sont menées sur des jeux de données variés (données simulées et réelles), dont les caractéristiques et paramètres sont soigneusement documentés, tout comme les configurations des algorithmes utilisés dans nos travaux.

Pour évaluer les performances, nous utilisons diverses métriques de classification standard telles que le taux de vrais positifs, le taux de faux positifs, la précision, le rappel, la F-mesure et les matrices de confusion lorsque leur calcul est pertinent et réalisable. Chacune de nos propositions est comparée aux approches de l'état de l'art sur des jeux de données standardisés, permettant d'identifier les améliorations en termes de performance, de robustesse et d'efficacité computationnelle. Une attention particulière est accordée à l'optimisation des méthodes développées afin d'en assurer la scalabilité et l'applicabilité dans des contextes réels, où la quantité de données génomiques peut être considérable.

Dans un souci de reproductibilité et de transparence, nous mettons à disposition de la communauté scientifique une documentation rigoureuse de toutes les étapes, de la collecte des données à l'évaluation des modèles, ainsi que le code source des implémentations, les jeux de données utilisés et les résultats détaillés sur des plateformes en libre accès. Cette démarche vise à renforcer la crédibilité du projet et à stimuler la collaboration scientifique.

0.4 Contributions

Au cours de cette thèse, trois nouvelles approches méthodologiques ont été développées et deux applications concrètes mises en œuvre, conduisant à cinq articles scientifiques publiés ou en préparation dans des conférences et journaux internationaux de bioinformatique. Ces travaux ont également été présentés lors de communications orales et par affiches dans différentes conférences. Les contributions principales se déclinent comme suit :

— **Chapitre 3**

  — Lebatteux, D., Remita, A. M., et Diallo, A. B. (2020). *KEVOLVE: a combination of genetic algorithm and ensemble method to classify viruses with variability*. *In 2020 NeurIPS Workshop : Learning Meaningful Representations of Life.* (Article publié et accepté pour présentation par affiche)

  — Lebatteux, D., et Diallo, A. B. (2021). *Combining a genetic algorithm and ensemble method to improve the classification of viruses*. *In 2021 IEEE International Conference On Bioinformatics And Biomedicine (BIBM) (pp. 688-693).* (Article publié et accepté pour présentation orale)

— **Chapitre 4**

  — Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., et Diallo, A. B. (2022). *Algorithme d'identification des variations de signatures génomiques au sein des séquences virales*. Dans *27èmes Rencontres de la Société Francophone de Classification (SFC).* (Résumé publié et accepté pour présentation orale)

  — Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., et Diallo, A. B. (2022). *KANALYZER: a method to identify variations of discriminative k-mers in genomic sequences*. *In 2022 IEEE International Conference On Bioinformatics And Biomedicine (BIBM) (pp. 757-762).* (Article publié et accepté pour présentation orale)

— **Chapitre 5**

  — Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., et Diallo, A. B. (2024). *Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures*. *In Plos one, 2024, vol. 19, no 1, p. e0296627.* (Article publié)

— **Chapitre 6**

  — Lebatteux, D., Corso, F., Soudeyns, H., Boucoiran, I., Gantt, S., et Diallo, A. B. (2025). *Assessing Discriminative k-mers and Variations in Viral Subspecies with the KmerSignificance Score*. (Article en préparation)

— **Chapitre 7**

  — Corso, F., Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S., et Diallo, A. B. (2025). *Characterization of the genetic variation in the gene encoding the GPCR UL33 : a tool for HCMV pathogenesis*. (Article en préparation, co-premiers auteurs)

CHAPITRE 1
CONCEPTS FONDAMENTAUX

Ce chapitre présente les concepts fondamentaux nécessaires à la compréhension des contributions et des approches développées dans cette thèse, en complément des notions introduites dans le chapitre précédent. Notre recherche se situe à la convergence de quatre disciplines majeures : la biologie, la virologie, l'informatique et l'apprentissage automatique. Cette intersection disciplinaire est essentielle pour aborder la problématique centrale de notre travail : l'identification et l'analyse des signatures génomiques pour la classification des séquences virales. L'intégration de ces différentes perspectives scientifiques nous permet de développer des approches innovantes en bioinformatique, combinant les connaissances biologiques des virus avec des méthodes avancées d'analyse computationnelle.

1.1 Définitions de biologie et de virologie

1.1.1 L'atome et la molécule

L'atome est l'unité constitutive de base de toute matière dans l'Univers et sur Terre. Il est composé d'un noyau central constitué de protons et de neutrons, entouré par un nuage d'électrons Flowers *et al.* (2019). Par exemple, l'atome d'hydrogène est le plus simple des atomes, constitué d'un seul proton dans son noyau et d'un électron en orbite autour de celui-ci. Une molécule est un ensemble d'atomes liés entre eux par des liaisons chimiques, formant une unité stable et distincte Flowers *et al.* (2019). Par exemple, la molécule d'eau est composée de deux atomes d'hydrogène liés à un atome d'oxygène (H_2O).

1.1.2 Le nucléotide

Un nucléotide, dont la structure générale est illustrée dans la figure 1.1, est une molécule organique qui constitue l'unité de base des acides nucléiques (ADN et ARN). Il est composé de trois éléments essentiels Chaffey (2003) :

1. Un sucre à cinq atomes de carbone (pentose), qui est le désoxyribose pour l'ADN et le ribose pour l'ARN.
2. Un groupe phosphate.
3. Une base azotée, qui peut être l'adénine (A), la cytosine (C), la guanine (G) ou la thymine (T) pour

l'ADN. Dans l'ARN, l'uracile (U) remplace la thymine.

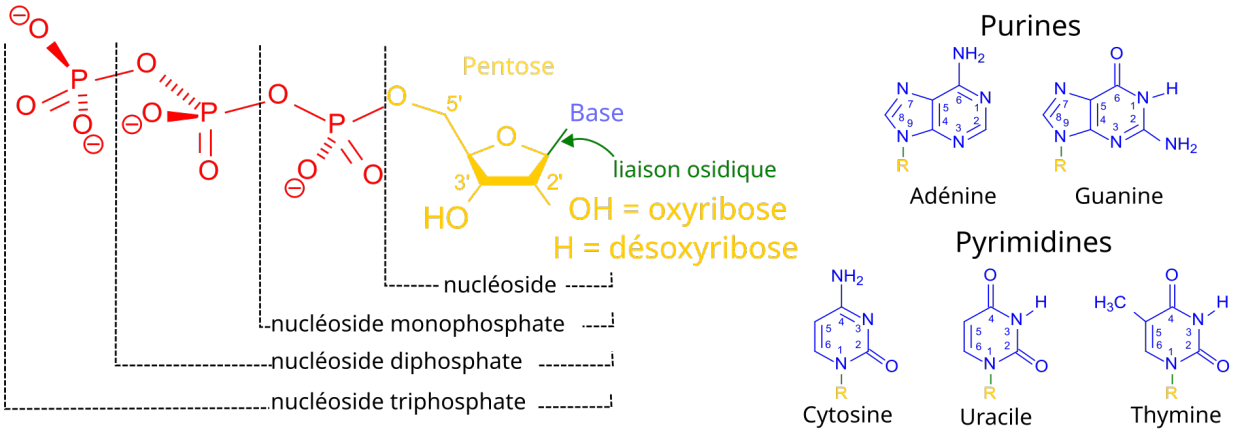


Figure 1.1 – Structure générale d'un nucléotide (Sjef, 2022, Wikimedia Commons, domaine public).

1.1.3 Les acides nucléiques : ADN et ARN

Les acides nucléiques, l'ADN (acide désoxyribonucléique) et l'ARN (acide ribonucléique), sont des macromolécules constituées de chaînes de nucléotides qui servent de support à l'information génétique chez tous les êtres vivants et les virus. Chez les organismes cellulaires, l'ADN, principalement localisé dans le noyau des cellules eucaryotes, contient les instructions génétiques nécessaires au développement et au fonctionnement des organismes. L'ARN, quant à lui, intervient dans divers processus biologiques, notamment la transcription et la traduction de l'information génétique, permettant ainsi la synthèse des protéines.

La structure de ces molécules diffère : l'ADN adopte une double hélice composée de deux brins complémentaires, tandis que l'ARN se présente généralement sous la forme d'une chaîne simple. Chez les virus, on distingue deux catégories : les virus à ADN, tels que le virus de l'hépatite B, qui utilisent l'ADN comme matériel génétique, et les virus à ARN, comme le HIV ou le virus de la grippe, qui utilisent l'ARN pour leur réplication.

1.1.4 La structure primaire des acides nucléiques

La structure primaire d'un acide nucléique correspond à la séquence linéaire des nucléotides qui composent la molécule, de son début à sa fin. Cette séquence est essentielle, car elle définit la composition exacte de la molécule d'ADN ou d'ARN, contient l'information nécessaire à la formation des structures secondaires et

tertiaires, et détermine la fonction biologique de la molécule.

Dans le contexte viral, la structure primaire revêt une importance particulière. Elle guide le processus de réplication virale, et ses mutations peuvent conduire à l'émergence de nouvelles variantes virales. De plus, cette structure influence directement la virulence et la capacité d'échappement des virus au système immunitaire.

L'analyse de la structure primaire permet d'identifier les variations de séquences, telles que les mutations, insertions ou délétions, de repérer les régions fonctionnelles importantes, comme les gènes ou les éléments régulateurs, et de mettre en évidence les similarités entre différentes séquences.

1.1.5 Les acides aminés, les codons et les protéines

Les acides aminés constituent les unités de base des protéines. Le code génétique universel comprend 20 acides aminés standard. Chaque acide aminé est codé par un codon, une séquence de trois nucléotides consécutifs dans l'ARN messager (ARNm). Le code génétique est redondant : plusieurs codons différents peuvent coder pour le même acide aminé, ce qui permet une certaine tolérance aux mutations silencieuses.

Les protéines sont des macromolécules constituées de chaînes d'acides aminés liés par des liaisons peptidiques. Elles assurent des fonctions essentielles dans les processus biologiques, notamment la catalyse enzymatique, le transport de molécules, la signalisation cellulaire et le maintien de la structure cellulaire.

La fonction d'une protéine dépend de sa structure tridimensionnelle complexe, qui s'établit progressivement selon quatre niveaux d'organisation. La structure primaire correspond à la séquence linéaire d'acides aminés, déterminée lors de la traduction de l'ARN messager. Les interactions locales entre acides aminés proches conduisent à la formation de la structure secondaire, caractérisée par des motifs réguliers comme les hélices alpha et les feuillets bêta. La structure tertiaire résulte du repliement global de la chaîne peptidique, stabilisé par diverses interactions entre acides aminés (liaisons hydrogène, ponts disulfures, interactions hydrophobes). Enfin, la structure quaternaire se forme par l'assemblage de plusieurs chaînes polypeptidiques en complexes fonctionnels.

La figure 1.2 illustre, dans sa partie supérieure, le dogme central de la biologie moléculaire décrivant le flux unidirectionnel de l'information génétique (ADN $\rightarrow$ ARN $\rightarrow$ Protéine), et dans sa partie inférieure, les quatre

niveaux d'organisation structurale des protéines.

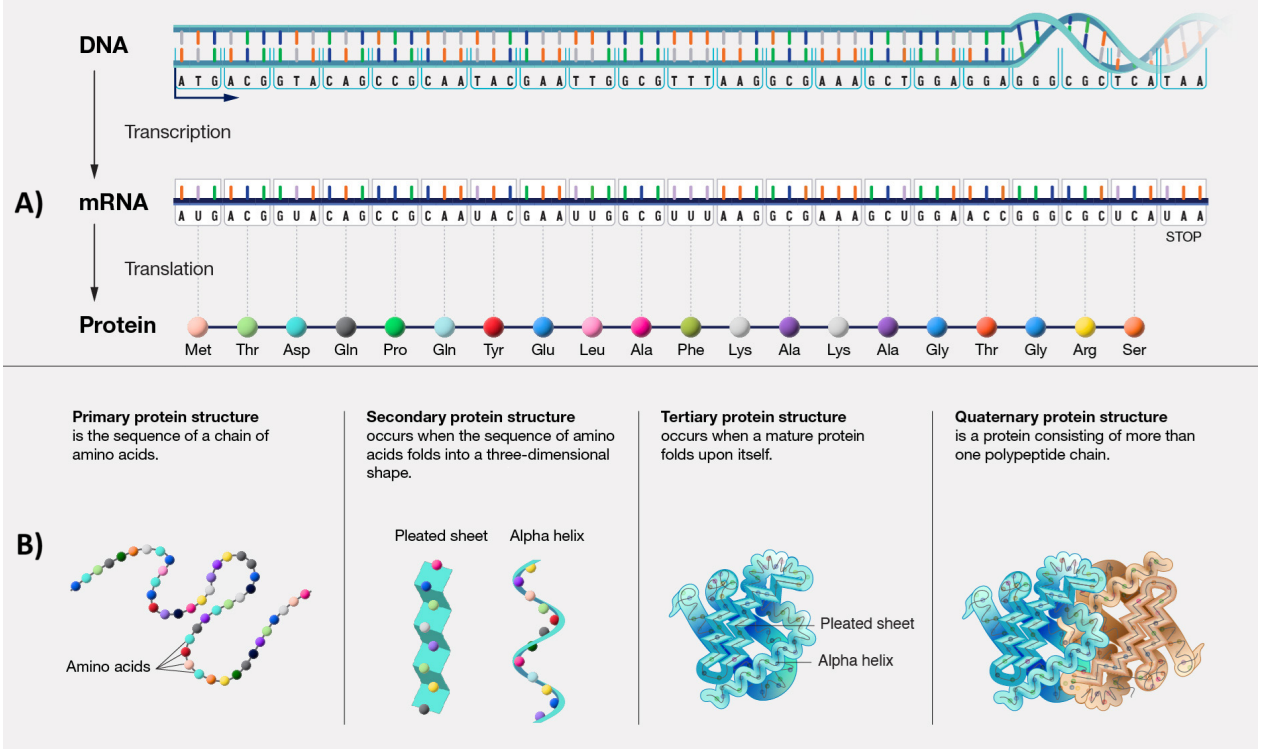


Figure 1.2 – Concepts fondamentaux en biologie moléculaire et structure des protéines.

La partie A illustre la théorie fondamentale de la biologie moléculaire : le flux de l'information génétique de l'ADN vers l'ARN, puis vers les protéines. La partie B présente les différents niveaux de structure des protéines. Source : `https://www.genome.gov/genetics-glossary`.

1.1.6 Les mutations ponctuelles

Une mutation ponctuelle est une modification d'une seule base nucléotidique dans une séquence d'ADN ou d'ARN. Comme illustré dans la figure 1.3, il existe trois types principaux de mutations ponctuelles :

- **Substitution** : remplacement d'une base par une autre, avec trois effets possibles :
  - **Mutation silencieuse** : le même acide aminé est produit.
  - **Mutation faux-sens** : production d'un acide aminé différent.
  - **Mutation non-sens** : formation d'un codon stop (UAA, UAG, UGA).
- **Insertion** : ajout d'une base dans la séquence.
- **Délétion** : suppression d'une base de la séquence.

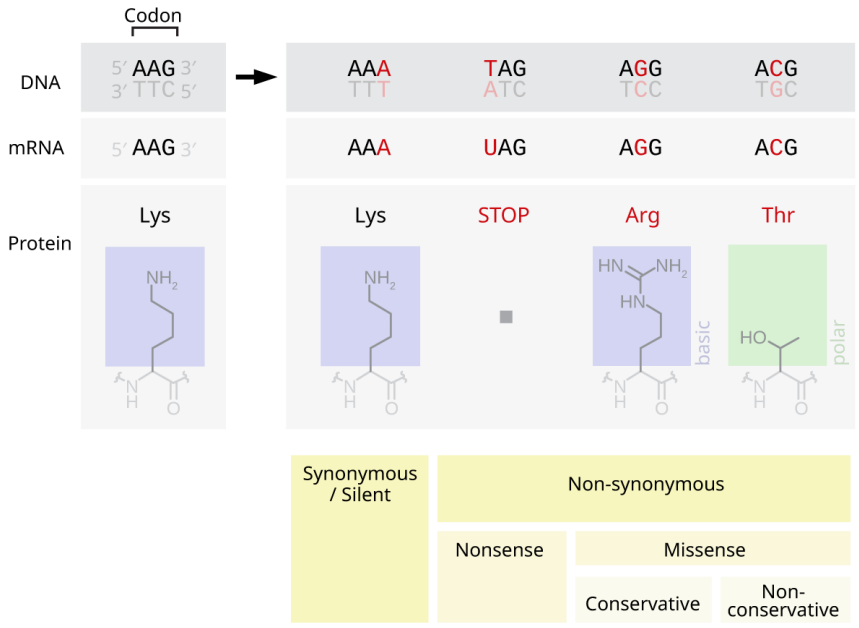


Figure 1.3 – Mutations ponctuelles et leurs impacts sur la séquence protéique (Jonsta247, 2022, Wikimedia Commons, domaine public).

Les insertions et délétions peuvent provoquer un décalage du cadre de lecture (*frameshift mutation*), modifiant la lecture des codons.

Par exemple :

Séquence originale :

```
ATG GCC ATT
```

Après insertion d'une base (A) :

```
ATG AGC CAT T...
```

Ce décalage modifie tous les codons suivants, altérant potentiellement la fonction de la protéine résultante.

1.1.7 Les variants génétiques

Les variants génétiques désignent des différences dans la séquence d'ADN pouvant survenir entre les individus d'une même population. Ces variations incluent des modifications allant d'un simple changement d'un nucléotide à des altérations impliquant plusieurs nucléotides, voire des segments chromosomiques entiers. Ces variants sont à la base de la diversité génétique et jouent un rôle fondamental dans l'évolution, la santé et les fonctions biologiques des organismes.

1.1.8 Homologie

L'homologie est un concept fondamental en biologie moléculaire qui décrit une relation évolutive entre des séquences biologiques. Deux séquences sont considérées comme homologues lorsqu'elles dérivent d'un ancêtre commun, partageant souvent des similarités en termes de séquence, de structure et de fonction. On distingue plusieurs types de relations homologues selon leur origine évolutive Fitch (1970) :

- **Orthologues** : Séquences issues d'un événement de spéciation, conservant généralement la même fonction dans différentes espèces.
- **Paralogues** : Séquences résultant d'une duplication génique, pouvant évoluer vers des fonctions distinctes au sein d'une même espèce.
- **Xénologues** : Séquences acquises par transfert horizontal de gènes entre espèces, souvent via des mécanismes tels que la conjugaison, la transduction ou la transformation.

Cette diversification des séquences homologues s'opère principalement via différents types de mutations. Parmi celles-ci, les mutations ponctuelles constituent l'unité fondamentale, modifiant un ou plusieurs nucléotides dans la séquence d'ADN ou d'ARN, ce qui peut entraîner des variations au niveau des protéines codées.

1.1.9 Les virus

Les virus sont des agents infectieux nanométriques qui parasitent les cellules des organismes vivants pour assurer leur réplication. Leur structure de base se compose d'un génome d'acide nucléique, qui peut être de l'ADN ou de l'ARN, entouré d'une capside protéique protectrice Howley *et al.* (2023). Certains virus possèdent également une enveloppe lipidique contenant des protéines virales.

Selon la définition de l'ICTV (2020), *un virus est un élément génétique mobile qui code au minimum pour une protéine majeure de sa capside, permettant l'encapsidation de son acide nucléique* Koonin *et al.* (2021).

Le cycle viral comprend six étapes principales Howley *et al.* (2023) :

1. **Attachement :** Les protéines virales de surface se lient aux récepteurs spécifiques de la membrane cellulaire hôte.
2. **Pénétration :** Le virus entre dans la cellule par endocytose ou fusion membranaire.
3. **Décapsidation :** La capside virale se désassemble dans le cytoplasme, libérant le matériel génétique viral.
4. **Réplication :** Le virus détourne la machinerie cellulaire (ribosomes, enzymes) pour répliquer son génome et synthétiser ses protéines.
5. **Assemblage :** Les génomes viraux et les protéines nouvellement synthétisés s'assemblent en virions fonctionnels par encapsidation.
6. **Libération :** Les virions quittent la cellule par lyse cellulaire ou bourgeonnement, permettant la propagation de l'infection.

1.1.10 La classification des virus

La classification des virus organise les virus dans un système taxonomique hiérarchique, permettant leur identification et leur comparaison systématique King *et al.* (2012).

Les principaux critères de classification virale sont :

- **Nature du matériel génétique** : ADN ou ARN, simple ou double brin.
- **Morphologie** : forme, taille et symétrie de la capside.
- **Structure génomique** : organisation et composition du génome.
- **Structure de la capside** : composition protéique et présence d'enveloppe.
- **Mode de réplication** : stratégie de multiplication virale.
- **Gamme d'hôtes** : types d'organismes infectés.
- **Pathogénicité** : capacité à causer des maladies.
- **Expression génique** : mécanismes de production des protéines.
- **Similarité génétique** : degré de ressemblance des séquences.

Les comparaisons de séquences constituent aujourd'hui le critère principal pour établir les relations taxonomiques entre virus Simmonds *et al.* (2017).

1.1.10.1 Le rôle de l'ICTV

Le Comité international de taxonomie des virus (ICTV) établit la classification officielle et la nomenclature des virus. Son système hiérarchique de classification est illustré dans la figure 1.4.

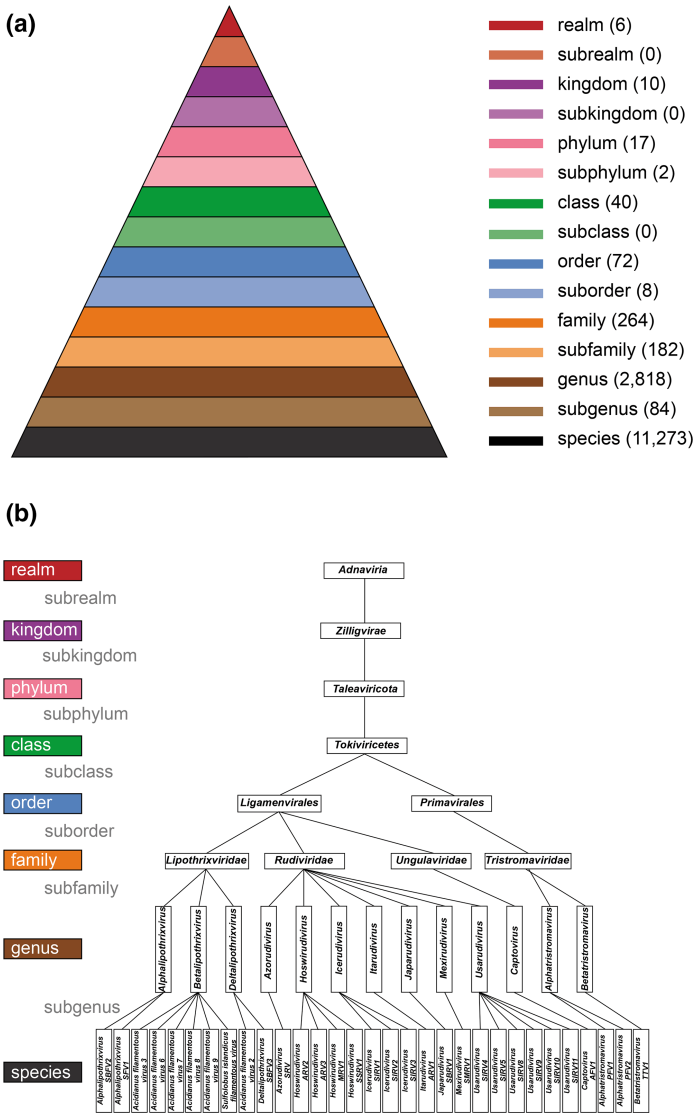


Figure 1.4 – Hiérarchie des rangs taxonomiques viraux de l'ICTV

La figure, extraite de Siddell *et al.* (2023), illustre dans la partie (a) la structure taxonomique complète des rangs taxonomiques viraux définis par l'ICTV, comprenant 15 niveaux avec le nombre de taxons indiqué entre parenthèses. La partie (b) présente un exemple de classification détaillée pour le règne *Adnaviria* Krupovic *et al.* (2021).

Cette classification repose sur une hiérarchie où les propriétés des taxons supérieurs sont partagées par leurs taxons inférieurs. Par exemple, dans l'ordre des *Picornavirales* Lefkowitz *et al.* (2018) :

- **Ordre** (*Picornavirales*) : virus non enveloppés à symétrie icosaédrique, ARN simple brin positif.
- **Famille** (*Picornaviridae*) : génome monocistronique unique.
- **Genre** (*Enterovirus*) : >42% d'identité protéique.
- **Espèce** (*Enterovirus C*) : >70% d'identité protéique, hôtes et récepteurs spécifiques.

En 2023, la taxonomie de l'ICTV comprenait 14 690 espèces virales réparties en différents niveaux taxonomiques, des règnes aux espèces (`https://ictv.global/taxonomy`).

1.1.10.2 La classification de Baltimore

La classification de Baltimore Baltimore (1971), illustrée dans la figure 1.5, organise les virus en fonction de leur type de matériel génétique et de leur stratégie de réplication. Elle repose sur deux critères principaux : la nature de leur génome, qu'il s'agisse d'ADN ou d'ARN, sous forme simple ou double brin, et leur méthode de production de l'ARN messager (ARNm).

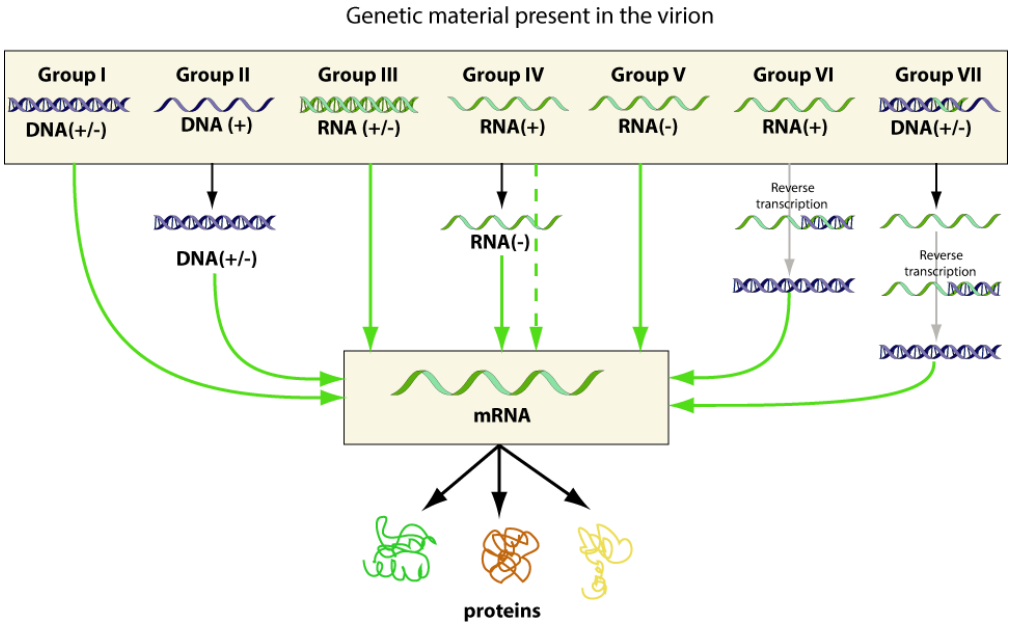


Figure 1.5 – Les sept groupes de la classification de Baltimore et leurs stratégies de réplication. Source : De Castro *et al.* (2024)

Les sept groupes sont :

1. **Virus à ADN double brin (ADNdb)** : transcription directe en ARNm *Ex.* : *Herpesviridae*.
2. **Virus à ADN simple brin (ADNsb)** : conversion en ADNdb avant transcription *Ex.* : *Parvoviridae*.
3. **Virus à ARN double brin (ARNdb)** : transcription par ARN polymérase virale *Ex.* : *Reoviridae*.
4. **Virus à ARN simple brin positif (ARNsb+)** : génome servant directement d'ARNm *Ex.* : *Picornaviridae*.
5. **Virus à ARN simple brin négatif (ARNsb-)** : synthèse d'ARNm complémentaire *Ex.* : *Orthomyxoviridae*.
6. **Virus à ARN+ avec transcriptase inverse** : conversion en ADN via transcriptase inverse *Ex.* : *Retroviridae* (HIV-1).
7. **Virus à ADN avec transcriptase inverse** : ADN partiellement double brin *Ex.* : *Hepadnaviridae* (virus hépatite B).

Cette classification est essentielle pour comprendre les mécanismes de réplication viraux et développer des stratégies antivirales ciblées.

1.2 Concepts et méthodes bioinformatiques

1.2.1 Processus de séquençage

Le séquençage est une technique permettant de déterminer l'ordre des nucléotides dans une molécule d'ADN ou d'ARN Schuster (2008). Cette étape cruciale révèle la structure primaire du matériel génétique des organismes et des virus. Les technologies de séquençage à haut débit (High-Throughput Sequencing, HTS) telles que Illumina Bentley *et al.* (2008), Nanopore Jain *et al.* (2016) ou encore PacBio Eid *et al.* (2009) permettent désormais d'analyser en parallèle des millions de séquences Goodwin *et al.* (2016). Ces avancées technologiques ont : accéléré l'acquisition des données génétiques, permis des analyses à grande échelle et révolutionné les domaines de la génomique et de la virologie. Les données générées par le HTS nécessitent ensuite un traitement bioinformatique et statistique approfondi pour leur interprétation.

1.2.2 Sous-séquences nucléotidiques (k-mers)

Les k-mers sont des sous-séquences contiguës de longueur k extraites d'une séquence biologique. Pour une séquence d'ADN ou d'ARN de longueur L, on peut extraire :

— $L - k + 1$ k-mers (en comptant les répétitions).

— Un maximum de 4^k k-mers distincts possibles (4 types de nucléotides).

Exemple pour la séquence "ACGT" :

— $k = 1$: A, C, G, T (4 mononucléotides).

— $k = 2$: AC, CG, GT (3 dinucléotides).

— $k = 3$: ACG, CGT (2 trinucléotides).

— $k = 4$: ACGT (1 tétranucléotide).

Les k-mers trouvent de nombreuses applications en bioinformatique, notamment dans l'assemblage de génomes, la classification virale, la détection de mutations, l'analyse métagénomique, l'optimisation de l'expression génique et le développement de vaccins Compeau *et al.* (2011); Solis-Reyes *et al.* (2018); Lebatteux *et al.* (2022); Wood et Salzberg (2014); Welch *et al.* (2009); Eschke *et al.* (2018).

1.2.3 Signatures génomiques

Les signatures génomiques sont des motifs ou caractéristiques distinctifs dans les séquences d'ADN qui permettent de caractériser des entités biologiques à différents niveaux d'organisation, des molécules aux organismes. Ces marqueurs moléculaires jouent un rôle fondamental dans la compréhension des systèmes biologiques, avec des applications allant de la recherche fondamentale à la médecine personnalisée. L'analyse bibliométrique de De la Fuente *et al*. De la Fuente *et al.* (2023) a mis en évidence les diverses interprétations de ce concept selon les domaines d'application.

1.2.3.1 Signatures associées aux phénotypes

Dans le contexte médical et génétique, les signatures génomiques désignent des profils moléculaires caractéristiques associés à des phénotypes spécifiques, comme des pathologies ou des réponses thérapeutiques. Ces signatures comprennent les signatures géniques (profils d'expression de gènes liés à des fonctions biologiques spécifiques), les signatures protéiques (ensembles de protéines caractéristiques d'états biologiques particuliers), les signatures mutationnelles (motifs de mutations associés à des pathologies) et les signatures immunitaires (profils d'expression caractérisant les réponses immunitaires).

1.2.3.2 Signatures spécifiques aux organismes

En génomique comparative, les signatures génomiques représentent des motifs d'ADN caractéristiques d'une espèce ou d'un groupe d'espèces. Ces signatures, basées sur des propriétés statistiques des séquences, permettent d'identifier les relations évolutives, de différencier les espèces et sous-espèces, et d'étudier l'évolution à travers l'analyse des motifs et fréquences de k-mers. Ces différentes formes de signatures génomiques constituent des outils essentiels pour l'analyse des séquences biologiques, combinant aspects fondamentaux et applications pratiques.

1.2.4 Alignement de séquence

L'alignement de séquence est une méthode bioinformatique fondamentale permettant de comparer des séquences biologiques (ADN, ARN, protéines). Cette technique consiste à superposer les séquences en préservant leur ordre, maximiser les correspondances entre les composants et insérer des "*gaps*" représentant les insertions ou délétions. Les alignements permettent ainsi d'identifier des régions de similarité, des relations fonctionnelles et des liens évolutifs entre les séquences.

Deux types principaux d'alignements sont couramment utilisés en bioinformatique. L'alignement par paire (*pairwise alignment*) compare deux séquences et peut être réalisé de manière locale, en recherchant des régions similaires, ou globale, en comparant les séquences sur toute leur longueur. L'alignement multiple (*multiple alignment*), quant à lui, permet la comparaison simultanée de plusieurs séquences pour identifier les régions conservées.

1.2.4.1 Alignement local

L'alignement local identifie les régions de similarité les plus significatives entre deux séquences, sans nécessiter un alignement complet. L'algorithme de Smith-Waterman Smith *et al.* (1981) utilise la programmation dynamique, une technique d'optimisation qui décompose le problème en sous-problèmes plus simples et mémorise leurs solutions pour éviter les calculs redondants, permettant ainsi de trouver l'alignement optimal des régions similaires.

Algorithme de Smith-Waterman :

1. **Initialisation :** Pour deux séquences $A = a_1 \dots a_n$ et $B = b_1 \dots b_m$, créer une matrice $S_{(n+1)\times(m+1)}$ avec :

$$S(i, 0) = 0 \quad \text{pour } i = 0, \dots, n$$

$$S(0, j) = 0 \quad \text{pour } j = 0, \dots, m$$

2. **Remplissage :** Pour chaque cellule $S(i, j)$:

$$S(i,j) = \max \begin{cases} 0 & \\ S(i-1, j-1) + s(a_i, b_j) & \text{(correspondance)} \\ S(i-1, j) + d & \text{(insertion)} \\ S(i, j-1) + d & \text{(délétion)} \end{cases}$$

avec le score de substitution :

$$s(a_i, b_j) = \begin{cases} +2 & \text{si } a_i = b_j \\ -1 & \text{si } a_i \neq b_j \end{cases}$$

et d la pénalité de gap (typiquement $d = -2$)

3. **Traceback :** Partir du score maximal $S(i^*, j^*)$ et retracer jusqu'à rencontrer un zéro.

Complexité :

- Temporelle : $O(nm)$
- Spatiale : $O(nm)$

Pour les grandes séquences, des heuristiques comme BLAST Altschul *et al.* (1990); Lu *et al.* (2024) sont utilisées pour accélérer la recherche.

1.2.4.2 Alignement global

Contrairement à l'alignement local, l'alignement global cherche à aligner les séquences sur toute leur longueur. Cette approche, implémentée par l'algorithme de Needleman-Wunsch Needleman et Wunsch (1970), utilise également la programmation dynamique pour trouver l'alignement optimal.

Algorithme de Needleman-Wunsch :

1. **Initialisation :** Pour deux séquences $A = a_1 \dots a_n$ et $B = b_1 \dots b_m$, créer une matrice $S_{(n+1)\times(m+1)}$ avec :

$$S(i,0) = i \times d \quad \text{pour } i = 0, \dots, n$$

$$S(0,j) = j \times d \quad \text{pour } j = 0, \dots, m$$

où d est la pénalité de gap.

2. **Remplissage :** Pour chaque cellule $S(i,j)$:

$$S(i,j) = \max \begin{cases} S(i-1,j-1) + s(a_i, b_j) & \text{(correspondance)} \\ S(i-1,j) + d & \text{(insertion)} \\ S(i,j-1) + d & \text{(délétion)} \end{cases}$$

avec le score de substitution :

$$s(a_i, b_j) = \begin{cases} +1 & \text{si } a_i = b_j \\ -1 & \text{si } a_i \neq b_j \end{cases}$$

3. **Traceback :** Partir de $S(n,m)$ et retracer le chemin optimal.

Complexité :

- Temporelle : $O(nm)$
- Spatiale : $O(nm)$

1.2.4.3 Alignement multiple

L'alignement multiple de séquences (Multiple Sequence Alignment, MSA) étend l'alignement par paire à plus de deux séquences en identifiant simultanément les régions conservées au sein d'un ensemble de séquences biologiques. Cet outil fondamental en biologie computationnelle permet de réaliser diverses analyses, telles que l'analyse phylogénétique à travers la reconstruction d'arbres et l'identification des relations évolutives, la prédiction structurale grâce à la détection de motifs conservés et à la modélisation par homologie, l'identification fonctionnelle par la détection des sites actifs et de liaison, ainsi que l'annotation génomique pour localiser les gènes et les éléments conservés.

1.2.4.3.1 Défis de l'alignement multiple

L'alignement multiple présente une complexité exponentielle liée au nombre de séquences à aligner. Pour n séquences de longueur moyenne l, le nombre total d'alignements possibles est de l'ordre de $O(l^n)$. Cette complexité rend les approches exactes basées sur la programmation dynamique impraticables au-delà de 3-4 séquences Wang et Jiang (1994). Des méthodes heuristiques ont été développées pour rendre l'alignement multiple réalisable en pratique.

1.2.4.3.2 Méthodes progressives

Les méthodes progressives construisent l'alignement multiple à travers une série d'alignements par paires, en commençant par les séquences les plus similaires puis en incorporant séquentiellement les plus divergentes. Cette approche heuristique s'appuie sur l'hypothèse que la conservation des alignements optimaux entre séquences proches minimise la propagation d'erreurs. Le programme **ClustalW** Thompson *et al.* (1994) implémente cette stratégie en trois phases : calcul des distances entre paires de séquences, construction d'un arbre guide, et réalisation des alignements progressifs suivant la topologie de l'arbre, avec optimisation par matrices de substitution et pénalités de gaps position-spécifiques.

1.2.4.3.3 Méthodes itératives

Les méthodes itératives améliorent l'alignement multiple en réévaluant et en affinant successivement l'alignement initial, tentant ainsi de surmonter les erreurs qui peuvent survenir dans les alignements initiaux des méthodes progressives. Le programme **MUSCLE** Edgar (2004) met en œuvre un algorithme itératif rapide, réalisant plusieurs tours d'alignement et d'optimisation pour converger vers un alignement de haute qualité, tandis que **MAFFT** Katoh *et al.* (2002) combine des techniques de calcul rapide des distances avec des étapes de raffinement successif.

1.2.4.3.4 Méthodes basées sur la consistance

Les méthodes basées sur la consistance visent à optimiser la cohérence entre les alignements par paire et l'alignement multiple final. Le programme **T-Coffee** Notredame *et al.* (2000) exploite une bibliothèque d'alignements par paire pondérés pour guider l'alignement multiple, ce qui améliore la qualité et la fiabilité de l'alignement final. **ProbCons** Do *et al.* (2005), quant à lui, emploie des modèles probabilistes pour estimer

la fiabilité des alignements et maximiser la consistance globale.

1.2.4.3.5 Modèles probabilistes

Les modèles probabilistes, tels que les modèles de Markov cachés (HMM), permettent d'attribuer des probabilités aux différentes combinaisons de gaps, de correspondances et de non-correspondances, afin de déterminer les alignements multiples les plus probables et de détecter des motifs conservés. Le programme **HMMER** Durbin *et al.* (1998) emploie ces modèles HMM pour construire des profils de séquences (aussi appelés profile HMM) et ainsi détecter des homologues distants, souvent impossibles à repérer par des méthodes d'alignement directes.

1.2.4.3.6 Complexité et performance

La complexité computationnelle des méthodes heuristiques est généralement acceptable pour un grand nombre de séquences. Cependant, il existe un compromis entre vitesse d'exécution et qualité d'alignement : les méthodes progressives offrent une rapidité d'exécution mais sont sensibles aux erreurs initiales, tandis que les méthodes itératives et de consistance fournissent une meilleure précision au prix d'un temps de calcul plus important. Malgré ces avancées méthodologiques, l'alignement multiple reste un défi, particulièrement pour les séquences très divergentes ou en grand nombre. La présence de régions répétitives, de réarrangements génomiques et de variations dans les taux d'évolution complexifie également la tâche d'alignement.

1.2.4.4 Séquences consensus

Une séquence consensus est une séquence représentative construite à partir d'un alignement multiple en identifiant, à chaque position, le nucléotide ou l'acide aminé le plus fréquent Durbin *et al.* (1998). Cette séquence permet de mettre en évidence les positions conservées et les variations communes tout en facilitant l'identification des variants rares et l'interprétation des données génomiques.

1.2.5 Matrices de substitution

Les matrices de substitution décrivent la fréquence à laquelle un caractère dans une séquence (nucléotide ou protéine) se transforme en un autre au cours de l'évolution. Ces matrices, exprimées sous forme de log-

odds, quantifient la probabilité de trouver deux caractères alignés en fonction de leur divergence évolutive présumée Zvelebil et Baum (2007).

Dans les algorithmes d'alignement, le score $s(a_i, b_j)$ entre résidus a_i et b_j est déterminé par ces matrices. Plus le score est élevé, plus la probabilité que ces résidus soient alignés et dérivés d'un ancêtre commun est importante. Pour l'ADN/ARN, les matrices sont relativement simples du fait des quatre nucléotides possibles (A, T/U, G, C). Exemple avec la matrice DNAfull (également connu sous le nom de EDNAFULL et NUC4.4) :

Table 1.1 – Matrice de substitution DNAfull pour les nucléotides de l'ADN

	A	C	G	T
A	+5	-4	-4	-4
C	-4	+5	-4	-4
G	-4	-4	+5	-4
T	-4	-4	-4	+5

Pour les séquences protéiques, les matrices de substitution sont plus complexes en raison des 20 acides aminés et de leurs propriétés physico-chimiques diverses. Les deux familles principales sont PAM et BLOSUM.

— **Matrices PAM** : Développées par Dayhoff et al.Dayhoff *et al.* (1978), les matrices PAM (Point Accepted Mutation) sont fondées sur le concept de "mutations acceptées" par la sélection naturelle. Elles utilisent un modèle évolutif global basé sur des séquences proches et sont indexées selon la distance évolutive (par exemple, PAM250 pour des séquences divergentes et PAM30 pour des séquences plus proches).

— **Matrices BLOSUM** : Les matrices BLOSUM (BLOcks SUbstitution Matrix) Henikoff et Henikoff (1992) se caractérisent par l'utilisation de blocs d'alignements conservés et une approche empirique ne reposant pas sur un modèle évolutif. Elles sont indexées selon un seuil d'identité, comme BLOSUM62, construite à partir de séquences partageant au maximum 62% d'identité.

La matrice BLOSUM62 est aujourd'hui largement utilisée dans les programmes d'alignement comme BLAST Altschul *et al.* (1997). D'autres matrices spécialisées ont également été développées, telle la matrice MIYATA Miyata *et al.* (1979) qui se base sur les propriétés physico-chimiques des acides aminés (masse moléculaire,

polarité, hydrophobicité, volume). Le choix de la matrice appropriée reste crucial pour obtenir des alignements significatifs et dépend principalement du niveau de similarité entre les séquences à analyser et du type d'homologie recherché (proche ou distante).

1.2.6 Matrices de similarité et de dissimilarité

Les matrices de similarité et de dissimilarité quantifient les relations entre séquences biologiques alignées, constituant ainsi un outil fondamental pour l'analyse des alignements multiples, le clustering et l'inférence phylogénétique.

1.2.6.0.1 Matrices de similarité

Une matrice de similarité $S_{N\times N}$ pour N séquences mesure leur proximité en se basant sur différentes approches tel que : les scores d'alignement (somme des scores des matrices de substitution et pénalités de gaps), les mesures de distance entre séquences (pourcentage d'identité, distance de Hamming ou de Levenshtein) et la similarité physico-chimique basée sur les propriétés des acides aminés.

1.2.6.0.2 Matrices de dissimilarité

Une matrice de distance $D_{N\times N}$ quantifie la divergence entre séquences, soit par transformation directe des similarités ($D_{ij} = 1 - \frac{S_{ij}}{\max(S)}$), soit via des modèles évolutifs comme Jukes-Cantor Jukes et Cantor (1969) ou Kimura Kimura (1980) qui corrigent pour les substitutions multiples.

	A	B	C	D
A	0	0.2	0.5	0.7
B	0.2	0	0.4	0.6
C	0.5	0.4	0	0.3
D	0.7	0.6	0.3	0

Table 1.2 – Matrice de dissimilarité entre quatre séquences.

Dans cette matrice 0 indique des séquences identiques et 1 une divergence maximale. Par exemple, la valeur 0.7 entre A et D indique une forte divergence entre ces séquences.

1.2.7 Mesures de distance

Les mesures de distance quantifient mathématiquement les différences entre séquences biologiques, constituant des outils essentiels pour l'analyse comparative, la phylogénie et le clustering.

1.2.7.0.1 Distances générales

Soient deux vecteurs x et y de dimension n, par exemple des fréquences de k-mers, avec $x = (x_1, x_2, \ldots, x_n)$ et $y = (y_1, y_2, \ldots, y_n)$. Nous pouvons appliquer la distance de Minkowski, une généralisation des distances euclidiennes englobant plusieurs cas particuliers selon la valeur du paramètre p :

— **Distance de Minkowski** :

$$D(x, y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^p\Big)^{\frac{1}{p}}$$

— **Distance de Manhattan** (p = 1) :

$$D_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

— **Distance euclidienne** (p = 2) :

$$D_{\text{Euclidienne}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

— **Distance de Chebyshev** (p = ∞) :

$$D_{\text{Chebyshev}}(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|$$

1.2.7.0.2 Distances pour séquences biologiques

Pour les séquences biologiques, trois types de distances sont particulièrement pertinents :

— **Distance de Hamming** : quantifie le nombre de positions différentes entre deux séquences alignées :

$$H(S_1, S_2) = \sum_{i=1}^{n} \delta(S_{1i}, S_{2i})$$

où $\delta(a, b) = 0$ si $a = b$, et 1 sinon.

— **Distance de Levenshtein** : mesure le nombre minimal d'opérations d'édition (insertions, suppressions, substitutions) nécessaires pour transformer une séquence en une autre.

— **Distances évolutives** : estiment le taux de substitutions par site en corrigeant pour les substitutions multiples :

  — *Jukes–Cantor* : corrige les substitutions multiples en supposant une probabilité égale de substitution pour tous les nucléotides :

$$D = -\frac{3}{4}\ln\Big(1 - \frac{4p}{3}\Big)$$

  où p est la proportion de sites différents.

  — *Kimura-2-parameter* : prend en compte la distinction entre transitions (A↔G, C↔T) et transversions (purine↔pyrimidine), reflétant leurs probabilités différentes de substitution :

$$D = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$$

  où P est la proportion de transitions et Q la proportion de transversions entre deux séquences.

1.2.8 Phylogénie

La phylogénie étudie les relations évolutives entre espèces ou séquences génétiques, permettant de retracer leur histoire évolutive depuis un ancêtre commun. Ces relations sont visualisées sous forme d'arbres phylogénétiques, comme illustré dans la Figure 1.6 pour l'évolution du SARS-CoV-2.

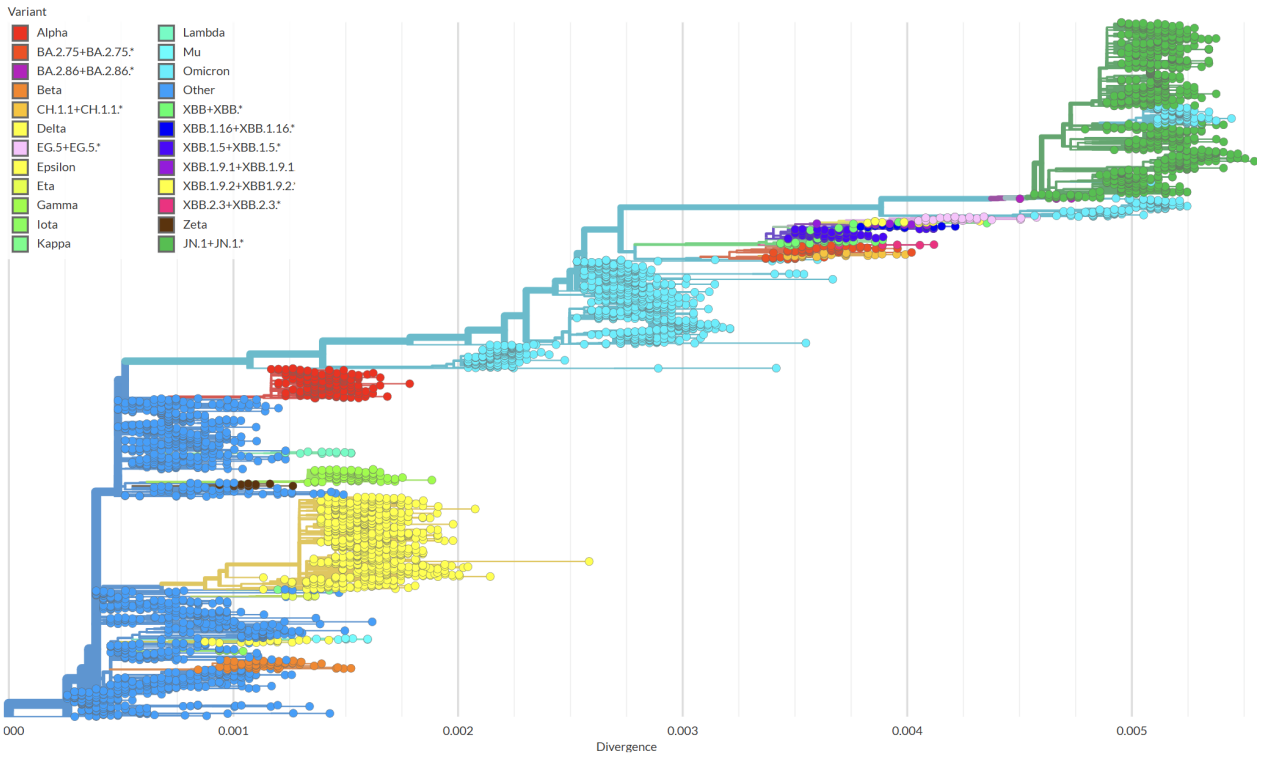


Figure 1.6 – Arbre phylogénétique montrant l'évolution mondiale du SARS-CoV-2 et de ses variants.

La figure illustre un arbre phylogénétique représentant l'évolution du SARS-CoV-2 et de ses variants à travers le monde. Elle est basée sur l'analyse de 4 063 génomes collectés entre 2019 et 2024, mettant en évidence la diversification du virus au cours de la pandémie. Source : `https://gisaid.org/phylodynamics/global/fiocruz/`.

Un arbre phylogénétique est une structure ramifiée où :

- Les **nœuds internes** représentent les ancêtres communs hypothétiques.
- Les **branches** indiquent les distances évolutives.
- Les **feuilles** correspondent aux séquences ou organismes actuels.
- La **racine** (si présente) désigne l'ancêtre commun le plus ancien.

La construction d'arbres phylogénétiques repose sur trois grandes catégories de méthodes : les méthodes basées sur les distances, les méthodes basées sur les caractères et les méthodes probabilistes.

1.2.8.1 Méthodes basées sur les distances

Ces méthodes utilisent une matrice de distances entre paires de séquences. Par exemple, la méthode **UPGMA** Sokal et Michener (1958) (Unweighted Pair Group Method with Arithmetic Mean) construit un arbre enraciné en supposant un taux d'évolution constant (horloge moléculaire) et regroupe récursivement les séquences par moyenne arithmétique des distances. La méthode **Neighbor-Joining** Saitou et Nei (1987), quant à elle, construit un arbre non enraciné sans supposer de taux d'évolution constant, en minimisant la longueur totale de l'arbre tout en optimisant le critère de vraisemblance minimale.

1.2.8.2 Méthodes basées sur les caractères

Ces méthodes utilisent directement les nucléotides ou acides aminés des séquences alignées. Par exemple, la méthode de **maximum de parcimonie** Fitch (1971) recherche l'arbre nécessitant le minimum de changements évolutifs, supposant que l'évolution suit le chemin le plus simple (principe d'économie), et analyse chaque site indépendamment. Le **maximum de vraisemblance** Felsenstein (1981) évalue la probabilité des données pour chaque arbre possible en utilisant un modèle d'évolution moléculaire explicite, et sélectionne l'arbre maximisant la vraisemblance des données.

1.2.8.3 Méthodes probabilistes

Les méthodes probabilistes incluent l'inférence bayésienne et les approches de vraisemblance maximale avancées. Elles estiment les arbres phylogénétiques en modélisant explicitement le processus évolutif et en utilisant des méthodes statistiques pour évaluer la qualité des arbres.

1.2.8.3.1 Inférence bayésienne

Les méthodes bayésiennes appliquent l'inférence bayésienne pour estimer la probabilité postérieure des arbres phylogénétiques, en tenant compte de l'incertitude sur les paramètres du modèle d'évolution. Ces méthodes utilisent des algorithmes de Monte Carlo par chaînes de Markov (MCMC) pour explorer l'espace des arbres possibles Huelsenbeck *et al*. (2001). Des logiciels tels que *MrBayes* Ronquist et Huelsenbeck (2003) implémentent ces approches.

1.2.8.3.2 RAxML (Randomized Axelerated Maximum Likelihood)

RAxML Stamatakis (2006) est un outil d'inférence phylogénétique reposant sur le principe de la vraisemblance maximale, spécialement conçu pour l'analyse de grands ensembles de données. Il implémente des algorithmes hautement optimisés via l'utilisation d'heuristiques efficaces pour la recherche d'arbres de vraisemblance maximale, intègre des options de calcul parallèle pour traiter des données volumineuses et prend en compte divers modèles de substitution, y compris des modèles hétérogènes.

1.2.9 Mesures de corrélation

Les coefficients de corrélation permettent d'évaluer la relation (ou la dépendance) entre deux variables quantitatives. Ils fournissent une valeur (généralement comprise entre -1 et $+1$) indiquant dans quelle mesure deux variables X et Y sont liées. Les coefficients de corrélation peuvent être utilisés pour analyser les relations entre certaines propriétés biophysiques des acides aminés (par exemple, leurs distances hydrophobiques ou leurs volumes moléculaires) et les valeurs des matrices de substitution (comme BLOSUM ou PAM).

1.2.9.1 Corrélation de Pearson

Le coefficient de corrélation de Pearson mesure l'intensité et la direction d'une relation linéaire entre deux variables X et Y. Le coefficient de Pearson nécessite des variables quantitatives (intervalle ou ratio), suppose une relation approximativement linéaire, et est sensible aux valeurs extrêmes (outliers). Il est défini par :

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives.

Interprétation :

- $r_{XY} = +1$: relation linéaire parfaite croissante.
- $r_{XY} = -1$: relation linéaire parfaite décroissante.
- $r_{XY} = 0$: absence de relation linéaire.

1.2.9.2 Corrélation de Spearman

Le coefficient de corrélation de Spearman (ρ) est une mesure non paramétrique basée sur les rangs des données plutôt que sur leurs valeurs brutes. Il est robuste aux valeurs extrêmes et peut détecter des relations monotones non nécessairement linéaires. Pour un échantillon de taille n, on remplace chaque x_i par son rang $R(x_i)$, et chaque y_i par son rang $R(y_i)$, puis on applique la formule de Pearson sur ces *rangs* :

$$\rho = \frac{\sum_{i=1}^{n} \big(R(x_i) - \overline{R(x)}\big)\,\big(R(y_i) - \overline{R(y)}\big)}{\sqrt{\sum_{i=1}^{n} \big(R(x_i) - \overline{R(x)}\big)^2 \;\sum_{i=1}^{n} \big(R(y_i) - \overline{R(y)}\big)^2}}$$

où $\overline{R(x)}$ et $\overline{R(y)}$ sont les moyennes des rangs.

Interprétation :

- $\rho = +1$: relation monotone parfaite croissante.
- $\rho = -1$: relation monotone parfaite décroissante.
- $\rho = 0$: absence de relation monotone.

1.2.9.3 Information Mutuelle

L'information mutuelle (MI) est une mesure issue de la théorie de l'information qui quantifie la dépendance statistique entre deux variables, y compris les relations non linéaires. Elle mesure la réduction d'incertitude sur une variable lorsque l'autre est connue. Elle est définie par :

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

où $p(x,y)$ est la distribution de probabilité jointe et $p(x)$, $p(y)$ sont les distributions marginales.

Interprétation :

- $I(X;Y) > 0$: existence d'une dépendance (plus la valeur est élevée, plus la dépendance est forte).
- $I(X;Y) = 0$: indépendance statistique des variables.
- Pas de limite supérieure fixe (peut être normalisée entre 0 et 1).

1.2.10 Algorithmes génétiques

Les algorithmes génétiques (AG) sont une classe d'algorithmes d'optimisation et de recherche inspirés des principes de la génétique et de la sélection naturelle Holland (1992). Ils utilisent des mécanismes comme la mutation, le croisement et la sélection pour résoudre des problèmes complexes d'optimisation.

1.2.10.1 Processus général d'un algorithme génétique :

Les AG simulent l'évolution naturelle pour trouver des solutions optimales, comme illustré dans la figure 1.7 extrait de Yu *et al.* (2022). Les principales composantes sont :

- **Population initiale :** Ensemble de solutions potentielles (chromosomes) générées aléatoirement ou à partir de connaissances préalables, codées sous forme appropriée (binaire, réelle, permutation).
- **Fonction d'évaluation (*fitness*) :** Évalue la qualité de chaque individu par rapport au problème à résoudre, permettant de comparer les solutions.
- **Sélection :** Choix des individus les plus aptes pour la reproduction (sélection par roulette, tournoi, rang).
- **Croisement (*crossover*) :** Échange de matériel génétique entre individus pour créer de nouveaux descendants (croisement à un point, deux points, uniforme).
- **Mutation :** Modifications aléatoires pour maintenir la diversité génétique et éviter la convergence prématurée.
- **Remplacement :** Formation de la nouvelle génération par remplacement partiel ou total (élitisme, remplacement des moins aptes).
- **Critère d'arrêt :** Condition de fin (nombre maximal de générations, convergence, seuil de performance).

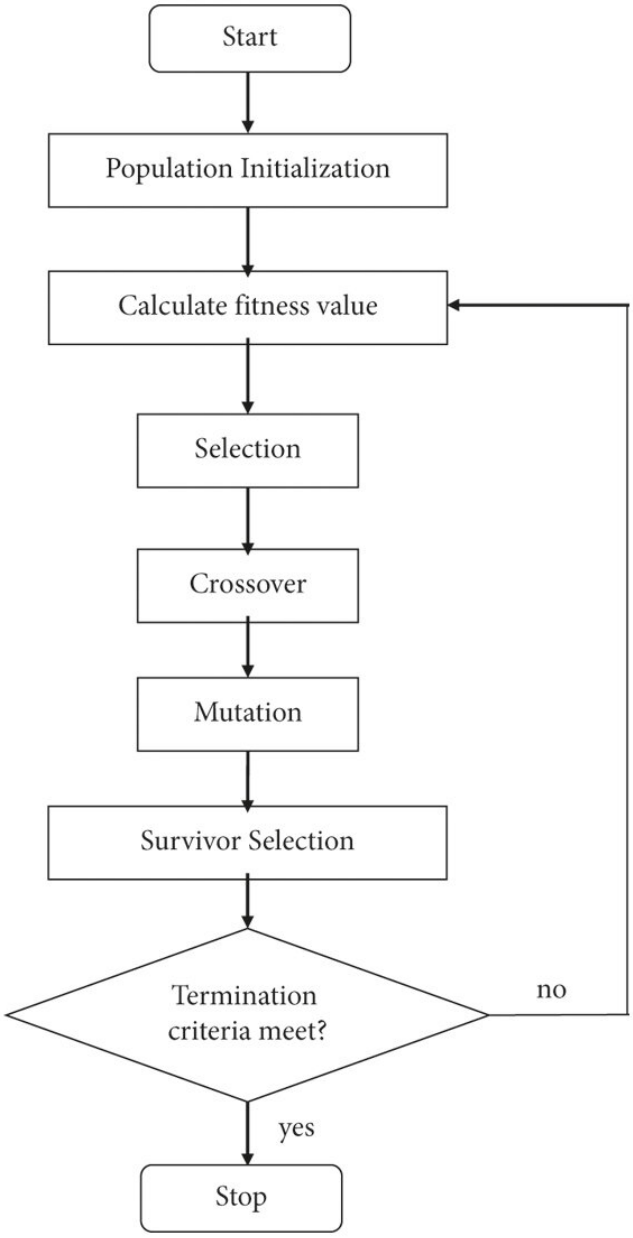


Figure 1.7 – Schéma général d'un algorithme génétique. Source : Yu *et al.* (2022)

1.2.10.2 Avantages et limitations :

Les algorithmes génétiques présentent plusieurs avantages majeurs : une robustesse face aux optima locaux, une grande flexibilité d'application et la possibilité de parallélisation des calculs. Cependant, elles comportent aussi certaines limitations, notamment une convergence qui peut être lente, une sensibilité importante au paramétrage initial, et l'absence de garantie d'atteindre l'optimalité globale.

1.3 Notions d'apprentissage automatique

1.3.1 Définition de l'apprentissage automatique

L'apprentissage automatique (machine learning) est une branche de l'intelligence artificielle qui développe des algorithmes permettant aux machines d'apprendre à partir de données et d'améliorer leurs performances sans programmation explicite Mitchell et Mitchell (1997). Il se décline en quatre approches principales selon le type d'apprentissage :

- **Apprentissage supervisé :** Apprentissage à partir d'exemples étiquetés pour prédire des sorties sur de nouvelles données (classification, régression).
- **Apprentissage non supervisé :** Découverte de structures dans des données non étiquetées (clustering, réduction de dimensionnalité, détection d'anomalies).
- **Apprentissage semi-supervisé :** Utilisation combinée de données étiquetées et non étiquetées pour améliorer l'apprentissage quand l'étiquetage est coûteux.
- **Apprentissage par renforcement :** Apprentissage par interaction avec un environnement dynamique, optimisant des récompenses reçues selon les actions effectuées.

1.3.2 Algorithmes d'apprentissage supervisé

1.3.2.1 Arbres de décision

Les arbres de décision sont des modèles prédictifs où chaque nœud interne représente un test sur une caractéristique, chaque branche un résultat possible, et chaque feuille une prédiction Quinlan (1986). Les arbres de décision présentent l'avantage d'être facilement interprétables et de gérer différents types de données sans prétraitement particulier, mais souffrent d'une certaine instabilité face aux variations des données et d'un risque de surapprentissage pouvant limiter leur précision.

La construction d'un arbre de décision implique les étapes suivantes :

1. **Sélection de la caractéristique optimale :** À chaque nœud, choisir la caractéristique qui sépare le mieux les données selon une mesure d'impureté, telle que :

— **L'entropie** mesure le degré d'incertitude ou d'impureté dans un ensemble de données. Pour un ensemble de données S avec c classes, l'entropie $H(S)$ est définie comme :

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

où p_i est la proportion d'instances appartenant à la classe i dans l'ensemble S.

— **Le gain d'information** $IG(S, A)$ mesure la réduction de l'entropie obtenue en divisant l'ensemble S selon la caractéristique A. Il est calculé comme :

$$IG(S, A) = H(S) - \sum_{v \in \text{Val}(A)} \frac{|S_v|}{|S|} H(S_v)$$

où Val(A) est l'ensemble des valeurs possibles de la caractéristique A et S_v est le sous-ensemble de S où la caractéristique A prend la valeur v.

— L'indice de Gini est une autre mesure d'impureté utilisée pour la classification. Pour un ensemble S, l'indice de Gini $G(S)$ est défini comme :

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2$$

où p_i est la proportion d'instances de la classe i dans S.

2. **Division récursive :** Diviser les données en sous-ensembles basés sur la caractéristique sélectionnée et les valeurs correspondantes. Répéter le processus pour chaque sous-ensemble jusqu'à ce qu'un critère d'arrêt soit atteint.
3. **Arrêt de la division :** La division s'arrête lorsque l'un des critères suivants est satisfait :
   — Les nœuds sont purs (toutes les instances appartiennent à la même classe).
   — Aucun gain d'information significatif n'est obtenu en divisant davantage.
   — Une profondeur maximale de l'arbre est atteinte.
   — Le nombre minimum d'instances requis pour une division n'est pas atteint.
4. **Élagage (*pruning*) :** Optionnellement, réduire la taille de l'arbre en supprimant les branches peu significatives pour éviter le surapprentissage (*overfitting*). L'élagage peut être effectué de manière préventive (pré-élagage) ou après la construction de l'arbre (post-élagage).

1.3.2.2 Forêts aléatoires

Les forêts aléatoires (*Random Forests*) constituent une méthode d'ensemble (*ensemble method*) qui étend le concept des arbres de décision en combinant plusieurs arbres pour améliorer la robustesse et la performance prédictive Breiman (2001). Cette approche repose sur deux principes fondamentaux : le **bagging** (*bootstrap aggregating*) et la **sélection aléatoire des caractéristiques**.

Le processus de construction d'une forêt aléatoire comprend trois étapes principales :

1. **Échantillonnage bootstrap :** Pour un jeu de données d'entraînement S, la méthode génère B échantillons bootstrap indépendants $\{S^{(b)}\}_{b=1}^{B}$ par tirage avec remise. Chaque échantillon $S^{(b)}$ sert à construire un arbre de décision $h^{(b)}(x)$ distinct.
2. **Construction des arbres avec sélection aléatoire :** Lors de la construction de chaque arbre, à chaque nœud :
   — Un sous-ensemble aléatoire de m caractéristiques est sélectionné parmi les p caractéristiques disponibles ($m \leq p$).
   — La division optimale est déterminée uniquement à partir de ce sous-ensemble, introduisant ainsi une diversité supplémentaire entre les arbres.
3. **Agrégation des prédictions :** Pour une nouvelle instance x, la prédiction finale dépend de la nature du problème :
   — En classification, un vote majoritaire est appliqué :

   $$\hat{C}(x) = \text{mode}\left\{h^{(b)}(x)\right\}_{b=1}^{B}$$

   où $\hat{C}(x)$ représente la classe prédite.
   — En régression, la moyenne des prédictions est calculée :

   $$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} h^{(b)}(x)$$

   où $\hat{f}(x)$ représente la valeur prédite.

Les forêts aléatoires permettent d'estimer l'importance des caractéristiques en mesurant la réduction moyenne de l'impureté ou l'augmentation de l'erreur de prédiction lorsque les valeurs d'une caractéristique sont permutées. Cette capacité d'auto-évaluation s'appuie sur deux mécanismes complémentaires.

1. **Réduction moyenne de l'impureté :** pour une caractéristique X_j, l'importance peut être calculée comme :

$$\text{Importance}(X_j) = \frac{1}{B}\sum_{b=1}^{B}\left(\Delta I^{(b)}(X_j)\right)$$

où $\Delta I^{(b)}(X_j)$ représente la réduction de l'impureté (par exemple, l'indice de Gini) attribuable à X_j dans l'arbre b. Cette mesure reflète la contribution moyenne de chaque caractéristique à la performance prédictive du modèle.

2. **Erreur Out-of-Bag (OOB) :** La construction de chaque arbre sur un échantillon bootstrap implique qu'environ $\frac{2}{3}$ des instances d'origine sont utilisées pour l'entraînement. Les instances restantes (environ $\frac{1}{3}$), non incluses dans l'échantillon bootstrap, sont qualifiées d'*out-of-bag*. Ces instances permettent d'estimer l'erreur de généralisation du modèle en calculant l'erreur OOB :

$$\text{OOB Error} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_{\text{OOB},i})$$

où L désigne une fonction de perte appropriée (par exemple, l'erreur 0-1 en classification), y_i représente la vraie étiquette de l'instance i, et $\hat{y}_{\text{OOB},i}$ correspond à la prédiction agrégée pour l'instance i en utilisant uniquement les arbres pour lesquels cette instance était OOB.

Cette approche OOB fournit une estimation non biaisée de l'erreur de généralisation, et permet également d'évaluer l'importance des caractéristiques en mesurant comment leur permutation aléatoire affecte l'erreur OOB.

1.3.2.3 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support (*Support Vector Machines*, SVM) constituent une classe d'algorithmes d'apprentissage supervisé particulièrement efficaces pour la classification et la régression Boser *et al.* (1992); Cortes (1995). Leur principe fondamental repose sur la recherche d'un hyperplan optimal qui maximise la séparation géométrique entre différentes classes de données, comme illustré dans la figure 1.8 García-Gonzalo *et al.* (2016).

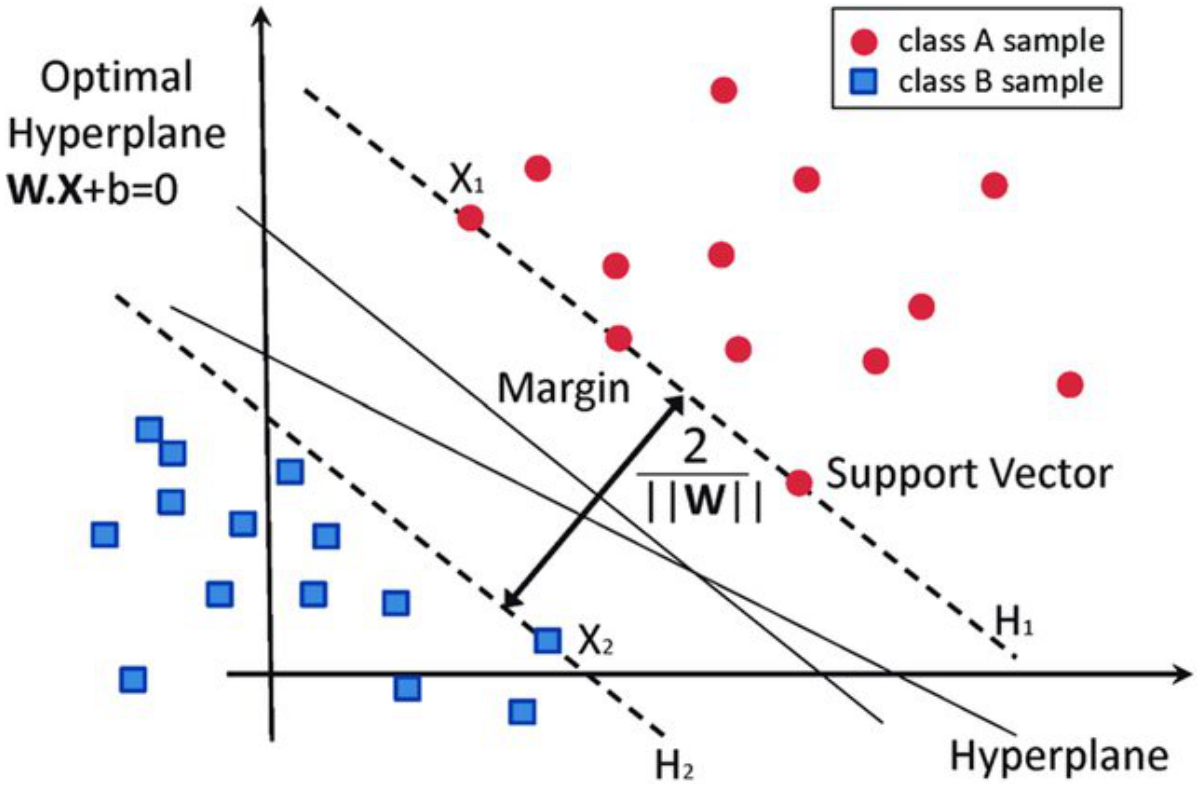


Figure 1.8 – Principe de fonctionnement d'un SVM.

L'hyperplan optimal sépare les deux classes (cercles rouges et carrés bleus) en maximisant la marge, définie comme la distance minimale entre l'hyperplan optimal et les vecteurs de support les plus proches de chaque classe. Les lignes en pointillés H_1 et H_2 passent par les vecteurs de support. Source : García-Gonzalo *et al.* (2016)

Dans le cas le plus simple d'une classification binaire, le SVM cherche à :

— Établir une frontière de décision linéaire (**hyperplan**) séparant l'espace des caractéristiques en deux régions.

— Maximiser la **marge** de séparation, définie comme la distance minimale entre l'hyperplan et les points d'apprentissage les plus proches.

— Identifier les **vecteurs de support**, qui sont les points d'apprentissage situés sur les marges et déterminant entièrement la frontière de décision.

1.3.2.3.1 Formulation mathématique des SVM

Les SVM reposent sur une formulation mathématique qui peut être décomposée en plusieurs éléments fondamentaux.

1. **Hyperplan séparateur**

   Pour un problème de classification binaire avec un ensemble d'entraînement

   $$\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$$

   où $\mathbf{x}_i \in \mathbb{R}^p$ représente le vecteur de caractéristiques du i-ème échantillon et $y_i \in \{-1, +1\}$ son étiquette de classe, un hyperplan séparateur dans $\mathbb{R}^p$ est défini par :

   $$f(\mathbf{x}) = \mathbf{w}^\top \cdot \mathbf{x} + b = 0$$

   avec $\mathbf{w} \in \mathbb{R}^p$ le vecteur normal à l'hyperplan et $b \in \mathbb{R}$ le biais.

2. **Maximisation de la marge**

   Le principe fondamental des SVM consiste à déterminer l'hyperplan maximisant la marge géométrique entre les classes. Cette marge est définie par deux hyperplans parallèles :

   $$\mathbf{w}^\top \cdot \mathbf{x} + b = +1 \quad \text{et} \quad \mathbf{w}^\top \cdot \mathbf{x} + b = -1$$

   La distance entre ces hyperplans, appelée marge, vaut $\frac{2}{\|\mathbf{w}\|}$. Les points situés sur ces hyperplans sont les **vecteurs de support**, qui déterminent entièrement la solution. Maximiser la marge équivaut à minimiser $\|\mathbf{w}\|$ ou de manière équivalente $\frac{1}{2}\|\mathbf{w}\|^2$.

3. **Problème d'optimisation**

   Dans le cas idéal de données linéairement séparables, le problème se formule comme :

   $$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimiser}} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{tel que} \qquad y_i\big(\mathbf{w}^\top \cdot \mathbf{x}_i + b\big) \geq 1, \quad \forall i = 1, \ldots, n \end{aligned}$$

   La contrainte garantit que chaque échantillon se trouve du côté correct de sa marge respective.

4. **Résolution**

   Le problème d'optimisation des SVM est résolu en utilisant les multiplicateurs de Lagrange, qui permettent de transformer le problème primal en un problème dual plus simple à résoudre.

   Formulation du Lagrangien : Le Lagrangien du problème primal s'écrit :

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big[y_i(\mathbf{w}^\top \cdot \mathbf{x}_i + b) - 1 \big]$$

   où $\alpha_i \geq 0$ sont les multiplicateurs de Lagrange.

   Conditions d'optimalité : Les conditions de Karush-Kuhn-Tucker (KKT) conduisent à :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

   Problème dual : En substituant ces conditions, le problème dual devient :

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximiser}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^\top \cdot \mathbf{x}_j) \\ \text{tel que} \quad & \alpha_i \geq 0, \quad \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

   Solution et vecteurs de support : La solution du problème primal se reconstruit à partir de la solution duale par :

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

   Les points $\mathbf{x}_i$ pour lesquels $\alpha_i > 0$ sont les vecteurs de support, satisfaisant :

$$\alpha_i \big[y_i(\mathbf{w}^\top \cdot \mathbf{x}_i + b) - 1 \big] = 0$$

1.3.2.3.2 Extension non linéaire : l'astuce du noyau

Dans de nombreuses applications, les données ne sont pas linéairement séparables dans leur espace d'origine $\mathbb{R}^p$. L'astuce du noyau (*kernel trick*) permet de contourner cette limitation en projetant implicitement les données dans un espace de dimension supérieure où une séparation linéaire devient possible.

— **Principe de la transformation :** On considère une transformation non linéaire :

$$\phi : \mathbb{R}^p \to \mathcal{H}, \quad \mathbf{x} \mapsto \phi(\mathbf{x})$$

où $\mathcal{H}$ est un espace de Hilbert de dimension potentiellement infinie. Dans cet espace, la fonction de décision devient :

$$f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b$$

où $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ désigne le produit scalaire dans $\mathcal{H}$.

— **Astuce du noyau :** L'innovation majeure réside dans l'utilisation d'une fonction noyau K qui calcule directement le produit scalaire dans $\mathcal{H}$ sans expliciter la transformation ϕ :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$$

Pour être valide, un noyau doit satisfaire les conditions de Mercer :

— Être symétrique : $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$

— Générer une matrice de Gram semi-définie positive.

— **Formulation duale non linéaire :** Le problème dual se reformule en remplaçant les produits scalaires par la fonction noyau :

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximiser}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{tel que} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, n \end{aligned}$$

où $C > 0$ est un paramètre de régularisation contrôlant le compromis entre la maximisation de la marge et la tolérance aux erreurs de classification.

— **Principaux noyaux et leurs caractéristiques :** Les SVM peuvent utiliser différentes fonctions noyaux pour transformer implicitement les données dans un espace de plus grande dimension où une séparation linéaire devient possible. Les principaux noyaux et leurs caractéristiques sont :

— **Noyau linéaire :**

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \cdot \mathbf{x}_j$$

Équivalent à un SVM linéaire dans l'espace d'origine.

— **Noyau polynomial :**

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^\top \cdot \mathbf{x}_j + r)^d$$

avec $\gamma > 0$, $r \geq 0$ et $d \in \mathbb{N}^*$. Capable de modéliser des frontières non linéaires via des combinaisons polynomiales des caractéristiques.

— **Noyau RBF (Gaussien) :**

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

avec $\gamma > 0$. Particulièrement efficace pour les données non linéaires, projette implicitement dans un espace de dimension infinie.

— **Noyau sigmoïde :**

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^\top \cdot \mathbf{x}_j + r)$$

avec $\gamma > 0$ et $r \geq 0$. Simule un réseau de neurones à une couche cachée.

Le choix du noyau et l'optimisation de ses hyperparamètres sont cruciaux et dépendent généralement de la nature des données et du problème à résoudre.

1.3.2.4 Estimation de probabilités via la méthode de *Platt scaling*

Les SVM, initialement conçus pour la classification binaire (-1 ou $+1$), peuvent être étendus pour fournir des probabilités associées à leurs prédictions. La méthode de *Platt scaling* Platt *et al.* (1999) propose de transformer la sortie réelle du SVM ($f(\mathbf{x})$) en une probabilité calibrée via une fonction logistique.

1.3.2.4.1 Principe de la calibration

La probabilité d'appartenance à la classe $+1$ est modélisée par une fonction sigmoïde :

$$P(y = +1 \mid f(\mathbf{x})) = \frac{1}{1 + \exp(A\, f(\mathbf{x}) + B)},$$

où les paramètres A et B sont estimés sur un ensemble de validation indépendant ou par validation croisée. Pour un échantillon i, la probabilité prédite est notée :

$$\widehat{p}_i = \frac{1}{1 + \exp(A\, f(\mathbf{x}_i) + B)}.$$

1.3.2.4.2 Optimisation des paramètres

Les paramètres A et B sont estimés en minimisant la log-vraisemblance négative :

$$\min_{A,B} \left\{ -\sum_{i=1}^{n} \left[t_i \, \ln(\widehat{p}_i) + (1 - t_i) \, \ln(1 - \widehat{p}_i) \right] \right\},$$

où t_i représente la cible modifiée pour l'échantillon i. Pour éviter les problèmes numériques liés aux probabilités extrêmes, Platt propose l'encodage suivant :

$$t_i = \begin{cases} \frac{N_{+1}+1}{N_{+1}+2} & \text{si } y_i = +1, \\ \\ \frac{1}{N_{-1}+2} & \text{si } y_i = -1, \end{cases}$$

avec N_{+1} et N_{-1} le nombre d'échantillons de chaque classe. Cette régularisation est particulièrement utile en présence d'échantillons mal classés ou pour les cas où les probabilités sont proches de 0 ou 1.

1.3.2.4.3 Procédure d'estimation

L'algorithme se décompose en quatre étapes :

1. **Apprentissage SVM :** Entraîner un SVM sur l'ensemble d'apprentissage pour obtenir $f(\mathbf{x})$.
2. **Collecte des scores :** Calculer $f(\mathbf{x}_i)$ pour chaque exemple de validation.
3. **Encodage des cibles :** Transformer les étiquettes y_i en cibles modifiées t_i.
4. **Optimisation :** Estimer A et B en minimisant la log-vraisemblance négative via des méthodes d'optimisation comme la descente de gradient ou les algorithmes quasi-Newton (L-BFGS).

Une fois les paramètres A et B estimés, cette calibration transforme les scores SVM en probabilités interprétables, facilitant la prise de décision en contexte incertain et permettant une comparaison équitable avec d'autres modèles probabilistes.

1.3.3 Algorithmes de clustering

Les algorithmes de clustering (ou *regroupement non supervisé*) visent à partitionner un ensemble de données en sous-groupes homogènes appelés **clusters**, de sorte que les objets au sein d'un même cluster présentent une plus forte similarité entre eux qu'avec les objets des autres clusters Rokach et Maimon (2005). Ces méthodes s'inscrivent dans le cadre de l'*apprentissage non supervisé*, ne nécessitant aucune information préalable sur l'appartenance des données à des classes.

Parmi les différentes approches de clustering, on distingue le clustering partitionnel, illustré par l'algorithme des k-means, le clustering hiérarchique, le clustering basé sur la densité, ainsi que le clustering spectral, chacune offrant des méthodes spécifiques pour regrouper les données en fonction de leurs caractéristiques.

1.3.3.1 Clustering partitionnel : l'algorithme k-means

L'algorithme des k**-means** Lloyd (1982) constitue l'une des méthodes de clustering les plus populaires grâce à sa simplicité conceptuelle et son efficacité computationnelle. Son objectif est de minimiser la variance intra-cluster en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{C}_1,\dots,\mathbf{C}_k} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_j} \left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|^2$$

où :

- k est le nombre prédéfini de clusters
- $\mathbf{C}_j$ représente le j-ème cluster
- $\boldsymbol{\mu}_j = \frac{1}{|\mathbf{C}_j|} \sum_{\mathbf{x}_i \in \mathbf{C}_j} \mathbf{x}_i$ est le centroïde du cluster $\mathbf{C}_j$
- $\|\cdot\|$ désigne la norme euclidienne

1.3.3.1.1 Algorithme

L'algorithme k-means procède de manière itérative :

1. **Initialisation** : Sélection de k centroïdes initiaux $\{\boldsymbol{\mu}_j^{(0)}\}_{j=1}^{k}$, soit aléatoirement, soit selon une stratégie d'initialisation (par exemple, k-means++).
2. **Étape d'affectation** : À l'itération t, chaque point est assigné au cluster dont le centroïde est le plus proche :

$$\mathbf{C}_j^{(t)} = \{\mathbf{x}_i : \|\mathbf{x}_i - \boldsymbol{\mu}_j^{(t-1)}\| \leq \|\mathbf{x}_i - \boldsymbol{\mu}_l^{(t-1)}\|, \forall l \neq j\}$$

3. **Étape de mise à jour** : Les centroïdes sont recalculés comme la moyenne des points de leur cluster :

$$\boldsymbol{\mu}_j^{(t)} = \frac{1}{|\mathbf{C}_j^{(t)}|} \sum_{\mathbf{x}_i \in \mathbf{C}_j^{(t)}} \mathbf{x}_i$$

4. **Convergence** : Les étapes 2 et 3 sont répétées jusqu'à ce qu'un critère d'arrêt soit satisfait (convergence des centroïdes ou nombre maximal d'itérations atteint).

1.3.3.1.2 Propriétés et limitations

L'algorithme k-means présente une convergence garantie vers un minimum local en un nombre fini d'itérations, avec une complexité de $O(nkdi)$ pour n échantillons, k clusters, d dimensions et i itérations. Ses principales limitations sont : la nécessité de spécifier le nombre de clusters a priori, la sensibilité à l'initialisation et l'hypothèse de clusters sphériques de densité homogène, le rendant inadapté aux distributions complexes ou de densités variables.

1.3.3.2 Clustering hiérarchique

Le **clustering hiérarchique** construit une hiérarchie de partitions représentée sous forme d'un arbre (*dendrogramme*). Cette approche repose sur deux stratégies principales de construction :

— **Approche agglomérative** (*bottom-up*) : Cette méthode commence avec des clusters unitaires (chaque donnée constitue un cluster) et fusionne itérativement les clusters les plus proches jusqu'à obtenir un cluster unique englobant toutes les données.
— **Approche divisive** (*top-down*) : À l'inverse, cette méthode débute par un cluster unique contenant toutes les données, puis procède par divisions successives jusqu'à atteindre des clusters unitaires ou satisfaire un critère d'arrêt prédéfini.

1.3.3.2.1 Mesures de distance inter-clusters

La distance entre clusters peut être définie selon plusieurs critères, chacun ayant des implications spécifiques sur la forme et la structure des clusters :

— **Single linkage** (liaison minimale) :

$$d(\mathbf{C}_1, \mathbf{C}_2) = \min_{\substack{x \in \mathbf{C}_1 \\ y \in \mathbf{C}_2}} d(x, y)$$

Cette méthode tend à créer des clusters allongés, souvent influencés par un effet de chaînage.

— **Complete linkage** (liaison maximale) :

$$d(\mathbf{C}_1, \mathbf{C}_2) = \max_{\substack{x \in \mathbf{C}_1 \\ y \in \mathbf{C}_2}} d(x, y)$$

Elle favorise des clusters compacts et de taille similaire, en limitant l'étendue des distances maximales.

— **Average linkage** (liaison moyenne) :

$$d(\mathbf{C}_1, \mathbf{C}_2) = \frac{1}{|\mathbf{C}_1|\,|\mathbf{C}_2|} \sum_{x \in \mathbf{C}_1} \sum_{y \in \mathbf{C}_2} d(x, y)$$

Cette approche propose un compromis entre la liaison minimale et maximale, en tenant compte de toutes les paires de distances.

— **Critère de Ward** :

$$d(\mathbf{C}_1, \mathbf{C}_2) = \frac{|\mathbf{C}_1|\,|\mathbf{C}_2|}{|\mathbf{C}_1| + |\mathbf{C}_2|} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$$

où $\boldsymbol{\mu}_i$ est le centroïde du cluster $\mathbf{C}_i$. Ce critère minimise l'augmentation de l'inertie intra-cluster à chaque fusion, favorisant des partitions homogènes.

1.3.3.2.2 Méthodes de mise à jour des distances

Lors de la fusion de deux clusters $\mathbf{C}_\alpha$ et $\mathbf{C}_\beta$, la distance entre le cluster résultant $\mathbf{C}_{\alpha \cup \beta}$ et un autre cluster $\mathbf{C}_\gamma$ peut être calculée selon plusieurs approches :

— **UPGMA (Unweighted Pair Group Method with Arithmetic Mean).** La distance est pondérée par la taille des clusters :

$$d(\mathbf{C}_{\alpha \cup \beta}, \mathbf{C}_\gamma) = \frac{|\mathbf{C}_\alpha|\; d(\mathbf{C}_\alpha, \mathbf{C}_\gamma) + |\mathbf{C}_\beta|\; d(\mathbf{C}_\beta, \mathbf{C}_\gamma)}{|\mathbf{C}_\alpha| + |\mathbf{C}_\beta|}$$

— **WPGMA (Weighted Pair Group Method with Arithmetic Mean).** Dans cette méthode, chaque cluster est considéré avec un poids égal :

$$d(\mathbf{C}_{\alpha\cup\beta}, \mathbf{C}_\gamma) = \frac{d(\mathbf{C}_\alpha, \mathbf{C}_\gamma) + d(\mathbf{C}_\beta, \mathbf{C}_\gamma)}{2}$$

1.3.3.2.3 Propriétés

Les algorithmes de clustering hiérarchiques procèdent en construisant une hiérarchie de clusters, généralement représentée sous forme de **dendrogramme**. Ce dernier présente les caractéristiques suivantes :

— Les **feuilles** représentent les points initiaux (objets ou observations individuelles).

— Les **nœuds internes** illustrent les fusions successives entre clusters.

— Une **coupe horizontale** dans le dendrogramme permet d'obtenir une partition en k clusters, en fonction du niveau de granularité souhaité.

Le clustering hiérarchique présente une complexité de $O(n^2 \log n)$, ce qui le rend particulièrement adapté aux ensembles de données de taille modérée. Son principal avantage réside dans le fait qu'il ne nécessite pas de spécifier le nombre de clusters à l'avance. Cependant, il souffre de certaines limites, notamment le caractère irréversible des fusions successives et un coût computationnel élevé, le rendant peu optimal pour les grands jeux de données.

1.3.3.3 Autres approches notables

1.3.3.3.1 Clustering basé sur la densité : DBSCAN

Le **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*) Ester *et al.* (1996) identifie les clusters comme des régions de haute densité séparées par des zones de faible densité. Cette méthode est particulièrement efficace pour détecter des formes complexes de clusters et pour identifier naturellement les points aberrants (*outliers*).

1.3.3.3.2 Spectral clustering

Cette approche utilise la structure spectrale du graphe de similarité des données, en procédant à une réduction dimensionnelle à l'aide des vecteurs propres avant d'effectuer le partitionnement. Elle est particulièrement utile pour des données non linéairement séparables.

1.3.3.4 Synthèse des méthodes de clustering

Le choix d'une méthode de clustering dépend principalement de la structure intrinsèque des données (densité, forme des clusters), de l'objectif de l'analyse (partition stricte, hiérarchie, détection d'anomalies) et des contraintes computationnelles (taille des données, complexité). Une approche de clustering rigoureuse combine souvent plusieurs méthodes et valide les résultats via des indices quantitatifs (Silhouette, Calinski-Harabasz, Davies-Bouldin).

1.3.4 Évaluation de l'apprentissage

L'évaluation de l'apprentissage consiste à mesurer la performance des modèles de prédiction en utilisant différentes métriques. Pour illustrer quelques-unes des métriques les plus répandues, considérons un ensemble de données composé de P instances positives et N instances négatives. Les prédictions effectuées par un modèle sur cet ensemble peuvent être résumées dans une matrice de confusion (Tableau 1.3).

Table 1.3 – Matrice de confusion pour un problème de classification binaire.

		Classe réelle	
		Positive ($+$)	Négative ($-$)
Classe prédite	Positive ($+$)	**Vrai positif (VP)**	**Faux positif (FP)**
	Négative ($-$)	**Faux négatif (FN)**	**Vrai négatif (VN)**

À partir de cette matrice de confusion, plusieurs métriques de performance peuvent être calculées :

— **Précision (*Precision*)** : également appelée *valeur prédictive positive*, elle correspond à la proportion des instances prédites positives qui sont effectivement positives.

$$\text{Précision} = \frac{VP}{VP + FP} \tag{1.1}$$

— **Rappel (*Recall*)** : également appelé *taux de vrais positifs* ou *sensibilité*, il désigne la proportion des instances positives correctement prédites par rapport au nombre total d'instances positives réelles.

$$\text{Rappel} = \frac{VP}{VP + FN} \tag{1.2}$$

— **Exactitude (*Accuracy*)** : elle désigne la proportion de prédictions correctes (vrais positifs et vrais

négatifs) parmi le nombre total d'instances.

$$\text{Exactitude} = \frac{VP + VN}{VP + VN + FP + FN} \tag{1.3}$$

— **Score F1 (*F1-score*)** : également appelé *F-mesure*, il correspond à la moyenne harmonique entre la précision et le rappel, fournissant un équilibre entre ces deux métriques.

$$\text{Score F1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{1.4}$$

Ces métriques permettent d'évaluer les performances d'un modèle sous différents aspects :

— La **précision** indique la fiabilité des prédictions positives du modèle.
— Le **rappel** mesure la capacité du modèle à identifier toutes les instances positives.
— L'**exactitude** donne une vue globale de la performance, mais peut être trompeuse en cas de classes déséquilibrées.
— Le **score F1** est particulièrement utile lorsque les classes sont déséquilibrées, car il prend en compte à la fois la précision et le rappel.

Il est important de choisir les métriques appropriées en fonction du contexte et des objectifs de l'application. Par exemple, dans le dépistage médical, un haut rappel (sensibilité) est souvent privilégié pour minimiser les faux négatifs, tandis que dans le filtrage de spam, une haute précision peut être plus souhaitable pour éviter les faux positifs.

1.3.5 Validation croisée

La validation croisée est une méthode statistique utilisée pour évaluer les performances des modèles en divisant un ensemble de données en k partitions disjointes, appelées *folds*, de tailles approximativement similaires Stone (1974). Cette méthode consiste à réaliser des processus d'entraînement et de test k fois de manière itérative. À chaque itération, k-1 partitions sont utilisées pour entraîner le modèle, tandis que la partition restante sert d'ensemble de test, comme illustré à la figure 1.9, extraite de Müller et Guido (2016). À l'issue de cette procédure, k scores de performance sont obtenus, un pour chaque itération. La moyenne de ces scores fournit une estimation du biais de la performance du modèle, tandis que l'écart type permet d'évaluer sa variance.

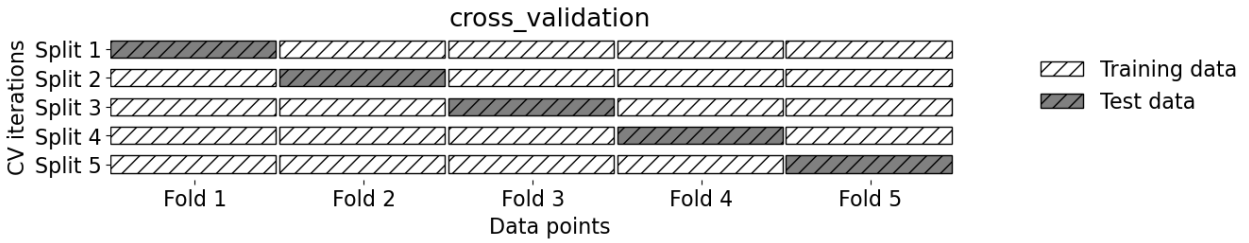


Figure 1.9 – Schéma de validation croisée 5.

La validation croisée permet d'évaluer la capacité de généralisation d'un modèle sur des données indépendantes et réduit le risque d'overfitting en utilisant efficacement l'ensemble des données disponibles. La validation croisée stratifiée assure que la proportion des classes dans chaque sous-ensemble reste similaire à celle de l'ensemble des données, ce qui est particulièrement important lorsque les classes sont déséquilibrées. Une variante courante est la validation croisée en *leave-one-out*, où (k) est égal au nombre total d'instances, chaque itération utilisant une seule instance pour le test et le reste pour l'entraînement.

CHAPITRE 2
OBJECTIF DE RECHERCHE ET REVUE DE LITTÉRATURE

2.1 Objectifs et hypothèses de recherche

Face aux défis majeurs que pose l'analyse des séquences biologiques virales pour leur classification, l'objectif général cette thèse est de proposer des approches novatrices pour l'identification et l'analyse de signatures génomiques dans les séquences virales, afin d'améliorer leur classification et d'approfondir la compréhension de leur biologie. Bien que notre étude porte initialement sur l'ensemble des espèces virales définies par l'ICTV, elle cible prioritairement des virus présentant une forte variabilité nucléotidique, un impact mondial significatif et un niveau taxonomique infra-espèce, conformément aux recommandations du comité d'évaluation du projet. Parmi ces virus, nous mettons principalement l'accent sur :

1. **Le HIV-1**, principal responsable du syndrome d'immunodéficience acquise (SIDA), connu pour sa grande diversité génétique et son impact majeur sur la santé publique Taylor *et al.* (2008).
2. **Le SARS-CoV-2**, agent causal de la COVID-19, caractérisé par l'émergence rapide de nouveaux variants ayant des répercussions importantes en santé publique Koyama *et al.* (2020).
3. **Le cytomégalovirus humain (HCMV)**, qui constitue l'infection congénitale la plus répandue et la principale cause de surdité, de déficits cognitifs et de troubles visuels chez l'enfant Gantt *et al.* (2016). Ce virus est au cœur des travaux de notre laboratoire et du Centre de recherche du CHU Sainte-Justine, où ce projet est co-supervisé.

Nos travaux se déclinent en trois objectifs spécifiques, chacun associé à une hypothèse de recherche :

— **Objectif O1 : Concevoir des méthodes efficaces pour identifier des signatures génomiques discriminantes et construire des modèles de prédiction basés sur ces signatures.**

S'appuyant sur nos travaux antérieurs (Lebatteux *et al.*, 2019), qui ont montré qu'un ensemble minimal de signatures génomiques (*k*-mers) peut distinguer diverses classes virales, nous formulons l'hypothèse suivante :

Hypothèse H1 : Il est possible de développer des algorithmes capables de détecter plusieurs sous-

ensembles minimaux de *k*-mers optimisant la discrimination entre classes virales. Intégrer ces *k*-mers dans des modèles d'apprentissage supervisé permettrait de générer des systèmes de prédiction performants à grande échelle (p. ex. SARS-CoV-2) et d'améliorer la classification pour des virus à forte variabilité (p. ex. HIV-1).

Cet objectif met l'accent sur l'identification de *k*-mers discriminants et leur intégration dans des modèles robustes. Les défis incluent la gestion de vastes volumes de données, la résistance au bruit de séquence et l'adaptabilité à différents contextes viraux, des virus à évolution rapide (p. ex. SARS-CoV-2) jusqu'aux virus hautement variables (p. ex. HIV-1, HCMV).

— **Objectif O2 : Développer des méthodes pour identifier les variations des *k*-mers discriminatifs et en dériver des informations biologiques au sein des sous-espèces virales.**

Les *k*-mers discriminants identifiés (O1) peuvent présenter des variations (substitutions nucléotidiques, etc.) entre sous-types ou lignées virales. La caractérisation de ces variations, ainsi que leur position sur le génome, est cruciale pour en comprendre la portée biologique.

Hypothèse H2 : Les *k*-mers discriminants sont liés à des changements génomiques intervenant dans des régions essentielles (p. ex. gènes clés), pouvant affecter la fonction ou la dynamique des protéines virales par des substitutions d'acides aminés.

Cette étude approfondie requiert le développement d'outils d'alignement local ciblé et l'évaluation des impacts fonctionnels des variations observées.

— **Objectif O3 : Élaborer un système de notation reflétant la pertinence discriminative et biologique des signatures génomiques et de leurs variations.**

Après avoir identifié les signatures génomiques (*k*-mers) discriminantes (O1) et caractérisé leurs variations (O2), nous proposons de développer un score d'importance évaluant à la fois le pouvoir discriminant (classification multiclasse) et la valeur biologique (impact sur la dynamique des protéines virales). Plus précisément, ce système de notation inclura :

1. Une composante destinée à générer un score discriminant à partir d'un cadre d'apprentissage

supervisé gérant la classification multiclasse.

2. Une composante de signification biologique évaluant :
   (i) l'importance fonctionnelle des régions codantes (modèles supervisés sur bases de données protéiques virales) ;
   (ii) l'impact potentiel des substitutions d'acides aminés sur la dynamique des protéines (différences de propriétés biophysiques).

Hypothèse H3 : Un score intégrant ces deux dimensions (discriminante et biologique) constituera un indicateur *in silico* fiable pour prioriser les mutations d'intérêt et guider les validations expérimentales (*in vitro*, *in vivo*). Cette approche pourrait également contribuer à définir de nouveaux génotypes ou clades pour des virus ou gènes encore peu caractérisés, comme *UL33* du HCMV.

2.2 Revue de littérature

2.2.1 Approches pour la classification et l'analyse basées sur les séquences virales

Historiquement, la classification et l'analyse des séquences virales reposent sur deux grandes catégories de méthodes :

1. **Approches basées sur l'alignement de séquences** : Ces méthodes consistent à aligner deux ou plusieurs séquences biologiques (nucléotidiques ou protéiques) afin de comparer leurs éléments constitutifs. L'alignement met en évidence et quantifie les modifications évolutives, telles que les insertions, les délétions et les substitutions, tout en identifiant les régions conservées. L'analyse des similarités qui en résulte fournit des informations essentielles sur les relations fonctionnelles, structurales et évolutives entre les séquences (Zielezinski *et al.*, 2017).
2. **Approches indépendantes de l'alignement de séquences** (*alignment-free*) : Ces méthodes s'affranchissent de la nécessité d'établir une correspondance positionnelle explicite entre les séquences. Elles reposent sur des caractéristiques intrinsèques, comme la fréquence des sous-séquences (méthodes basées sur les mots ou k-mers) ou le contenu informationnel global (méthodes basées sur la théorie de l'information). Les descripteurs ainsi obtenus servent de fondement pour diverses analyses statistiques, calculs de distances et approches d'apprentissage automatique, permettant de comparer et de classer les séquences (Zielezinski *et al.*, 2019).

La figure 2.1 adapté de Chan et Ragan (2013) illustre schématiquement la différence fondamentale entre une approche basée sur l'alignement multiple (A) et une approche indépendante de l'alignement (B) dans la reconstruction phylogénétique. Dans cet exemple, quatre séquences homologues (1, 2, 3 et 4), dont la phylogénie de référence est connue (partie supérieure), comportent plusieurs régions conservées représentées par différentes couleurs (bleu, jaune, vert, rouge et gris). Les séquences 2 et 4 présentent des réarrangements dans les régions jaune et verte par rapport aux séquences 1 et 3.

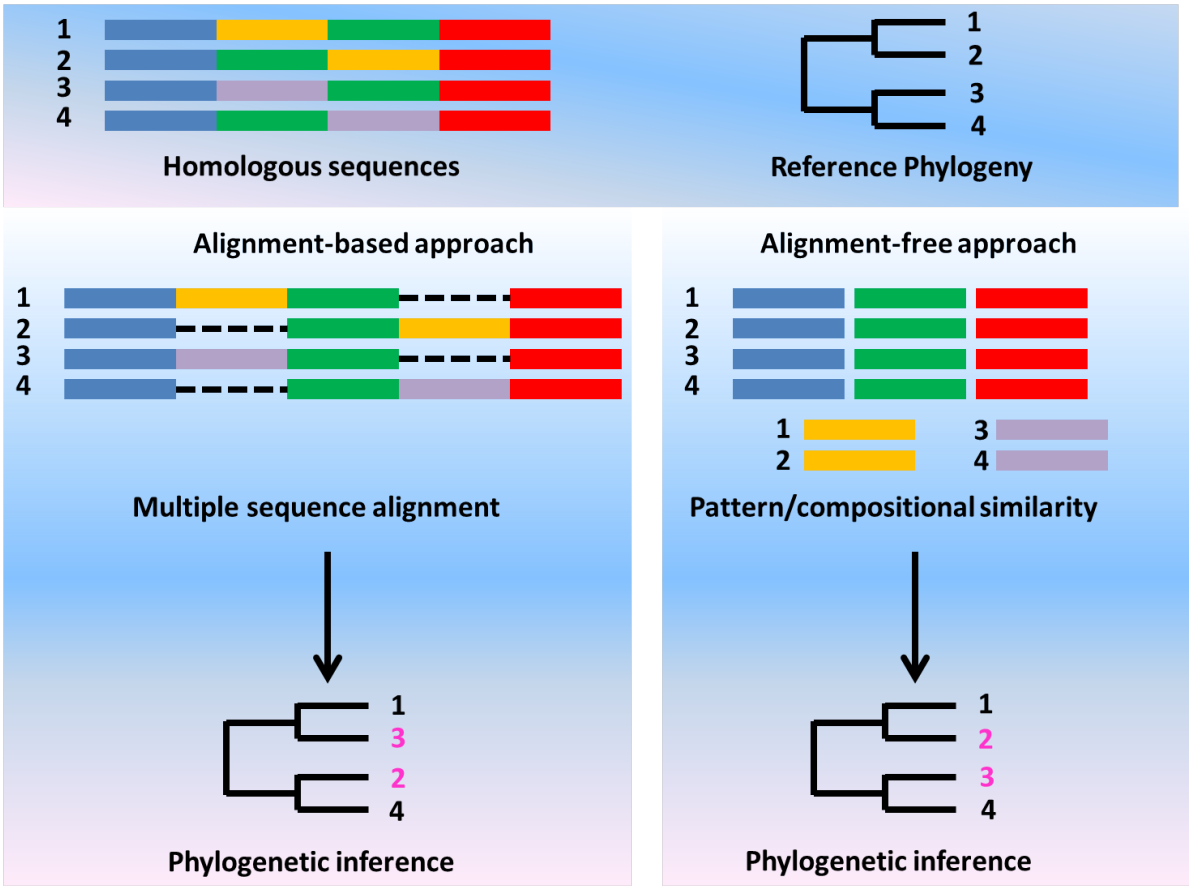


Figure 2.1 – Comparaison des approches basées sur l'alignement et indépendantes de l'alignement pour la reconstruction phylogénétique. Source : Wikimedia Commons - Licence GFDL 1.2+

L'alignement multiple (A) se trouve perturbé par ces réarrangements, nécessitant l'introduction de *gaps* (représentés par des lignes pointillées) et conduisant à une phylogénie qui s'écarte de la référence. À l'inverse, l'approche indépendante de l'alignement (B), qui repose sur l'analyse de signatures de séquences et de propriétés globales, reconstruit une phylogénie fidèle à la référence et démontre une robustesse accrue face aux réarrangements.

2.3 Approches basées sur l'alignement

Les méthodes basées sur l'alignement, définies précédemment (section 2.2.1), constituent historiquement le socle de l'analyse comparative et de la classification des séquences virales King *et al.* (2011). Ces méthodes se déclinent en trois catégories principales :

2.3.1 Recherche de similarité par alignement local

L'algorithme de Smith-Waterman (Smith *et al.*, 1981), avec une complexité en $O(mn)$ où m et n représentent les longueurs des séquences comparées, constitue la référence pour l'alignement local. Cette approche se concentre sur la détection de régions similaires entre séquences, sans nécessiter l'analyse de la séquence complète. Elle s'avère particulièrement adaptée à la recherche de motifs conservés et à la détection d'homologies partielles. Dans le contexte de la classification virale, une séquence à classifier est comparée à une base de données de référence, puis un taxon lui est assigné en fonction du degré de similarité observé, généralement évalué via l'*e-value* ou le *bit score*. Bien que Smith-Waterman soit l'algorithme exact de référence, plusieurs heuristiques ont été développées pour faciliter la recherche dans de vastes bases de données. **BLAST** (*Basic Local Alignment Search Tool*) (Altschul *et al.*, 1997) offre un compromis entre rapidité et sensibilité grâce à son système de «mots graines» (*seed words*), qui identifie d'abord de courtes séquences exactement conservées avant d'étendre l'alignement. **USEARCH** (Edgar, 2010) optimise la recherche en triant les séquences selon leur nombre de mots communs avec la requête, permettant d'identifier rapidement les meilleurs alignements parmi les premiers candidats. **DIAMOND** (Buchfink *et al.*, 2015, 2021) introduit une approche de double indexation qui traite simultanément les séquences requêtes et références, combinée à l'utilisation de graines espacées, offrant une vitesse supérieure à BLASTX tout en maintenant une sensibilité comparable. Dans le domaine spécifique du typage viral, des outils dédiés ont été développés. Le **NCBI subtyping tool** (Rozanov *et al.*, 2004) emploie une approche par fenêtres glissantes, comparant successivement chaque segment aux séquences de référence via BLAST pour identifier les génotypes et détecter les événements de recombinaison. Le **Stanford HIV-db** (Gifford *et al.*, 2006) se spécialise dans l'analyse du HIV-1 en assignant les sous-types via le calcul d'une distance entre les gènes *PR* et *RT* de la séquence requête et les séquences consensus des différents lignages viraux, tout en intégrant une analyse des mutations de résistance. Ces solutions spécialisées, malgré une sensibilité légèrement inférieure à Smith-Waterman, accélèrent significativement l'identification et la caractérisation des séquences virales.

2.3.2 Calcul de distance par alignement global

L'algorithme de Needleman-Wunsch (Needleman et Wunsch, 1970), avec une complexité en $O(mn)$, constitue la méthode de référence pour l'alignement global des séquences. Cette approche optimise la correspondance sur l'intégralité des séquences via une matrice de programmation dynamique, pénalisant les insertions et délétions. Elle est particulièrement adaptée à la comparaison de séquences de taille comparable, mais peut donner des résultats moins pertinents en présence de différences de longueur importantes ou de réarrangements majeurs. Pour la classification des virus, plusieurs outils exploitent ce principe avec des approches complémentaires. **PASC** (*Pairwise Sequence Comparison*) (Bao *et al.*, 2014) adapte sa stratégie d'alignement selon la taille des génomes, utilisant l'algorithme de Needleman-Wunsch pour les virus de moins de 32 kb et l'algorithme de Hirschberg pour les plus grands génomes, permettant d'établir des seuils taxonomiques spécifiques basés sur la distribution des identités de séquences. **DEmARC** (*DivErsity pArtitioning by hieRarchical Clustering*) (Lauber et Gorbalenya, 2012) se distingue en utilisant les distances évolutives par paires (*PEDs*) calculées à partir d'alignements multiples de protéines conservées, une approche ayant démontré son efficacité notamment pour les *Picornaviridae* et *Filoviridae*. Des outils plus spécialisés répondent à des besoins spécifiques. **VIRIDIC** (*Virus Identification for Distance Calculation*) (Moraru *et al.*, 2020) adapte le calcul des distances aux particularités des génomes de bactériophages via un alignement BLASTN optimisé et un calcul de similarité intergénomique prenant en compte la longueur totale des génomes. Pour faciliter la prise de décision taxonomique, **SDT** (*Sequence Demarcation Tool*) (Muhire *et al.*, 2014) standardise le calcul des identités génétiques par paires en gérant spécifiquement les problèmes d'alignement multiples et de gaps, fournissant des visualisations conformes aux critères de l'ICTV.

2.3.3 Placement phylogénétique

Dans cette troisième catégorie, il s'agit de positionner une séquence inconnue au sein d'un contexte phylogénétique déjà établi, afin de tenir compte de l'information évolutive pour la classification. Dans un premier temps, un arbre de référence est construit à partir d'un alignement multiple de séquences connues, souvent obtenu à l'aide de programmes tels que **MAFFT** (Katoh *et al.*, 2002), **MUSCLE** (Edgar, 2004) ou **ClustalW** (Thompson *et al.*, 1994). Ensuite, l'analyse phylogénétique d'une nouvelle séquence peut se faire selon deux stratégies :

- la reconstruction d'un nouvel arbre à partir de l'alignement multiple enrichi de la nouvelle séquence ;
- ou le placement direct de la séquence inconnue sur l'arbre déjà construit.

De nombreux outils ont été développés pour mettre en œuvre ces principes, chacun répondant à des be-

soins spécifiques :

- **Pour l'analyse taxonomique générale : VICTOR** (Meier-Kolthoff et Göker, 2017) classifie les virus procaryotes en utilisant des distances intergenomiques et une approche de clustering ; **mPTP** (Kapli *et al.*, 2017) applique un modèle de coalescence pour la délimitation d'espèces ; **ViCTree** (Modha *et al.*, 2018) intègre arbres phylogénétiques et métriques de similarité pour l'analyse de groupes monophylétiques.
- **Pour les virus spécifiques : REGA** (De Oliveira *et al.*, 2005; Abecasis *et al.*, 2010; Pineda-Peña *et al.*, 2013) combine analyse phylogénétique et arbres de décision pour le typage du HIV-1 et la détection de recombinants ; **PhyloPlace** (Hraber *et al.*, 2008) utilise un indice de branchement (*branching index*) combinant distances et phylogénie pour quantifier objectivement l'appartenance d'une séquence aux sous-types du HIV-1 et du HCV ; **APCluster** (Fischer *et al.*, 2018) applique le clustering par affinité pour la classification intra-spécifique du virus de la rage ; **MyCoV** (Wilkinson *et al.*, 2020) analyse une région conservée de 387 nucléotides du gène *RdRp* des coronavirus pour leur classification au niveau du sous-genre via une approche phylogénétique bayésienne.
- **Pour l'optimisation du placement : SCUEAL** (Kosakovsky Pond *et al.*, 2009) optimise le placement par maximum de vraisemblance avec détection de recombinaisons ; **pplacer** (Matsen *et al.*, 2010) implémente une approche bayésienne pour le placement phylogénétique probabiliste.
- **Pour la surveillance épidémiologique : Nextclade** (Aksamentov *et al.*, 2021) et **UShER** (Turakhia *et al.*, 2021) utilisent des arbres phylogénétiques incrémentaux pour le placement rapide et le suivi en temps réel des variants viraux.

Ces approches phylogénétiques, qui intègrent l'information évolutive dans le processus de classification, se sont révélées particulièrement efficaces tant pour la surveillance épidémiologique que pour la classification taxonomique des virus. Leur utilisation lors de la pandémie de COVID-19 a notamment démontré leur pertinence pour le suivi en temps réel des variants viraux émergents.

2.3.4 Limitations des approches basées sur l'alignement pour l'analyse des génomes viraux

Malgré leur grande utilité, les méthodes basées sur l'alignement présentent des limitations significatives pour l'analyse des séquences virales (Bonham-Carter *et al.*, 2014; Zielezinski *et al.*, 2017, 2019; Mayne *et al.*, 2024) :

- **Hypothèse de colinéarité** : Les alignements reposent sur l'hypothèse que les séquences partagent

un ordre ancestral commun. Cette condition est fréquemment violée chez les virus, qui subissent de nombreux réarrangements génomiques et événements de recombinaison au cours de leur évolution.

— **Divergence importante** : Les virus présentent des taux de mutation particulièrement élevés, conduisant à une accumulation rapide de différences dans leurs séquences. Au-delà d'un certain seuil de divergence génétique, les alignements deviennent peu fiables et peuvent produire des résultats erronés.
— **Complexité computationnelle** : L'alignement, en particulier multiple, nécessite des ressources informatiques considérables en temps et en mémoire. Cette limitation devient critique face à l'augmentation exponentielle du nombre de séquences générées par le séquençage à haut débit, et s'applique également à la comparaison simultanée par paires de centaines de génomes viraux.
— **Dépendance aux paramètres** : La qualité des alignements dépend fortement du choix des matrices de substitution et des pénalités d'écart (*gap penalties*). L'absence de standardisation universelle de ces paramètres, qui varient selon la famille virale étudiée, complique la comparaison entre différentes analyses.

Face à ces contraintes, de nouvelles approches dites «indépendantes de l'alignement» ont émergé. Ces méthodes s'avèrent particulièrement adaptées à l'étude des génomes viraux hautement divergents et au traitement des jeux de données massifs issus des technologies de séquençage modernes. La section suivante détaillera ces approches alternatives qui permettent de contourner les limitations intrinsèques aux méthodes basées sur l'alignement.

2.4 Approches indépendantes de l'alignement de séquences (alignment-free)

Les approches dites *alignment-free*, définies précédemment (section 2.2.1), présentent plusieurs avantages qui les rendent particulièrement adaptées à l'ère du séquençage à haut débit. Leur complexité algorithmique, généralement linéaire Bonham-Carter *et al.* (2014) et dépendant uniquement de la longueur de la séquence requête, permet une analyse efficace des génomes complets Bernard *et al.* (2019). Ces méthodes se montrent également plus robustes face aux événements de recombinaison et de réarrangements génomiques Vinga (2014a). Elles s'avèrent particulièrement pertinentes dans les cas où la faible conservation des séquences rend l'alignement traditionnel peu fiable.

De manière générale, on distingue deux grandes catégories parmi les méthodes *alignment-free* (Vinga et Almeida, 2003; Vinga, 2014b; Zhang *et al.*, 2017; Bernard *et al.*, 2019) :

1. **Méthodes fondées sur l'information contenue dans les séquences complètes :** elles s'appuient sur des principes de théorie de l'information afin de calculer la similarité ou la dissimilarité entre deux séquences.
2. **Méthodes basées sur la fréquence des mots (words / k-mers) :** elles décomposent les séquences en motifs de longueur fixe (k-mers) pour générer des profils de composition comparables.

2.4.1 Approches fondées sur la théorie de l'information

Les méthodes issues de la théorie de l'information évaluent la quantité d'information partagée entre deux séquences biologiques, considérées comme de simples chaînes de symboles (nucléotides ou acides aminés). Cette perspective permet d'appliquer des concepts fondamentaux tels que la complexité et l'entropie à l'évaluation de la similarité ou de la dissimilarité entre séquences.

2.4.1.1 Complexité de Kolmogorov

La *complexité de Kolmogorov* d'une séquence se définit comme la longueur de sa description la plus concise. Par exemple, la suite `TTTTTTTT` peut être décrite par « huit T consécutifs », tandis qu'une séquence plus variée, telle que `AGTGCCTA`, nécessite la mention explicite de chaque symbole. En pratique, la détermination de cette description minimale étant difficile, on utilise des algorithmes de compression (p. ex. `zip`, `gzip`) ou des méthodes spécialisées pour estimer la complexité (Li *et al.*, 2008).

Pour comparer deux séquences x et y (Figure 2.3) extraite de (Zielezinski *et al.*, 2017), on procède généralement ainsi :

1. Calculer la complexité de chaque séquence de manière indépendante : $c(x)$ et $c(y)$
2. Calculer la complexité de leur concaténation : $c(xy)$

L'algorithme de *Lempel-Ziv* (Otu et Sayood, 2003), fréquemment utilisé pour estimer ces quantités, parcourt la séquence de gauche à droite en détectant progressivement les motifs inédits. Les distances entre séquences s'obtiennent ensuite à partir de $c(x)$, $c(y)$ et $c(xy)$, en exploitant deux principes :

— Pour des séquences similaires, $c(xy)$ est proche de $\max c(x), c(y)$

— Pour des séquences très différentes, $c(xy)$ tend vers $c(x) + c(y)$

Ces propriétés permettent de définir des distances de compression normalisées, offrant une estimation à la fois robuste et efficace de la similarité entre séquences biologiques.

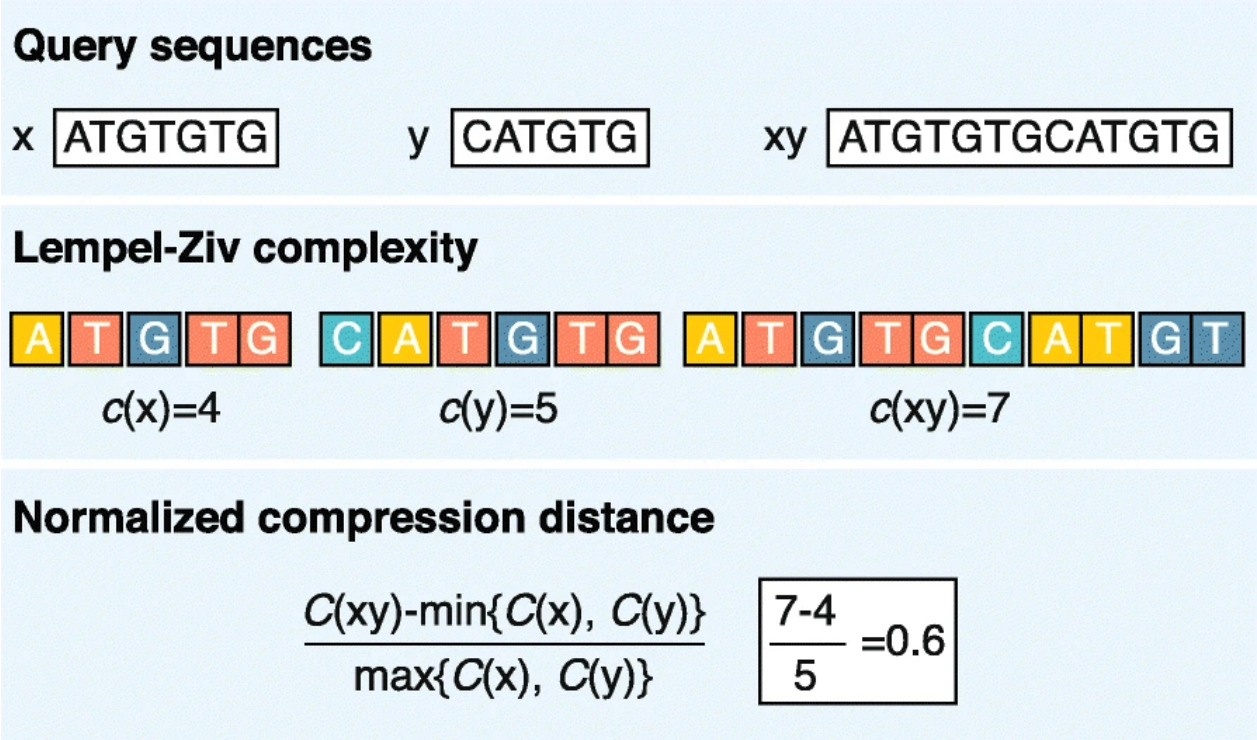


Figure 2.2 – Exemple de calcul *alignment-free* : comparaison de deux séquences d'ADN à l'aide d'un critère issu de la théorie de l'information.

2.4.1.2 Entropie (Shannon) et mesures associées

L'*entropie de Shannon*, concept fondateur en théorie de l'information, mesure l'incertitude ou la diversité symbolique d'une séquence (Shannon, 1948). Ainsi, une suite très monotone (par exemple, la répétition du même nucléotide) présente une entropie faible, tandis qu'une répartition plus homogène des nucléotides se traduit par une entropie élevée. Il est important de souligner que les notions de complexité et d'entropie, bien que méthodologiquement différentes, sont étroitement liées : en effet, une séquence de faible complexité (p.ex. `TTTTTTTT`) affichera également une entropie plus faible qu'une séquence plus hétérogène (p.ex. `ACCTGATGT`).

La *divergence de Kullback–Leibler* (KL) (Kullback et Leibler, 1951) constitue une généralisation permettant de comparer deux distributions de symboles issues de séquences différentes. À titre d'exemple, considérons deux courtes séquences :

— Séquence $x = \texttt{ATGTGTG}$

— Comptes (C_1^x) : A(1), T(3), G(3), C(0)

— Fréquences (p_1^x) : A(0,14), T(0,43), G(0,43), C(0)

— Séquence $y = \texttt{CATGTG}$

— Comptes (C_1^y) : C(1), A(1), T(2), G(2)

— Fréquences (p_1^y) : C(0,17), A(0,17), T(0,33), G(0,33)

La divergence KL se calcule comme :

$$D_{KL}(P||Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \tag{2.1}$$

ce qui, dans l'exemple ci-dessus, aboutit à :

$$D_{KL}(P||Q) = 0.14 \log_2(\frac{0.14}{0.17}) + 0.43 \log_2(\frac{0.43}{0.33}) + 0.43 \log_2(\frac{0.43}{0.33}) = 0.24 \tag{2.2}$$

La valeur obtenue (0.24) indique une divergence relativement faible entre les deux distributions. Plus cette valeur est proche de zéro, plus les distributions sont similaires, tandis qu'une valeur élevée signale des distributions très différentes.

Ces mesures issues de la théorie de l'information constituent des outils précieux pour caractériser et comparer des séquences biologiques. L'entropie permet de quantifier la complexité locale d'une séquence, tandis que la divergence KL offre un moyen robuste de mesurer la dissimilarité entre les distributions de symboles de différentes régions génomiques.

2.4.1.3 Variantes et robustesse

De nombreuses variantes de la divergence KL et d'autres mesures d'entropie ont été développées, offrant des possibilités de comparaison plus fines entre distributions symboliques. Par ailleurs, l'emploi de pseudo-comptes est fréquent en analyse de séquences biologiques pour éviter les problèmes liés à des fréquences nulles (p. ex. symboles absents d'une séquence, mais présents dans une autre) et augmenter la robustesse des mesures.

Dans l'ensemble, les approches fondées sur la théorie de l'information connaissent un succès grandissant dans l'analyse de séquences, aussi bien pour l'étude globale des génomes que pour la détection locale de régions fonctionnelles (Vinga, 2014b).

2.4.2 Méthodes basées sur la fréquence des mots (words / *k*-mers)

Les méthodes basées sur la fréquence des mots représentent une autre approche clé pour l'analyse de séquences sans alignement. Les *k*-mers sont des sous-séquences contiguës de longueur k (p. ex. 3-mers pour des triplets). Leur distribution, souvent appelée signature génomique, se révèle spécifique d'une espèce ou d'un ensemble d'espèces (De la Fuente *et al.*, 2023).

Le schéma général de comparaison des signatures génomiques fondées sur les *k*-mers est illustré en Figure 2.3 et comprend les étapes suivantes :

1. **Génération des k-mers** : À partir d'une séquence donnée, on énumère tous les motifs de longueur k. Par exemple, pour les séquences `ATGTGTG` et `CATGTG` avec $k = 3$, les triplets `ATG`, `TGT`, `GTG` et `CAT` sont extraits.
2. **Calcul des fréquences** : On compte les occurrences de chaque *k*-mer dans la séquence, ce qui produit un profil de fréquences caractéristique.
3. **Comparaison des signatures** : Les profils de fréquences (k-mers) de deux séquences sont ensuite comparés, soit via des mesures de distance (euclidienne, Manhattan, corrélation, etc.), soit à l'aide d'approches d'apprentissage automatique (clustering, classifieurs supervisés, etc.). Une distance faible entre deux profils signale une similarité élevée des séquences.

Historiquement, Karlin et ses collaborateurs furent parmi les premiers, en 1995, à introduire la notion de « signature génomique », reposant sur la distribution de mots au sein des génomes (Kariin et Burge, 1995). Ils ont ainsi montré que des organismes proches partagent des distributions de *k*-mers similaires, reflétant des liens évolutifs ou taxonomiques. Par la suite, ce concept a été étendu à des *k*-mers de plus grande taille (Gao *et al.*, 2017).

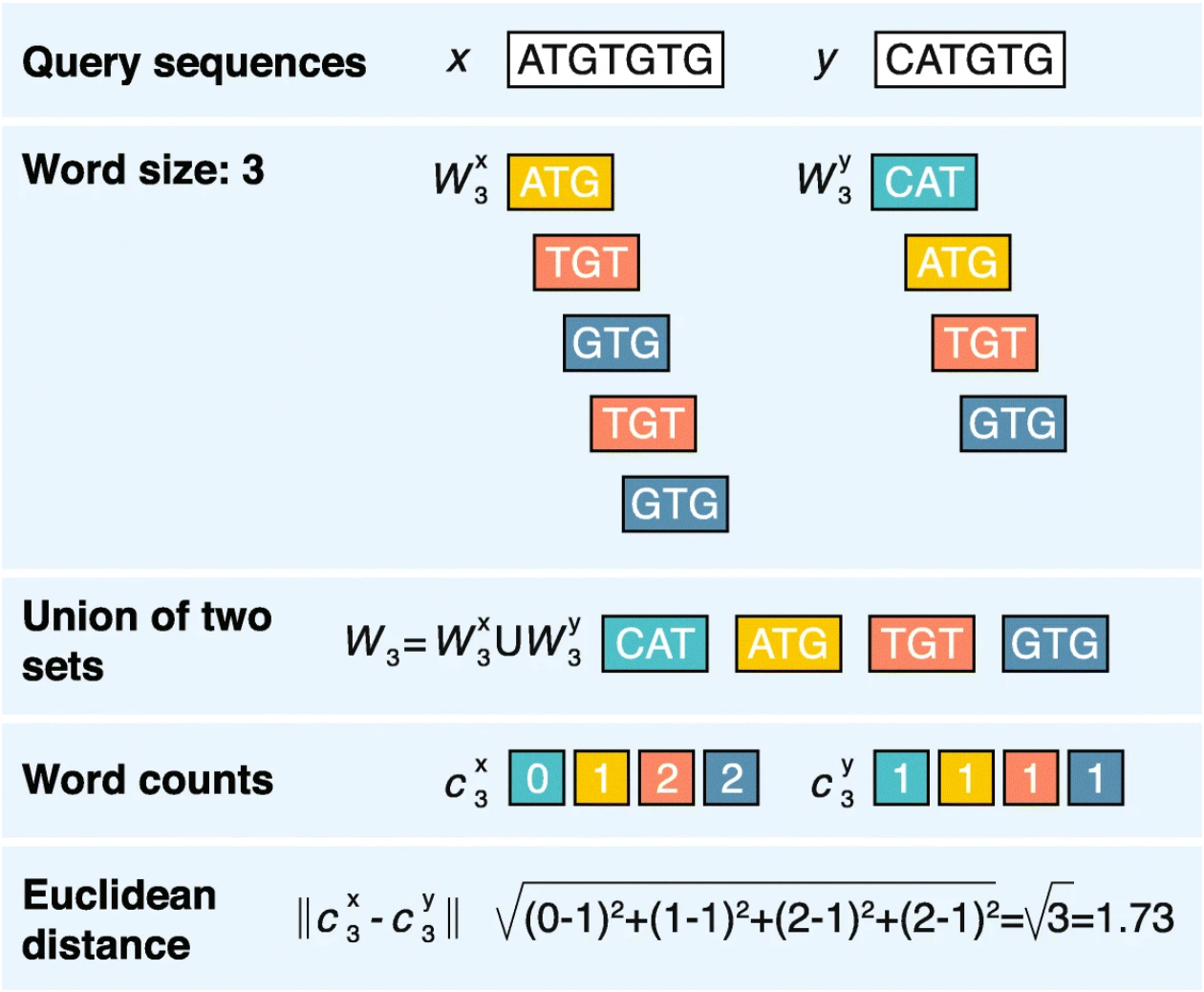


Figure 2.3 – Calcul sans alignement de la distance basée sur les mots entre deux séquences d'ADN à l'aide de la distance euclidienne.

2.4.3 Méthodes basées sur les *k*-mers dans le contexte de la classification virale

Les méthodes basées sur les *k*-mers se sont imposées comme un pilier de la classification virale *alignment-free*, à la fois pour l'assignation rapide de séquences en métagénomique et pour le sous-typage fin de virus d'intérêt clinique. L'identification de k-mers discriminants, via des analyses statistiques ou des méthodes de sélection de caractéristiques, ajoute une dimension essentielle : elle optimise la classification en isolant les éléments les plus pertinents sur le plan discriminatif, tout en préservant l'efficacité et la scalabilité des analyses et en facilitant l'interprétation des résultats.

2.4.3.1 Classification virale *k*-mer dans un contexte métagénomique

La métagénomique se définit comme l'étude de l'ensemble du matériel génétique présent dans un échantillon complexe (environnemental, clinique, etc.), sans nécessiter l'isolement ou la culture des organismes qui le composent. Dans ce cadre, où il s'agit d'identifier les séquences virales et de les classer aux rangs taxonomiques (espèce ou supérieurs), la rapidité et la précision de la classification sont cruciales pour analyser d'importants jeux de données issus du séquençage à haut débit.

Parmi les outils de référence, on peut citer :

- **VirFinder** (Ren *et al.*, 2017) : combine des signatures basées sur les *k*-mers et un modèle d'apprentissage automatique (régression logistique) pour identifier les séquences virales au sein d'échantillons métagénomiques. Ne reposant pas sur une comparaison directe avec des bases de données de référence, il se montre particulièrement efficace à la fois pour l'identification de séquences courtes ou de contigs fragmentés ($\sim$ 1 kb) et pour la détection de virus potentiellement inconnus.
- **CLARK** (*CLAssifier based on Reduced K-mers*) (Ounit *et al.*, 2015) : construit une base de données de *k*-mers uniques pour chaque taxon. Pour l'analyse d'une nouvelle séquence, l'algorithme recherche les *k*-mers correspondants dans cette base, puis applique l'algorithme du plus proche ancêtre commun (*LCA*) pour déterminer l'assignation taxonomique la plus probable.
- **Kraken** (Wood et Salzberg, 2014; Wood *et al.*, 2019) : il emploie également une base de données de *k*-mers, optimisée par des structures de type *trie* ou *hash* pour accélérer la recherche. Sa base de données étendue, incluant de nombreuses séquences virales, bactériennes et eucaryotes, permet une classification rapide et précise, même pour de petites lectures (*reads*).

2.4.3.2 Approches basées sur les k-mers pour le sous-typage viral

Au-delà de la simple détection, la caractérisation fine (ou « sous-typage ») des virus permet d'identifier les différents génotypes, clades ou lignées circulantes de rang inférieur à celui de l'espèce. Il s'agit là d'un élément critique en surveillance épidémiologique, pour la conception de vaccins et pour la personnalisation des stratégies thérapeutiques (Hemelaar, 2013). Plusieurs approches basées sur les k-mers se sont ainsi démarquées :

- **COMET** (*COntext-based Modeling for Expeditious Typing*) : s'appuie sur des modèles de Markov d'ordre variable bâtis à partir des fréquences de *k*-mers obtenues de séquences de référence ali-

gnées. Pour chaque position d'une séquence inconnue, l'algorithme calcule une log-vraisemblance, puis un arbre de décision détermine le sous-type. Principalement développé pour le HIV-1, COMET distingue tant les sous-types que les recombinants circulants.

— **CASTOR** : innove en s'appuyant sur les *k*-mers caractérisant des signatures RFLP (*Restriction Fragment Length Polymorphism*) de type II. Deux métriques principales sont calculées : (1) le nombre de sites de coupure par enzyme et (2) la moyenne quadratique des longueurs de fragments. Ces variables alimentent ensuite des classifieurs d'apprentissage automatique, avec une forte efficacité démontrée sur divers virus (HBV, HPV, HIV, etc.).

— **KAMERIS** : utilise les fréquences de 6-mers pour générer un vecteur de représentation des séquences, suivi d'une réduction de la dimensionnalité par décomposition en valeurs singulières (SVD), permettant d'obtenir une représentation plus compacte (environ 10% de la taille initiale). La classification finale s'appuie sur un SVM linéaire, offrant un équilibre entre performance et efficacité computationnelle.

— **Covidex** : application plus récente centrée sur SARS-CoV-2, reposant sur des profils de fréquences de *k*-mers intégrés dans un *Random Forest*. Cette approche *open source* a atteint une précision de 96,56 % pour discriminer 1 437 lignées Pango du SARS-CoV-2 à la fin de l'année 2021, illustrant l'intérêt de ces méthodes pour un suivi à grande échelle de l'évolution virale.

2.4.4 Identification de *k*-mers discriminants

Au-delà des outils de sous-typage, l'identification de *k*-mers discriminants vise à sélectionner uniquement les sous-séquences les plus pertinentes pour la classification. Cette réduction de dimension permet d'accélérer les calculs, de diminuer la complexité et, dans certains cas, de fournir une meilleure interprétation des résultats en mettant en évidence les motifs les plus déterminants d'un point de vue taxonomique ou clinique.

— **MISSEL** (*Motif Identification in Sequences by Supervised Evolutionary Learning*) Fiscon *et al.* (2016) s'appuie sur un algorithme génétique supervisé pour repérer des motifs discriminants dans les génomes viraux. Il exige toutefois un alignement préalable des jeux de données. Appliqué notamment à l'influenza, aux polyomavirus et aux rhinovirus, MISSEL a affiché une précision de 90 % à 100 %, soulignant la valeur de ces motifs dans divers contextes épidémiologiques ou taxonomiques.

— **CASTOR-KRFE** (*K-mer Recursive Feature Elimination*) associe l'utilisation de SVM linéaires à une éli-

mination récursive des caractéristiques, dans le but de déterminer la longueur k qui forme un ensemble minimal de *k*-mers optimisant les performances de classification dans un contexte donné. Validé par validation croisée et appliqué sur des bases de données virales (HIV-1, HCV, Ebola et Dengue), CASTOR-KRFE illustre une robustesse sur plusieurs familles virales, tout en réduisant significativement la charge de calcul durant la prédiction (Lebatteux *et al.*, 2019).

— **Autres approches de sélection de motifs discriminants :** des outils établis tels que **MEME** (*Multiple EM for Motif Elicitation*) (Bailey *et al.*, 2010) et **STREME** (*Simple Tool for Rapid Exact Motif Elicitation*) (Bailey, 2021) permettent également de repérer des motifs discriminants dans les séquences virales. MEME emploie l'algorithme EM (Expectation-Maximization) pour découvrir des motifs sans a priori structurel, tandis que STREME s'appuie sur une approche exacte et rapide fondée sur des tests statistiques rigoureux. Conçues principalement pour des scénarios de classification binaire, ces méthodes nécessitent toutefois des adaptations de type « un contre tous » (*one-versus-rest*) pour traiter efficacement la nature multiclasses de la classification virale.

2.4.5 Limitations actuelles des approches basées sur la fréquence des mots

Malgré leurs apports significatifs, les approches basées sur les *k*-mers présentent plusieurs limitations importantes :

1. **Sensibilité à la variabilité génétique :** Ces approches montrent leurs limites face à la grande diversité des virus, particulièrement pour les séquences divergentes ou les classes virales sous-représentées (Lebatteux et Diallo, 2021).
2. **Contraintes computationnelles :** La dimension des tables de *k*-mers (4^k combinaisons) peut nécessiter des ressources considérables (>480 Go) pour les analyses métagénomiques (Liu *et al.*, 2024). Même s'il est possible d'utiliser des serveurs de calcul haute performance, leur mise en place n'est pas toujours adaptée aux biologistes ou virologues, car elle requiert une expertise informatique spécialisée.
3. **Interprétation biologique limitée :** L'absence de mécanismes permettant d'évaluer la pertinence biologique des motifs discriminants complique la mise en relation avec des caractéristiques biologiques interprétables (Solis-Reyes *et al.*, 2018).

Ces limitations motivent le développement d'approches novatrices visant à sélectionner des sous-ensembles pertinents de *k*-mers plutôt que d'exploiter l'ensemble des combinaisons possibles (4^k), tout en assurant une meilleure intégration des aspects biologiques et évolutifs à travers des modèles prédictifs robustes.

Le chapitre suivant introduit KEVOLVE, une approche combinant un algorithme génétique avec des méthodes d'apprentissage automatique pour identifier des sous-ensembles minimaux de k-mers discriminants et construire des modèles prédictifs pour les séquences virales.

CHAPITRE 3
COMBINING A GENETIC ALGORITHM AND ENSEMBLE METHOD TO IMPROVE THE CLASSIFICATION OF VIRUSES

3.1 Abstract

Genomic features identification is an important step toward machine learning (ML) derived classifiers. Designing robust feature identification methods can increase the performance of ML algorithms and reduce high dimensional data. The Classification of genomes from nucleotides sequences often requires an exponential feature space according to the size of the genomes. These feature spaces should also capture the abundant mutations affecting the genomic sequences over time of several complex viruses such as the human immunodeficiency virus (HIV). One way of overcoming such challenges could be to design an accurate and efficient algorithm to capture a bag of minimal subsets of genomic features based on k-mers that could represent the most likely alternative and evolutionary space. In this article, we introduce Kevolve, a new method based on a Genetic Algorithm (GA) including a ML kernel to extract a bag of minimal subsets of genomic features maximizing a given classification score threshold. Kevolve is coupled with an ensemble prediction model based on support vector machines (SVMs) and is applied on the classification of HIV genomic sequences. Subsets of genomic features identified reduce the size of initial feature matrices by 99% and models based on them outperform state-of-the-art HIV predictors. The results also show that Kevolve is less sensitive to high mutation rates of the virus. The source code is available at : `https://github.com/bioinfoUQAM/Kevolve`

Keywords : Genomic features identification, Virus classification, Machine learning, Alignment-free, Genetic Algorithm

3.2 Introduction

Genomic feature identification is an important process to nucleotide sequence classification using ML. It is a feature selection problem, associated with data representation, ML algorithms performances, reducing dimensionality of data and making models more understandable Li *et al.* (2018). These issues are defined as the task of identifying a subset of features that could discriminate between classes. Kira et Rendell (1992). Identification of optimal subset of features is a complex problem due to the large size of the involved search space where the number of potential solutions is 2^n-1 and n is the number of features Dash et Liu (1997).

In alignment-free genomic classification, k-mers (nucleotide subsequences of length k) are mainly used for sequence representation Bonham-Carter *et al.* (2014); Solis-Reyes *et al.* (2018); Lebatteux *et al.* (2019). This yields 4^k possible number of features conducting to computational challenges, where 4 corresponds to the cardinality of the nucleotide alphabet $\mathcal{A} = \{A, C, G, T\}$. These k-mers based features are defined as genomic signatures which are a species-specific pattern known to be pervasive throughout the genome Randhawa *et al.* (2020). Identifying relevant subsets of genomic signatures based on k-mers can allow the discovery of motifs that are associated with several roles Chiu *et al.* (2020) and improving genomic and metagenomic classification Storato et Comin (2020).

In a previous study, an alignment-free method was proposed to identify a set of genomic signatures based on k-mers Lebatteux *et al.* (2019). It highlights the importance of a minimal set of features to discriminate between groups of viral genomic sequences. In parallel, it has been shown a way of using GA for the processing of high-dimensional data in order to extract sets of features based on k-mers Thomas *et al.* (2019). However, these studies are limited to one minimal feature set based on k-mers, instead of exploring the suboptimal sets of feature space. This limitation can constitute a major drop of accuracy when mutations are present within the genome and barely represented in the identified features. Therefore, in the context of classification of viruses including those with high mutation rates, it is important to develop a method allowing the identification of bag of genomic feature subsets based on k-mers with minimal size satisfying a performance threshold. The identified subsets of features could then be used to build a robust predictive model taking into account mutation and genetic recombination.

In this article, we present Kevolve : a method to classify nucleotide sequences taking into account several alternatives of suboptimal genomic features based on k-mers in order to be robust to mutations. It combines evolutionary processes and ML to, respectively, extract bag of minimal relevant subsets of features and classify virus genomes. Kevolve was assessed with the HIV genomes (a high mutation rate virus). The article is organized as follows. We first detail the units of Kevolve. Then, we describe the data sets and the assessment processes. Finally, the results are discussed and followed by a conclusion.

3.3 Methods

3.3.1 Overview of the method

Kevolve is designed to address the problem of classifying genomic sequence while taking into account a bag of optimal subsets of genomic features based on k-mers. Kevolve implementation is based on two main units : 1) an identification unit that provides subsets of features that are minimal and likely to provide the best performance metrics ; and 2) a prediction unit that applies an ensemble classifier using the subsets of features.

The Algorithm 1 of Kevolve addresses the sub-problem of identifying bag minimal of feature sets while maximizing a given performance metrics through a cross-validation evaluation. It initializes its search with reduced subsets of features and evolves these subsets using GA. This process relies on a progressive increment of the subset size and an adjustment of the probability of selecting a given set of features. This mechanism is combined with an efficient configuration of mutation and crossover processes Mirjalili (2019). We compute a fitness function to identify the optimal sets of features. This function is derived from the assessment of a ML model in order to compute a given performance metric (e.g. F1-score, Area Under the Curve, ...). Once the bag of feature subsets is identified, it is provided as input to the algorithm 2 to build an ensemble classifier.

3.3.2 Kevolve : genomic feature identification unit

Algorithm 1 takes as inputs : a matrix X of size ($n\_samples, nTotalFeatures$) characterizing a set of raw genomic data $\mathcal{D}$, where $n\_samples$ is the number of sequences $\in \mathcal{D}$ and $nTotalFeatures$ is the number of genomic features based on k-mers computed from $\mathcal{D}$. This set of k-mers is taken as the variable $features$. The target values y are represented by a $1d$ array where the i-th entry corresponds to the target of the i-th sample (row) of X ; $nSubsets$ is the number of feature subsets to be generated at each iteration ; $nFeatures$ is the number of features that compose each subset and is induced to evolve during the execution of the algorithm ; $nIterations$ and $nSolutions$ define stopping criteria for the algorithm search. The $nMutations$ and $nCrossovers$ referring to the probabilities that mutation (random substitution of a feature by another within a subset) and crossover (exchange of some features between two subsets) phenomena occur in a bag of subsets, respectively. Finally, $score$ is used to specify the minimum score that a subset of features must achieve during the assessment to be a solution.

Algorithm 1: Identification unit

Input : X : feature matrix representing the sequences
y : vector of target variables relative to X
$features$: features based on the k-mers composing X
$nSubsets$: number of feature subsets to generate
$nFeatures$: number of features per subset
$nSolutions$: number of solutions to stop the search
$nIterations$: number of maximum generations of subsets
$nMutations$: probability of mutation
$nCrossovers$: probability of crossover
$score$: score to be achieved to define a solution
Output: $solutions$: list of subsets returned by the algorithm

1 **Begin**
2 $solutions \leftarrow \emptyset$
3 $objective \leftarrow False$
4 $temporaryBagSubsets \leftarrow \emptyset$
5 $weights \leftarrow initialWeights(features)$
6 $X \leftarrow varianceThreshold(X, threshold = 0.01)$
7 **while** n ***in*** $nIterations$ ***or*** $solutions.length < nSolutions$ **do**
8 $bagSubsets \leftarrow generateBagSubsets(nSubsets, features, nFeatures, weights)$
9 $bagSubsets \leftarrow bagSubsets + temporaryBagSubsets$
10 $scores \leftarrow fitnessCalculation(X, y, bagSubsets)$
11 $solutions \leftarrow checkSolutions(scores, score, bagSubsets)$
12 $objective, nFeatures \leftarrow checkObjective(solutions, nFeatures)$
13 $selection \leftarrow selection(scores, bagSubsets)$
14 $weights \leftarrow updateWeights(weights, selection)$
15 $selection \leftarrow crossover(selection, nCrossovers)$
16 $selection \leftarrow mutation(selection, nMutations)$
17 $temporaryBagSubsets \leftarrow selection$
18 **end**
19 **End**

The algorithm starts with the initialization of variables to default values from the line 1 to 6. It assigns the initial value 1 for of each feature (k-mer) to $weights$ table (line 5). These weights will influence the probability of a feature being selected during the bag of subsets generation step (line 8). Next the prepossessing step removes quasi-constant features (with 99% of similarity) at line 6. Then, the main loop (lines 7 to 18), generates the bag of subsets that evolves through several iterations until converging to a valid generation of solutions. The first step generates a bag of feature subsets (lines 8 and 9). At initial generation, the bag is composed of $nSubsets$ subsets formed by $nFeatures$ k-mers. Each subset is represented by a feature index vector and $nFeatures$ will be tune automatically during the iterations of the algorithm (line 12). In the next generation, the bag of feature subsets ($bagSubsets$) retains several optimal subsets from the previous generation and inserts new ones. During iterations, k-mers selection is impacted by the update of the associated $weights$ (line 14).

For each subset, the $fitnessCalculation$ function consisting in stratified cross-validation evaluation of a model is applied. The fitness scores returned (line 10) are the weighted and unweighted $F1$-$score$. The model consists of : the derived submatrix of X composed only of the columns corresponding to the k-mers of the subset under assessment, the y vector of target variables relative to X and a given supervised ML

algorithm. Due to their robust performance in previous viral genomic classification tasks Lebatteux *et al*. (2019), we opted for linear SVM as ML algorithm. After the evaluation step, if one or more sets have an associated score higher than $score$, they are included in the table of $solutions$. If the algorithm finds a solution for the first time, the variable $objective$ is set to True to limit the search to solutions composed of $nFeatures$ (line 12). If no solution is identified, $nFeatures$ is incremented to increase the size of the next generation subsets.

The next step is the selection of subsets for the next iteration (line 13). The 50% of the subsets with the highest fitness score will belong to the next generation. The weights are then increased so that the probability of the most relevant k-mers will be featured in the subsequent subsets. Then comes the uniform crossover (line 15) and mutation (line 16) processes where $nCrossovers$ and $mutation$ are respectively set to 0.2 and $1/nFeatures$. At the end of each iteration (line 17), the $selection$ is saved in the variable $temporaryBagSubsets$ to be part of the next generation. This one is completed by new subsets generated from the updated parameters (line 8 and 9). Finally, if stopping criterion ($nIterations$ or $nSolutions$) is reached, the bag of k-mers sets in $solutions$ is returned.

3.3.3 Kevolve : model building/prediction unit

The building/prediction unit (Algorithm 2) takes as input the $solutions$ of the identification unit : the training samples matrix X_train, the associated target vector y_train and ultimately X_test which is a testing samples matrix.

Algorithm 2: Model building/prediction unit

Input : $solutions$: feature set list
svm : supervised machine learning algorithm
X_train : train samples matrix
y_train : train target variable
X_test : test samples matrix
Output: $ensembleModel$: prediction model
y_pred : predicted target variable

1 **Begin**
2 $ensembleModel \leftarrow \emptyset$
3 **foreach** $solution \in solutions$ **do**
4 $model \leftarrow svm.fit(X\_train,\ y\_train,\ solution)$
5 $ensembleModel.add(model)$
6 **end**
7 $y\_pred \leftarrow \emptyset$
8 **foreach** $x \in X\_test$ **do**
9 $y\_pred \leftarrow ensembleModel.predict(x)$
10 **end**
11 **End**

The first step is the building of the SVM-based ensemble model (line 3 to 6). For each subset of features $\in$

$solutions$, a linear SVM model is trained and embedded in the ensemble model. For each SVM model, class membership probabilities are computed using the SVM probability output method Wu *et al.* (2004). Once the ensemble model is trained, the test instances can be predicted (lines 7 to 10). For each instance i, each SVM s of the ensemble model will generate a probability density P_{i_s} of belonging to the different classes for which the model was trained. The predicted class c_i associated to an instance i will be the one with the highest sum of probabilities $p_{i_{s,c}}$, such that $c_i = \arg\max_c(\sum_s p_{i_{s,c}})$. The predicted classes are then added to the table y_pred (line 9). Finally, the algorithm returns the trained prediction model ($ensembleModel$) and the list of predictions y_pred

3.4 Datasets

To assess Kevolve, we address the problem of typing and subtyping of HIV. It is responsible for acquired immunodeficiency syndrome (AIDS) and represents a pandemic of unprecedented genetic and geographical complexity due to several phenomena including : high mutation and recombination rates Taylor *et al.* (2008). Accurate classification of HIV unknown strains from sequenced fragment still constitute a challenge to the scientific community. We focus on HIV-1 classification at both complete genomes (CGs) and fragments covering pol gene having a minimum size of $1\,000$ bp and maximum size of $2\,100$ bp. In order to perform a complete study, we included in the training set all subtypes with at least 4 instances available on : `https://www.hiv.lanl.gov/` (LANL). An initial training set of 380 complete HIV genomes (CG train set), and a second one composed of $1\,160$ pol fragments (pol train set) were collected. The requirements of 4 instances minimum per class for a tool based on ML increases the difficulty of building a powerful prediction model. Kameris fixed during their evaluation the lower bound (18 instances) of the required instances for a class to be considered Solis-Reyes *et al.* (2018). This choice implies the evaluation of only 6 subtypes in their comparative study. Similarly, the analyses performed in Pineda-Peña *et al.* (2013) and Struck *et al.* (2014) involve 15 and 16 subtypes respectively. Although the selected classes represent the majority of the HIV universe, the literature agrees that these types of classes are predominantly well predicted by the prediction tools Solis-Reyes *et al.* (2018); Struck *et al.* (2014); Pineda-Peña *et al.* (2013). Therefore, it is interesting to study the behavior of these tools in the dominant classes as well as in the underrepresented ones.

Subsequently, we built a CG test set consisting of $8\,302$ CGs with an average length of $8\,943$ nucleotides, covering 24 HIV subtypes (A1, B, C, D, F1, F2, G, H, J, 01_AE, 02_AG, 04_cpx, 06_cpx, 07_BC, 08_BC, 11_cpx, 12_BF, 13_cpx, 14_BG, 18_cpx, 24_BG, 29_BF, 35_AD and 42_BF). We also designed a pol test set with 27,739

pol fragments with an average length of $1\,363$ nucleotides covering 30 HIV subtypes (A1, A2, B, C, D, F1, F2, G, H, 01_AE, 02_AG, 03_AB, 06_cpx, 07_BC, 08_BC, 09_cpx, 11_cpx, 12_BF, 13_cpx, 14_BG, 18_cpx, 19_cpx, 20_BG, 24_BG, 25_cpx, 31_BC, 35_AD, 37_cpx, 42_BF and 47_BF). For the test sets, we limited the number of instances of the overrepresented classes to $5\,000$, to ensure a representative set of instances while limiting the imbalance effect of the majority classes against the minority classes. Finally, only complete, curated and well-annotated sequences were extracted from LANL.

3.5 Evaluation

3.5.1 Identification of a bag of minimal subsets features

First, we studied the ability of Kevolve to extract minimal subsets of genomic features that maximize classification performance. From the CG train set, consisting of 36 subtypes, the occurrences of k-mers of length $k \in [1, 10]$ were used as features, yielding 449 550 features and a search space of $2^{449\,550}$ for this dataset. From pol train set consisting of 30 HIV subtypes, the occurrences of k-mers of length $k \in [1, 7]$ were used as features to characterize the sequences. This represents $17\,264$ features, with a search space of $2^{17\,264}$ potential subsets of features. The boundaries of the k-mers range were set based on the identified optimal lengths in Lebatteux *et al.* (2019). The initial bag of subsets was set to $nSubsets = 2\,500$ and $nFeatures = 3$. The parameters $nIterations$ and $score$ were respectively set 250 and 0.999 for CGs. For pol fragments, $score = 0.985$ and $nIterations = 350$.

3.5.2 Benchmark study

The identified bag subset features and the training/test sets are then given as input to algorithm 2 to build the prediction models and make the predictions. The returned predictions are then used in a comparative study with the most popular HIV prediction tools. The first tool included in the comparative study is COMET (COntext-based Modeling for Expeditious Typing). It is an alignment-free typing method for HIV-1 and other viruses Struck *et al.* (2014) inspired by the Partial Matching compression algorithm. The second tool is the latest version (3.46) of REGA, the reference predictor for HIV genomic sequences. It combines phylogenetic analyses with bootscanning methods for the genetic subtyping of HIV-1 sequences Pineda-Peña *et al.* (2013). The last evaluated tool is Kameris, an alignment-free method based on k-mers using a supervised learning approach. Kameris had shown outperforming prediction performance over reference tools on several subtypes of HIV-1 Solis-Reyes *et al.* (2018). Consistent with the best performance obtained by Kameris in their

evaluation Solis-Reyes *et al.* (2018), we set its k-mers length to 7. For REGA, COMET and Kevolve, the training sequences are independent of the test sequences. Regarding Kameris, its pipeline only integrates the automatic download of the LANL training sequences which may have led to an overlap between its training set and the test sets. This bias could imply a risk of overfitting that should be considered when interpreting the results.

3.5.3 Performance metrics

The performance metric used in the fitness function of the GA to evaluate the sets of features in the generated bag of subsets is the $F1\text{-}score$ defined as follows : $F1\text{-}score = (Precision * Recall)/(Precision + Recall) * 2$, where $Precision = TP/(TP+FP)$ and $Recall = TP/(TP+FN)$. TP, TN, FP, and FN are respectively the number of true positive, true negative, false positive and false negative predictions. In the comparative study, for each class i, we computed the percentage of correct classification (also called recall or sensitivity) given by the formula : $\%\_correct_i = n_{true}/n_i * 100$, where n_{true} is the number of instances correctly predicted and n_i is the total number of instances present in class i. This choice is consistent with the comparative study performed in Solis-Reyes *et al.* (2018). Moreover, REGA, COMET and Kameris can predict an instance to be unassigned (considered as a misprediction), yielding complete information on TP but incomplete for FP. In addition, for each prediction tool we specify the percentage of correct prediction on a weighted average basis, to reflect a performance score according to the natural distribution of HIV subtypes : $\%\_correct\_total_{weighted} = n\_total_{true}/n_{total} * 100$ where n_total_{true} is the number of instances correctly predicted and n_{total} is the total number of instances present in the dataset. We also compute the unweighted average to remove the dominance effect of the majority subtypes in order to highlight the ability of prediction tools on minority subtypes prediction : $\%\_correct\_total_{unweighted} = \sum_{i=0}^{n} \%\_correct_i/n$ where n is the number of classes.

3.5.4 Synthetic mutation

For the purpose of assessing the behaviour and robustness of the predictive tools against the mutations that may occur, we have set up an evaluation framework inspired by Struck *et al.* (2014). For each sequence of the datasets, a new synthetic sequence is generated with a percentage of noise. The noise involves a random replacement of nucleotides by others in each dataset. Several synthetic data sets were generated with mutation occurrences of 0.5%, 1%, 3% and 5%. The selected percentages provide coverage of *in vivo* HIV mutation rates Taylor *et al.* (2008). We are aware that this type of mutation, assuming a uniform random

distribution, is not representative of biological reality. However, this choice reflects the breaking of the biological structure of the sequences in order to increase the difficulty of its prediction by the classifiers.

3.6 Results and Discussion

3.6.1 Identification of minimal feature sets

For CG train set, after 250 iterations, the Kevolve identification unit identified $1\,634$ feature subsets of size $nFeatures = 96$, satisfying a performance in terms of weighted and unweighted F1-score ≥ 0.999 during a cross-validation evaluation. Kevolve succeeded to identify from a universe of $2^{449\,550}$ potential subsets of features several subsets representing less than 1% of the size of the original matrix while maintaining a performance very close to maximal value of 1. For the pol train set, Kevolve identified 64 feature subsets of size $nFeatures = 243$ in the space of $2^{17\,264}$ possibilities after 350 iterations. The identified sets reduce the original matrix size by almost 99% while maintaining a performance greater than 0.985 in terms of weighted and unweighted $F1\text{-}score$ during a cross-validation evaluation. In summary, Kevolve shows its ability to extract several subsets of features in a high-dimensional space. These subsets considerably reduce the dimension of the original feature matrix while providing the necessary information to discriminate between the different classes.

3.6.2 Comparative study

To validate the previously identified feature subsets, they are given as input to the algorithm 2 to build the ensemble prediction models. The model for CGs is built by combining the CG training set and 100 subsets with the lowest similarity among the $1\,634$ identified by Kevolve. The model for pol fragments is built using the training set pol and the 64 subset of identified features. Then, these models are assessed by predicting the CG test set and the pol test set. The prediction results obtained by Kevolve are compared to those of REGA, COMET and Kameris. The A2 and 03_AB subtypes are not included in the Kameris evaluation because their model has not been trained to predict these classes.

3.6.3 Prediction of complete genomes

The overall prediction of the CG test set ($8\,302$ CGs) shows for all prediction tools, a weighted correct classification score above 95% for each percentage of mutation (Fig.3.1). The best performances are achieved by Kevolve, which is close to 100% for all mutation rates. Kameris obtains results close to Kevolve for mutation

rates from 0% to 1%. At 3% mutation, the score drops by about 1%. At 5% mutation Kameris underperforms REGA with a score of 95.7%. This decrease is explained by prediction errors for subtypes B, C and 01_AE. Regarding REGA, it shows stable performance for mutation rates from 0% to 1%, obtaining scores above 99%. A slight drop occurs at 3% mutation, reaching 97.9% correct classification at 5% mutation. This is explained by some B and C subtypes predicted as unassigned or B/C recombinants. Finally, COMET has the lowest performance, about 96% correct prediction, for all mutation rates. Misprediction occurs in the same subtypes as Kameris, which are mostly predicted as unassigned.

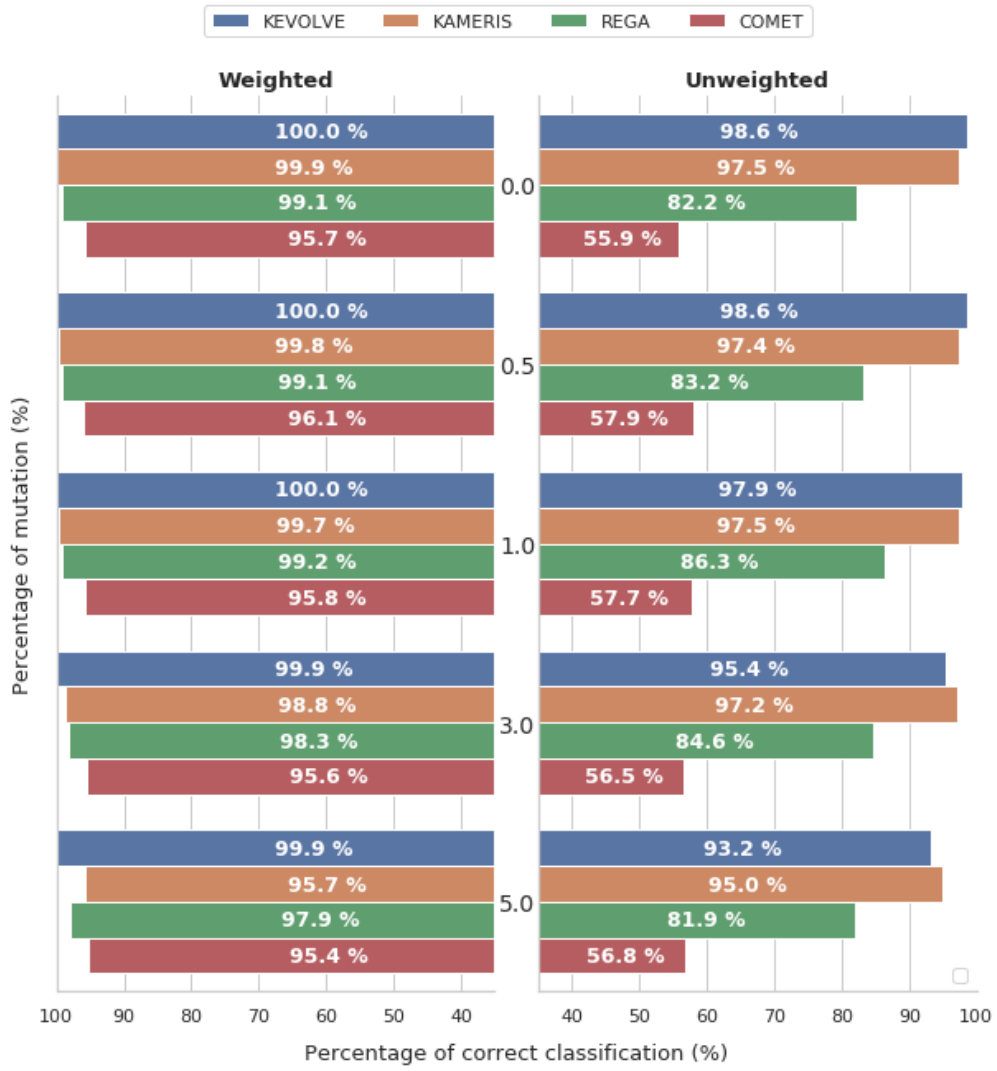


Figure 3.1 – Trends of correct classification of prediction tools based on mutation rate variation in the $8\,302$ HIV complete genomes

Regarding the unweighted correct performances, we highlight a significant decrease of three predictors

(Fig.3.1). Kevolve and Kameris performed the best with scores above 97% for mutation rates from 0% to 1%. From 3% mutation, the performance of Kevolve is about 1.8% lower than Kameris. This is explained by prediction errors of the recombinants 06_cpx, 12_BF, 14_BG and 29_BF predicted as pure B subtype. For Kameris, the decrease in performance is associated with recombinants of 01_AE, 06_cpx, 13_cpx and 18_cpx sometimes predicted as unassigned or subtype B. REGA performance is also negatively impacted (between 82.2% and 86.3%). It often predicts the F2 subtype as F1 or recombinant F. The recombinants 06_cpx, 12_BF and 29_BF are also sometimes predicted as other recombinant subtypes. Finally, COMET had the lowest performance (about 56% for the different mutation rates). COMET is inefficient in predicting F2, J, 04_cpx and 18_cpx subtypes. In addition, several recombinants 06_cpx, 07_BC, 08_BC, 11_cpx, 12_BF, 13_cpx, 14_BG, 29_BF, and 42_BF are predicted as unassigned.

3.6.4 Prediction of pol fragments

Regarding the weighted prediction results of the $27\,739\,pol$ fragments presented in Fig. 3.2, Kevolve gets the best scores, followed by REGA with correct prediction percentages above 97% and 96% respectively for the different mutation rates. Kevolve errors are associated with a minority of B subtypes predicted as recombinant B and some misprediction of recombinant 01_AE. For REGA, the prediction errors involved subtypes C, D, G, 02_AG, 07_BC and 12_BF sometimes predicted as recombinants or unassigned. COMET achieves a correct classification of 94.1% at 0% mutation and drops to 83.9% at 5% mutation. This decrease is mainly due to many 02_AG recombinants predicted as subtype A or G. A significant number of 07_BC recombinants are also predicted to be subtype C. Finally, Kameris has the lowest correct prediction performance (91.3% to 0% mutation and 68.7% to 5%). Kameris predicted many subtypes A1 as AD, B as BF, C as AC or BC, D as AD, 01_AE as B and 02_AG as G.

With regard to the unweighted results presented in Fig.3.2, we notice for all predictors a decrease in the score compared to the weighted performance. Kevolve achieves the best scores (over 96.5% for mutation rates below 1%). From 3% mutation, Kevolve's performance begins to drop to reach 90.5% at 5% mutation. This decrease is explained by incorrect predictions for the recombinants 09_cpx, 14_BG, 24_BG and 47_BF sometimes predicted as B, D or G. REGA has the second best performance with a strong resistance to mutation insertion. It obtained a score of 89% at 0% mutation and 87.1% at 5% mutation. The decrease in performance is mainly explained by the subtypes A2 predicted as A1 or unassigned, F2 predicted as F1, 12_BF predicted as B or unassigned and 14_BG predicted as G.

Figure 3.2 – Trends of correct classification of prediction tools based on mutation rate variation in the $27\,739$ HIV pol fragments.

Finally, the lowest scores were obtained by COMET and Kameris with correct predictions of 73.3% and 77% respectively for a mutation rate of 0% and 53.7% and 49.5% for a mutation rate of 5%. Both, have difficulty predicting recombinants 06_cpx, 09_cpx, 13_cpx, 14_BG, 18_cpx, 19_cpx, 20_BG, 24_BG, 35_AD, 37_cpx and 42_BF which are frequently predicted as pure or unassigned subtypes.

3.6.5 Overall results summary

In general, Kevolve performs better with mutation rate insertion and REGA shows stable performance in this evaluation framework. This robustness is supported by the fact that REGA is based on sequence alignments

that will accurately capture a very large part of the genome information. However, this type of approach remains costly in terms of temporal complexity. Kevolve, Kameris and COMET can predict several sequences per second. For REGA, it requires an average of 2 min 45 sec to predict a HIV CG. COMET performs well in predicting pure HIV subtypes. However, it has difficulties in predicting minority recombinants, which negatively impacts the non-weighted scores. Kameris obtains high performances for the prediction of CGs and shows a decrease in scores for the prediction of pol fragments.

Kameris and Kevolve use prediction models based on linear SVM with L2 regularization and features based on k-mers. However, Kameris shows very weak resistance to the mechanism of adding mutations within sequences. The notable difference between Kameris and Kevolve is how the k-mers are selected and used. Kameris authors used a complete matrix of features based on one size of k-mers and performed a dimensionality reduction using truncated singular value decomposition to reduce the vectors to 10% of the average number of non-zero entries of the feature vectors Solis-Reyes *et al.* (2018). For Kevolve, we chose to identify few discriminant k-mers subsets rather than performing a feature vector compression. According to the results of the comparative study, this feature selection method seems to be more robust to noise and generalizes better than Kameris dimensionality reduction based method. Finally, the prediction of the 14_BG pol fragment caused a significant problem for the prediction tools. Its pol region is assimilated to a pure G genome and the recombination with B pure genome chain occurs in the env region of the genome (`https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html`). The pol region is theoretically insufficient to correctly classify such a type of recombinant. Despite this, Kevolve was able to correctly predict such a type of fragment in 95% of cases before random mutation insertion. We anticipate that over time, mutations may have produced signatures associated with subtype 14_BG within the pol fragments.

3.7 Conclusion

We propose Kevolve, a method coupling GA with ML that identifies bag of minimal subsets of genomic features based on k-mers. The identified feature subsets are used to build ensemble prediction models based on SVMs. We assessed the models in a comparative study against the gold standard HIV sequence prediction tools. The results show that Kevolve outperforms the state-of-the-art methods in HIV prediction (REGA, COMET and Kameris). Kevolve performs better in minority subtypes such as recombinants. In addition, Kevolve highlights a robust and steady correct prediction when accounting for mutation rates from 0.5% to 5%. Kevolve generalize well within several viral datasets such as hepatitis B/C, Coronavirus, Dengue

virus and Ebolavirus (not shown in this article). Even though the prediction tools provide an excellent performance in classifying pure subtypes, an effort is still required to improve the prediction capacity of the recombinant viruses. These minority subtypes of HIV are often not considered in studies or go unnoticed by the single use of weighted performance metrics. Actually, we are extending Kevolve to improve the precision on subset identification as well as search speed. Finally, we also plan to add an extension to extract a biological meaning to the identified k-mers.

3.8 Bilan et perspectives

Ce chapitre a présenté la conception de KEVOLVE, une méthode permettant d'identifier des k-mers discriminants pour la classification des séquences virales. Le caractère discriminant de ces k-mers provient des changements qu'ils ont subis au cours de l'évolution, tels que des substitutions, des délétions ou des insertions, modifiant leur forme selon les différentes classes de séquences virales. Le chapitre suivant s'intéresse au développement d'une approche permettant de caractériser ces variations de k-mers afin de mieux comprendre leur capacité discriminative. Dans le cas des régions codantes, cette analyse permettra également d'évaluer l'impact de ces variations sur les séquences protéiques qu'elles encodent.

CHAPITRE 4

KANALYZER : A METHOD TO IDENTIFY VARIATIONS OF DISCRIMINATIVE k-MERS IN GENOMIC SEQUENCES

4.1 Abstract

Discriminative k-mers are unique genomic regions that characterize a given viral family, genus, species, or variant. Most existing algorithms for identifying discriminative k-mer sets are limited to returning raw sub-sequences. However, to explain the discriminative properties of a given k-mer for specific taxonomic groups of viruses, it is important to identify the variations (nucleotide sequences derived from an initial k-mer having undergone one or more nucleotide changes) of this k-mer that occur in other groups of viruses. These variations as well as their frequencies of occurrence, their genomic location and their potential influence on biological functions represent important insights to understand the classification process. In this article, we introduce KANALYZER, a novel algorithm to identify variations of discriminative k-mers and associated information according to viral taxonomy. The algorithm was assessed to identify k-mer variations in both simulated and real viral sequence sets. In these evaluations, KANALYZER correctly and quickly identified over 95% of the variations and associated information. KANALYZER algorithm is integrated directly into CASTOR-KRFE discriminative k-mers identification tool pipeline. The source code, detailed results and data to reproduce the experiments are available at : `https://github.com/bioinfoUQAM/CASTOR_KRFE`.

Keywords : Discriminative k-mer, Variation identification, Viral sequences, Pairwise alignment, Parallel computing.

4.2 Introduction

Identifying and functional characterization of discriminative k-mers is an important topic in viral sequence study. Discriminative k-mers are defined as unique genomic regions that allow the identification of the family, genus, and species of a viral organism Bui et Wei (2020). Discriminative k-mers are essential in classifying genomic and metagenomic sequences Ounit *et al.* (2015). They can support genomic analysis to discriminate between distinct groups of variants Voichek et Weigel (2020), explain their evolutionary history Randhawa *et al.* (2020), elucidate the structure of viral quasispecies Lau *et al.* (2021), and help to characterize functional properties of viral gene products Slezak *et al.* (2020). Several approaches exist to identify

relevant discriminative k-mers. These approaches includes : 1) identifying discrepant regions Fiscon *et al*. (2016) from multiple sequence alignment ; 2) machine learning and use feature selection methods to identify minimal sets of k-mers Lebatteux *et al*. (2019); Lebatteux et Diallo (2021) ; 3) performing statistical testing to select sub-sequence based on a p-value Bailey (2021) ; 4) building discriminating k-mers databases based on a hash table by scanning the reference genomes and storing the k-mers that are genome-specific Ounit *et al*. (2015). All these approaches only give an output of raw discriminative k-mers. However, from a biological perspective, it is essential to understand the discriminative properties of a k-mer for specific taxonomic groups of viruses. Suppose a k-mer presents a discriminative potential for a specific viral species. In that case, the presence of variations of this k-mer (nucleotide sequences derived from an initial k-mer having undergone one or more nucleotide changes) in other viral species must be assessed. Hence, to understand the classification process, it is essential to identify the variations derived from initial discriminative k-mers, the viral species that are associated with these variations, their frequency of occurrence, their genomic localization, and their potential influence on biological function. To address this issue, we have designed KANALYZER, an algorithm combining search functions based on pairwise alignment and parallel computing. This algorithm identifies discriminative k-mers variations and associated information according to a set of viral nucleotide sequences. In this article, we first introduce the KANALYZER algorithm and the problem it attempts to address. In a second step, we describe the assessment framework, including a first evaluation based on simulated sequences and a second one based on actual sets of viral sequences. Third, we discuss the results obtained from the different evaluations. Finally, we highlight the contribution of our approach and the perspectives for improvement.

4.3 Methods

4.3.1 Problem statement

Given a set of nucleotide sequences D belonging to different classes (i.e., sub-species), a set of discriminative k-mers S associated with D, and an annotated reference sequence A related to the organism of sequences belonging to D, identify :

- the variations of the discriminative k-mers belonging to S in the sequences of D,
- the genomic location of the discriminative k-mers / variations,
- the potential amino acid change occurring if located in a protein coding region,
- and the frequency according to the sequence class.

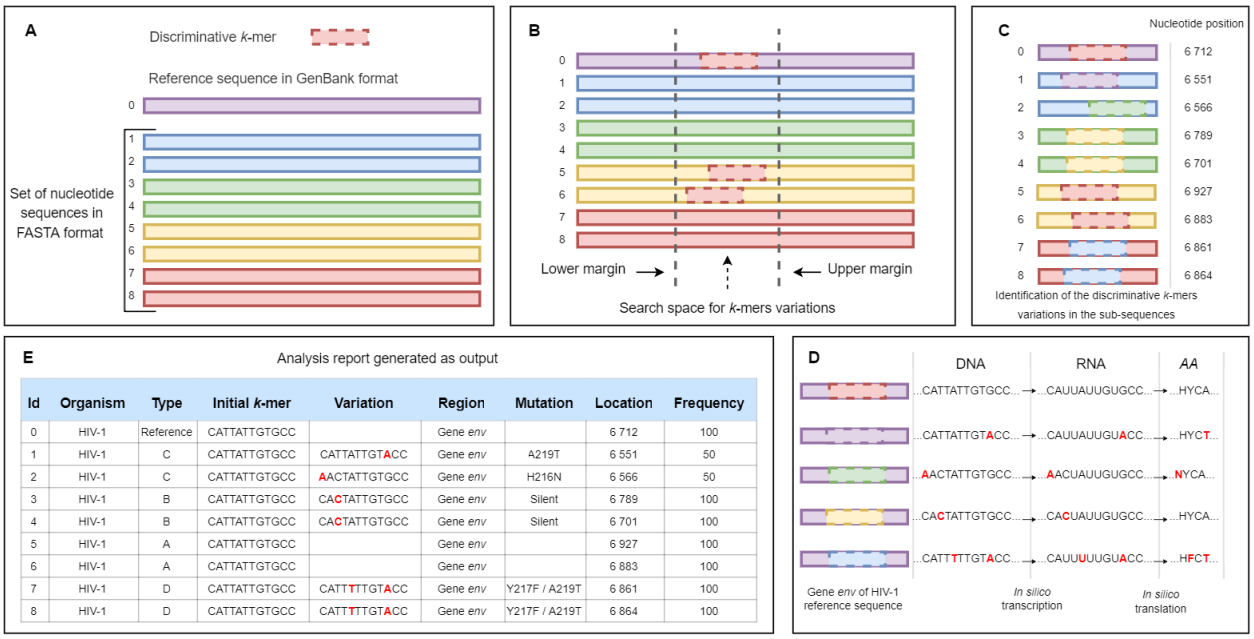
The used data should respect the following properties : A dataset D is composed of a set of whole genomes.

A sequence s of length l is defined as a succession of characters (A, C, G and T), referring to the different nucleotides that compose this sequence. Each sequence is associated with a class c. A discriminative k-mer α related to s is a sub-sequence of s of length inferior to l. A variation β of α is a sequence derived from α that has experienced at least one nucleotide substitution, insertion or deletion. The problem is then summarized as follows : given a set of sequences D, a set of discriminative k-mers S and an annotated reference sequence A related to the organism of the sequences belonging to D, how to identify $\forall\ \beta$ variations and associated information $\in D$.

4.3.2 Overview of the algorithm

The KANALYZER algorithm comprises of five main steps illustrated in Fig. 4.1. The first step (Fig. 4.1A) is to read the different input files. The second step (Fig. 4.1B) is to identify the sequences that perfectly match each discriminative k-mer based on a counting function that returns the results in a hash table data structure. This hash table has for keys the unique identifiers of sequences, and for value, a second hash table that stores the initial discriminative k-mers and the nucleotide position associated with its sequence (in case of a perfect match). The reference sequence contained in the GenBank file is also included in the dataset with the « Sequence reference » class. Then, using the average position of the perfect matches, to which a lower and upper margin is added, the algorithm determines a search space that will be used to identify the variations. Using a subspace increases the quality of the alignment by focusing on an area of interest and decreases the spatial and temporal complexity by analyzing a much shorter string than the complete sequence.

The third step (Fig. 4.1C) identifies the variations of the sequences not associated with the initial discriminative k-mer. A function will perform a local alignment using the Smith-Waterman algorithm based on dynamic programming Smith *et al.* (1981). This alignment is performed with the initial discriminative k-mer and a subsequence defined by the search space calculated in the previous step. The returned variation is the one that, when overlapping with the initial k-mer, maximizes and respects a threshold alignment score. In this step, a first alignment with parameters promoting the occurrence of nucleotide substitutions is performed. If no solution is returned, a second alignment favoring the appearance of deletions/insertions is performed. If no alignment satisfies the performance threshold, the function will return « Unassigned ». The results are returned in the hash table built in step two. The fourth step (Fig. 4.1D) identifies the information associated with the different variations. This step uses the annotation of the reference sequence coupled



Id	Organism	Type	Initial *k*-mer	Variation	Region	Mutation	Location	Frequency
0	HIV-1	Reference	CATTATTGTGCC		Gene *env*		6 712	100
1	HIV-1	C	CATTATTGTGCC	CATTATTGTACC	Gene *env*	A219T	6 551	50
2	HIV-1	C	CATTATTGTGCC	AACTATTGTGCC	Gene *env*	H216N	6 566	50
3	HIV-1	B	CATTATTGTGCC	CACTATTGTGCC	Gene *env*	Silent	6 789	100
4	HIV-1	B	CATTATTGTGCC	CACTATTGTGCC	Gene *env*	Silent	6 701	100
5	HIV-1	A	CATTATTGTGCC		Gene *env*		6 927	100
6	HIV-1	A	CATTATTGTGCC		Gene *env*		6 883	100
7	HIV-1	D	CATTATTGTGCC	CATTTTTGTACC	Gene *env*	Y217F / A219T	6 861	100
8	HIV-1	D	CATTATTGTGCC	CATTTTTGTACC	Gene *env*	Y217F / A219T	6 864	100

Figure 4.1 – Pipeline analysis of a discriminative k-mer relative to a set of HIV-1 sequences with the KANALYZER tool.

A) Reading of input files consisting of : a list of discriminative k-mers to be analyzed in FASTA format (only one is used in this example), a set of nucleotide sequences (colors represent the different classes of sequences) and a reference sequence in GenBank format relative to the studied organism. B) Identification of perfectly matched sequences with the discriminative k-mer and determination of the search space for the identification of the variations. C) Identification of the variations and its starting nucleotide position in the genome. D) Analysis of the potential influence of different variations at the amino acid level. E) Generation of the analysis report containing the information associated to each sequence.

with the Biopython library functions Cock *et al.* (2009). In the case of a coding region, its corresponding nucleotide sequence is retrieved from the GenBank file. Then each identified variation is inserted in the place of the one contained in the reference sequence. *In silico* transcription and translation processes are performed to generate the amino acid sequence. This sequence is then compared to the reference sequence in the GenBank file using a function based on the Needleman-Wunsch global alignment algorithm Needleman et Wunsch (1970) to identify potential amino acid changes.

Finally, the information about the genomic region and potential mutations are saved in the hash table. The last step generates an information report for each initial discriminative k-mer based on the hash table formed during the different phases of the algorithm and containing the information shown in Fig. 4.1E. In addition, to optimize the computational speed, steps B, C, and D of KANALYZER are performed in parallel

computing by implementing the functions of the Joblib library (`https://joblib.readthedocs.io`).

4.4 Evaluation

4.4.1 Synthetic data

To conduct a comprehensive assessment of KANALYZER, we designed a first evaluation framework inspired from Struck *et al.* (2014) and based on synthetic sequences. We generated 27 datasets, each composed of 10,000 random sequences following an uniform distribution of nucleotides. An initial position and discriminative k-mer is generated randomly for each dataset. The k-mer is inserted into each dataset sequence at the initial position with a random variation of $\pm$ 2.5%. Then, each sequence undergoes a synthetic variation process that randomly mutates a proportion of nucleotides from the input sequence. The dataset and the initial k-mer are given as input to KANALYZER to identify the variations occurring on the discriminative k-mer in the different sequences. In this evaluation, we assessed the following parameters : the length of the sequences : 5,000, 15,000 and 100,000 nucleotides / the length of the discriminative k-mers : 12, 15 and 30 nucleotides / the percentage of synthetic variation : 1%, 5% and 10% (See Table 4.2).

4.4.2 Viral sequences

We also assessed KANALYZER with five heterogeneous viral datasets, including viruses with high mutation rates and large-scale impact on the worldwide population. For the first dataset, which focuses on hepatitis B virus (HBV), from the HBVdb database Hayer *et al.* (2013), we randomly selected 300 complete genomes divided into six genotypes (A, B, C, D, E and F). The second dataset comprises of human immunodeficiency virus type 1 (HIV-1), from the Los Alamos sequence database (`www.hiv.lanl.gov`). We randomly sampled 300 whole genomes from 6 of the major clades (i.e., A, B, C, D, F, and G). The third dataset is SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) taken from the study Lebatteux *et al.* (2024). We randomly selected 250 full genomes covering the so-called « variant of concern » (i.e., Alpha, Beta, Gamma, Delta and Omicron). The fourth dataset that deals with dengue virus (DENV), from the ViPR database Pickett *et al.* (2012), we randomly downloaded 200 complete genomes to cover the four major DENV serotypes. Finally, the last dataset is hepatitis C virus (HCV) from the ViPR database.We randomly collected several complete genomes to cover the six major HCV genotypes. General information about the datasets used in the evaluations are summarized in Table 4.1.

4.4.3 Discriminative k-mers

Since KANALYZER is integrated into the CASTOR-KRFE discriminative k-mers identification tool Lebatteux *et al.* (2019), we designed a complete analysis pipeline of a user and applied the following evaluation process. First, we provide as input to CASTOR-KRFE each viral dataset with the parameters k-min = 10, k-max = 20 and T = 0.99. The parameters k-min and k-max are the minimum and maximum lengths of k-mers that CASTOR-KRFE can identify. T is the percentage performance threshold to be maintained when trying to reduce the number of k-mers during an internal evaluation by 5-fold stratified cross-validation. Then, we used the discriminative k-mers identification option of CASTOR-KRFE, which returned for each viral dataset a set of discriminative k-mers containing the information to maximize the separation between the classes of viruses in each set.

Table 4.1 – Viral dataset

Organism	**Virus group**	**Average sequence length**	**Classification**	**Number of instances**	**Number of classes**
Hepatitis B Virus (HBV)	dsDNA-RT	3,209	Genotypes	300 [50]	6
Human Immunodeficiency viruses-1 (HIV-1)	ssRNA-RT	9,106	Subtypes	300 [50]	6
Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)	(+)ssRNA	29,764	Variants	250 [50]	5
Dengue Virus (DENV)	(+)ssRNA	10,660	Serotypes	200 [50]	4
Hepatitis C Virus (HCV)	(+)ssRNA	9,450	Genotypes	269 [19-50]	6

The main values in the column « Number of instances » indicate the total number of instances in the dataset, and the one in brackets specifies the number of instances per class. Abbreviations in the « Virus group » column : (+) : positive sense, ss : single-stranded, ds : double-stranded, DNA : deoxyribonucleic acid, RNA : ribonucleic acid, RT : Reverse transcription. Percentage of missing nucleotides $<$ 1% for each sequence.

4.4.4 Variation identification and validation

The sets of viral sequences and the associated discriminative k-mers were given as input to KANALYZER to identify the variations relative to the different virus sub-species. To allow KANALYZER to extract the information associated with the identified variations, the following reference sequences in GenBank format have been used : HBV (strain ayw) genome (Accession number : NC_003977.2), HIV-1 clade B, complete genome (Accession number : NC_001802.1), SARS-CoV-2 isolate Wuhan-Hu-1, complete genome (Accession number : NC_045512.2), DENV-1, complete genome (Accession number : NC_001477.1) and HCV genotype 1, complete genome (Accession number : NC_004102.1). Then to confirm the accuracy of the variations identified by KANALYZER, we computed multiple alignments with refinement on each viral dataset using the MUSCLE algorithm Edgar (2004). The identified variation was then confirmed by scanning the overlapping columns with the initial discriminative k-mers. Concerning the extracted information associated with the variations, these were validated by analyzing the reference sequence associated with each virus using the SnapGene software (`https://www.snapgene.com`). For the performance metrics, we computed the total percentage of correctly identified k-mers. This metric is high if KANALYZER correctly identifies the discriminative k-mer or the variation in its associated sequence. Second, to highlight the main utility of KANALYZER, we computed a metric focusing only on the percentage of k-mers variation correctly identified. It removes the influence of the initial discriminative k-mers more easily identifiable. Its values are high when KANALYZER correctly identifies the variation in its associated sequence, information associated with the reference virus, as well as its potential impact on the amino acid sequence. These two metrics were also used to measure the performance of synthetic data evaluation. The detailed results for viral dataset are summarized in table 4.3.

4.5 Results and Discussion

4.5.1 Synthetic data

For all analyzed sequences (270,000), KANALYZER correctly identified more than 96% of k-mers (including the initial unaltered discriminative k-mers and those modified during the synthetic variation process). More than 150,000 of k-mers had undergone synthetic variations. The number of variations in the dataset increased proportionally to the variation rate and the length of the discriminating k-mers. Indeed, a 30-mer with a synthetic rate variation of 10% is more likely to undergo a nucleotide change than a 12-mer with a synthetic variation rate of 1%. KANALYZER identified over 95% of k-mers that have undergone variation.

Table 4.2 – Performance of KANALYZER on synthetic dataset

Length of sequences	Length of discriminative *k*-mers	Variation rate	Correctly identified *k*-mers	Correctly identified variations	Execution time (s)
5,000 (Short)	12	1%	99.98% (9998/10000)	99.83% (1170/1172)	15.26
		5%	99.38% (9938/10000)	98.61% (4390/4452)	19.53
		10%	95.37% (9537/10000)	93.61% (6788/7251)	24.84
	21	1%	99.99% (9999/10000)	99.95% (1971/1972)	17.66
		5%	99.74% (9974/10000)	99.61% (6600/6626)	28.18
		10%	96.31% (9631/10000)	95.87% (8557/8926)	33.47
	30	1%	100.0% (10000/10000)	100.0% (2604/2604)	19.96
		5%	99.07% (9907/10000)	98.81% (7712/7805)	34.93
		10%	87.97% (8797/10000)	87.45% (8383/9586)	42.01
15,000 (Medium)	12	1%	99.9% (9990/10000)	99.12% (1124/1134)	21.51
		5%	98.73% (9873/10000)	97.25% (4483/4610)	43.53
		10%	93.68% (9368/10000)	91.21% (6558/7190)	59.40
	21	1%	100.0% (10000/10000)	100.0% (1870/1870)	26.79
		5%	99.65% (9965/10000)	99.47% (6620/6655)	64.08
		10%	96.01% (9601/10000)	95.52% (8509/8908)	81.78
	30	1%	99.99% (9999/10000)	99.96% (2616/2617)	35.05
		5%	98.97% (9897/10000)	98.69% (7776/7879)	83.40
		10%	88.20% (8820/10000)	87.69% (8409/9589)	104.57
100,000 (Large)	12	1%	99.66% (9966/10000)	97.04% (1116/1150)	87.90
		5%	96.31% (9631/10000)	91.78% (4119/4488)	297.84
		10%	87.69% (8769/10000)	82.91% (5972/7203)	448.76
	21	1%	99.87% (9987/10000)	99.30% (1834/1847)	144.09
		5%	99.47% (9947/10000)	99.20% (6542/6595)	453.40
		10%	95.05% (9505/10000)	94.45% (8423/8918)	572.95
	30	1%	99.85% (9985/10000)	99.43% (2599/2614)	200.19
		5%	98.62% (9862/10000)	98.22% (7617/7755)	540.04
		10%	87.23% (8723/10000)	86.63% (8275/9552)	715.99
Average			96.91% (261669/270000)	95.99% (142637/150968)	156.19

Mainly, the lowest results (between 86% and 89%) were obtained for synthetic variation rates of 10% coupled with 30-mers. These results highlight the limitations of KANALYZER in identifying k-mers that have undergone more than three nucleotide substitutions. However, it is important to note that with these parameters, the frequencies of non-assignment by KANALYZER (considered as errors) were 91%, 90%, and 86%, respectively, for the short, medium, and large sequences. The lowest results (82.91%) were obtained for identifying k-mers variations of length 12 with a variation rate of 10% within the large sequences. We suppose that sequences of this length increase the probability of containing several sub-sequences close to the initial discriminative k-mers. This configuration makes it difficult for KANALYZER to identify the correct k-mers variation, unlike the k-mers of length 21, where the results for the identification of k-mers are $>$ 94%.

Regarding the execution time, KANALYZER can analyze (best-case scenario) 10,000 short sequences in 15 seconds (about 667 sequences/second) with a percentage of variation of 1% and k-mers of length 12. In the worst case, the execution time is 716 seconds for 10,000 large sequences (about 14 sequences/second) with 10% variations and k-mers of length 30. The execution time is bound by the variation percentage, the length of the sequences, and the length of the k-mers. A high percentage of variation increases the number of k-mers variations in the dataset. To identify an initial discriminative k-mer in a sequence s, the complexity is O(n), where n is the length of s. This complexity corresponds to the counting function (Fig. 4.1B). Nevertheless, to identify a k-mer variation, the complexity is O(n*m), where m is the length of the variation. This complexity corresponds to the pairwise alignment algorithm (Fig. 4.1C). By increasing the number of variations of k-mers, we increase the number of operations with a complexity of O(n*m). Therefore, the execution time grows when the length of the sequences (n) and the length of the k-mers (m) increases. All experiments were performed on an MSI GL65 Leopard laptop with a RAM of 16 GB DDR4 2666 MHz and a processor Intel Core i7-10750H.

4.5.2 Viral sequences

The discriminative k-mers identified by CASTOR-KRFE for each viral dataset are presented in the column « Initial discriminative k-mers » of Table 4.3. For the HBV, HIV-1, SARS-CoV-2, DENV, and HCV datasets, CASTOR-KRFE identified respectively five 20-mers (id 1 to 5), seven 18-mers (id 6 to 12), three 20-mers (id 13 to 15), two 20-mers (id 21 to 22) and five 15-mers (id 23 to 27). These k-mers coupled with CASTOR-KRFE prediction models based on support vector machines allowed to discriminate the classes of viruses contained in each dataset with an f1-score close to 0.99 during the internal evaluation process. In order to also assess the deletion cases in a relevant framework, we manually included five discriminative k-mers (id 16 to 20) covering the known 156-157 deletion of SARS-CoV-2 spike protein associated with the Delta variant (`https://gvn.org/covid-19/delta-b-1-617-2/`) and the Δ211 deletion specific to the Omicron variant (`https://gvn.org/covid-19/omicron-b-1-1-529/`). These choices allow us to consider cases where deletions are located at the beginning, the middle, and the end of the discriminative k-mer.

Table 4.3 – Performance of KANALYZER on viral dataset

Dataset	Initial discriminative k-mers (id)	Indentified CDS	Correctly identified k-mers	Correctly identified variations
HBV	AGCAATCCTCTGGGATTCTT (1)	Gene P	98.34% (296/301)	98.00% (245/250)
		Gene S	98.34% (296/301)	98.00% (245/250)
	CTATAGACCACCAAATGCCC (2)	Gene C	100% (301/301)	100% (184/184)
	GACCACCAGTTGGATCCAGC (3)	Gene P	80.73% (243/301)	71.29% (144/202)
		Gene S	80.73% (243/301)	71.29% (144/202)
	GCCTGTGCCAAGTGTTTGCT (4)	Gene P	100% (301/301)	100% (106/106)
	TGCTGTACAAAACCTACGGA (5)	Gene P	99.67% (300/301)	99.61% (256/257)
		Gene S	99.67% (300/301)	99.61% (256/257)
HIV-1	TGAAGGGGCAGTAGTCAT (6)	Gene gag-pol	100% (301/301)	100% (126/126)
	TCTGGCATAGTGCAACAG (7)	Gene env	100% (301/301)	100% (292/292)
	TCTTTGGATGGGTTATGA (8)	Gene gag-pol	99.67% (300/301)	99.63% (269/270)
	GGCTATAGAGGCTCAACA (9)	Gene env	100% (301/301)	100% (277/277)
	GAAAGGACCAGCAAAACT (10)	Gene gag-pol	100% (301/301)	100% (161/161)
	CTTTGGAAAGGACCAGCA (11)	Gene gag-pol	100% (301/301)	100% (236/236)
	ACATTATTGTGCCCCAGC (12)	Gene env	100% (301/301)	100% (236/236)
SARS-CoV-2	CTAATTCTCCTCGGCGGGCA (13)	Gene S	100% (251/251)	100% (150/150)
	CGAACTTCTCCTGCTAGAAT (14)	Gene N	100% (251/251)	100% (194/194)
	CTTCCAAAATCATAACCCTC (15)	Gene ORF3a	100% (251/251)	100% (50/50)
	AGTTCAGAGTTTATTCTAGT (16)	Gene S	100% (251/251)	100% (56/56)
	GGAAAGTGAGTTCAGAGTTT (17)	Gene S	100% (251/251)	100% (56/56)
	TTGGATGGAAAGTGAGTTCA (18)	Gene S	100% (251/251)	100% (56/56)
	CGCCTATTAATTTAGTGCGT (19)	Gene S	100% (251/251)	100% (30/30)
	TAAGCACACGCCTATTAATT (20)	Gene S	100% (251/251)	100% (30/30)
DENV	AAACGCGAGAGAAACCGCGT (21)	Gene polyprotein	100% (201/201)	100% (150/150)
	CATGGAAGCTGTACGCATGG (22)	No CDS	99.50% (200/201)	99.00% (99/100)
HCV	AATTTGGGCGTGCCC (23)	No CDS	95.93% (259/270)	93.21% (151/162)
	GACTTCGGAGCGGTC (24)	Gene polyprotein	97.78% (264/270)	97.22% (210/216)
	GCACGAATCCTAAAC (25)	Gene polyprotein	97.78% (264/270)	96.18% (151/157)
	GCGCCGGCGGGGGCG (26)	Gene polyprotein	81.85% (221/270)	80.00% (196/245)
	TTGATGAGATGGAGG (27)	Gene polyprotein	98.52% (266/270)	98.31% (232/236)
	HBV Average		94.68% (2,280/2,408)	92.51% (1,580/1,708)
	HIV-1 Average		99.96% (2,265/2,266)	99.94% (1,597/1,598)
	SARS-CoV-2 Average		100% (2,008/2,008)	100% (622/622)
	DENV Average		99.75% (401/402)	99.60% (249/250)
	HCV Average		94.37% (1,274/1,350)	92.52% (940/1,016)
	Total Average		97.56% (8,228/8,434)	96.09% (4,988/5,191)

The results of the discriminative k-mers analysis for the viral datasets by KANALYZER are shown in Table 4.3. The sets of sequences allowed the identification of 8,434 discriminative k-mers in total, of which 5,191 are

variations. Regarding identifying all k-mers, KANALYZER correctly identified 97.6% of them. Focusing only on the variations, the performance was slightly higher than 96%. Performance for HIV-1, SARS-CoV-2, and DENV sets was ≈100%. Regarding HBV and HCV, identification of variations decreased to ≈92.5%. For HBV, these results are explained by the case of the discriminative k-mer 3, where KANALYZER failed to identify the associated variation of the F genotype, implying a performance of 71%. Analysis with multiple alignments revealed that k-mer 3 had undergone five substitutions for this genotype, including three successive substitutions at the end. In this situation, the solution returned by the function based on pairwise alignment did not satisfy the minimum score threshold. Therefore, the algorithm preferred to return an unidentified solution rather than an error. Furthermore, we note that for HBV, KANALYZER generated two output reports for the discriminative k-mers 1, 3, and 5. Indeed these k-mers are located in genomic regions that code for both the P and S genes. As we will see in the following subsection, a variation in this region causes different changes in each of the genes that are involved. Concerning HCV, the performance decreases is associated with k-mer 26, where KANALYZER identified only 80% of the variations. This percentage results from this k-mer being localized in the NS5B region and exposed to high nucleotide variability. Many sequences predominantly associated with genotype 1 have undergone five substitutions from the original k-mer. For ≈70% of the sequences, KANALYZER returned no solution.

The information extracted from the KANALYZER output reports highlight that k-mer 1 was specific to the D genotype of HBV. Among variations of interest that were identified is ACCAATCCTCTGGGATTTTT, which is present in 90% of E genotype sequences. This variation has undergone two substitutions as compared to the initial k-mer, located in a region coding for two different genes. This variation leads to the substitutions Q188H and L194F for the P protein and S8T for the S protein. k-mer (5) (TGCTGTACAAAACCTACGGA) was observed in almost all genotype A and B sequences. Compared to the variation occurring in the reference sequence (associated with genotype D), it involves substitutions Q484K and F486Y in the P protein and S143T in the S protein. Genotype C is characterized by the TGCTGTACAAAACCTTCGGA variation that leads to the Q484K substitution in the P protein and a silent mutation in the S protein. Genotype E is associated with the TGCTGTTCAAAACCTTCGGA variation that results in the Y483F and the Q484K substitution in the P protein and the T140S substitution in the S protein. At last, the F genotype is mainly marked by the TGCTGTTCCAAACCCTCGGA variation, which leads to the Y483F and F486L substitutions in the P protein and the T140S substitution in the S protein.

Focusing on HIV-1, we observed that k-mer 11 was present in all gag-pol genes of the clade B genomes. For clades A, D, and G, the main variation present was ATTTGGAAAGGACCAGCA, which led to the L1381I substitution. Clade C was mainly characterized by the ATTTGGAAAGGACCAGCC variation that involved the same substitution. In clade F, the GTTTGGAAAGGACCAGCA variation was present in 86% of the genomes, leading to the L1381V substitution. The rest of the discriminative k-mers in the KANALYZER reports revealed many variations that allowed the assignment of genomic sequences to the different HIV-1 clades. However, they involved often substitutions that were silent at the amino acid level.

For SARS-CoV-2, k-mer 13 was present in 100% of the Beta and Gamma variants. The Alpha variant was characterized by the variation CTAATTCTCATCGGCGGGCA, which encodes the P681H substitution in the spike protein. The Delta variant was associated with the CTAATTCTCGTCGGCGGGCA variation, which leads to the P681R substitution. The last variation identified was CTAAGTCTCATCGGCGGGCA, which is specific to the Omicron variant and leads to the N679K and P681H double substitution. k-mer 14 is occurring in the gene encoding the N protein in the Alpha, Gamma, and Omicron variants. The nucleotide change compared to the reference sequence of the Delta variant (GGAACTTCTCCTGCTAGAAT) leads to the G204R substitution. The last variation derived from k-mer 14 was GGAATTTCTCCTGCTAGAAT, which is specific to the Beta variant and leads to the T205I substitution. k-mer 15 was present in ORF3a of all the variants except for Omicron, which has the CTTCCAAAATCATAACTCTC and TTTCCAAAATCATAACCCTC variations encoding a silent mutation and the A59V substitution in 86% and 12% of the sequences, respectively. For k-mers 16, 17, and 18 that were included manually, KANALYZER correctly identified all the variations ——GAGTTTATTCTAGT, GGAAAGTG——GAGTTT, and TTGGATGGAAAGTG—— and extracted the information about the double deletion 156/157 and the R158G substitutions in the $spike$ protein of the Delta variant and some Omicron sequences. Finally, for k-mers 19 and 20, KANALYZER accurately identified the variations and information associated with all Omicron variant sequences that involved the $\triangle$211 deletion and the L212I substitution in the spike protein.

For DENV, k-mer 21 was specific to serotypes 2 and 4. Compared to the AAACGCGCGAGAAACCGCGT variation present in the reference genome and all serotype 1 sequences, it encodes the A19E substitution. Serotype 3 is characterized by the AAACGCGTGAGAAACCGCGT variation that leads to the A19V substitution. k-mer 24 is present in all HCV genotype 4 genomes. Compared to the GACTTCCGAGCGGTC variation present in the reference genome and the majority of genotype 1 sequences, it induces the P53R substitution. k-mer 26 is characteristic of genotype 3. Compared to the GCTACAGCGGGAG variation present in the reference genome, it undergoes four nucleotide changes that lead to the Y2975A and S2976G substitutions. Finally,

k-mer 27 is characteristic of genotype 5. Compared to the variation associated with the reference genome TCGATGAGATGGATAG, it involves two nucleotide changes that do not impact the amino acid sequence.

The execution times of KANALYZER to identify variations and generate information reports associated with each dataset are 12 seconds for HBV, 46 seconds for HIV-1, 16 seconds for SARS-CoV-2, 28 seconds for DENV, and 117 seconds for HCV. This yields an average of 39 sequences per second for all the datasets combined. HCV shows a higher execution time than the other datasets. This is because the $\approx$9,600 nucleotide long HCV RNA genome is translated into a single polyprotein which is then processed by host and viral proteases to give rise to individual HCV gene products. The same situation is observed with DENV (single polyprotein $\approx$3,392 amino acids) which involves a higher execution time than HBV, HCV and HV-1 with only two discriminative k-mers to analyze. The difference in execution time of about four times higher for HCV than DENV is explained by the fact that these datasets involved 250 and 1,016 variations to analyze. All experiments were performed using the same hardware and configuration described above.

4.6 Conclusion

This article introduced KANALYZER, an algorithm based on pairwise alignment and parallel computing. This algorithm identifies the variations of discriminative k-mer and associated information in nucleotide sequences. KANALYZER was first evaluated to identify the variations of discriminative k-mer in several synthetic sequence sets. In this evaluation, where we varied several parameters, KANALYZER correctly identified $\approx$96% of the $>$150,000 variations present in the different datasets. Second, KANALYZER was assessed on sets of heterogeneous viral sequences in terms of virus groups, taxonomic classification, genome length/type, and genomic variability. The results showed that KANALYZER successfully identified $>$97% of the total signatures associated with the different sequences and $>$96% of the variations. KANALYZER also generated reports containing information associated with the multiple variations, such as associated virus classes, frequency of occurrence, genomic location, and impact on amino acid sequences in case of location in protein coding regions. The evaluations highlighted the limitation of our algorithm in identifying variations involving more than three substitutions from the original discriminative k-mers. To overcome this problem, we plan to include a function to amplify the signals of the initial k-mer and optimize the alignment score threshold function by including specific substitution matrices for penalty management. We believe that KANALYZER presents a valuable contribution to the biological and bioinformatics community, allowing researchers to perform analysis of discriminative k-mers associated with different groups of sequences qui-

ckly and accurately. The results of these analyses can help to complement discriminative k-mer sets for genomic and metagenomic sequence classification algorithms, understand virus speciation and evolution, and highlight mutations that may influence the biological properties of pathogens, including virulence and host range.

4.7 Bilan et perspectives

Ce chapitre a présenté la conception de KANALYZER, une méthode permettant d'identifier et de caractériser les variations des k-mers discriminants qui surviennent à travers les différentes classes de séquences virales. Le chapitre suivant présentera une application de KEVOLVE et KANALYZER à un large ensemble de séquences du SARS-CoV-2, un virus particulièrement bien documenté en raison de la surveillance mondiale intensive durant la pandémie de COVID-19. Cette étude de cas comportera deux volets : premièrement, une comparaison de KEVOLVE avec des outils statistiques spécialisés dans l'identification de motifs discriminants ; deuxièmement, une analyse avec KANALYZER des motifs identifiés par KEVOLVE afin d'extraire leurs variations et d'évaluer leur correspondance avec les mutations décrites dans la littérature.

CHAPITRE 5

MACHINE LEARNING-BASED APPROACH KEVOLVE EFFICIENTLY IDENTIFIES SARS-COV-2 VARIANT-SPECIFIC GENOMIC SIGNATURES

5.1 Abstract

Machine learning was shown to be effective at identifying distinctive genomic signatures among viral sequences. These signatures are defined as pervasive motifs in the viral genome that allow discrimination between species or variants. In the context of SARS-CoV-2, the identification of these signatures can assist in taxonomic and phylogenetic studies, improve in the recognition and definition of emerging variants, and aid in the characterization of functional properties of polymorphic gene products. In this paper, we assess KEVOLVE, an approach based on a genetic algorithm with a machine-learning kernel, to identify multiple genomic signatures based on minimal sets of k-mers. In a comparative study, in which we analyzed large SARS-CoV-2 genome dataset, KEVOLVE was more effective at identifying variant-discriminative signatures than several gold-standard statistical tools. Subsequently, these signatures were characterized using a new extension of KEVOLVE (KANALYZER) to highlight variations of the discriminative signatures among different classes of variants, their genomic location, and the mutations involved. The majority of identified signatures were associated with known mutations among the different variants, in terms of functional and pathological impact based on available literature. Here we showed that KEVOLVE is a robust machine learning approach to identify discriminative signatures among SARS-CoV-2 variants, which are frequently also biologically relevant, while bypassing multiple sequence alignments. The source code of the method and additional resources are available at : `https://github.com/bioinfoUQAM/KEVOLVE`.

5.2 Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV-2) is the etiological agent of coronavirus disease 2019 (COVID-19). This highly infectious coronavirus was first identified in December 2019 in Wuhan, China Zhu *et al.* (2020). It belongs to the betacoronavirus genus, which includes SARS-CoV-1 and Middle East respiratory syndrome-related coronavirus (MERS-CoV) Gorbalenya *et al.* (2020). The genome of SARS-CoV-2 is a single-stranded RNA molecule composed of approximately 30,000 nucleotides. The nucleotide sequence identity of SARS-CoV-2 with SARS-CoV-1 and MERS-CoV is 79.5% and 50%, respectively Lee *et al.* (2020); Lu *et al.* (2020). The SARS-CoV-2 genome encodes 29 different proteins, including 16 nonstructural proteins,

4 structural proteins, and 9 accessory proteins (see Fig 5.1 adapted from Gordon *et al.* (2020)). The N (nucleocapsid) protein contains the viral RNA genome, while the S (spike), E (envelope), and M (membrane) proteins together form the viral envelope Kandeel *et al.* (2020). SARS-CoV-2 exhibits a notably high mutation rate, with numerous mutations—particularly in the spike gene—correlated to increased SARS-CoV-2 transmission rates Toyoshima *et al.* (2020), augmented fusogenic and pathogenic properties of the virus Saito *et al.* (2022), as well as the emergence of new variants that could diminish the efficacy of existing COVID-19 vaccines and antibody-based therapies Koyama *et al.* (2020).

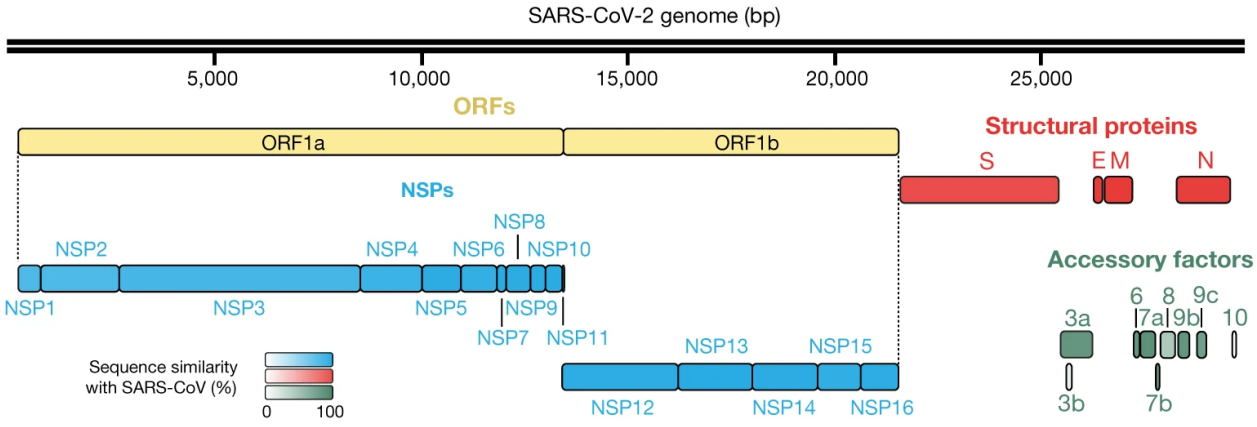


Figure 5.1 – SARS-CoV-2 genome organization.

Four structural proteins (red), 16 non-structural proteins (NSPs; blue), and 9 accessory factors (green) are shown. ORFs (open reading frames; yellow) 1a and 1b encode polyproteins. The protein sequence similarity with SARS-CoV homologues (when homologues exist) is depicted by the color intensity.

Given its rapid rate of evolution, it is important to be able to efficiently identify genomic signatures that can distinguish between different variants of SARS-CoV-2 and highlight potential functional changes. These signatures, also known as species- or variant-specific motifs that are prevalent throughout the viral genome Randhawa *et al.* (2020), can contribute to taxonomic Lopez-Rincon *et al.* (2021) and phylogenetic Bauer *et al.* (2020) studies to differentiate distinct groups of variants, provide insight into their evolutionary history Randhawa *et al.* (2020), help to understand the structure of the viral quasispecies Lau *et al.* (2021), and facilitate mechanistic studies to determine the functional basis of variant-specific differences in virulence Slezak *et al.* (2020). To identify discriminative motifs, or genomic signatures, among different groups of biological sequences, the traditional approach is to compute multiple sequence alignments using tools such as MUSCLE Edgar (2004), Clustal W/X Larkin *et al.* (2007), or MAFFT Katoh *et al.* (2019). These alignments are then analyzed to identify divergent genomic regions that constitute the discriminative motifs. However, multiple alignment approaches have significant limitations when applied to viral genomes Slezak *et al.*

(2020).

First, alignment-based approaches are generally computationally and time-intensive, making them less well suited for dealing with large viral sequence datasets that are increasingly available Bernard *et al.* (2019). In fact, computing an accurate multi-sequence alignment is an NP-hard problem with $(2N)!/(N!)^2$ possible alignments for two sequences of length N Lange (2002), which means that in some cases, the alignment cannot be solved within a realistic time frame or involves significant compromise in accuracy Katoh *et al.* (2019). Even with dynamic programming, the time requirement is on the order of the product of the lengths of the input sequences Eddy (2004). Second, alignment algorithms assume that homologous sequences consist of a series of more or less conserved linearly arranged sequence segments. However, this assumption, named collinearity, is often questionable, especially for RNA viruses Zielezinski *et al.* (2017). This is because RNA viruses show extensive genetic variation due to high mutation rates, as well as high frequencies of genetic recombination, horizontal gene transfer, and gene duplication, leading to the gain or the loss of genetic material Duffy *et al.* (2008). Finally, performing multiple alignments often requires adjusting several parameters, such as substitution matrices, deviation penalties, and thresholds for statistical parameters, which are dependent on prior knowledge about the evolution of the compared sequences Zielezinski *et al.* (2017). However, the adjustment of these parameters is sometimes arbitrary and requires a trial-and-error approach, and research has shown that small variations in these parameters can significantly impact the quality of alignments Wong *et al.* (2008).

To address the limitations of discriminative motif identification using multiple sequence alignment, specialized statistical-based tools were developed, such as MEME Bailey *et al.* (1994, 2015). MEME has a discriminative mode Bailey *et al.* (2010) that identifies enriched motifs that distinguish a primary set of sequences from a control set. Other MEME tools were also developed, including STREME Bailey (2021), the most powerful tool for discovering motifs in sequence datasets. STREME uses a generalized suffix tree and evaluates motifs using a statistical test that compares the enrichment of matches to the motif in the primary set of sequences to the control set Bailey (2021).In recent years, a series of machine-learning techniques were developed and widely used in the field of genomics, and were proven to be highly effective for solving complex and large-scale data analysis problems Libbrecht et Noble (2015). For example, the CASTOR study Remita *et al.* (2017) demonstrated the usefulness of machine learning models coupled with restriction fragment length polymorphism (RFLP) signatures for classifying viral genomic sequences, achieving f1-scores $\geq$ 0.99 for predicting hepatitis B virus and human papillomavirus genomes. However, these signatures were

found to have limitations in predicting human immunodeficiency viruses (HIV) sequences, resulting in an f1-score $\leq$ 0.90. To address this issue, the KAMERIS study Solis-Reyes *et al.* (2018) used k-mers (nucleotide subsequences of length k) to characterize the sequences provided to the learning model. To reduce the exponential number of features (4^k) associated with k-mers, KAMERIS applied truncated singular value decomposition for dimensionality reduction, but this transformation affected the ability to identify and analyze relevant features identified by the machine-learning model for discriminating between groups of sequences.

In response to this challenge, CASTOR-KRFE Lebatteux *et al.* (2019) was developed as a method for identifying minimal sets of genomic signatures based on minimal sets of k-mers to discriminate among multiple groups of genomic sequences. During cross-validation evaluations covering a wide range of viruses, CASTOR-KRFE successfully identified minimal sets of motifs, which when combined with supervised learning algorithms, resulted in average f1-scores $\geq$ 0.96 Lebatteux *et al.* (2019). However, this study was limited to identifying the optimal set of motifs, rather than exploring suboptimal sets in the feature space, which can be a major limitation when dealing with viral sequences with high genomic diversity or when attempting to infer biological functions based on the identified motifs. To overcome this limitation, KEVOLVE Lebatteux et Diallo (2021) was developed as a new method that uses a genetic algorithm incorporating a machine-learning kernel to identify multiple minimal subsets of discriminative motifs. A preliminary comparative study on HIV nucleotide sequences showed that the KEVOLVE-identified motifs allowed for the construction of models that outperformed specialized HIV prediction tools Lebatteux et Diallo (2021). In the context of the COVID-19 pandemic, this paper assessed the performance of KEVOLVE in a comparative study with several reference tools (MEME, STREME, and CASTOR-KRFE) for identifying discriminative motifs in the genomes of SARS-CoV-2 variants. The identified motifs were then analyzed using the new KEVOLVE extension (KANALYZER) to extract the associated information, and this information, which is discussed in light of the available literature to highlight the potential biological functions of the sequences/motifs in question.

5.3 Materials and methods

To assess the accuracy of KEVOLVE in identifying discriminative motifs, we conducted a comparative study with specialized tools. This involved using each tool to identify a subset of discriminating motifs in a set of training sequences of SARS-CoV-2 variants. These sets of motifs were designed to provide genomic signatures specific to each variant. In a second step, we used these signatures and a supervised learning algorithm

to fit a prediction model on the training sequences. Then, we evaluated the quality of the signatures by predicting the trained models on a large test set of unknown sequences. Finally, we used KANALYZER, the latest extension of KEVOLVE, to analyze the variant-discriminative motifs identified by KEVOLVE and assess their potential functional impact based on their location in the genome, as previously described in the literature.

5.3.1 Discriminative motif identification tools

We first evaluated KEVOLVE Lebatteux et Diallo (2021), a machine learning method based on a genetic algorithm for identifying multiple minimal sets of k-mers to discriminate nucleotide sequences. KEVOLVE takes as input a set of labeled nucleotide sequences and a parameter k, which corresponds to the length of the k-mers used to represent the sequences in an occurrence matrix. KEVOLVE starts by using a meta-transformer to remove k-mers with low discriminative contribution based on importance weights assigned by a linear Support Vector Machine (SVM). Then, the genetic algorithm begins its search by initializing several subsets (chromosomes) composed of a reduced set of k-mers (genes). Each chromosome is evaluated in a cross-validation process where prediction models are trained and tested on nucleotide sequences represented by the genes in the chromosome. The chromosomes with the best scores are then subjected to mutation/crossover processes. The mutation process involves randomly substituting a gene with another within a chromosome, and the crossover process involves exchanging genes between different chromosomes. In addition, the genes in the best chromosome have an increased probability of being selected in the next iteration. The next generation is then composed of the best current chromosomes and new chromosomes, which are generated based on the updated probability of selection. This process is repeated and coupled with a progressive increase in chromosome size until a stopping criterion is met (number of iterations or performance score of the solutions). The detailed KEVOLVE pseudo code is available in the original article Lebatteux et Diallo (2021), and the algorithm code can be accessed in the GitHub repository.

The second tool we evaluated was CASTOR-KRFE Lebatteux *et al.* (2019), an alignment-free machine learning approach for identifying a set of genomic signatures based on k-mers to discriminate between groups of nucleic acid sequences. The core of CASTOR-KRFE is based on feature elimination using SVM (SVM-RFE). It identifies the optimal length of k to maximize classification performance and minimize the number of features, providing a solution to the problem of identifying the optimal length of k-mers for genomic sequence classification Zhang *et al.* (2017). The third tool we evaluated was MEME (discriminative mode) Bailey *et al.* (2010), a tool from the MEME suite Bailey *et al.* (2015) specialized in motif identification. MEME takes two

sets of sequences as input and identifies enriched motifs that discriminate the primary set from the control set. By default, MEME assumes that all positions in the sequences have an equal chance of being a motif site. However, in discriminative mode, the algorithm uses additional information such as sequence conservation, nucleosome positioning, and negative examples to compute a measure of the probability that a discriminative motif starts at each position in each sequence Bailey *et al.* (2010).This measure, called "position specific prior" (PSP), is then used to guide the sequence motif discovery algorithm in the primary set, resulting in motifs that are more likely to discriminate it from the control set Narlikar *et al.* (2007). MEME also allows for the specification of a potential motif distribution type to improve the sensitivity and quality of the motif search. There are two available options in discriminative mode : zero or one occurrence per sequence (ZOOPS), where MEME assumes that each sequence may contain at most one occurrence of each motif, and one occurrence per sequence (OOPS), where MEME assumes that each sequence in the dataset contains exactly one occurrence of each motif. The last tool we evaluated was STREME Bailey (2021), which was found to be more accurate, sensitive, and thorough than several widely used algorithms in a recent comparative study Bailey (2021). STREME's algorithm uses a data structure called a generalized suffix tree and evaluates motifs using a one-sided statistical test of the enrichment of matches to the motif in a primary set of sequences compared to a control set. STREME assumes that each primary sequence may contain ZOOPS of the motif, but the discovery of the motif will not be negatively affected if a primary sequence contains more than one occurrence.

5.3.2 Dataset

To set up the most comprehensive evaluation framework possible, we built a dataset of 334,956 SARS-CoV-2 genomes representing the different variants defined by the World Health Organization (WHO) with at least 100 available sequences. The sequences for this dataset, covering variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Kappa (B.1.617.1), Epsilon (B.1.427/B.1.427), Iota (B.1.526), Eta (B.1.525), Lambda (C.37), and Omicron (B.1.1.529/BA.x), were downloaded on November 1, 2022 from the NCBI database Sayers *et al.* (2022) using their command line data download tool (`https://www.ncbi.nlm.nih.gov/datasets/docs/v2/how-tos/virus/get-sars2-genomes/`). We only included complete genomes with high coverage (less than 1% missing nucleotides) in our dataset (Table 5.1), and the list of accession ids for the sequences used in our different datasets is available on our GitHub repository.

Table 5.1 – Genomic sequence dataset of SARS-CoV-2 variants.

WHO Label	Pango Lineage	Number of sequences
Alpha	B.1.1.7	175,212
Beta	B.1.351	695
Gamma	P.1	8,129
Delta	B.1.617.2	9,408
Kappa	B.1.617.1	127
Epsilon	B.1.427/B.1.429	14,674
Iota	B.1.526	19,274
Eta	B.1.525	716
Lambda	C.37	428
Omicron	B.1.1.529/BA.x	106,293
Total number of sequences		334,956

5.3.3 Benchmarking

We assessed the performance of the different tools to identify discriminative motifs using an established approach Lebatteux *et al*. (2019). We performed a repeated K-fold evaluation 100 times with a different randomization at each repetition. For each iteration, 2,500 sequences were used to form a training set and the rest (332,456) were used as a testing set. In the training set, the variants were represented by 250 sequences, with the exception of Kappa, which was represented by 100 sequences due to the low number of available sequences. Alpha and Omicron were each represented by 350 and 300 sequences, respectively, due to the large number of available sequences. At each iteration, the training sets were given as input to each tool to identify the motifs that discriminate the sequences of the variants. The identified motifs, along with the training sequences, were used to train a machine-learning algorithm (linear-SVM). Indeed, linear SVMs are one of the most commonly used approaches in the classification of viral genomes, including SARS-CoV-2 Ahmed et Jeon (2022). They have also shown robustness when combined with k-mer occurrence vectors to represent sequences Lebatteux et Diallo (2021). The ability to exploit the weights assigned to characteristics (based on k-mers in our case) makes them particularly interesting for highlighting regions of interest in viral genomes. This model was then used to predict the test set, and different performance metrics were calculated. For each iteration, we computed the unweighted average of precision, recall, and f1-score. By computing each metric as an unweighted average, we avoided the dominance effect of prevalent variants, as demonstrated in Equations 5.1, 5.2, and 5.3.

$$\text{precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i} \tag{5.1}$$

$$\text{recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i} \tag{5.2}$$

$$\text{f1-score} = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{5.3}$$

The distributions of the different performance metrics for each tool are illustrated through violin plots in Fig 5.2.A-C. In addition, to visualize the prediction by class more specifically, we computed the average confusion matrix with its standard deviation for each tool (Fig 5.3.A-E). Finally, Fig 5.2.D illustrates the average number of unique motifs identified by each tool during the hundred iterations to train the prediction models.

5.3.4 Identification of discriminating motifs and tool settings

In the identification phase of the discriminative motifs, we set the length of the motifs to k = 9 for two reasons. First, this length is consistent with other studies that have used k-mers for viral sequence classification Zhang *et al.* (2017); Lebatteux *et al.* (2019); Randhawa *et al.* (2020). Second, the selection of a multiple of 3 is consistent with the codon size, and as we use sliding windows with a step of 1 to calculate the number of k-mers, encompassing all reading frames, we believe this method facilitates the capture of potential amino acid-level mutations. For KEVOLVE, we set the following search parameters : $n\_chromosomes$ = 100 (the number of chromosomes generated at each iteration), and n_genes = 1 (the number of genes composing the chromosome in the first generation). Initiating with a unitary instance allows KEVOLVE to ascertain the optimal size during its search process since this is unknown, and the training sets vary throughout the evaluation. The stopping criterion parameters were set at $n\_iterations$ = 1000 and $n\_solutions$ = 10. We utilized the default crossover and mutation rates from a previous study Lebatteux et Diallo (2021) for these parameters. For CASTOR-KRFE, we set the performance threshold to be maintained while reducing the number of features to T = 0.99.

To evaluate MEME, considering its limitation to take as input a binary set, we implemented the following process : for each variant v in the training set V, we selected all sequences belonging to v to form the primary set and used the remaining sequences in V to form the control set. We then applied MEME to discover motifs that discriminated the primary set from the control set. This process was repeated for each variant v in order to build a set of motifs that could discriminate each variant from the others. This set of motifs was used to train a model and predict the testing set in the same configuration as CASTOR-KRFE and KEVOLVE. Both the ZOOPS and OOPS options were evaluated for the associated distribution site parameters. Additionally, to strongly characterize the different groups of sequences, we performed experiments to discover 10 motifs of width 9 for each variant. This choice allows us to theoretically characterize each training set with 100 motifs, assuming there are no duplicates. We applied the same iterative process for identifying motifs to STREME. As mentioned previously, STREME does not require an input parameter for the motif distribution type and handles this automatically. Moreover, considering the number of experiments involved in evaluating the tools of the MEME suite because of their limitation to not handle multi-class sequences, it was not feasible to perform it on their web platform. To handle this, we set up virtual Linux environments where we installed the MEME suite version 5.5.0 with all the necessary dependencies for its functioning. Then several Shell/Python scripts were developed to run the different experiments and process the output files to extract the identified motifs. Finally, we specified that for the tools that identify multiple sets of motifs (KEVOLVE, MEME and STREME), the union of the motifs is used to represent the sequences through the feature matrix at each iteration.

5.3.5 Analysis of the biological significance of the motifs identified by KEVOLVE

To broaden the utility of KEVOLVE beyond identifying discriminative motifs and building prediction models for nucleotide sequences, we developed KANALYZER Lebatteux *et al.* (2022). KANALYZER is an extension of KEVOLVE that uses pairwise alignment and parallel computing. It takes as input a reference sequence in GenBank format, a list of nucleotide sequences labeled by their classes related to the organism of the reference sequence, and a list of discriminative motifs associated with the studied sequences. KANALYZER aims to understand the reasons behind a motif's discriminatory potential by identifying the variations associated with it within different groups of variants. A variation is defined as a nucleotide sequence derived from an initial k-mer that has undergone one or more nucleotide changes. KANALYZER generates a report for each motif as output, containing information on their variations that occur in the different nucleotide sequences, their genomic localization, their frequencies of appearance according to the different types of variants, and

the resulting mutations at the amino acid level in the case of coding regions. In this study, we used the KANALYZER extension to extract information associated with the discriminative motifs identified by KEVOLVE. The information was derived from the 334,956 sequences we collected and used the SARS-CoV-2 reference sequence NC_045512.2 (Wuhan-Hu-1 isolate, complete genome) for analysis.

5.4 Results and Discussion

5.4.1 Prediction performances

Initially, we examined the number of discriminative motifs identified by each tool, as summarized in Fig 5.2.D. CASTOR-KRFE identified the lowest average number of motifs at 10 per iteration, which is minimally constrained by the number of classes in the input dataset. KEVOLVE, MEME ZOOPS, and MEME OOPS identified an average of 55, 60, and 84 motifs, respectively. Finally, STREME identified the highest average number of motifs at 107 per iteration, including several degenerate motifs that were converted into classical motifs. The predictive performance of the models based on the motifs identified by each tool is shown in Fig 5.2.A, 5.2.B, and 5.2.C in terms of precision, recall, and f1-score, respectively.

KEVOLVE performed the best, with an average score of 0.99 across all metrics. The associated confusion matrix (Fig 5.3.A) for KEVOLVE indicates that misclassifications sometimes occur, with Kappa sequences being incorrectly predicted as Delta and Omicron in 4.8% and 2.6% of cases, on average. For Lambda variants, approximately 3.1% of the sequences were incorrectly predicted as Omicron. STREME models, which are based on approximately twice as many motifs as KEVOLVE, yielded the second-best predictions with an average performance of 0.96, 1.00, and 0.98 for precision, recall, and f1-score, respectively.The associated confusion matrix for STREME (Fig 5.3.B) revealed some limitations for Lambda sequences, with more than 12.5% of them being incorrectly predicted as Alpha, Delta, Epsilon, or Omicron, on average. There was also an average of 9% of Beta sequences that were misclassified in a similar manner as Lambda sequences. Like KEVOLVE, STREME models had difficulty predicting certain Kappa variant sequences ($\approx$ 9% on average), with many of them being incorrectly assigned as Delta.

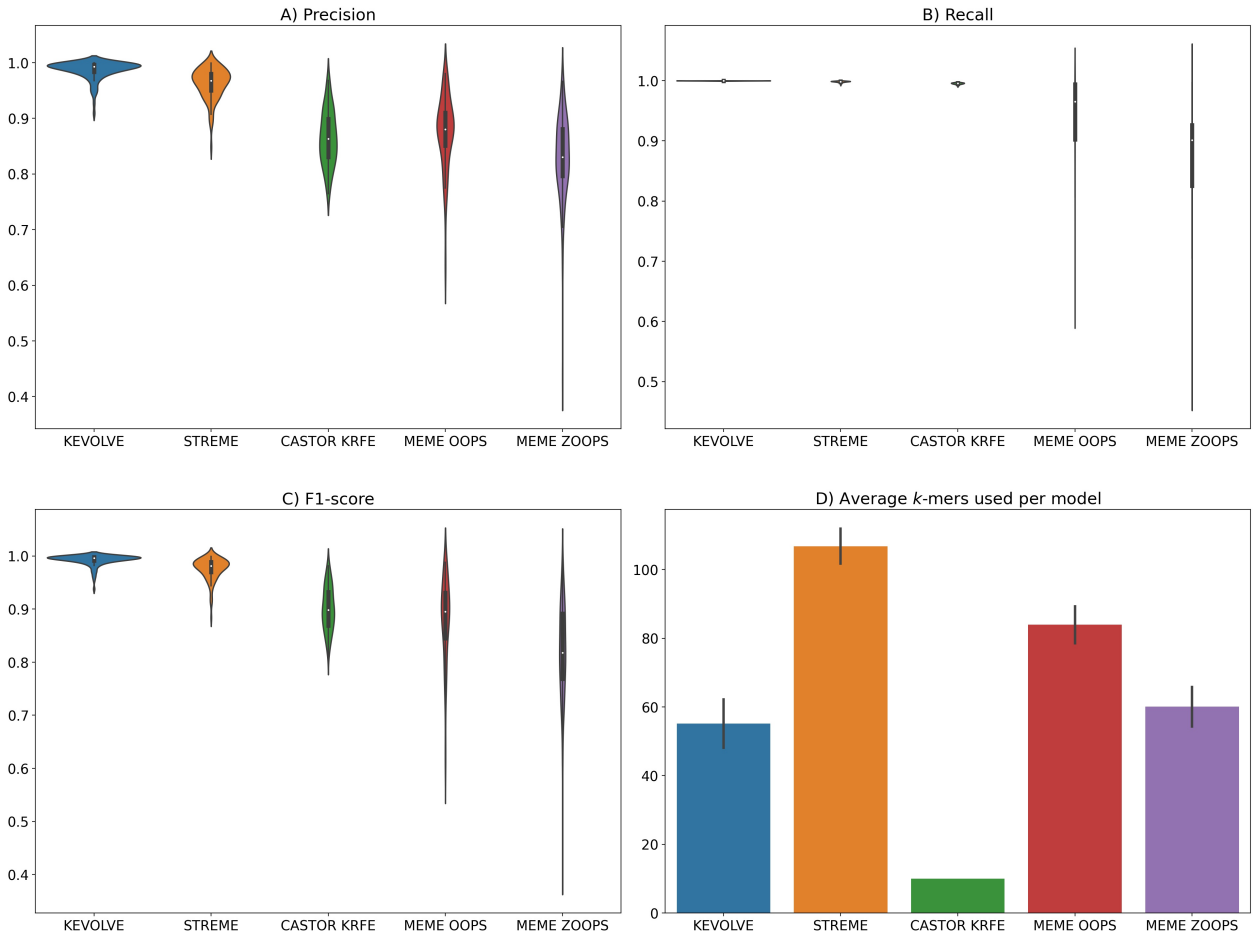


Figure 5.2 – Results of the comparative study.

A-C) The violin plots illustrate the distributions of the performance metrics, including Precision, Recall, and F1-score, obtained for the test set predictions during the cross-validation evaluation of 100 iterations. D) The bar plot depicts the average number of motifs identified by each approach to build their prediction model. The black vertical bar indicates the standard deviation.

The CASTOR-KRFE method had an average precision of 0.86, a recall of 1.00, and an f1-score of 0.90. The confusion matrix for the CASTOR-KRFE method (Fig 5.3.C) indicates that it shares the same challenges as the KEVOLVE and STREME methods in inaccurately classifying some Kappa variant sequences, with 19% of these sequences being incorrectly predicted as Alpha, 17% as Delta, and 11% as Epsilon, on average. There were also limitations in the classification of Lambda variants, with nearly 29% of the sequences being incorrectly assigned to Omicron. In addition, more than 12% of the Beta variant sequences were incorrectly assigned to Alpha, on average, and 17% of the Eta variant sequences were incorrectly assigned to Alpha.

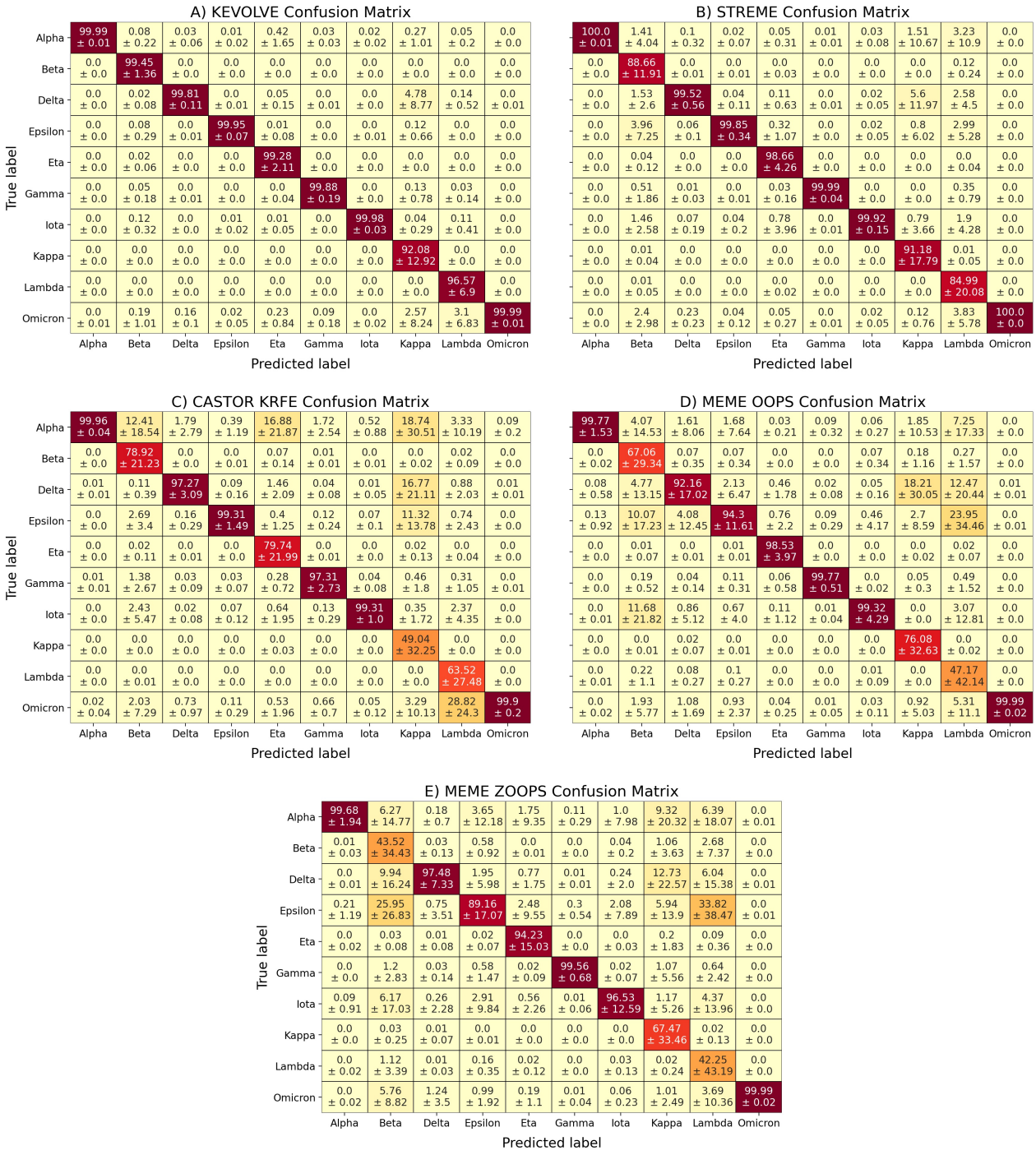


Figure 5.3 – Results of the comparative study.

E-I) The confusion matrices represent the average prediction performance as a function of the different variants for each tool over the 100 iterations. Each cell shows the average percentage of the assigned instance in the top value, and the standard deviation in the bottom value.

The MEME OOPS and MEME ZOOPS models showed the poorest prediction performance, with average precisions of 0.97 and 0.83, average recalls of 0.94 and 0.88, and average f1-scores of 0.88 and 0.82, respectively. Both models frequently made classification errors with Lambda variants, which were often incorrectly predicted to be Epsilon. Beta variants were sometimes incorrectly predicted to be Epsilon, Delta, or Iota, and Kappa variants were often incorrectly predicted to be Delta. More detailed results can be seen in the confusion matrices shown in Fig 5.3.D and 5.3.E.

5.4.2 Biological significance of KEVOLVE-identified motifs

To extract biological information related to the motifs identified by KEVOLVE, we first combined all the motifs identified during different iterations. We used these motifs to represent all 334,956 sequences in our dataset and trained a SVM model. We ranked the motifs based on their discriminant contribution, as determined by the importance weights assigned by the model. We subsequently used KANALYZER to analyze the top 50 non-overlapping, non-redundant motifs (with regards to highlighted mutations), which encompass at least the first third of the most discriminating motif set according to the SVM-assigned weights. This analysis was conducted alongside the full set of sequences and the reference sequence NC_045512.2. The results are summarized in Table 2. In cases where KANALYZER did not produce results for a specific motif, we assumed that it was located in a genomic region with high nucleotide variability (e.g., near residues 203-205 of the nucleocapsid protein Johnson *et al.* (2022)) or involved numerous successive deletions (e.g., the large 9-base SGF deletion in OR1ab Tamanaha *et al.* (2022)). To improve the signal for these motifs, we extended them to 30 nucleotides based on a consensus sub-sequence from the genomes where they were initially present. These extended motifs (ID 3, 8, 12, 18, and 38 in Table 2) were then analyzed using KANALYZER like the others. As shown on Table 2, the majority of the identified motifs were located in the coding regions of structural proteins, particularly the S protein. These motifs tended to involve missense mutations, which can have significant impacts on the infectivity, tropism, and pathogenesis of the virus even when few changes are involved Zhu *et al.* (2021).

Table 5.2 – Mutational landscape of the motifs identified by KEVOLVE.

ID	REFERENCE *K*-MERS	LOCATIONS	VARIATIONS	AMINO ACID CHANGE	VARIANTS
1	CTGGAAAGA	S	CTGGAAATA	K417N	Beta (98%) / Omicron (90%)
			CTGGAACGA	K417T	Gamma (99%)
2	AATTGCTAT	M	AATTGCTAC	I82T	Delta (98%) / Eta (99%)
			AATTGCTAG	I82S	Kappa (96%)
3	AGTTGGATGGAAAGTGAGTTCAGAGTTTAT	S	AGTTGGATGGAAAGTG−−−GAGTTTAT	Del156-157 / R158G	Delta (92%)
			AGTTGTATGGAAAGTGAGTTCAGAGTTTAT	W152C	Epsilon (98%)
			AGTTGGATGAAAAGTGAGTTCAGAGTTTAT	E154K	Kappa (79%)
4	ACCCACTAA	S	ACCCACTTA	N501Y	Alpha (99%) / Beta (98%) / Gamma (99%) / Omicron (97%)
5	GCTAGAAAA	ORF8	GCTATAAAA	R52I	Alpha (99%)
6	CAAACTAAA	None	CAAACTATA	No CDS	Epsilon (99%)
			CAAACTTAA		Lambda (99%) / Omicron (99%)
7	CCTCGGCGG	S	CATCGGCGG	P681H	Alpha (99%) / Omicron (99%)
			CGTCGGCGG	P681R	Delta (99%) / Kappa (97%)
8	CCAGGCAGCAGTAGGGGAACTTCTCCTGCT	N	CCAGGCAGCAGTAAACGAACTTCTCCTGCT	R203K / G204R	Alpha (94%) / Lambda (96%) / Omicron (98%)
			CCAGGCAGCAGTAGGGGAATTTCTCCTGCT	T205I	Beta (98%) / Epsilon (99%) / Eta (98%)
			CCAGGCAGCAGTATGGGAACTTCTCCTGCT	R203M	Delta (97%) / Kappa (92%)
			CCAGGCAGCTCTAAACGAACTTCTCCTGCT	R203K / G204R	Gamma (96%)
			CTAGGCAGCAGTAGGGGAACTTCTCCTGCT	P199L	Iota (70%)
			CCAGGCAGCAGGAGGGGAACTTCTCCTGCT	S202R	Iota (27%)
9	CAACCAGAA	S	TAACCAGAA	T19I	Omicron (71%)
			GAACCAGAA	T19R	Delta (96%)
			CAAACAGAA	T20N	Gamma (98%)
10	TTCAGAGCG	ORF3a	TTCATAGCG	Q57H	Beta (98%) / Epsilon (99%) / Iota (98%)
11	CTTGGTGCA	S	TTTGGTGCA	L699F	Beta (100%) / Iota (71%)
12	TTGGTTCCATGCTATACATGTCTCTGGGAC	S	TTGGTTCCATGCTA−−TCTCTGGGAC	Del69 / Del70	Alpha (97%) / Omicron (6%)
			TTGGTTCCATGTTA−−TCTCTGGGAC	A67V / Del69-70	Eta (98%) / Omicron (6%)
13	AAATGCACC	N	AAATGGACC	A12G	Eta (99%)
			AAATGCACT	P13L	Iota (28%) / Lambda (97%) / Omicron (99%)
14	TTACGCAAT	ORF1ab	CTACGCAAT	L3201P	Iota (99%) / Lambda (99%)
15	TGTATAGAT	S	GGTATAGAT	L452R	Delta (98%) / Epsilon (99%) / Kappa (100%) / Omicron (5%)
			AGTATAGAT	L452Q	Lambda (99%)
16	GTTGCAGCC	5' UTR	TTTGCAGCC	No CDS	Beta (6%) / Delta (98%) / Kappa (100%)
17	CCACTGAGA	S	CCATTGAGA	T95I	Delta (20%) / Kappa (88%) / Iota (99%) / Omicron (27%)
18	CATTTTTGGGTGTTTATTACCACAAAAACA	S	CATTTTTGGGTGT−TTACCACAAAAACA	Del144	Alpha (98%) / Eta (99%)
			CATTTTTGGATGTTTATTACCACAAAAACA	G142D	Delta (62%) / Kappa (69%) / Omicron (70%)
			CATTTTTGG−−−ACCACAAAAACA	Del142-144 / Y145D	Omicron (27%)
19	AGATCAGTT	ORF7a	AGATCAGCT	V82A	Delta (94%) / Kappa (100%)
20	CTAAGAGGT	S	CTACGAGGT	K77T	Delta (52%)
			TTAAGAGGT	T76I	Lambda (98%)
21	AGGAATCAC	ORF1ab	GGGAATCAC	K6711R	Delta (53%)
			GGGAAGCAC	K6711R / S6713A	Kappa (94%)
22	TTAATCTTA	S	TTAATTTTA	L18F	Beta (33%) / Gamma (99%)
23	ATATCCTTT	S	ATATCCTTG	S982A	Alpha (99%)
			ATATCTTTT	L981F	Omicron (27%)
24	GACTCAGAC	S	GACTCACAC	Q677H	Eta (98%)
			TACTCAGAC	Q675H	Lambda (8%)
25	AACTTCAAG	S	AACTTCAAA	D950N	Delta (96%)
			AACTCCAAG	Silent	Kappa (19%)
26	AATGATCCA	S	AATTATCCA	D138Y	Gamma (98%)
			AATCATCCA	D138H	Lambda (5%)
27	TACACCAAA	N	TACACCGAA	Silent	Eta (99%)
28	CACAACTGT	ORF8	CATAACTGT	T11I	Iota (99%)
29	CTAATTCTC	S	CTAAGTCTC	N679K	Omicron (99%)
30	AGAGTTCCT	E	AGAGTTCTT	P71L	Beta (99%)
31	CAATGGAAC	M	GAATGGAAC	Q19E	Omicron (96%)
32	GCTCCAATT	S	GCTCCAAAT	N969K	Omicron (99%)
33	AGACATTGC	S	AGACATTGA	A570D	Alpha (99%)
34	AAAGTGGAA	ORF1ab	AAATTGGAA	K1655N	Beta (99%)
35	GTTGGACCT	S	GTTGGACCC	F888L	Eta (99%)
36	TGTTTTTCT	S	TGTTTTTTT	L5F	Iota (99%)
37	AAAATATCT	ORF1ab	ACAATATCT	K1795Q	Gamma (99%)
38	ACTAGTTTGTCTGGTTTTAAGCTAAAAGAC	ORF1ab	ACTAGTTTG−−−AAGCTAAAAGAC	Del3675–3677	Alpha (99%) / Beta (95%) / Gamma (99%) / Eta (99%) / Iota (99%) Lambda (99%) / Omicron (71%)
			ACTAG−−−TTTTAAGCTAAAAGAC	Del3674–3676	Omicron (28%)
39	GTCAACCAA	S	GTCAACCAT	Q954H	Omicron (99%)
40	CTTACTGTT	S	CTTAATGTT	T859N	Lambda (99%)
41	GTACATCGA	ORF8	GTGCATCGA	Y73C	Alpha (99%)
42	AGAAAAGTA	ORF1ab	AGAAAAATA	Silent	Eta (99%)
43	GTCTCTAGT	S	GTCTCTATT	S13I	Epsilon (98%)
44	ATCATAACC	ORF3a	ATCATAACT	Silent	Omicron (99%)
45	ATCTCAGAT	ORF1ab	ATCTCATAT	D5584Y	Epsilon (98%)
46	GGTTCATCC	ORF3a	GGTTCACCC	S253P	Gamma (98%)
47	AACTCGTCT	5' UTR	AACTCTTCT	No CDS	Beta (99%)
48	CCAACCCAC	S	CCGACCCAC	Q498R	Omicron (97%)
49	CCTTTCTGC	ORF7b	CCTTTCTGT	Silent	Omicron (99%)
50	AAGGAAGAC	N	AAGGAAGGC	D63G	Delta (96%)

The ID column is used to reference motifs and their associated information within the text. The REFERENCE *K*-MERS column comprises the motifs in their original form as seen in the reference sequence NC_045512.2. The LOCATIONS column pinpoints the genomic region where the motifs reside. The VARIATIONS column illustrates the changes stemming from the initial motifs that transpire across different sequences. The AMINO ACID CHANGE column details the distinct amino acid level mutations induced by the variations. The VARIANTS column represents the percentage of variations' occurrence within different groups of variants. The nucleotides subject to mutations are highlighted by underlining. All data is sourced from the comprehensive SARS-CoV-2 dataset (334,956 sequences).

Motif 1, located in the S glycoprotein, is an interesting example. It has a variation present in Beta variants and in 90% of Omicron variants that involves the K417N mutation. A second variation of motif 1, found in Gamma variants, involves the K417T mutation. Both mutations occur in the receptor binding domain (RBD) of S protein, which plays a crucial role in viral infection by interacting with the host ACE2 cell surface receptor. According to published reports, these mutations may potentially decrease binding ACE2 Starr *et al.* (2020) and facilitate immune escape Barton *et al.* (2021). In contrast to the K417N/T mutations, the N501Y substitution found in the RBD-ACE2 interface was shown to result in one of the largest increases in ACE2 affinity conferred by a single RBD mutation Starr *et al.* (2020). This substitution, which is associated with the variation of motif 4, is present in several different variants, including Alpha, Beta, Gamma, and Omicron. According to Nelson et *al*. Nelson *et al.* (2021), the additional presence of the E484K mutation can further enhance virus binding to ACE2, while the presence of the K417N substitution can stabilize this binding. The combination of these mutations may result in the emergence of a mutant, whith the potential to evade host immune responses Nelson *et al.* (2021). In addition, tests in individuals who received the Moderna or Pfizer-BioNTech SARS-CoV-2 vaccines suggest that the presence of the K417N, N501Y, and E484K mutations may result in a small but significant reduction in viral neutralization, potentially impacting the effectiveness of these vaccines against certain variants Wang *et al.* (2021).

KEVOLVE highlighted several other notable mutations in the S protein, including the P681H and P681R substitutions. P681H is present in the sequences of both Alpha and Omicron variants, and its proximity to the furin protease cleavage site is thought to increase the cleavage of the S protein, potentially contributing to the rapid transmission of these variants Desingu *et al.* (2022). This mutation was suggested to enhance SARS-CoV-2 infectivity Zuckerman *et al.* (2021). The P681R substitution, which is highly conserved in the Delta and Kappa variants, appears to be associated with enhanced fusogenicity and pathogenicity Saito *et al.* (2022). The Omicron variant is distinguished by the N679K substitution, which is associated with motif 28 and also located near the furin cleavage site Kannan *et al.* (2022). When combined with P681H, both substitutions allow for the inclusion of basic amino acids near the furin cleavage site, facilitating the partition of the S protein into S1 and S2 subunits and enhancing virus fusion and infection He *et al.* (2021). Among other notable mutations in the S protein, KEVOLVE identified the double Del156-157 and R158G substitution (highlighted by motif 3), which are located in the N-terminal domain (NTD) of the protein and are unique to the Delta variant. These mutations, known as vaccine breakthrough mutations Muttineni *et al.* (2022), may potentially contribute to enhanced transmissibility or reduced sensitivity to pre-existing neutralizing antibodies Fan *et al.* (2021).

Motif 3 also allowed the identification of the W152C and E154K mutations, which are present in more than 98% of Epsilon variants and ≈ 80% of Kappa variants. The W152C mutation, in particular, is correlated with the S13I mutation associated with motif 43, which together have important biological consequences that may allow immune evasion Zhang *et al.* (2021). According to Zhang *et al.* (2021), mass spectrometry and structural studies showed that the S13I and W152C mutations resulted in a complete loss of neutralization for 10 of 10 NTD-specific monoclonal antibodies, due to the remodeling of the NTD antigenic supersite by the shift of the signal peptide cleavage site and the formation of a new disulfide bond. Other examples of mutations that affect the ability of SARS-CoV-2 to bind to specific antibody molecules (antigenicity) include the L18F, T19R/I, and T20N substitutions, which are highlighted by motifs 9 and 22. L18F is found in the Gamma variant and in ≈ 35% of Beta genomes. T19R and T19I are present in 96% of Delta variants and 71% of Omicron variants, respectively, while T20N is a Gamma-specific mutation. Epitope binding of 41 NTD-specific monoclonal neutralizing antibodies (mAbs) identified six antigenic sites, one of which, termed the "NTD supersite", is recognized by all known NTD-specific mAbs and consists of residues 14-20, 140-158, and 245-264 Harvey *et al.* (2021). The mutations associated with motifs 9 and 22 therefore include substitutions close to these antigenic regions of the NTD, including L18F, which is known to reduce neutralization by some antibodies McCallum *et al.* (2021). A last example of motif located in the S protein identified by KEVOLVE that involves major impacts on the characteristics of SARS-CoV-2 is motif 14. A first variation of this motif, present in Delta, Epsilon, Kappa, and a minority of Omicron variants (5%), involves the L452R substitution. Located in the spike RBD which interacts directly with ACE2, this mutation was shown to increase spike stability, viral infectivity, viral fusogenicity, and viral replication Motozono *et al.* (2021). The L452Q substitution, which is present in the Lambda variant, appears to be correlated with the T76I mutation associated with motif 20. These specific mutations are major contributors to the increased infectivity of the Lambda variant compared to other variants Kimura *et al.* (2022).

Regarding the mutations of interest associated with the motifs identified by KEVOLVE outside the S protein are I82T and I82S which are located in the M protein. M protein is highly conserved with low mutation rates and is a key element in virion morphogenesis and assembly, facilitating the release of viral particles from host cells and enhancing glucose transport during replication Thakur *et al.* (2022). The I82T mutation, found in Delta and Eta variants, was suggested to enhance viral replicative fitness by altering cellular glucose uptake Shen *et al.* (2021). The I82S mutation, which is currently unique to Kappa, has not yet been well studied for its effects on SARS-CoV-2 Singh *et al.* (2022). Motif 8, located in the highly immunogenic and abundantly expressed N protein, is a last relevant example of a motif associated with mutations of interest. KANALYZER's

analysis of this motif has identified variations in that region that involve P199L, S202R, R203K/M, G204R, and T205I, at least one of which is found in every major natural variant Syed *et al.* (2021). The R203K/G204R mutation, which is present in the majority of Alpha, Gamma, Lambda, and Omicron variants, was shown to confer replication advantages likely related to ribonucleocapsid (RNP) assembly, and to be associated with increased infectivity, adaptability, and virulence of SARS-CoV-2 Wu *et al.* (2021). The R203M mutation, present in Delta and Kappa, as well as the S202R mutation present in $\approx$ 27% of Iota variants, were shown to increase viral infectivity by $\approx$ 50-fold Syed *et al.* (2021). Addition of the P199L mutation (present in $\approx$ 70% of Iota variants) to S202R and R203K/M increases transmissibility by four to seven times and enhances luciferase activity, which is positively correlated with the more efficient assembly of virus-like particles and more effective mRNA delivery Syed *et al.* (2021). Overall, the highly variable region of residues 203-205 in the N protein of SARS-CoV-2, which includes the T205I substitution specific to Beta, Epsilon, and Eta, was associated with increased replication and pathogenicity Johnson *et al.* (2022). The motif analysis reports generated by KANALYZER and the accession numbers of the sequences used in our study are available on our GitHub directory (`https://github.com/bioinfoUQAM/KEVOLVE`). In addition, all identified mutations were manually confirmed using resources found at `https://covdb.stanford.edu/variants/` and `https://covariants.org/`.

5.4.3 Motifs identified by KEVOLVE/KANALYZER as genomic signature of SARS-CoV-2 variants

In the comparative study, we used KEVOLVE to identify motifs that discriminate between different classes of SARS-CoV-2 variants. We then selected the top 50 non-overlapping and non-redundant motifs determined by the importance weights assigned by the model. These 50 motifs were subsequently input into KANALYZER to characterize and identify their variations within the different SARS-CoV-2 variant groups (Column "VARIATIONS" of Table 2)." In total, we obtained 125 motifs and their associated variations, which are represented in the form of a cluster map (Fig 5.4). This map illustrates the frequency of absence/presence of each motif across different SARS-CoV-2 variants. Although these motifs were identified by KEVOLVE from a training subset of 2,500 sequences, the frequencies shown in Fig 5.4 are computed from the entire dataset of 334,956 sequences. By examining the columns, it is possible to identify different profiles and clusters of absence/presence of motifs specific to various variants. For example, Omicron has a cluster of 7 motifs that are unique to this variant (located in the lower left of the cluster map), with the exception of the ATCATAACT motif, which is also present in Iota.

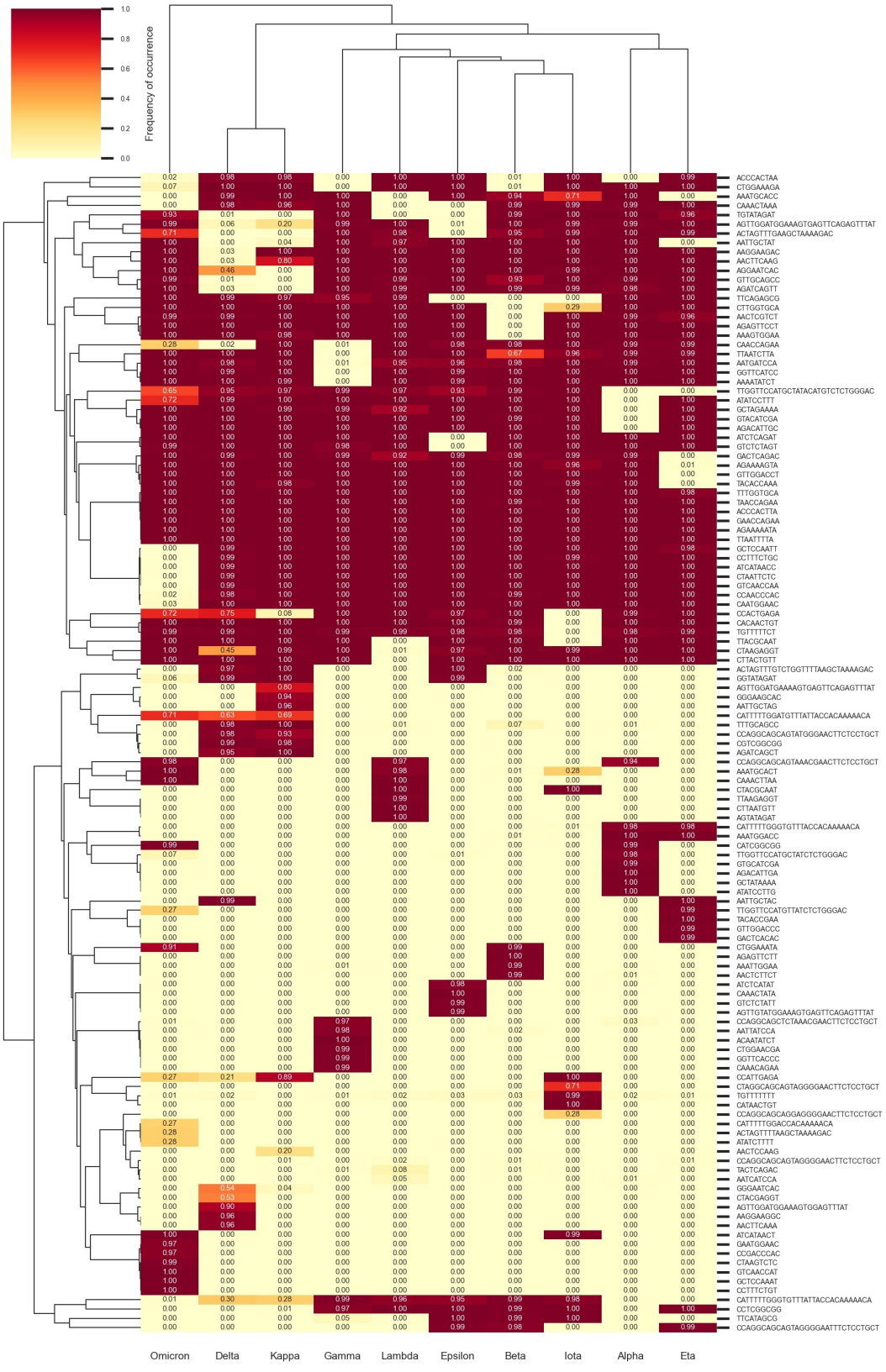


Figure 5.4 – Cluster map of motif occurrence frequency according to SARS-CoV-2 variants

Towards the middle of the cluster map, we can see a second cluster of 7 motifs that appear in all variants except Omicron. These two Omicron-specific clusters contribute to its distance from the other SARS-CoV-2 variants. In summary, this figure illustrates KEVOLVE's ability to identify motifs in temporally conserved regions starting with a limited set of sequences and to generalize to a larger dataset of sequences collected since the start of the COVID-19 pandemic. The identified motifs provide genomic signatures that can be used to generate peptide or oligonucleotide libraries for rapid and accurate detection of listed pathogens with tools such as VirScan Xu et Tian (2015) or to design specific primer sets for the classification of SARS-CoV-2 variants with artificial intelligence McMillen *et al.* (2022). These approaches, which use models built from a restricted number of motifs and sequences, can efficiently classify large sets of sequences, which is crucial during major viral outbreaks where swift identification of the virus' taxonomic classification and genomic sequence origin is necessary for effective strategic planning, containment, and treatment Randhawa *et al.* (2020).

In addition, the identified genomic signatures, along with the reports generated by KANALYZER, provide valuable insights that can help understand the viral evolution and transmission, the mechanisms through which the virus causes disease, and the development of treatments and vaccines. These approaches, which use models built from a restricted number of motifs and sequences, can efficiently classify large sets of sequences, which is crucial during major viral outbreaks where swift identification of the virus' taxonomic classification and genomic sequence origin is necessary for effective strategic planning, containment, and treatment Randhawa *et al.* (2020).

5.4.4 Perspective and Future Directions

For future work, we believe it would be insightful to explore comparisons with approaches that have been developed concurrently and exhibit similarities. One such tool is CouGaR-g, recently published, which introduces an approach using a deep learning model (convolutional neural networks) to classify SARS-CoV-2 sequences represented by frequency chaos game representation Avila Cartes *et al.* (2023). In their study, CouGaR-g demonstrated strong performance with an accuracy exceeding 96% for a test set comprising 19,146 SARS-CoV-2 sequences divided into 11 clades. The authors also utilize saliency maps to highlight relevant k-mers and further demonstrate their association with known marker variants. It could, therefore, be beneficial to conduct experiments to compare the impact of sequence representation, the influence of the choice of machine learning model (especially to investigate performance on GISAID clades such as GR,

GRY, or O where CouGaR-g's performance was lower), or to assess the overlap and differences in the k-mers identified as significant in correlation with known marker variants.

Another pertinent comparative analysis would consider Nextclade Aksamentov *et al.* (2021) in classifying viral sequences with significant nucleotide divergence. Nextclade conducts pairwise alignments of viral genomes against a reference sequence, discerns mutations, and employs mutational distances to ascertain the nearest match within a phylogenetic framework, thereby designating the query sequence to a closely related clade Aksamentov *et al.* (2021). While both Nextclade and KEVOLVE demonstrate robustness in SARS-CoV-2 sequence classification, KEVOLVE may offer superior performance for viruses exhibiting substantial nucleotide divergence, such as HIV—with divergence rates between subtypes ranging from 25 to 35% Hemelaar *et al.* (2006) and hepatitis C virus (HCV), where genotypic differences reach 31 to 33% at the nucleotide level Simmonds *et al.* (2005). Notably, Nextclade is tailored for rapid alignment of sequences with less than 10% divergence Aksamentov *et al.* (2021), a scenario less applicable to the broad variability seen in HIV or HCV. KEVOLVE employs k-mer occurrence vectors for sequence representation and a SVM for prediction, a methodology previously validated for robust viral classification across diverse divergence levels Solis-Reyes *et al.* (2018); Lebatteux *et al.* (2019); Lebatteux et Diallo (2021). The sensitivity of k-mer occurrences, as opposed to mutations at specific positions, is particularly advantageous for sequences with elevated rates of nucleotide divergence Zielezinski *et al.* (2017). Furthermore, the SVM framework provides a nuanced approach by relating a set of training sequences to their features and assigning weights based on their discriminative value—unlike Nextclade's distance-based classification. This feature weighting proves instrumental in prioritizing mutations for analysis. Nextclade and KEVOLVE each have their place in the genomics toolbox, with specific scenarios where they can distinguish themselves.

Finally, although our current approach and all those mentioned above operate within a closed classification framework, which is limited to the classes defined by the training sequence dataset, we plan to extend it to an open classification context. To achieve this, we propose a strategy to calculate the distance between each new sequence and the existing genomic signature profiles, generated in the cluster map (Fig 5.4). By using an appropriate distance threshold, we can identify sequences that are significantly distant from known signatures, potentially indicating a new variant. Thresholds can be determined by leveraging the knowledge of distances between genomic signature profiles of different known variants. This method could be based on distance metrics such as Euclidean distance, Manhattan distance, or even a normalized distance based on k-mer similarity. Furthermore, to make our approach more flexible and adaptable to new variants, we

could also implement an incremental learning mechanism. In this way, each time a new variant is identified above a certain support threshold, the associated sequences could be integrated into the initial training set, and the model would be retrained to account for this new information. This would allow our model to learn and progressively adjust its parameters based on the newly encountered sequences. This approach could facilitate the detection of new variants and enable regular model updates with the integration of new sequences associated with emerging variants.

5.5 Conclusion

In this study, we compared the performance of the machine learning-based tools KEVOLVE and CASTOR-KRFE with statistical tools specialized in identifying discriminative motifs in unaligned sequence sets for the classification of SARS-CoV-2 variants. Overall, the models based on the motifs identified by KEVOLVE outperformed the models based on the motifs identified by the statistical tools, while using a lower number of motifs. Models based on STREME motifs achieved the second-best performance (slightly below KEVOLVE), but these models require the use of twice as many motifs. The drop in performance was mainly due to prediction errors for Beta, Kappa, and Lambda variants. CASTOR-KRFE obtained the third-best performance with models based on 10 times fewer motifs than STREME, as the tool only identifies a single subset of motifs, unlike the others. The prediction errors of the CASTOR-KRFE models are associated with the same variants as those of STREME, but they are more pronounced. Finally, the weakest performances were associated with the MEME OOPS/ZOOPS models, with many more errors for the same variants than STREME and CASTOR-KRFE. This study also demonstrated that KEVOLVE and CASTOR-KRFE are able to handle multi-class sets, rather than being limited to binary sets like some other tools. This is an important advantage when analyzing organisms such as SARS-CoV-2, which are constituted of multiple classes of viral variants.

Subsequently, we analyzed the motifs identified by KEVOLVE using KANALYZER, a new extension based on pairwise alignment and parallel computing. This analysis allowed us to identify variations of the discriminative motifs in different classes of SARS-CoV-2 variants, including their frequency, genomic localization, and mutation at the amino acid level. This analysis, performed on all 334,956 sequences belonging to the 10 major variant classes defined by the WHO, showed that the majority of the motifs identified by KEVOLVE were located in structural proteins, with a particular focus on the S protein. The motifs and variations identified were linked to known mutations previously reported in the literature, which are assumed to affect key characteristics of the virus such as infectivity, pathogenicity, tropism, transmission, and evolution. In

conclusion, this study demonstrates the utility of KEVOLVE as a robust tool for identifying discriminative motifs of SARS-CoV-2 variants. These motifs provide genomic signatures that can be used to construct oligonucleotide libraries or to build artificial intelligence models for rapid and accurate pathogen detection. Furthermore, KANALYZER allows the analysis of motifs identified by KEVOLVE, providing valuable insights into the biological properties of viruses and viral gene products that serve as targets for the development of vaccines or antiviral therapy

5.6 Bilan et perspectives

Ce chapitre a présenté une application de KEVOLVE et KANALYZER sur un large ensemble de séquences du SARS-CoV-2. Cette étude a démontré la capacité de KEVOLVE à identifier des motifs basés sur les k-mers présentant un potentiel discriminant supérieur à ceux identifiés par des outils statistiques spécialisés, tout en surmontant leur limitation au contexte binaire. L'analyse avec KANALYZER des k-mers identifiés par KEVOLVE a révélé que leurs variations sont associées majoritairement à des mutations documentées dans la littérature, mutations connues pour affecter des caractéristiques clés du virus telles que l'infectiosité et la pathogénicité. Le chapitre suivant introduira un système de notation permettant de hiérarchiser les ensembles de k-mers et leurs variations au sein des sous-espèces virales. Ce score intègrera deux dimensions complémentaires : une dimension discriminative, évaluant la capacité d'un ensemble de k-mers et leurs variations à distinguer les sous-espèces virales, et une dimension biologique prenant en compte (i) l'importance fonctionnelle de leur localisation génomique dans les régions codantes et (ii) l'impact potentiel des substitutions d'acides aminés sur la dynamique protéique.

CHAPITRE 6

INTRODUCING THE *KMERSIGNIFICANCE SCORE* : A HEURISTIC APPROACH TO ASSESSING DISCRIMINATIVE k-MERS AND THEIR VARIATIONS IN VIRAL SUBSPECIES

6.1 Abstract

High-throughput sequencing has dramatically expanded the viral sequences catalog, intensifying the challenge of classifying and understanding the global virome. Within this context, discriminative k-mers, representing unique genetic patterns, have emerged as essential markers for classifying viral subspecies or strains. K-mer variations, defined as nucleotide sequences derived from a reference k-mer that have undergone substitutions, insertions, or deletions, provide crucial insights into subspecies classification, evolution, functional elements, and disease transmission. While existing methods can identify discriminative k-mers, they often fail to capture the biological significance of variations, limiting our understanding of their impact on viral functions and pathogenicity. We introduce the *KmerSignificance Score* (KSS), a novel heuristic scoring system that evaluates the relevance of discriminative k-mers and their variations within viral subspecies. The KSS considers both the discriminative power of k-mers through a machine learning framework evaluation and their biological significance by integrating two key factors : (i) the functional importance of genomic location within coding regions and (ii) the potential impacts of amino acid substitutions on protein dynamics, assessed through differences in amino acid biophysical properties. We validated the KSS against experimental evidence and published literature across three viral systems with distinct clinical significance : SARS-CoV-2 polyproteins (1a and 1b) and structural proteins (spike, membrane, envelope, and nucleocapsid) critical for viral entry and assembly, HIV-1 gag-pol and env genes associated with drug resistance and immune evasion, and HCMV UL73, UL55, and US28 genes involved in viral tropism and pathogenesis. Our analysis demonstrates strong correlation with known functional mutations while identifying novel potentially significant variations. KSS prioritizes the most informative k-mers, enhancing results from other analytical methods and offering new perspectives on k-mer significance. By providing comprehensive insights on the impact of mutations, KSS is a valuable tool for understanding genetic determinants of viral functions and pathogenicity, with broad applications in virome research.

Keywords : Scoring System, Discriminative k-mers, Viral Classification, Key Mutations, Machine Learning.

6.2 Introduction

Advancements in high-throughput sequencing technologies have profoundly transformed our understanding of viral diversity and evolution, generating an abundance of viral sequences Simmonds *et al.* (2017). While these developments offer significant opportunities, they also present unique challenges for the comprehensive classification and characterization of the global virome. Traditional taxonomic classification methods, which rely heavily on phenotypic properties, are increasingly inadequate due to the expanding volume and complexity of data Simmonds et Aiewsakun (2018). Alignment-based methods, while powerful for closely related sequences, struggle to capture distant evolutionary relationships and are computationally expensive for large datasets Zielezinski *et al.* (2017). Consequently, alignment-free approaches, particularly those utilizing k-mers—short, contiguous sequences of nucleotides—have become essential for taxonomic classification and in-depth characterization of viral sequences Wood et Salzberg (2014); Ren *et al.* (2017); Lebatteux *et al.* (2019); Remita et Diallo (2019); Lebatteux et Diallo (2021); Raju *et al.* (2022). These approaches offer both computational efficiency and sensitivity in detecting genomic patterns across diverse viral sequences. K-mers capture unique features of a virus's genetic sequence, enabling researchers to identify distinctive patterns that differentiate specific subspecies or strains. Furthermore, variations of discriminative k-mers, defined as nucleotide sequences derived from an initial k-mer that have undergone one or more nucleotide modifications (substitutions, insertions, or deletions) Lebatteux *et al.* (2022), provide valuable insights into subspecies classification, revealing critical biological information including evolutionary relationships, functional elements, and mechanisms of disease transmission Lebatteux *et al.* (2024).

However, existing methods that evaluate and identify discriminative k-mers present several limitations in their ability to fully characterize k-mers and their variations, from both discriminative and biological perspectives. From a discriminative standpoint, current approaches face multiple methodological challenges. Most methods rely on univariate feature ranking and top-n selection approaches using statistical measures (e.g., information gain, mutual information, chi-square test, or ANOVA F-test) applied to k-mer sequence characterization matrices Yu et Liu (2004); Yousef *et al.* (2017); Alam et Chowdhury (2020); Orozco-Arias *et al.* (2021). While these approaches are computationally efficient with linear time complexity, they present three major limitations : (i) the difficulty in comparing features across different classification contexts, as discrimination score thresholds achieved in top-n selections may vary significantly between contexts, (ii) the inability to effectively eliminate redundant features, as similar relevance rankings lead to the selection of highly correlated features providing redundant discriminative information Yu et Liu (2004), and (iii) the

failure to capture potentially significant discriminative combinations of k-mers and their variations, as patterns are evaluated independently Dong *et al*. (2023). Moreover, many methods were initially designed for binary classification scenarios Bailey *et al*. (2010); Huggins *et al*. (2011); Yu *et al*. (2015); Bailey *et al*. (2021), necessitating the use of one-versus-rest adaptations when applied to the inherently multiclass nature of viral classification problems Lebatteux *et al*. (2024).

The biological significance of k-mer variations presents an equally important yet distinct challenge. While current methods successfully identify discriminative patterns, they lack mechanisms to evaluate their biological relevance Lebatteux *et al*. (2019); Alam et Chowdhury (2020); Juneja *et al*. (2022). This limitation is particularly critical given that k-mers and their variations can have significant biological implications, as their distribution and frequency patterns can reveal crucial information about genomic characteristics, including regulatory elements, conserved domains, structural variations, and potential functional motifs Moeckel *et al*. (2024). These methodological constraints significantly limit our ability to understand how specific mutations associated with k-mers and their variations influence viral functions, pathogenicity, and evolution. A more comprehensive analytical framework is needed to effectively integrate both discriminative power and biological significance in the evaluation of k-mer variations in viral sequences.

To address these limitations, we introduce the *KmerSignificance Score* (KSS), a novel heuristic scoring system that evaluates both the discriminative power and biological significance of k-mers and their variations. The KSS combines two complementary components :

1. A discriminative component that uses a supervised machine learning framework to generate context-independent importance scores, enabling direct comparisons of k-mer relevance across different viral classification scenarios while handling multiclass classification challenges.
2. A biological significance component that assesses : (i) the functional importance of genomic locations within coding regions, using supervised machine learning models trained on viral protein database data, and (ii) the potential impact of amino acid substitutions on protein dynamics, quantified through differences in amino acid biophysical properties derived from substitution matrices.

This integrated approach provides a standardized evaluation of k-mer discriminative significance while prioritizing variations in functionally important genomic regions, thereby guiding the strategic selection of targets for experimental in *vitro* and in *vivo* analyses.

To validate the robustness and versatility of our methodology, we systematically evaluated the KSS across

multiple viral systems, each chosen for their well-documented genomic variations and distinct clinical significance :

1. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiological agent of coronavirus disease 2019 (COVID-19) pandemic, with particular focus on the polyproteins (1a and 1b) and structural proteins, including the spike protein (S), membrane protein (M), envelope protein (E), and nucleocapsid protein (N), which play critical roles in viral entry, assembly, and host immune response Lamers et Haagmans (2022).
2. Human immunodeficiency virus type 1 (HIV-1), the retrovirus responsible for the HIV/AIDS pandemic, focusing on the *gag-pol* and *env genes*, which harbor well-documented mutations associated with drug resistance and immune evasion Collier *et al*. (2019).
3. Human cytomegalovirus (HCMV), the most common congenital viral infection and a leading cause of childhood hearing loss, cognitive deficits, and visual impairment Gantt *et al*. (2016), analyzing the UL73, UL55, and US28 genes associated with viral tropism and pathogenesis.

Through comprehensive comparison with published literature and expert assessments across these diverse viral systems, we demonstrate the KSS's effectiveness in identifying discriminative and biologically significant sequence variations. The methodology presented here represents a significant advancement in viral genome analysis, offering a standardized framework that bridges the gap between computational pattern recognition and biological relevance in viral classification and pathogenesis studies. This integrated approach has broad implications for viral surveillance, vaccine development, and the understanding of viral evolution and adaptation.

6.3 Materials and Methods

6.3.1 Overview of the *KmerSignificance Score*

The *KmerSignificance Score* (KSS) is a heuristic scoring system designed to assess the discriminative and biological significance of k-mers and their variations within viral subspecies. This system specifically targets k-mers and their variations from the coding regions of viral nucleotide sequences. The extraction of these k-mer variations is performed in two steps :

— **Pairwise sequence alignment :** Using Nextalign to align query sequences against a reference sequence Aksamentov *et al*. (2021). Nextalign is based on banded pairwise Smith-Waterman alignment with an affine gap cost (Smith *et al*., 1981). Nextalign takes into account genome annotations speci-

fying coding regions to make the gap-opening penalty dependent on the reading frame, allowing it to select the most biologically meaningful gap placement among otherwise equivalent alignments.

— **Variation identification :** Application of the KANALYZER algorithm (Lebatteux *et al.*, 2022) on the obtained alignment to scan and identify k-mers, their variations, and associated amino acid mutations.

For each set of identified k-mer variations, the KSS computes three preliminary scores :

1. **Discriminative score :** Quantifies the ability of k-mers and their variations to distinguish between different viral subspecies or strains using a supervised machine learning framework.
2. **Protein score :** Assesses the functional importance of the proteins where k-mer variations are localized. This score is derived from a machine learning predictive model that integrates protein functional information sourced from the UniProt database (Bateman *et al.*, 2024) and other relevant resources.
3. **Mutational score :** Evaluates the biological impact of mutations associated with k-mer variations, relying on a substitution matrix that focuses on the conservation of the biophysical properties of amino acid changes.

The KSS is derived by integrating these three scores to assess the overall significance of the k-mer variations in their biological and taxonomic contexts.

6.3.2 Input of the *KmerSignificance Score*

The computation of the *KmerSignificance Score* (KSS) requires several key inputs for analysis. The primary input, denoted as $\mathcal{D}$, consists of a collection of viral nucleotide sequences categorized by their respective classes and stored in FASTA format. For sequence alignment, three reference sequence files are required : the reference sequence in FASTA format, its GenBank annotation file, and the gene annotation file in GFF3 format. Additionally, two parameters must be specified : k, which defines the length of k-mers to be identified, and a threshold t, which determines the minimum frequency at which a k-mer must be present in at least one class to be considered for analysis. Using these inputs, $\mathcal{D}$ undergoes alignment and subsequent identification of k-mers, their variations, and associated amino acid mutations. These sets of k-mers form the foundation for the KSS computation, as detailed in the following sections.

6.3.3 Discriminative score

The first of the KSS is the discriminative importance of the sets of k-mers and their variations. From each set of extracted k-mers, we generate a feature set $\mathcal{F}$ consisting of the k-mers and their variations present in $\mathcal{D}$. From this, we construct a presence/absence matrix X of dimensions (m, n), where m represents the number of sequences in $\mathcal{D}$ and n denotes the number of k-mers and variations in $\mathcal{F}$. Each element $X_{i,j}$ in the matrix is assigned a value of one if k-mer j is present in sequence i, and zero if it is absent. The corresponding target classifications are represented by a vector y of length m, where each entry corresponds to the target classification of the respective sequence in X.

Using X and y, we construct and evaluate a supervised learning model based on Support Vector Machine (SVM) with a linear kernel, which has proven effective for viral sequence classification using k-mer representation vectors (Solis-Reyes *et al.*, 2018). The model evaluation is performed through k-fold stratified cross-validation, where in each fold, a subset of data instances is designated as the training set (X_{train}, y_{train}), while the remainder forms the test set (X_{test}, y_{test}). The performance is assessed using a novel metric : the Root Mean Square Rank-Weighted F1 Score (RMS Rank-Weighted F1), which is computed through the following steps :

1. **Class-wise F1 score calculation :** For each class c, the F1 Score $F1_c$ is calculated as :
$$F1_c = 2 \cdot \frac{\text{precision}_c \times \text{recall}_c}{\text{precision}_c + \text{recall}_c}$$
where precision_c and recall_c represent the precision and recall for class c respectively.
2. **F1 score ranking :** The F1 Scores are sorted in descending order to assign ranks.
3. **Weight calculation :** Each F1 Score is assigned a weight w_r based on its rank r :
$$w_r = \frac{n}{r^{\left(\frac{\log(n)}{2}\right)}}$$
where n is the total number of classes.
4. **Weighted sum computation :** The weighted sum of squared F1 Scores is calculated as follows :
$$\text{Sum}_w = \sum_{r=1}^{n} w_r \cdot F1_r^2$$
5. **Final score computation :** The RMS Rank-Weighted F1 Score is determined as :
$$\text{RMS Rank-Weighted F1} = \sqrt{\frac{\text{Sum}_w}{\sum_{r=1}^{n} w_r}}$$

6. **Performance level assignment :** The RMS Rank-Weighted F1 Score is categorized into five discrete levels (1-5) using the thresholds [0.55, 0.65, 0.75, 0.85], where level 1 represents scores below 0.55, and level 5 represents scores above 0.85.

The RMS Rank-Weighted F1 Score evaluates the discriminative power of k-mers and their variations in viral sequences by providing a score between 0 and 1. This metric is effective in both scenarios : datasets with few classes where a single set of k-mers and their variations can theoretically discriminate between classes, and datasets with numerous classes where k-mers and their variations at a given position alone cannot discriminate all classes. This allows for comparable scores across these different scenarios through two key relaxation mechanisms : 1) the Root Mean Square calculation, which balances the impact of high-performing and low-performing classes by reducing the penalty for classes that are harder to discriminate, and 2) the assignment of weights inversely proportional to class ranks based on F1 Scores inspired by techniques discussed in (Vovk et Wang, 2020), which allows the metric to focus on the most discriminative classes while still considering the overall classification performance.

6.3.4 Protein score

The second component of the KSS evaluates the functional importance of viral proteins containing the identified k-mers and their variations. This assessment is crucial as variations in proteins critical to viral functions can significantly impact viral replication, pathogenicity, and immune evasion (Uversky et Longhi, 2012), while variations in less essential or redundant proteins may have minimal effect on the viral life cycle (Betts et Russell, 2003).

The protein functional importance is quantified using a machine learning model trained on data annotated by our team. This model assigns an importance score from 1 to 5 based on comprehensive protein attributes including : evidence level of protein existence, molecular function, biological process involvement, cellular localization, protein family membership, domain and motif presence, pathway associations, protein-protein interactions, structural information, post-translational modifications, drug target status, and supporting literature evidence (see Table 6.1).

Table 6.1 – Decision support criteria for the assignment of viral protein importance classes

Criteria	Very low importance (Grade 1)	Low importance (Grade 2)	Moderate importance (Grade 3)	High importance (Grade 4)	Very high importance (Grade 5)
Protein existence	Uncertain or predicted protein. No evidence at transcript or protein level.	The protein existence is probable because of clear orthologs in closely related species.	There is at least experimental evidence at the transcription level (confirmed by RT-PCR or Northern blots).	Clear experimental evidence of the protein existence (confirmed by mass spectrometry, X-ray crystallography experiments, NMR, protein-protein interaction, detection by antibodies such as Western blot, etc.).	
Molecular function	Unknown	The molecular functions are not clearly associated with viral dynamics (e.g., chaperone, multifunctional enzyme, signal transduction inhibitor). May be associated with a small number of molecular functions.	The molecular functions could influence viral dynamics (e.g., receptor, protease, protein kinase inhibitor). May be associated with a moderate number of molecular functions.	Important molecular functions that can favor the replication and survival of the virus (e.g., cyclin, cytokine, DNA invertase). Associated with several molecular functions.	Molecular functions playing crucial roles in the replication and survival of the virus. Can favor immune escape responses and/or could be associated with specific diseases/pathologies (e.g., polymerase, superantigen, mitogen). Associated with several molecular functions.
Biological process	Unknown	The biological processes are not clearly associated with viral dynamics (e.g., viral process, cell shape, sensory transduction). May be associated with a small number of biological processes.	The biological processes could influence viral dynamics (e.g., host-virus interaction, apoptosis). May be associated with a moderate number of biological processes.	Important biological processes that can favor the replication and survival of the virus (e.g., cytokinin signaling pathway, mRNA processing, host mRNA suppression). Associated with several biological processes.	Biological processes playing crucial roles in the replication and survival of the virus (e.g., inhibition of innate immune response, virus entry into host cell, viral release). Associated with several biological processes.
Cellular component	Unknown	Associated with a small number of unspecific compartments (e.g., cytoplasm, host endoplasmic reticulum, host cytoplasm).	Found in cellular components which could influence the replication, formation, or survival of the virus (e.g., virion, membrane). Associated with a moderate number of cellular components.	Found in components which can favor replication, formation, or survival of the virus (e.g., capsid, exosome, nucleus). Associated with several cellular components.	Found in components which are strongly associated with the replication, formation, and survival of the virus (e.g., DNA-directed RNA polymerase, T-cell receptor, viral assembly complex). Associated with several cellular components.
Protein families, domains, regions, and motifs	Unknown	At least one domain or region is identified.	A moderate number of regions or domains are identified.	A high number of regions or domains are identified.	
Biological pathways and processes	Unknown		Associated with a small number of biological pathways.	Associated with a moderate number of biological pathways.	Associated with several biological pathways.
Protein-protein interaction	Unknown	Few interactions with other proteins.		Interactions with a large number of proteins.	
Knowledge and study of 3D structure	Unknown		3D structure is partially known, studied, or predicted.	Several 3D structures well studied.	
Post-translational modifications (PTMs)	Unknown	At least one PTM is identified.	A moderate number of PTMs are identified.	A high number of PTMs are identified.	
Target for existing drugs	Not targeted by any drug.			Targeted by a small number of drugs.	Targeted by a large number of drugs.
Associated references	Described by only one main reference.	Described by a small number of main references.	Supported by a moderate number of main references.	Supported by several main and cross-references.	Widely studied and supported by a large number of main and cross-references.
Annotation score	Annotation score equal to 1.	Annotation score between 1 and 3.	Annotation score equal to or greater than 3.		Annotation score equal to 5.

To assess protein functional importance, we leverage the UniProt API to collect comprehensive protein metadata. The UniProt database provides detailed annotations and links to multiple resources including InterPro, Gene Ontology, BioGRID, IntAct, PDB, DrugBank, and Reactome (Bateman *et al.*, 2024; Paysan-Lafosse *et al.*, 2023; Aleksander *et al.*, 2023; Oughtred *et al.*, 2021; Orchard *et al.*, 2014; Burley *et al.*, 2023; Wishart *et al.*, 2018; Jassal *et al.*, 2020). This information is used to build feature vectors representing each protein's functional characteristics.

The importance scoring model, based on Random Forest classification (Breiman, 2001), was trained on 750 viral protein instances. Each instance was represented by a uniform feature vector derived from the collected metadata and classified by our team into five importance levels based on criteria outlined in Table 6.1. For instance, proteins with minimal evidence and limited functional understanding receive low scores (level 1), while those essential for viral replication or targeted by antiviral drugs receive high scores (level 5).

For each protein associated with a coding region where k-mer variations are located, we extract the taxonomy ID and genomic region from the GenBank reference sequence file to identify the corresponding viral protein through the UniProt API. The retrieved metadata is then processed by our model to assign an importance score (1-5) based on the protein's functional importance.

6.3.5 Mutational score

The third component of the KSS assesses the impact of mutations resulting from nucleotide changes in the k-mer variations on the biophysical properties of the involved amino acids. Amino acid properties and positions critically influence numerous biological processes ; mutations can lead to subtle or drastic alterations in protein function or disease manifestation (Betts et Russell, 2003). Amino acids vary in size, charge, and polarity, encompassing categories such as acidic, basic, aromatic, or aliphatic side chains.

Mutations are classified into two types :

- **Conservative replacements :** An amino acid is substituted with another that has similar properties. This type of replacement is expected to rarely result in dysfunction of the corresponding protein.
- **Radical replacements :** An amino acid is substituted with another that has different properties. This can lead to changes in protein structure or function, potentially resulting in phenotypic changes, sometimes pathogenic.

Substitutions of an amino acid with another of similar properties are less likely to destabilize the protein

or induce significant functional changes than substitutions with one of distinct properties (Betts et Russell, 2003; Rudnicki *et al.*, 2014; Das et Roy, 2021). Indeed, substitution frequencies correlate much more strongly with the overall chemical differences between exchanging residues than with the minimum base changes between their codons (Grantham, 1974).

Quantitative assessment of the impact of residue changes is facilitated by substitution matrices, which correlate strongly with amino acid properties (Tomii et Kanehisa, 1996; Rudnicki *et al.*, 2014). Recognized matrices such as the Point Accepted Mutation (PAM) matrix and the BLOcks SUbstitution Matrix (BLOSUM) are derived from large sets of aligned sequences, documenting the frequency of specific substitutions and their expected occurrences by chance (Dayhoff *et al.*, 1978; Henikoff et Henikoff, 1992). Higher values in these matrices indicate substitutions that occur frequently in nature, suggesting their biological favorability, while lower scores indicate less favorable substitutions between amino acids with distinct biophysical properties.

A substitution matrix M with dimensions 20×20, such as BLOSUM62, can be represented as follows :

$$M = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,19} & a_{1,20} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,19} & a_{2,20} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{19,1} & a_{19,2} & \cdots & a_{19,19} & a_{19,20} \\ a_{20,1} & a_{20,2} & \cdots & a_{20,19} & a_{20,20} \end{bmatrix} \tag{6.1}$$

where a_{ij} represents the substitution value between residues i and j.

To calculate the mutational score of a set of k-mers/variations, we first normalize M to range between 0 and 1 :

$$M_{\text{scaled}} = \frac{M - \min(M)}{\max(M) - \min(M)} \tag{6.2}$$

For each set of k-mers and their variations, we extract a list L of amino acid mutation pairs resulting from nucleotide changes in the k-mer variations, as reported by KANALYZER (e.g., $L = [(\mathsf{R}, \mathsf{K}), (\mathsf{R}, \mathsf{M}), (\mathsf{T}, \mathsf{I})]$). We identify the smallest value V, corresponding to the substitution least likely to occur naturally and potentially most disruptive :

$$V = \min_{(a_i, b_i) \in L} (M_{\text{scaled}}(a_i, b_i)) \tag{6.3}$$

The mutation score is then calculated as mutationScore $= 1 - V$, representing the most consequential biophysical alteration among the mutations provided. To classify the levels of impact, disturbance scores are categorized using thresholds [0.2, 0.4, 0.6, 0.8], assigning scores from 1 to 5 to reflect the range of disturbance significance.

6.3.6 Calculation and output of the *KmerSignificance Score* framework

The *KmerSignificance Score* (KSS) for each set of k-mer variations is computed as the arithmetic mean of its three component scores : discriminative score, protein score, and mutational score. The framework outputs a JSON file that, for a given dataset, contains detailed information about the analyzed coding regions. For each nucleotide position within these regions, the file includes the reference sequence k-mer, a list of extracted variations found in the analyzed sequences according to the defined threshold, the associated amino acid changes, the three component scores (discriminative, protein, and mutational scores), and the final KSS value.

6.3.6.1 Evaluation of discriminative power using real viral sequence data

The discriminative aspect was evaluated using viral datasets selected for their established classification schemes and significant public health impact. These viruses were chosen based on three criteria : (1) well-documented genotype/phenotype relationships, (2) substantial sequence data availability, and (3) clinical relevance.

First, we analyzed two RNA viruses with distinct evolutionary patterns and classification systems :

1. **Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)**, sourced from the NCBI Viral Genomes Resource Hatcher *et al.* (2017). This virus represents a model of recent emergence with rapid evolutionary dynamics and global health significance due to its role in the COVID-19 pandemic.
2. **Human Immunodeficiency Virus type 1 (HIV-1)**, obtained from the Los Alamos HIV Database (`https://www.hiv.lanl.gov/`). HIV-1 serves as a model for high genetic diversity and well-characterized subtype classification, with decades of molecular epidemiology data.

The selection of these two viruses enables us to evaluate whether the discriminative k-mer variations identified by our framework correspond to established mutation patterns that characterize viral classes in current state-of-the-art classification systems.

To validate our approach on a distinct viral system, we analyzed select genes from a DNA virus :

3. **Human betaherpesvirus 5 (Human Cytomegalovirus, HCMV)**, retrieved from the NCBI Viral Genomes Resource Hatcher *et al.* (2017).

The purpose of including this HCMV dataset is twofold. First, it serves to demonstrate the robustness of KSS when applied to less-studied viral models with different genomic architectures. Second, by applying the mutational criteria established in Arav-Boger *et al.* (2002); Chou et Dennison (1991); Pignatelli *et al.* (2001), we aim to validate historical genotyping classifications using the substantial number of sequences now available in current databases.

For each dataset of SARS-CoV-2 and HIV-1, we collected multiple instances selected randomly to represent the different classes of each virus, based on the availability of sequences, aiming to balance the different classes and allowing a maximum of 1% missing nucleotides. Regarding the HCMV datasets, we collected all complete sequences of the genes US28, UL55, and UL73 available on the NCBI Viral Genomes Resource. Details regarding each dataset are summarized in Table 6.2.

Table 6.2 – Viral sequence dataset

Viral organism	Classification type	Analyzed coding genes	Average sequence length	Number of instances [min-max]	Number of classes
Severe acute respiratory syndrome coronavirus 2	Variants (PANGO lineages)	ORF1a / ORF1b / S / E / M / N	29762 ± 70	3124 [19-100]	34
Human immunodeficiency virus 1	Subtypes	gag / gag-pol / env	8958 ± 303	910 [21-100]	14
Human betaherpesvirus 5 (UL55)	Genotypes	UL55	2722 ± 3	367 [33-190]	4
Human betaherpesvirus 5 (UL73)	Genotypes	UL73	413 ± 4	524 [43-102]	7
Human betaherpesvirus 5 (US28)	Genotypes	US28	1065 ± 1	423 [25-159]	6

For sequence analysis with our framework, we fixed $k = 9$ for extracting k-mers and their variations by scanning sequences without overlap, consistent with previous virus classification studies Solis-Reyes *et al.* (2018); Lebatteux *et al.* (2024). We set a threshold $t = 0.25$ defining the minimum frequency required for a k-mer in at least one class to be retained for analysis. The discriminative score of each k-mer and its variations was then calculated as described in Section 6.3.3. To evaluate our RMS Rank-Weighted F1 Score, we compared it with traditional F1-score variants (weighted, macro, and micro averages).

To investigate the relationship between discriminative power and sequence conservation, we computed

PhastCons and PhyloP scores Siepel *et al.* (2005). These conservation metrics were calculated using the PhastWeb platform Ramani *et al.* (2019), based on multiple sequence alignments of consensus sequences from each viral class. The alignments were generated using MAFFT version 7 Katoh et Standley (2013) with the parameter `-maxiterate 1000`.

6.3.7 Functional importance of genomic location

To build an optimal predictive model for protein importance, we implemented a comprehensive evaluation strategy based on supervised machine learning. We compared several machine learning algorithms : Random Forests Breiman (2001), SVM Cortes (1995), k-Nearest Neighbors (KNN) Fix (1985), Multinomial Naïve Bayes McCallum *et al.* (1998), and Multilayer Perceptrons (MLP) Rumelhart *et al.* (1986). For each algorithm, we implemented a systematic optimization pipeline that included :

— Testing of different preprocessing scaling methods.

— Hyperparameter tuning.

— Feature selection using recursive feature elimination.

Model optimization was performed through grid search with stratified 10-fold cross-validation on the training set defined in Section 6.3.4, evaluating F1 scores with both macro and weighted averages.

Robustness of the best performing model with its optimized configuration (hyperparameters, preprocessing methods, and selected features) was validated on an independent test set of 125 viral proteins. This test set, curated by an external researcher specialized in functional analysis, ensured balanced representation across functional importance categories with 25 instances per category. Using precision, recall, and F1 scores, supplemented by confusion matrices, we analyzed prediction patterns and evaluated model performance. Additionally, we evaluated model consistency using a comprehensive UniProt dataset of viral proteins (13,719 entries excluding bacteriophage-related proteins, collected 2024-07-02). The dataset was divided into two groups based on publication volume : the top 1% most-studied proteins and the remaining 99%. This stratification enabled testing the hypothesis that highly studied proteins would correlate with higher functional importance scores, while less-studied proteins would display a more uniform distribution of scores.

6.3.8 Impact of mutations on the biophysical properties of residues

To identify the substitution matrix that best correlates with biophysical property distances between amino acids for scoring mutations in the KSS framework, we first considered eight properties commonly emphasized in scientific literature, as described in Table 6.3.

Table 6.3 – Biophysical properties of amino acids.

Amino Acid	Side chain hydrophobicity	Polarity	Volumes of side chains	Fraction of accessible area lost	Isoelectric point	In–out propensity	Aromaphilicity
A	0,31	8,10	27,50	0,74	6,01	0,20	0,03
C	1,54	5,50	44,60	0,91	5,07	0,67	0,20
D	-0,77	13,00	40,00	0,62	2,77	-0,72	-0,03
E	-0,64	12,30	62,00	0,62	3,22	-1,09	0,05
F	1,79	5,20	115,50	0,88	5,48	0,67	0,58
G	0,00	9,00	53,2	0,72	5,97	0,06	0,00
H	0,13	10,40	79,00	0,78	7,59	0,04	0,45
I	1,80	5,20	93,50	0,88	6,02	0,74	0,20
K	-0,99	11,30	100,00	0,52	9,74	-2,00	0,10
L	1,70	4,90	93,50	0,85	5,98	0,65	0,13
M	1,23	5,70	94,10	0,85	5,74	0,71	0,33
N	-0,60	11,60	58,70	0,63	5,41	-0,69	0,20
P	0,72	8,00	41,90	0,64	6,48	-0,44	0,13
Q	-0,22	10,50	80,70	0,62	5,65	-0,74	0,30
R	-1,01	10,50	105,00	0,64	10,80	-1,34	0,75
S	-0,04	9,20	29,30	0,66	5,68	-0,34	0,13
T	0,26	8,60	51,30	0,70	5,87	-0,26	0,08
V	1,22	5,90	71,50	0,86	5,97	0,61	0,15
W	2,25	5,40	145,50	0,85	5,89	0,45	1,00
Y	0,96	6,20	117,30	0,76	5,66	-0,22	0,85

Side chain hydrophobicity : reflects the tendency of the side chain to repel water, influencing protein folding and stability in aqueous environments FAUCHÈRE *et al.* (1988). **Polarity :** characterizes the distribution of electric charges across amino acids, determining molecular interactions and solubility properties Grantham (1974). **Side chain volume :** quantifies the spatial occupation of amino acid side chains, impacting protein packing and conformational stability Krigbaum et Komoriya (1979). **Isoelectric point :** defines the pH at which an amino acid carries no net electrical charge, influencing protein behavior Zimmerman *et al.* (1968). **Fraction of accessible area lost :** measures the surface area buried during protein folding, reflecting conformational changes Rose *et al.* (1985). **Interior–exterior propensity :** quantifies amino acid preferences for protein interior versus surface locations, determining structural organization Miller *et al.* (1987). **Aromaphilicity :** measures the tendency for aromatic interactions between amino acids, contributing to protein structural stability Hirano et Kameda (2021). Additionally, we compute the **Minimum codon distance** Xia et Xia (2018), which represents the minimal number of nucleotide changes required between amino acid codons. This property, not shown in the table as it consists of a pairwise distance matrix between amino acids, is directly integrated into our KSS framework to evaluate mutational distances. For more details, please refer to the cited literature.

Then, we selected 30 substitution matrices from the AAindex database Kawashima *et al*. (1999), including BENNER, BLOSUM, PAM, and others, to determine the matrix that best encapsulates the correlation with biophysical property distances between amino acids. The correlation between each matrix and amino acid biophysical property distances was evaluated using both Spearman Spearman (1961) and Pearson correlation coefficients Pearson (1895), with particular emphasis on properties such as hydrophobicity, polarity, and side chain volume, which are extensively studied in the literature Betts et Russell (2003); Rudnicki *et al*. (2014); Das et Roy (2021).

Finally, we optimized the values of the best-performing matrix using a genetic algorithm. The optimization process consisted of several key steps :

- Generation of an initial population based on the previously identified best matrix.
- Evolution process including :
  - Mutations : randomly adding or subtracting values from specific matrix elements.
  - Crossovers : randomly exchanging corresponding amino acid values between matrices.
- Evaluation process : computing global correlation between generated matrices and amino acid properties.
- Selection process : retaining and duplicating the top 50% of each population for the next generation.

This optimization was performed over 1,000 iterations with a population size of 1,000 matrices per iteration, using mutation and crossover rates of 0.2 and 0.5, respectively. The matrix yielding the highest cumulative correlation scores was selected for calculating mutation impact scores in our KSS framework.

6.4 Results and Discussion

6.4.1 Discriminative importance of k-mers and variations

The results presented in Figure 6.1 demonstrate the distribution of discriminative nucleotide regions across the five viral datasets, assessed using the RMS Rank-Weighted F1 score. A primary finding is that highly discriminative regions (scores between 4 and 5) consistently correlate with variable regions identified by PhastCons and PhyloP across all datasets.

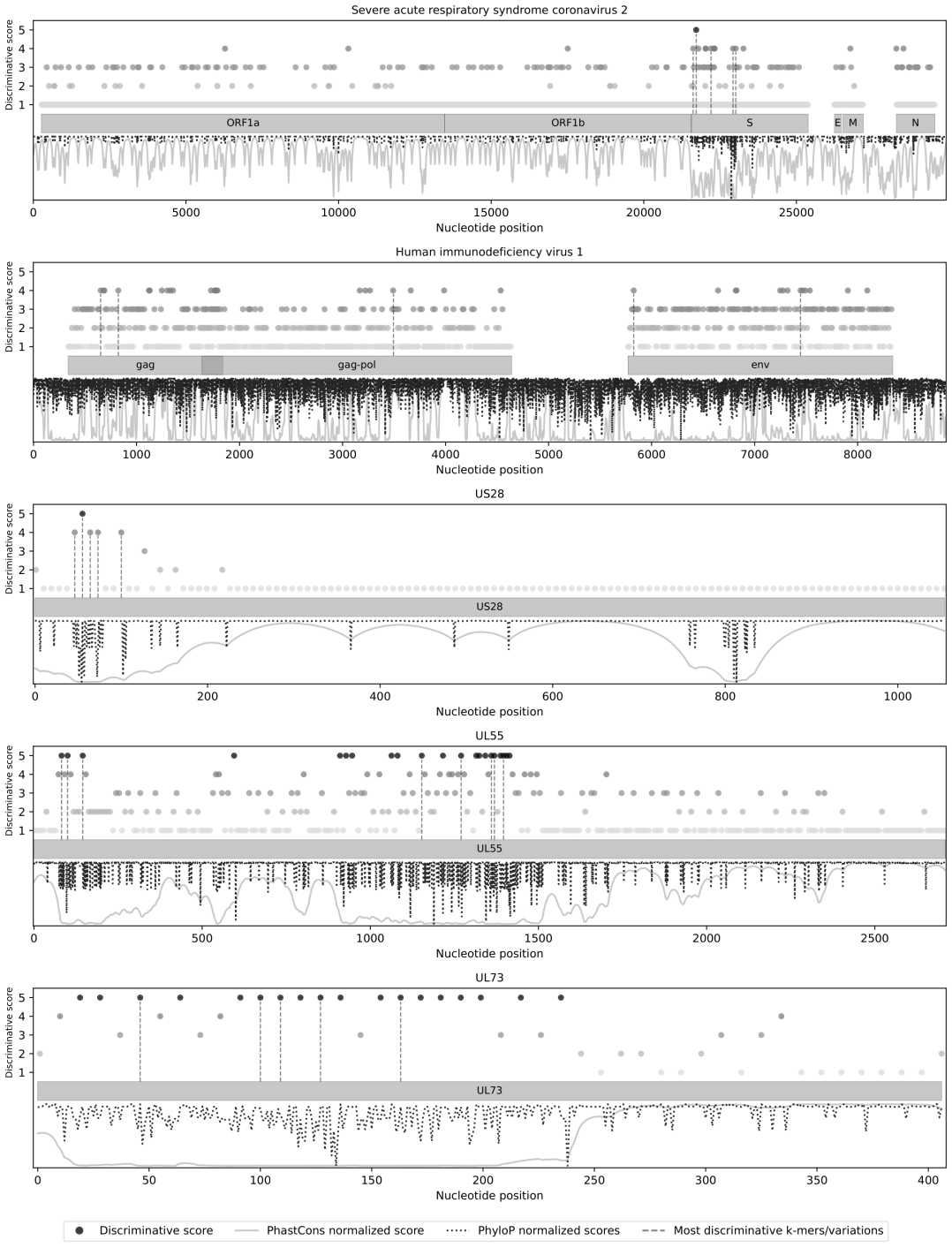


Figure 6.1 – Distribution of discriminative and variable nucleotide regions across viral datasets.

Each plot illustrates, for one of the five viral datasets, the distribution of nucleotide regions based on two metrics. The upper part shows discriminative importance assessed by the RMS Rank-Weighted F1 score, while the lower part displays genomic variability evaluated through PhastCons and PhyloP analyses. The middle section maps the studied coding regions, highlighting positions of highly discriminative regions for each viral organism.

In SARS-CoV-2, the most discriminative regions predominantly occur within the spike glycoprotein gene, which displays greater genetic variation compared to other structural proteins and genomic regions. These results align with established SARS-CoV-2 variant classification methods that primarily rely on structural proteins, particularly the spike protein Harvey *et al.* (2021), as documented in public databases such as CoVariants (`https://covariants.org/`). Low discriminative scores were observed in generally conserved regions, such as the ORF1ab Naqvi *et al.* (2020), while high discriminative potential was identified in known variable regions, including the N-terminal domain (NTD) and Receptor Binding Domain (RBD) containing its Receptor Binding Motif (RBM) Harvey *et al.* (2021), as well as variable regions in the nucleocapsid studied for genotyping Yin (2020).

For HIV-1, the most discriminative regions are distributed across the three studied coding regions (*gag*, *gag-pol*, and *env*). Specifically, highly discriminative variations are found within the Capsid (p24/CA) and Matrix (p17/MA) domains of the gag polyprotein, the Ribonuclease H (RNaseH) domain of the gag-pol polyprotein, and throughout the envelope glycoprotein (env). These results align with the fact that initially the *pol* region was favored for subtype assignment, although later the *env* region also proved to be rich in information for discriminating HIV-1 subtypes Peng (2024). HIV-1 exhibits higher nucleotide variability than the other organisms studied throughout its genome, especially at the 5'end of the *env* gene. This high variability explains why no maximum score was identified for HIV-1, where genomic variability within the same subtype reaches 15–20% Hemelaar *et al.* (2006).

Focusing on HCMV datasets, we observe that the most discriminative regions align with the least conserved regions identified by PhastCons and PhyloP. In US28, the discriminative regions are concentrated within the first 180 nucleotides. These findings, based on our analysis of 423 sequences, validate the results of Arav-Boger *et al.* (2002), who initially identified this region for genotype classification using only 61 sequences. For UL55, the majority of discriminative information is located in the middle of the gene, corresponding to its most variable region. This confirms the findings of Chou et Dennison (1991), who defined four UL55 genotypes based on 14 sequences. Our current analysis of 367 UL55 sequences not only validates these historical results but also identifies additional discriminative information in the 5' variable region for genotype classification. For UL73, the gene exhibits two distinct regions : a less conserved 5' region containing discriminative information that enables differentiation of the seven UL73 genotypes, and a more conserved 3' region. These discriminative regions, identified using 524 UL73 sequences, confirm the findings of Pignatelli *et al.* (2001), who initially characterized UL73 genotypes using only 43 sequences.

Our approach based on the RMS Rank-Weighted F1 score successfully distinguished between regions with varying discriminative importance, while this discrimination was not achievable using traditional F1 scores (micro, macro, and weighted). With traditional metrics, all nucleotide regions of SARS-CoV-2 and HIV-1 were assigned a minimum score of 1 (results available in the supplementary material folder of the GitHub repository). This limitation arises from the high number of classes in these datasets, where a single k-mer and its variations cannot effectively discriminate between all classes, thus necessitating the adaptation of these metrics to highlight discriminative regions.

6.4.2 Protein importance score

Through our comprehensive evaluation strategy of supervised machine learning algorithms, assessing preprocessing scaling methods, hyperparameter tuning, and feature selection for each algorithm, the Random Forest classifier (n_estimators = 250, criterion = "gini", max_features = None, bootstrap = True, random_state = 42) emerged as the best model, offering the most balanced performance on cross-validation of the training set. The pipeline incorporated MinMaxScaler for feature scaling and RFECV (Recursive Feature Elimination with Cross-Validation) for feature selection. The results associated with this model, illustrated in Figure 6.2A and B, show an average performance of 0.93 across all metrics (precision, recall, and f1-score). Extreme classes (1 and 5) achieved the highest performance with f1-scores of 0.96 and 0.97 respectively, while intermediate classes (2, 3, and 4) showed slightly lower performance with f1-scores of 0.91, 0.90, and 0.91. The confusion matrix reveals that classification errors occur only between adjacent classes, with misclassifications limited to scores one level higher or lower.

Validation on the independent test set (Figure 6.2D and E) maintained similar average performance metrics. Class 1 retained its f1-score of 0.96, while class 5 showed a slight decrease in performance, corresponding to improved accuracy in classes 2 and 3. The confusion matrix exhibits the same pattern of adjacent-class misclassifications observed in the training data. The model's consistent performance and pattern of limited misclassifications between neighboring classes suggest its reliability for estimating viral protein importance, particularly considering that the reference annotations in both training and test sets involve some degree of subjective interpretation among researchers.

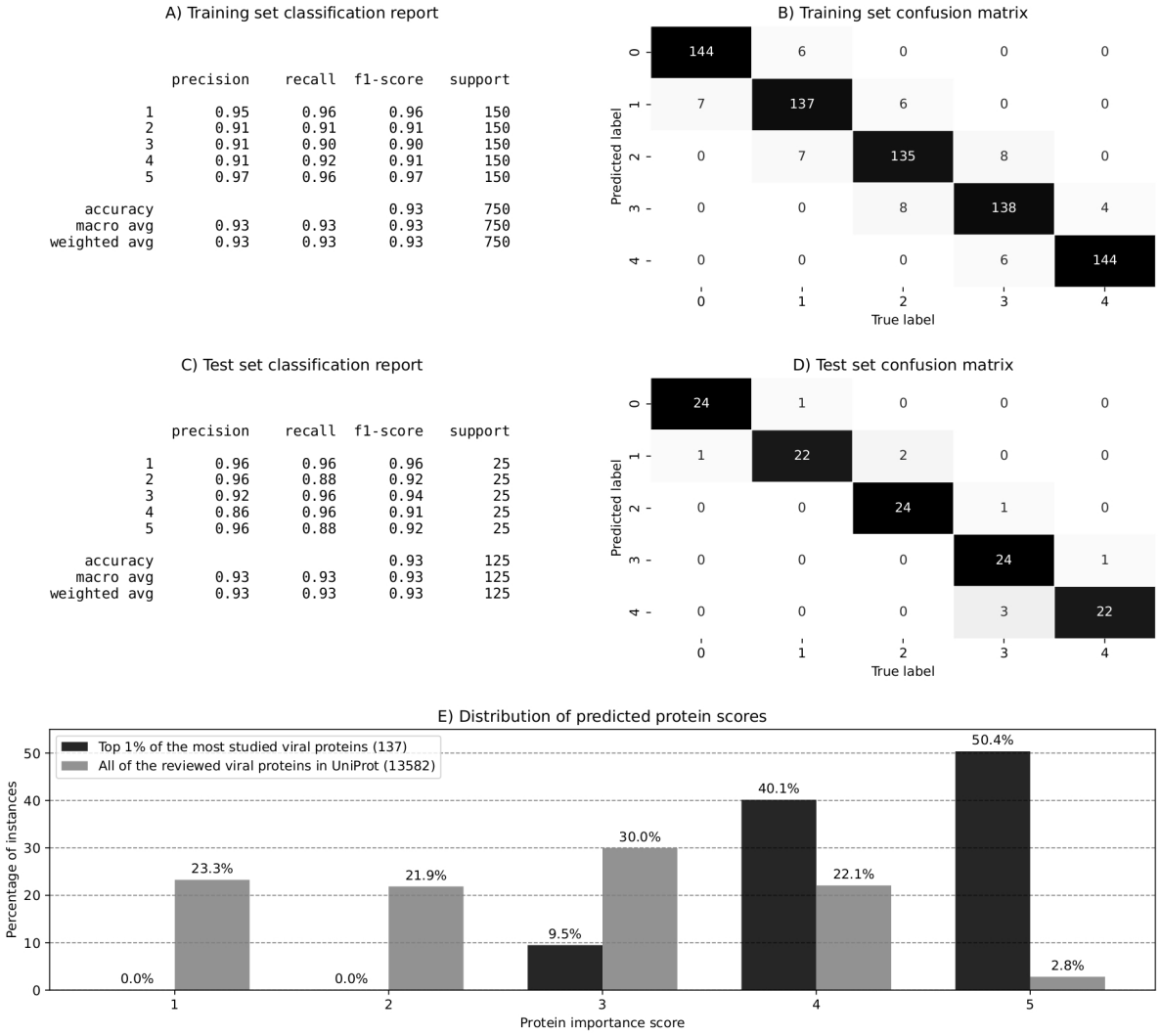


Figure 6.2 – Robustness and consistency of the protein importance score prediction model for viral proteins.

Panels A) and B) represent, respectively, the performance in terms of precision, recall, f1-score, and confusion matrix obtained by our model on the training set from a 10-fold stratified cross-validation evaluation. Panels C) and D) are similar to panels A) and B) but for the prediction on the test set. Panel E) represents the distribution of the predicted protein importance scores assigned by our model between the top 1% of the most studied viral proteins and the rest of the reviewed viral proteins available on UniProt.

Then, focusing on the prediction analysis of the complete UniProt viral database illustrated in Figure 6.2E, we observe distinct distributions between the top 1% most studied viral proteins and the remaining proteins. The top 1% were exclusively assigned functional importance scores between 3 and 5, with over 90%

receiving scores of 4 or 5. These results align with the expectation that highly studied viral proteins not only demonstrate greater functional significance but also benefit from extensive metadata across various databases. The remaining viral proteins show a more balanced distribution of scores from 1 to 4, with 21.9% assigned a score of 2 and 30% a score of 3. Notably, only 2.8% of these proteins received a score of 5. This distribution appears coherent, showing a relatively uniform spread across lower scores, while the highest score remains predominantly associated with the most studied viral proteins.

6.4.3 Impact of mutations on the biophysical properties of residues

Our analysis, depicted in Figure 6.3 (panels A, B, C, and D), evaluates the correlation between mutation impact scores from 30 substitution matrices and amino acid biophysical properties. Among existing matrices, the MIYATA substitution matrix Miyata *et al.* (1979), despite being based solely on volume and polarity, demonstrates the strongest correlations with all eight evaluated biophysical properties, particularly with hydrophobicity, polarity, and side chain volume. The BENNER matrices (BENNER22 and BENNER74) Bennet *et al.* (1994), derived from log-odds scores for evolutionary distances of 22-29 PAM and 74-100 PAM respectively Dayhoff (1978), show the second-highest correlation with residue biophysical properties. Interestingly, Grantham's distance matrix Grantham (1974), although based on composition, polarity, and molecular volume, ranks lower in overall correlation, despite showing strong correlations specifically with side chain volume and polarity. Similarly, the SNEATH matrix Sneath (1966), while developed using 134 categories of biological activity and chemical structure, displays weak correlations with our selected main biophysical properties. The commonly used BLOSUM Henikoff et Henikoff (1992) and PAM Dayhoff (1978) matrices prove less effective than MIYATA for quantifying biophysical changes.

Based on these results, we selected the MIYATA matrix for optimization using our genetic algorithm approach as described in Section 6.3.8. The optimization process yielded a new matrix (MIYATA_EVO) with an improved global Spearman correlation of 4.49, representing a 25% increase compared to the original matrix. Correlation improvements were observed for hydrophobicity (from 0.73 to 0.83) and polarity (from 0.74 to 0.78). However, we noted a significant decrease in correlation with side chain volume (from 0.48 to 0.28), a trade-off that appeared necessary to enhance correlations with other properties.

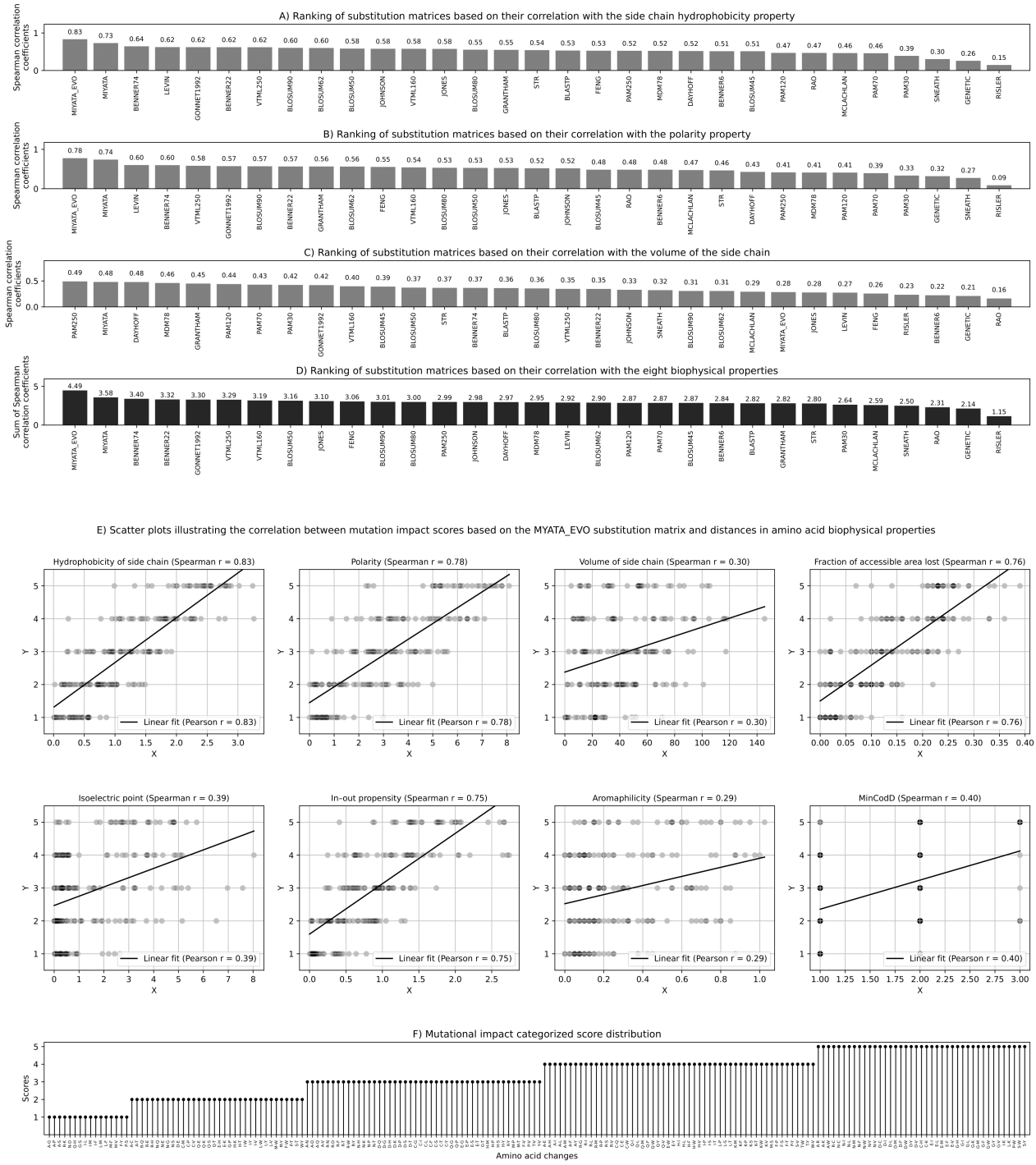


Figure 6.3 – Correlation between substitution matrices and amino acid biophysical property distances.

Panels A), B), C), and D) respectively illustrate the Spearman correlation rankings of the different substitution matrices with the properties of hydrophobicity, polarity, side chain volume, and the sum of the correlations with the 8 biophysical properties of amino acids. The scatter plots in panel E) represent the correlation between mutation impact scores based on the MIYATA_EVO substitution matrix (the matrix whose impact scores are most correlated with the distances between all amino acid biophysical properties) and the distances in amino acid biophysical properties. Panel F) illustrates the distribution of mutation impact scores based on the MIYATA_EVO substitution matrix.

Looking at panel E) in Figure 6.3, which shows the correlation between MIYATA_EVO matrix mutation impact scores and biophysical property distances across amino acid exchanges, we observe four strong correlations ($r \geq 0.75$) with hydrophobicity, polarity, fraction of accessible area lost, and in-out propensity. Additionally, four moderate correlations ($0.29 \leq r \leq 0.40$) were observed with side chain volume, isoelectric point, aromaphilicity, and minimum codon distance. Based on these results, we selected MIYATA_EVO for evaluating amino acid changes' importance regarding their impact on biophysical properties that influence some of protein function. Based on this matrix panel F) in Figure 6.3 illustrates the distribution of mutation impact scores across different amino acid substitutions, showing a balanced distribution of impact scores among possible amino acid changes.

6.4.4 Mutations of interest are associated with the most discriminative k-mers/variations identified by the KSS

Tables 6.5 and 6.4, extracted from the KSS report, demonstrate that the most discriminative k-mers/variations frequently correspond to non-synonymous mutations at the amino acid level. The Spike protein of SARS-CoV-2, one of the most studied glycoproteins, illustrates this correspondence. We identified the five most discriminative k-mers associated with multiple amino acid changes within two key regions : the N-terminal domain (NTD) including T19I/R, T20N, R21T, Q52R/H, D215G, Ins215EPE, L216F, and the Receptor Binding Motif (RBM) including L455S, F456L, E484A/Q/K, F486P/V/S Huang *et al.* (2020); Lan *et al.* (2020); Xia *et al.* (2020).

T19 mutations affect remdesivir interactions in Delta and Omicron strains, where T19I perturbs the interaction interface, while T19R decreases binding affinity Mahmood *et al.* (2022); Saifi *et al.* (2022). In the same NTD region, T20N introduces a new N-glycosylation site (N_{20}RT) in the Gamma strain, disrupting NTD antibody recognition and potentially enabling immune escape Pegg *et al.* (2024). Nearby, R21T, while not clearly impactful alone, appears to enhance viral entry into lung cells when combined with other mutations Hoffmann *et al.* (2021). Further along the protein sequence, D215G, characteristic of B.1.351 variant, affects neutralizing response to the D614G variant without impacting viral entry, antibody production, or T cell responses Peng *et al.* (2022). In the same region, L216F might alter the antigenicity of BA.2.86 Yang *et al.* (2023). Additionally, the insertion Ins215EPE, present in Omicron BA.1 but absent in BA.2 Venkatakrishnan *et al.* (2022); Gerdol *et al.* (2022); Greco et Gerdol (2022), reduces NTD monoclonal antibody binding and increases Spike expression Javanmardi *et al.* (2022).

Table 6.4 – Mutational landscape of k-mers/variations highlighted by the KSS in SARS-CoV-2 and HIV-1

VIRAL ORGANISM	GENE	POSITION	REFERENCE K-MER	VARIATION	AMINO ACID CHANGE	CLASS	DISCRIMINATIVE SCORE	MUTATIONAL SCORE	PROTEIN SCORE	K-MER SIGNIFICANCE SCORE
Severe acute respiratory syndrome coronavirus 2	S	55	ACAACCAGA	ACAAACAGA	T20N	P.1 (98%)	5	2	5	4,7
				ATAACTACA	T19I / R21T	BA.2.86.1 (100%) / JN.1 (99%) / JN.1.11.1 (100%)		3 / 3		
				ATAACCAGA	T19I	B.1.1.529 (28%) / BA.2.12.1 (100%) / BA.2 (99%) / BA.2.75 (99%) / BA.4 (100%) / BA.5 (100%) / (BQ.1 100%) / CH.1.1 (100%) / EG.5.1 (94%) / XBB (100%) / XBB.1.5 (100%) / XBB.1.16 (100%) / XBB.2.3 (97%)		3		
				AGAACCAGA	T19R	B.1.617.2 (98%)		3		
		154	CAGGACTTG	CAGGATTTG	Silent	BA.2.12.1 (93%)	5	0		4,0
				CGGGACTTG	Q52R	B.1.525 (99%)		2		
				CATGACTTG	Q52H	EG.5.1 (94%)		1		
		1360	AGATTGTTT	AGATCGCTT	L455S / F456L	JN.1.11.1 (98%)	5	1/ 4		4,7
				AGATCGTTT	L455S	JN.1 (98%)		4		
				AGATTGTTA	F456L	EG.5.1 (86%)		1		
		1450	GAAGGTTTT	GCAGGTCCT	E484A / F486P	XBB.1.5 (91%) / XBB.1.16 (100%) / XBB.2.3 (99%) / EG.5.1 (89%)	5	3 / 4		4,7
				GCAGGTTTT	E484A	BA.1 (98%) / BA.2 (100%) / BA.2.12.1 (100%) / BA.2.75 (87%)		3		
				GCAGGTGTT	E484A / F486V	BA.4 (96%) / BA.5 (99%) / BQ.1 (100%)		3 / 3		
				CAAGGTTTT	E484Q	B.1.617.1 (100%)		2		
				GCAGGTTCT	E484A / F486S	CH.1.1 (99%) / XBB (77%)		3 / 4		
				AAAGGTCCT	E484K / F486P	BA.2.86.1 (100%) / JN.1 (100%) / JN.1.11.1 (100%)		2 / 4		
				AAAGGTTTT	E484K	B.1.351 (99%) / B.1.525 (99%) / B.1.526 (62%) / B.1.621 (95%) / P.1 (97%)		2		
		640	CGTGATCTC	CGTGATTTC	L216F	BA.2.86.1 (100%) / JN.1 (100%) / JN.1.11.1 (100%)	5	1	5	4,3
				CGTGGTCTC	D215G	B.1.351 (96%)		3		
				CGTGAGCCAGAAGATCTC	-215E / -215P / -215E	BA.1 (95%) / B.1.1.529 (38%)		5 / 5 / 5		
Human immunodeficiency virus 1	env	55	ACCATGCTC	ATTATGCTC	T19I	63_02A6 (50%)	5	3	4	4,7
				CTCATGATC	T19L / L21I	11_cpx (90%)		3 / 1		
				———	T19- / M20- / L21-	A6 (89%) / 01_AE (56%)		5 / 5 / 5		
				GTCTTAGGC	T19V / M20L / L21G	08_BC (70%)		3 / 1 / 4		
				ATCTTAGGC	T19I / M20L / L21G	C (69%) / 08_BC (27%)		3 / 1 / 4		
				ACTTTGATC	M20L / L21I	01_AE (33%) / G (55%)		1 / 1		
				ATTTTGATC	T19I / M20L / L21I	20_BG (88%)		3 / 1 / 1		
				CTTTTATTC	T19L / M20L / L21F	F1 (77%)		3 / 1 / 1		
				ACTATGATC	L21I	A1 (25%) / 35_A1D (71%)		1		
		1675	ATTGAGGCG	ATCGAGGCA	Silent	11_cpx (33%)	5	0		3,7
				ATAGAGGCG	Silent	01_AE (94%) / G (78%) / C (94%) / D (92%) / 08_BC (93%) / 20_BG (100%)		0		
				ATAGAGGCT	Silent	02_AG (93%) / A1 (95%) / A6 (98%) / 35_A1D (95%) / 63_02A6 (100%)		0		
				ATTGAGGCA	Silent	11_cpx (33%)		0		
				ATACAGGCC	E560Q	O (93%)		2		
				ATTGAAGCG	Silent	F1 (88%)		0		
	gag	316	GAAGAGCAA	GAAGTACAA	E107V	01_AE (57%)	5	5	4	4,7
				AAGTTACAA	E106K / E107L	20_BG (100%)		2 / 5		
				GAAGTGCAA	E107V	63_02A6 (69%)		5		
				GAAGTAATG	E107V / Q108M	O (52%)		5 / 4		
				GAAGAACAA	Silent	C (82%) / D (82%) / F1 (42%) / 08_BC (93%)		0		
				GAAATACAA	E107I	02_AG (24%) / A1 (53%) / A6 (65%) / 35_A1D (86%) / 11_cpx (71%)		5		
		487	GCTTTCAGC	GGTTTTAAC	A163G / S165N	01_AE (75%)	5	1 / 2		3,7
				GCTTTTAGC	Silent	C (45%) / 02_AG (27%) / F1 (53%) / 08_BC (97%) / 63_02A6 (85%)		0		
				GCTTTCAGT	Silent	20_BG (88%)		0		
				GCCTTCAGT	Silent	G (70%)		0		
				GCCTTTAGC	Silent	F1 (42%)		0		
				GCCTTTAAC	S165N	O (100%)		2		
	gag-pol	3167	AGGGGTGCC	AGGTCTGCC	G946S	02_AG (70%) / F1 (67%)	5	1	5	4,7
				AAGGCCTCC	R945K / G946A / A947S	O (34%)		1 / 2 / 1		
				AGGACTGCC	G946T	C (61%) / 08_BC (80%) / 11_cpx (67%)		3		
				AAAAGTGCC	R945K / G946S	20_BG (96%)		1 / 1		
				AGGTCTGCT	G946S	A1 (62%) / 01_AE (78%) / 35_A1D (86%)		1		
				GGGTCTGCT	R945G / G946S	A6 (77%) / 63_02A6 (81%)		4 / 1		
				GGGTCTGCC	R945G / G946S	G (81%)		4 / 1		

The columns **VIRAL ORGANISM** and **GENE** provide information on the studied organism and the genes where the most significant k-mers/variations are located, respectively. The columns **REFERENCE k-MER** and **VARIATION** indicate the form of the k-mer in the reference sequence and in the other classes of sequences. The column **AMINO ACID CHANGE** lists the amino acid changes implied by the variations compared to the reference k-mers. The column **CLASS** shows the classes and their percentages associated with the different variations. The columns **DISCRIMINATIVE SCORE**, **MUTATIONAL SCORE**, **PROTEIN SCORE**, and **k-MER SIGNIFICANCE SCORE** indicate the different importance scores calculated by the KSS.

Several viruses code and express polyproteins that are cleaved into different functional proteins. In the case of HIV-1, the gag polyprotein is divided into six proteins : Matrix protein (MA, residues 1–132), Capsid protein (CA, residues 133–363), Spacer Peptide 1 (SP1, residues 364–377), Nucleocapsid protein (NC, residues 378–432), Spacer Peptide 2 (SP2, residues 433–448), and p6 protein (residues 449–500) Marie et Gordon (2022). Among the five best k-mers predicted by the KSS 6.5, mutations were identified in both MA and CA regions. In MA, mutations E106K, E107V/L/I, and Q108M were found Marie et Gordon (2022). While E107V/L/I and Q108M are not well studied, residues E107 and Q108 appear crucial for MA folding and stability when R76G is present Giagulli *et al.* (2017). The loss of negative charge from E107 may destabilize MA through a mechanism similar to SARS-CoV-2 Spike protein, where loss of glutamic acid at position 484 disrupts salt bridge formation with ACE2 Koehler *et al.* (2021).

In the CA region, mutations A163G and S165N were identified. CA's 231 amino acids are organized into three domains : N-terminal (residues 1-145), linker (residues 146-150), and C-terminal (residues 151-231) Marie et Gordon (2022). A163G and S165N correspond to residues 31 and 33 within the N-terminal domain. Studies show that A31G alone reduces infectivity, which can be partially restored by S33N Crawford *et al.* (2007). Independently, both G31 and N33 decrease HIV-1 replication Payne *et al.* (2014).

In the gag-pol polyprotein, mutations R945K/G, G946S/A/T, and A947S occur in the RNase H region. Within reverse transcriptase (RT), composed of subunits p51 (residues 588-1027) and p15 (residues 1028-1147), these mutations correspond to residues 358, 359, and 360 in the nucleic acid interface interaction region Sarafianos *et al.* (2001). During antiviral treatments, R358K, G359S, and A360I/V mutations emerge frequently, though G359A/T and A360S are not documented Delviks-Frankenberry *et al.* (2008); Lengruber *et al.* (2011). While R358K shows no antiviral resistance correlation, G359A provides azidothymidine protection without affecting viral replication or RNase H activity Brehm *et al.* (2007); Delviks-Frankenberry *et al.* (2007).

In the case of HCMV (Table 6.4), we focused our analysis on three promising viral genes for the development of vaccine or pharmacologic strategies : US28, gB, and gN. The five most discriminative k-mers of US28 are localized within the N-terminal domain (E18D/L, D19E/A/G, T21A, V24T, F25L). This region of 37 amino acids is exposed on the surface of the viral particle and infected cells Lee *et al.* (2017).

Table 6.5 – Mutational landscape of k-mers/variations highlighted by the KSS in HCMV

VIRAL ORGANISM	GENE	POSITION	REFERENCE K-MER	VARIATION	AMINO ACID CHANGE	CLASS	DISCRIMINATIVE SCORE	MUTATIONAL SCORE	PROTEIN SCORE	K-MER SIGNIFICANCE SCORE
Human betaherpesvirus 5	US28	55	GACGCGACT	GAAGCGACT	D19E	D (100%)	5	2	5	4,3
				GACGCGGCT	T21A	C (100%)		2		
				GCCGCGACT	D19A	A1 (100%)		3		
				GGAGCAACC	D19G	B1 (100%) / B2 (100%)		3		
		64	CCTTGTGTT	CCCTGTGTC	Silent	D (100%)	4	0		4,0
				CCTTGTACC	V24T	B1 (100%) / B2 (100%)		3		
				CCCTGTGTT	Silent	C (100%)		0		
		46	TACGATGAA	TATGATGAA	Silent	C (94%)	4	0		4,7
				TACGACGAT	E18D	D (100%)		1		
				TACGACCTT	E18L	B1 (100%) / B2 (100%)		5		
		73	TTCACCGAC	CTCACCGAC	F25L	D (100%)	4	1		3,3
				CTAACTGAC	F25L	C (100%)		1		
		100	CCAGTTACG	CCGGTCACG	Silent	B1 (100%)	4	0		3.0
				CCAGTCACG	Silent	B2 (28%) / D (91%) / C (70%)		0		
				CCGGTTACG	Silent	A1 (91%)		0		
	UL55	1369	AGAACCAAA	AGAACCAAG	Silent	3 (100%)	5	1	4	3,3
				AGAACCAGA	K459R	4 (100%)		1		
				AGGACCAGA	K459R	2 (100%)		1		
		1153	TCCGACTCT	TCCGACCCC	S387P	3 (92%)	5	1		3,3
				TCCGATTCC	Silent	4 (97%)		0		
				TCCGACTCC	Silent	2 (100%)		0		
		1270	ACTGGTGGT	ACAGGCGGT	Silent	3 (100%)	5	0		3,7
				AGCGGCGGT	T424S	2 (100%)		2		
				ACCGGCGGA	Silent	4 (97%)		0		
		1396	AATGCAACT	ACGACCACC	N466T / A467T	3 (99%)	5	3 / 2		4,0
				AATGTAACT	A467V	4 (100%)		3		
				AATACAACT	A467T	2 (100%)		2		
		1360	ACTCATAAT	ACTCATAGT	N456S	4 (100%)	5	2		4,7
				ACG—CGT	H455- / N456R	3 (100%)		5 / 3		
				ACTCAT—	N456-	2 (99%)		5		
		82	ACTCGTGGA	ACTTCTCATGCA	-29S / R29H / G30A	3 (81%)	5	5 / 2 / 2		4,7
				ACTTCCCATGCA	-29S / R29H / G30A	2 (80%) / 3 (11%)		5 / 2 / 2		
				TCTCGTGCA	T28S / G30A	4 (100%)		2 / 2		
		100	ACTCACAGT	ACCCACAAT	S36N	2 (52%)	5	2		4,0
				ACTCACAAT	S36N	2 (35%) / 3 (19%)		2		
				CATAATGGA	T34H / H35N / S36G	4 (94%)		3 / 2 /1		
				GCTCACAAT	T34A / S36N	3 (79%)		2 / 2		
		145	CGATCCGGT	CGGTCA—	G51-	2 (35%)	5	5		4,7
				CGGTCAATC	G51I	4 (97%)		5		
				CGGTCAGTT	G51V	3 (83%)		5		
				CGGTCAGTC	G51V	2 (52%) 3 (15%)		5		
	UL73	163	ACGACGACT	ACGATGAGT	T56M / T57S	3a (99%)	5	3 / 2	2	3,3
				ACGACGACC	Silent	3b (100%)		0		
				ACGGCGACA	T56A	4c (97%)		2		
				ACGACAACT	Silent	2 (72%)		0		
				GTGACAAGT	T55V / T57S	4b (96%)		3 / 2		
				ACGAGGAGC	T56R / T57S	4a (99%)		3 / 2		
		100	AAGTCTTCT	AAATCGTCCAGT	-35S	3a (100%)	5	5		4.0
				CCAAGTCCT	K34P / S36P	2 (100%)		4 / 1		
				CCTCCTAGT	K34P / S35P	3b (100%)		4 / 1		
				AGCTCTACT	K34S / S36T	4a (100%) / 4b (96%)		4 / 2		
				AGTTCTACT	K34S / S36T	4c (98%)		4 / 2		
		136	ACAGTTGCA	ACTGTTGCG	Silent	4c (99%)	5	0		3,3
				ACCAGCGTA	V47S / A48V	3b (100%)		3 / 3		
				ACGAGCGTA	V47S / A48V	2 (100%)		3 / 3		
				ACTACTGCG	V47T	4b (98%)		3		
				ACAACTGCA	V47T	3a (88%) / 4a (100%)		3		
		46	GTGGCAGAG	GCGGCA—	V16A / E18-	4b (84%)	5	3 / 5		4.0
				GCGGCG—	V16A / E18-	4b (16%) / 4a (100%)		3 / 5		
				GCGGCAGGG	V16A / E18G	3b (70%)		3 / 3		
				GCGGTAGGG	V16A / A17V / E18G	3b (30%)		3 / 3 / 3		
				GCAGCGGGG	V16A / E18G	3a (99%)		3 / 3		
				GTGGCA—	E18-	4c (100%)		5		
				GTAACAGGG	A17T / E18G	2 (100%)		2 / 3		
		109	GCTAGCGTA	ACTAGCGTA	A37T	1 (26%)	5	2		4.0
				CACACCTCA	A37H / S38T / V39S	4b (98%)		4 / 2 / 3		
				TCTAGCGTG	A37S	3a (99%)		1		
				CACGCCTCA	A37H / S38A / V39S	4c (99%)		4 / 2 / 3		
				TCTAGTGTA	A37S	3b (100%)		1		
				TCTAGTGTG	A37S	2 (100%)		1		
				CGCACCTTA	A37R / S38T / V39L	4a (100%)		5 / 2 / 2		

The columns **VIRAL ORGANISM** and **GENE** provide information on the studied organism and the genes where the most significant k-mers/variations are located, respectively. The columns **REFERENCE k-MER** and **VARIATION** indicate the form of the k-mer in the reference sequence and in the other classes of sequences. The column **AMINO ACID CHANGE** lists the amino acid changes implied by the variations compared to the reference k-mers. The column **CLASS** shows the classes and their percentages associated with the different variations. The columns **DISCRIMINATIVE SCORE**, **MUTATIONAL SCORE**, **PROTEIN SCORE**, and **k-MER SIGNIFICANCE SCORE** indicate the different importance scores calculated by the KSS.

Among Class A GPCRs, the N-terminus can participate in the recognition of specific ligands, as illustrated by the interaction between the CXCR4 and its ligands CXCL12 (SDF-1α) and vMIP-II (a chemokine from human herpesvirus 8) Zhou *et al.* (2001); Qin *et al.* (2015). A study based on the prediction of the 3D structure of US28 variants suggests that mutations in the N-terminal domain impact the binding of several chemokines. Specifically, E18L could destabilize the binding of CX3CL1 and CCL2 ; D19A could disturb the interaction with CCL2/5/13 ; T21A and F25L negatively impact the binding of CX3CL1 and CCL2/3/4/5/13 Waters *et al.* (2021). Furthermore, E18D/L and D19A/E/G were found in different studies and seem to be under positive selection Arav-Boger *et al.* (2002); Gong et Padhi (2011). These mutations could also impact viral immune evasion, as shown by a study where HCMV seronegative persons who received a hematopoietic stem cell transplant from HCMV-positive donors developed a B-cell response against epitopes E13FDYDEDATPCVFTD27 and T5TTAELTTEFDYDED19 Perez-Bercoff *et al.* (2014).

For gB, discriminative mutations cluster in the N-terminal domain (NTD) (-28S, T28S, R29H, G30A, T34H/A, H35N, G51I/V/-), Domain II (DII) (S387P, T424S), and around the furin cleavage site (H455-, N456S/R/-, K459T, N466T, A467T/V) Liu *et al.* (2021). While NTD's function remains unclear, G51I/V/- within antigenic domain II Pötzsch *et al.* (2011) may disrupt antibody recognition. S387P in DII might affect protein folding due to proline's unique conformational properties Morgan et Rubenstein (2013), potentially impacting this highly glycosylated, antibody-targeted domain Burke et Heldwein (2015). N456 and K459 mutations within the furin cleavage site (residues 456–459) appear positively selected, potentially affecting glycosylation profiles and optimizing viral protein processing Stangherlin *et al.* (2017). For gN, our identified mutations (V16A, A17V/T, E18G/- in the signal peptide ; K34P/S through T57S in the extracellular domain) lack direct functional evidence. However, specific subtypes (4a, 4b, and 4c) correlate with symptomatic congenital infections, suggesting potential influence on infection severity Pignatelli *et al.* (2010).

6.5 Conclusion

In this article, we introduced the KSS, a novel heuristic scoring system designed to assess the relevance of discriminative k-mers and their variations within viral subspecies. This integrates taxonomic relevance with biological significance by considering genomic locations and the resulting amino acid changes. Each component of the KSS can also be used independently for specific purposes in various studies. For example, it can be employed to measure and identify sets of k-mers/variations within subspecies of different organisms, which may be useful for identifying variable and discriminative regions or for constituting signatures

that enable rapid and precise identification of certain viral strains. Additionally, our approach is pertinent for measuring the impact of specific amino acid changes associated with identified k-mers/variations and for gaining insights into the functional importance of viral proteins. Our experiments on a large set of substitution matrices highlighted those exhibiting the strongest correlation with amino acid biophysical property distances, particularly the MIYATA and BENNER matrices. These findings can guide the selection of appropriate matrices in future studies involving amino acid substitutions. Furthermore, we propose MIYATA_EVO, an optimized version of the MIYATA matrix that offers improved correlation with amino acid biophysical property distances.

Moreover, our results have demonstrated consistency with the discriminative regions associated with different subtypes and variants of HIV-1 and SARS-CoV-2, which are widely studied and for which extensive reference data is available. By focusing on genes of interest (UL55, UL73, and US28) of a less-studied virus like HCMV, our study has confirmed the validity of the mutational criteria defined in previous research, while enabling analysis of much larger sets of annotated sequences. Finally, our results have shown that the k-mers/variations with the highest KSS scores enable the extraction of subsets of sequences that are both relevant for viral taxonomy discrimination and associated with regions of mutations of interest that may impact protein functions. This underscores the utility of the KSS in both taxonomic classification and functional genomics.

Looking ahead, while the KSS could be applied to other viral genomes or extended to non-viral organisms, potentially aiding in the discovery of novel biomarkers and therapeutic targets, several improvements are under development. These include optimizing the discriminative score calculation to better handle datasets with numerous viral classes, such as HIV-1 and SARS-CoV-2. Additionally, we are developing a Large Language Model-based approach for protein importance scoring that will provide explanations based on established biological criteria. This development will be evaluated using a comprehensive test set incorporating annotations from multiple external evaluators. We believe that our approach offers valuable tools for researchers in genomics and bioinformatics, contributing to the advancement of personalized medicine and epidemiological surveillance. The complete KSS framework, along with all datasets and code to reproduce our experiments, is available at : `https://github.com/bioinfoUQAM/KmerSignificanceScore`.

6.6 Bilan et perspectives

Ce chapitre a introduit le *KmerSignificance Score* (KSS), un système de notation permettant de hiérarchiser les ensembles de k-mers et leurs variations au sein des sous-espèces virales. Ce système est dérivé de trois sous-scores : 1) Un score discriminant évaluant la capacité des k-mers et leurs variations à distinguer les sous-espèces virales ; 2) Un score protéique prédisant l'importance fonctionnelle des protéines encodées par les régions où se trouvent les variations et 3) Un score mutationnel évaluant l'impact des changements d'acides aminés via une matrice de substitution basée sur la conservation des propriétés biophysiques. Le chapitre suivant présentera une application du KSS à travers une étude détaillée du gène *UL33* du HCMV, un récepteur couplé aux protéines G, démontrant l'utilité de ce système de notation pour caractériser les régions discriminatives et les mutations permettant de définir différents génotypes de ce gène.

CHAPITRE 7

CHARACTERIZATION OF THE GENETIC VARIATION IN THE GENE ENCODING THE GPCR UL33 : A TOOL FOR HCMV PATHOGENESIS

7.1 Abstract

Human cytomegalovirus (HCMV) UL33, a viral G protein-coupled receptor, exhibits remarkable sequence variability among HCMV's GPCRs, yet its role in viral pathogenesis remains poorly understood. This study investigated UL33's genetic diversity and its potential involvement in viral transmission through comprehensive analysis of 392 complete UL33 sequences, including mother-infant paired samples from Kenyan and Ugandan cohorts. Comparative analysis with other HCMV GPCRs (US27, US28, and UL78) confirmed UL33's hypervariability, displaying the highest Shannon entropy (0.17), percentage of variable sites (23.27%), and nucleotide diversity (0.08). Phylogenetic analysis of mother-infant transmission pairs revealed predominant transmission of dominant maternal variants to infants, suggesting their essential role in establishing primary infection. Using the *KmerSignificance Score* (KSS) framework, we identified five distinct UL33 genotypes (Toledo, AD169, Merlin, Towne, and C5), characterized by specific mutational profiles. Our analysis highlighted twelve sets of discriminative k-mers with significant amino acid changes (KSS $\geq$ 4), primarily located in N-terminal and extracellular domains. These mutations could potentially affect receptor glycosylation and receptor activation. Notably, these variable regions overlap with predicted MHC class I epitopes, suggesting a possible role in immune evasion through modulation of T-cell epitope presentation. These findings provide new insights into UL33's genetic diversity and suggest that specific UL33 variants may influence HCMV transmission and pathogenesis, opening new avenues for therapeutic intervention.

Keywords : Human cytomegalovirus (HCMV), UL33, G protein-coupled receptor, genetic diversity, viral transmission, mother-infant transmission, genotypes, k-mer analysis

7.2 Introduction

Human cytomegalovirus (HCMV) is responsible for what has been termed a "silent pandemic." The virus infects more than 90% of the global population, with seroprevalence varying considerably across geographical regions. While developed countries such as Canada report prevalence rates of 50-60%, these rates can approach 100% in developing regions Zuhair *et al.* (2019). HCMV transmission occurs through expo-

sure to infected biological fluids, including urine, saliva, breast milk, and blood, with primary infection typically requiring multiple exposure events Mayer *et al.* (2017). Following initial infection, HCMV establishes a chronic infection characterized by alternating latent and lytic phases. Although HCMV infection remains largely asymptomatic in immunocompetent individuals, it can cause severe complications and mortality in immunocompromised patients, particularly those with HIV co-infection or organ transplant recipients under immunosuppressive therapy Gompels *et al.* (2012); Alyazidi *et al.* (2018). HCMV infection presents significant risks during pregnancy. Vertical transmission from mother to fetus can occur through two distinct mechanisms : during acute primary infection or through reactivation of chronic infection. Congenital HCMV infection affects approximately 0.5% of all births, with 16% of infected infants developing long-term impairments including hearing loss and neurodevelopmental delays Boppana *et al.* (2013); Stagno *et al.* (1982); de Vries *et al.* (2011). Despite its substantial global health burden, no effective prophylactic therapy currently exists to prevent HCMV infection in the general population Hu *et al.* (2022).

Understanding HCMV pathogenesis, particularly the molecular mechanisms governing viral entry and propagation, is therefore crucial for developing effective vaccines and antiviral drugs. Mayer *et al.* (2017) proposed that successful infection of buccal cells requires specific HCMV genotypic signatures, with individual viral particles possessing distinct haplotypes. The HCMV genome contains 165 coding genes, including several that encode glycoproteins essential for viral entry. These glycoproteins form distinct complexes that interact with specific cellular receptors : the trimeric complex (gH, gL, gO) binds to PDGFRα and TGFβR3, the pentameric complex (gH, gL, UL128, UL130, UL131a) engages NRP2 and THBD, while gB interacts specifically with EGFR Fulkerson *et al.* (2020); Kschonsak *et al.* (2021, 2022). This diverse array of receptor interactions explains the virus's broad cell tropism and reveals multiple potential targets for antiviral intervention.

Among its glycoproteins, the HCMV genome encodes four distinct G protein-coupled receptors (GPCRs). These receptors belong to an ancient family of proteins found across animals and fungi, characterized by a conserved seven-transmembrane architecture and interaction with trimeric G proteins. GPCRs serve diverse biological functions in organisms : rhodopsin mediates phototransduction in the eye, muscarinic receptors regulate neurotransmission, Frizzled receptors control cell migration, and chemokine receptors such as CCR5 coordinate immune responses Bleul *et al.* (1997); Rosenbaum *et al.* (2009). HCMV's viral GPCRs share significant homology with chemokine receptors, with US28 being the most extensively characterized. US28 contributes to viral pathogenesis through multiple mechanisms : it facilitates viral entry by recognizing the chemokine fractalkine, promotes immune evasion by internalizing circulating chemokines Kledal *et al.*

(1998); Fraile-Ramos *et al.* (2003), and plays critical roles in viral latency, reactivation, and dissemination throughout the host Noriega *et al.* (2014); Crawford *et al.* (2019); Krishna *et al.* (2017).

In contrast to US28, UL33 remains less thoroughly investigated despite displaying intriguing features potentially linked to HCMV pathogenesis. Functional studies have shown that UL33 can partially complement its murine CMV (MCMV) ortholog M33, restoring viral replication in mouse salivary glands and promoting viral spread in human cell lines Farrell *et al.* (2011); van Senten *et al.* (2020); Krishna *et al.* (2021). Additionally, UL33 can inhibit cellular chemokine receptors, including CCR5 and CXCR4 Tadagaki *et al.* (2012). A distinctive feature of UL33 is its high sequence variability compared to US28 ; its coding gene represents one of the most hypervariable loci within the HCMV genome and undergoes frequent recombination events Lassalle *et al.* (2016). Considering both Mayer's infection model and UL33's remarkable hypervariability, we hypothesize that specific UL33 sequence signatures may facilitate viral transmission and promote the establishment of primary infection, potentially contributing to persistent infection.

This study had four main objectives. First, we aimed to compare UL33 sequence variability with that of other HCMV GPCRs, extending the initial observations of Lassalle *et al.* (2016) who reported UL33's hypervariability in their analysis of 20 complete genomes. Second, we investigated UL33 variants specifically associated with primary infection using mother-infant paired samples from our Kenyan cohort. Third, we examined whether these variants were unique to our cohort by analyzing additional HCMV sequences available in GenBank. Finally, we characterized the mutational profiles of these variants to define distinct UL33 genotypes.

7.3 Materials and methods

7.3.1 Sample collection and sequencing

Our study analyzed UL33 gene sequences compiled from three distinct sources. The first dataset comprised 21 consensus sequences from a Kenyan mother-infant cohort, with breast milk samples collected approximately one month before the first sustained viral load detection in infants. These samples underwent PCR amplification and Oxford Nanopore sequencing Kasianowicz *et al.* (1996). The second dataset included 12 sequences from a Ugandan cohort, where saliva samples were collected from both mothers and their infants (initial collection at 6 weeks postpartum, followed by 4-month intervals). These samples were processed through DNA extraction and viral enrichment before Illumina NextSeq sequencing Bentley *et al.* (2008)

and de novo assembly. The third dataset consisted of 359 sequences retrieved from GenBank Benson *et al*. (2018), excluding those with "UNVERIFIED" status, bringing our total dataset to 392 complete UL33 gene sequences. For comparative analysis of HCMV GPCRs variability, we additionally collected complete nucleotide sequences of US27 (n=366), US28 (n=446), and UL78 (n=364) from GenBank using identical selection criteria.

7.3.2 Sequence variability analysis

To compare the variability of UL33 with other HCMV GPCRs (US27, US28, and UL78), nucleotide sequences of each dataset were first aligned using MUSCLE Edgar (2004) with 100 iterations. Sequence variability was then quantified using three complementary metrics : mean Shannon entropy (H), percentage of variable sites (PVS), and nucleotide diversity (π).

The Shannon entropy at each position i was calculated as :

$$H(i) = -\sum_{b \in \{A,T,C,G\}} p_b(i) \log_2(p_b(i)) \tag{7.1}$$

where $p_b(i)$ represents the frequency of nucleotide b at position i, excluding gaps. The mean Shannon entropy H was then computed by averaging $H(i)$ across all positions in the alignment :

$$H = \frac{1}{L} \sum_{i=1}^{L} H(i) \tag{7.2}$$

where L is the total alignment length.

The percentage of variable sites (PVS) was computed as :

$$\text{PVS} = \frac{N_v}{L} \times 100 \tag{7.3}$$

where N_v represents the number of positions with variant nucleotides exceeding 1% frequency, and L is the total alignment length.

Nucleotide diversity (π) was calculated as :

$$\pi = \frac{\sum_{i,j} n_{ij}}{L \times \binom{N}{2}} \tag{7.4}$$

where n_{ij} represents the number of nucleotide differences between sequences i and j, L is the sequence length, and N is the total number of sequences in the alignment.

7.3.3 Identification of UL33 specific variants associated with primary infection

7.3.3.1 Variant prediction and consensus sequence generation

Analysis of the 21 Kenyan mother-infant paired samples was performed using the HaROLD software suite Venturini *et al*. (2022). The analysis pipeline included strand count generation, variant prediction, and refinement steps, using the UL33 gene from HCMV Merlin strain (GenBank accession : NC_006273.2) as reference. Consensus sequences were subsequently generated using Samtools and BCFtools Danecek *et al*. (2021). For each mother-infant pair, sequences were organized into family-specific datasets. Variants with frequencies exceeding 1% were extracted from HaROLD output files, and sequences containing more than 5% missing nucleotides were excluded from further analysis. Reference sequences from well-characterized HCMV strains (AD169 (FJ527563.1), Merlin (NC_006273.2), Toledo (GU937742.2), and Towne (FJ616285.1)) were included as controls for phylogenetic analysis.

7.3.3.2 Phylogenetic analysis

Family-specific sequences were aligned using MUSCLE Edgar (2004) with the same parameters as for variability analysis. Phylogenetic trees were constructed using maximum likelihood method implemented in RAxML Stamatakis (2006), employing the GTRGAMMA substitution model with 100 bootstrap replicates. Trees were annotated with variant frequencies extracted from HaROLD output files, emphasizing the relationship between majority variants in infant samples and their corresponding maternal sequences. Pairwise genetic distances between sequences within each family were computed from the multiple sequence alignments using the 'blastn' substitution model (supplementary materials).

7.3.4 Identification of UL33 genotypes and associated mutational profiles

7.3.4.1 Identification of amino acid mutation profiles

To identify UL33 genotypes and their characteristic mutations, we applied the *KmerSignificance Score* (KSS) pipeline (6) to analyze sequence variations across our dataset. This analysis began with codon-aware pairwise alignments of all sequences against the HCMV Merlin reference sequence, considering both nucleotide sequences and their protein translations along with gene annotation information. From these alignments, the pipeline performed a systematic scan to identify non-overlapping k-mers of length 9 nucleotides (default parameter) and their variations throughout the sequences. Each identified k-mer/variations was then

associated with its corresponding amino acid mutations based on the protein translation. The KSS framework evaluates each k-mer/variations set through four scoring components : 1) A discriminative score assessing the ability to distinguish between sequence classes. 2) A mutation score evaluating the potential impact of amino acid changes based on their biophysical properties. 3) A protein score based on functional importance and protein characteristics. 4) A global score combining the three previous scores (KSS). The scoring component of the KSS framework was applied only after UL33 clusters were identified, as it requires sequence datasets labeled by class.

The KSS framework output was then used to construct a presence-absence matrix capturing the mutational amino acid profile of each sequence. In this matrix format, sequences were represented as rows and amino acid mutations as columns, with binary values indicating the presence (1) or absence (0) of specific mutations in each sequence.

7.3.4.2 Clustering of amino acid mutation profiles

The presence-absence matrix of amino acid mutations was used to perform WPGMA hierarchical clustering Sokal et Michener (1958) based on pairwise Hamming distances Hamming (1950) between sequences. The Hamming distance, specifically designed to compare binary vectors, provides an intuitive measure of the differences between mutation profiles. For each pair of sequences, represented as binary vectors A and B, their Hamming distance was calculated as :

$$d_{\text{Hamming}}(A, B) = \frac{\sum_{i=1}^{n} \delta(a_i, b_i)}{n}$$

where $\delta(a_i, b_i) = 1$ if elements differ at position i, and 0 otherwise, and n is the total number of positions.

WPGMA was chosen for its ability to account for cluster sizes during merging, preventing bias towards larger clusters. When merging clusters S and T into a new cluster U, the distance to another cluster V is calculated as :

$$d(U, V) = \frac{n_S \, d(S, V) + n_T \, d(T, V)}{n_S + n_T}$$

where n_S and n_T are the sizes of clusters S and T, respectively. This approach ensures balanced representation of sequence groups regardless of their size. The resulting hierarchical clustering was visualized as a clustermap.

7.3.5 Analysis of mutations using the Kmer Significance Score

After identifying the UL33 clusters, sequences were labeled based on their associated reference sequences within the corresponding cluster. The KSS framework was then applied to identify the most significant k-mers/variations within the sequence dataset. This framework assigns categorical scores between 1 and 5 to each component (discriminative score, mutational score, and protein score), resulting in a global KSS that enables the characterization of discriminative regions and their associated mutations that define UL33 clusters. Using these labeled UL33 sequences, we performed multiple sequence alignments within each cluster using MUSCLE with previously described parameters. From these alignments, we generated consensus sequences characteristic of each cluster. These consensus sequences were then aligned and analyzed using PhastCons and PhyloP to calculate conservation scores Siepel *et al.* (2005). The resulting conservation scores were plotted against the KSS. This visualization allowed us to examine the relationship between discriminative and conserved regions across different UL33 domains and assess their correlation.

7.4 Results and Discussion

7.4.1 Sequence diversity analysis reveals distinct evolutionary patterns among HCMV GPCRs

The comparative analysis of genetic diversity metrics among the four HCMV GPCRs (Figure 7.1), conducted on 392 complete UL33 sequences, reveals marked heterogeneity in their evolutionary profiles and confirms previous observations by Lassalle *et al.* (2016) made on a smaller dataset of 20 complete sequences. UL33 particularly stands out with the highest values across all three metrics studied : a Shannon entropy of 0.17, a percentage of variable sites (23.27%), and nucleotide diversity (π) of 0.08.

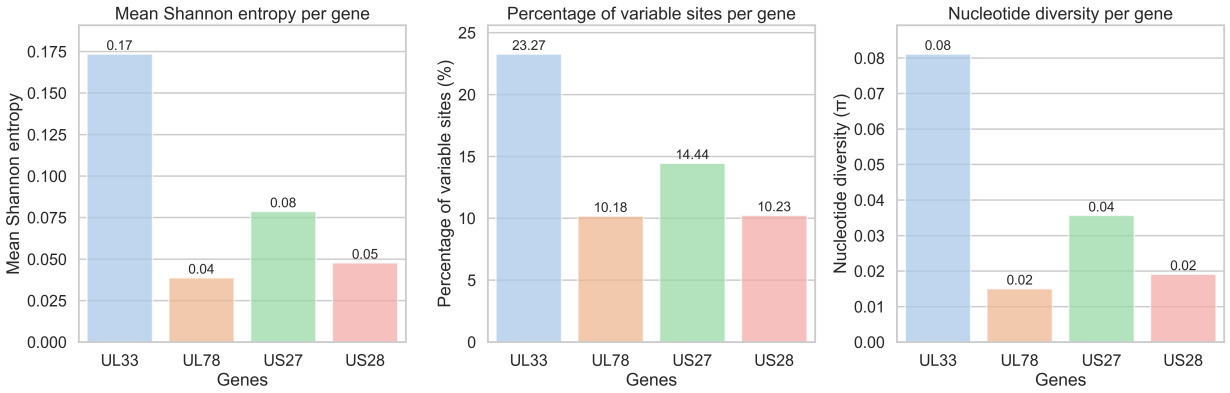


Figure 7.1 – Comparative analysis of genetic diversity metrics across HCMV GPCRs.

Bar plots showing the mean Shannon entropy (left), percentage of variable sites (middle), and nucleotide diversity π (right) for UL33, UL78, US27, and US28 genes.

In comparison, US27 shows intermediate values (Shannon entropy : 0.08, variable sites : 14.44%, π : 0.04), while UL78 (Shannon entropy : 0.04, variable sites : 10.18%, π : 0.02) and US28 (Shannon entropy : 0.05, variable sites : 10.23%, π : 0.02) exhibit the lowest diversity levels.

The increased variability of UL33 could reflect more intense selective pressure, possibly related to its role in immune evasion de Munnik *et al.* (2015). The structure of UL33 might also demonstrate greater tolerance for mutations without compromising functionality, a characteristic previously observed in other viral GPCRs Couty et Gershengorn (2005). Furthermore, this genetic diversity could contribute to the virus's ability to adapt to different cellular environments, thereby influencing its tropism Sijmons *et al.* (2015). Based on these observations regarding genetic variability, we hypothesize that specific UL33 signatures could drive primary infection establishment by facilitating viral transmission and pathogenesis.

7.4.2 Identification of UL33 specific variants associated with primary infection

To investigate the correlation between potential UL33 signatures and the establishment of a primary infection, we focused on studying HCMV transmission from mother to infant during breastfeeding. Analysis of the phylogenetic trees presented in Figure 7.2 reveals two distinct patterns in UL33 sequence variability between mothers and infants.

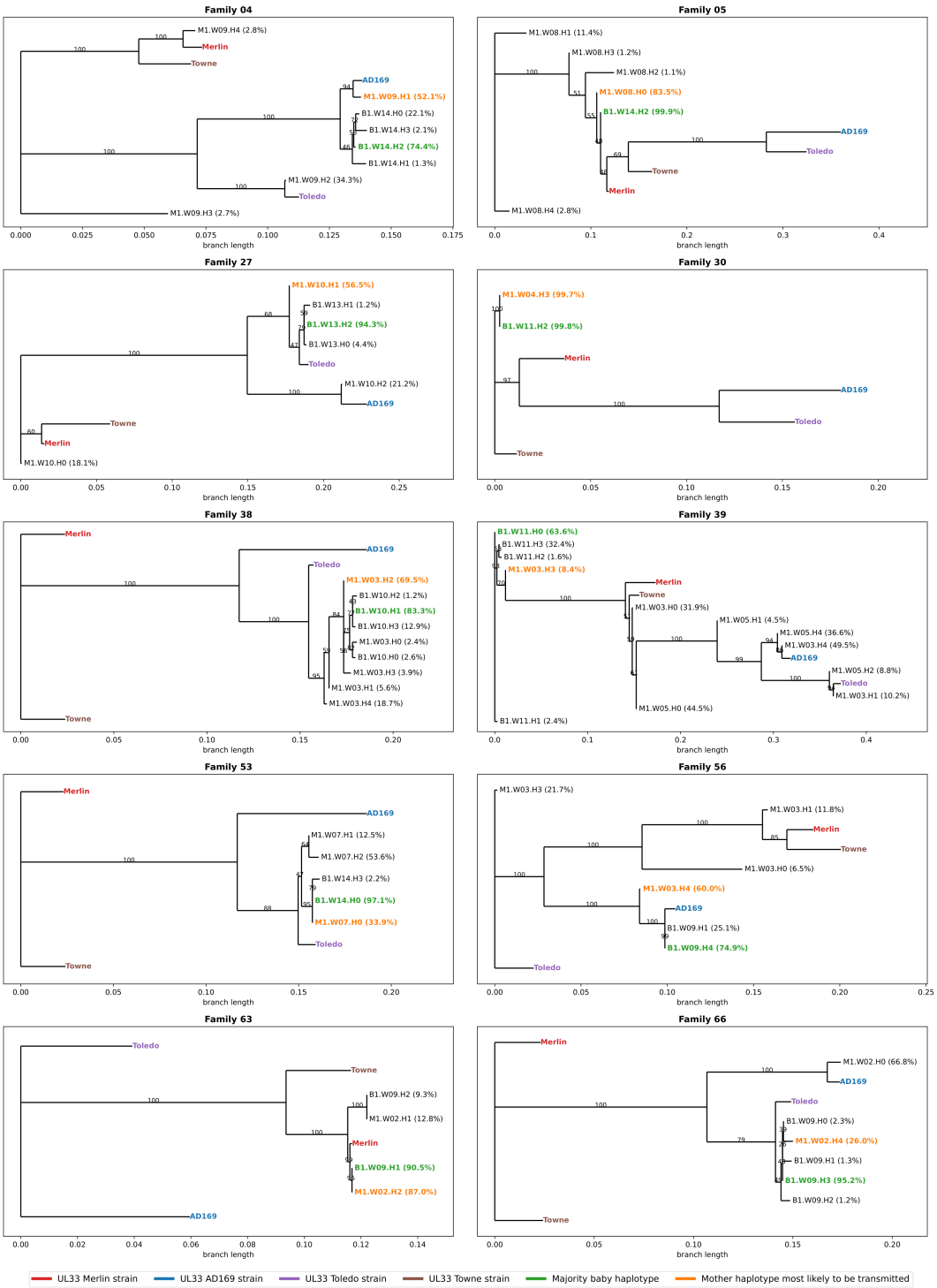


Figure 7.2 – Phylogenetic analysis of UL33 variants in mother-child transmission pairs.

Phylogenetic trees showing the evolutionary relationships between maternal variants (M), infant variants (B), and reference strains (AD169, Merlin, Toledo, and Towne) for each family. Scale bars represent the average number of nucleotide substitutions per sequence position.

First, we observe a higher sequence similarity among infant variants within each family, as evidenced by their close clustering in the phylogenetic trees, indicating limited nucleotide variability. This result is expected, given the recent nature of the infection in infants. According to the Mayer *et al.* (2017) model, the viral particle responsible for the infection has not yet had sufficient time to accumulate a significant number of mutations. In contrast, maternal sequences generally display greater nucleotide diversity, with variants showing larger phylogenetic distances between them across all family trees, due to chronic infection over time.

A second notable pattern emerges regarding viral transmission dynamics : in the majority of families, the predominant maternal variant exhibits the closest phylogenetic relationship to the infant variants. This pattern is particularly well illustrated in the first four families shown in Figure 7.2 (families 04, 05, 27, and 30), where the dominant maternal variant consistently clusters with the infant sequences, suggesting its probable role in horizontal transmission.

Interestingly, each mother-child transmission cluster appears to be associated with different reference sequences : family 04 clusters with the UL33 sequence from AD169, family 05 with Merlin, family 27 with Toledo, and family 30 with Towne. While we initially hypothesized that specific variant(s) might be predominantly associated with primary infection, our sample set does not support this assumption. These findings suggest the existence of mutational profiles that may favor persistent primary infection, rather than the predominance of a particular strain. However, given that this study includes only 10 families, larger-scale studies would be necessary to confirm these observations.

7.5 Identification of UL33 genotypes and associated mutational profiles

To establish a potential link between the mother-child transmission clusters identified in the previous analysis and to elucidate the evolutionary dynamics of UL33, we performed clustering of all available UL33 sequences based on their amino acid mutational profiles to identify key UL33 genotypic signatures.

The clustermap shown in figure 7.3 illustrates the mutational profiles of 70 amino acid mutations identified through KSS framework. This analysis encompasses 392 complete UL33 gene sequences, combining data from GenBank, consensus sequences from the Kenyan mother-child cohort, and sequences from our Ugandan cohort. The Merlin laboratory strain sequence was used as a reference to report mutations.

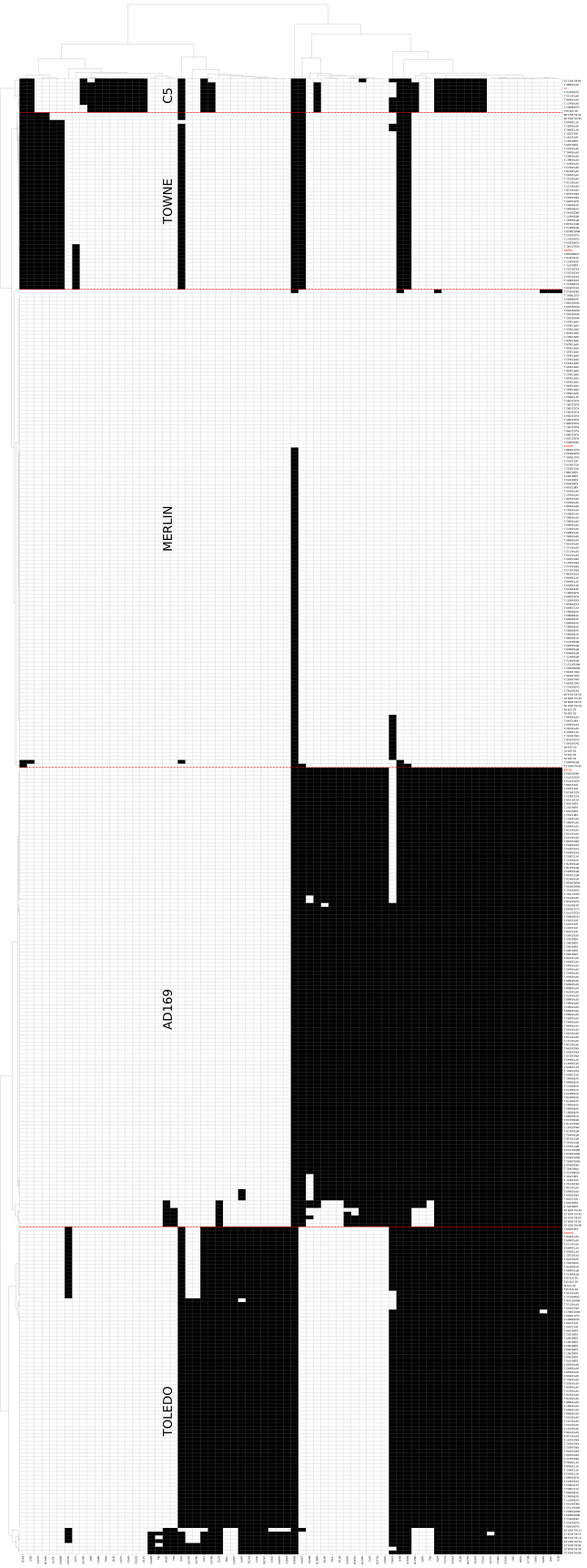


Figure 7.3 – Clustermap of the mutational landscape of UL33 across 392 complete sequences.

Black and white colors indicate the presence and absence of mutations respectively, with mutations labeled relative to the Merlin reference sequence. Analysis reveals five distinct clusters : (Toledo, AD169, Merlin, Towne and C5).

The hierarchical clustering analysis reveals two major clusters and one minor cluster, each defined by distinctive mutational profiles. The first major cluster, Toledo/AD169, is characterized by 24 shared mutations that distinguish it from the Merlin reference sequence. The second major cluster, Merlin/Towne, comprises sequences closely related to Merlin that show 0-2 mutations difference, and sequences associated with Towne that differ from Merlin sequences by 11 mutations. A third minor cluster (C5), characterized by the sequence (GU180094.1) identified by Deckers *et al.* (2009), emerges with an intermediate mutational profile between the two major clusters. The two main clusters further subdivide into distinct subclusters that correspond to the laboratory strain reference sequences (Toledo, AD169, Merlin, and Towne) previously identified in our mother-child transmission analysis.

In our phylogenetic analysis (Figure 7.2), family 39 exhibited a unique transmission cluster, not related with any control sequences. These results are consistent with those shown in Figure 7.3, where the maternal consensus sequence is associated with cluster C5, which was not included among the control sequences of the phylogenetic analysis. The infant consensus sequence from this family shows slight divergence from Merlin in both analyses.

These mutational profiles, based on all available UL33 sequences, support the characterization of five distinct genotypes of UL33 (Toledo, AD169, Merlin, Towne and C5) Deckers *et al.* (2009), including four major ones that appear to represent distinct evolutionary lineages of HCMV UL33. The existence of these well-defined genotypes suggests that UL33 evolution might be constrained by functional requirements, while still allowing for sufficient diversity to potentially modulate receptor function across different host environments Couty et Gershengorn (2005). The identification of an intermediate genotype (C5) further suggests ongoing evolutionary processes and possible recombination events in UL33, which could contribute to HCMV's adaptability during primary infection and transmission.

7.5.1 Analysis of significant k-mers and variations based on KSS

To identify regions responsible for UL33 genotype characterization and evaluate the significance of associated mutations, we applied the KSS framework. The output identified the most significant k-mers and their variations among the five UL33 genotypes previously defined. Twelve sets of k-mers/variations with KSS scores ≥ 4 were identified and are presented in Table 7.1. These variations demonstrate high discriminative power among UL33 genotypes and are associated with amino acid changes that could significantly impact protein dynamics.

Table 7.1 – Mutational landscape of the most significant k-mers/variations highlighted by the KSS in UL33

NUCLEOTIDE POSITION	DOMAIN	REFERENCE K-MER	VARIATION	CLASS	AMINO ACID CHANGE	MUTATIONAL SCORE	DISCRIMINATIVE SCORE	PROTEIN SCORE	K-MER SIGNIFICANCE SCORE
28	NTD	AACCGCAGT (99%) / Towne (30%)	—CGCAAC	Toledo (91%)	N10-	5	5	4	4.7
					S12N	2			
			CGCAACAAC	AD169 (93%)	N10R	3			
					R11N	3			
					S12N	2			
			ATCCGCAAT	C5 (100%)	N10I	5			
					S12N	2			
			AATCGCAGT	Towne (70%)	Silent	0			
37	NTD	ACCGATACT (98%)	ACCAGTACT	Towne (100%)	D14S	3	5	4	4.7
			ACTCCT—	AD169 (94%)	D14-	5			
					T15P	2			
			ACCACTACT	Toledo (100%)	D14T	3			
			ACCAGCACC	C5 (100%)	D14S	3			
55	NTD	AACATCACC (98%)	AACAGCACC	C5 (100%)	I20S	4	5	4	4.7
			AACAGTACC	Towne (100%)	I20S	4			
			AATGACACT	AD169 (100%) / Toledo (100%)	I20D	5			
73	NTD	ACGGAGCCG (99%) / Towne (74%)	ACGGGGCCG	Toledo (98%)	E26G	3	5	4	4.0
			ACGGAGACG	Towne (26%)	P27T	2			
			ACAGGGCCG	AD169 (99%)	E26G	3			
			ACTGAGCCG	C5 (100%)	Silent	0			
82	NTD	CTATCCGCC (99%) / Towne (100%)	CTATTCGCC	AD169 (100%)	S29F	5	5	4	4.7
			CTGTTCGCC	C5 (100%) / Toledo (99%)	S29F	5			
442	TM4	TATATAATT (100%) / Towne (98%)	TATACAATT	Toledo (99%)	I149T	4	5	4	4.3
			TATATGATT	AD169 (89%)	I149M	1			
			TACATAGTT	C5 (100%)	I150V	2			
451	TM4	TTGCTATTG (99%) / AD169 (100%) / Towne (100%)	TTGACGCTG	C5 (100%)	L152T	4	4	4	4.0
			TTGATATTG	Toledo (99%)	L152I	1			
523	ECD2	AATGGTACA (99%) / Towne (100%)	CACGATGCC	AD169 (98%)	N175H	3	5	4	4.3
					G176D	3			
					T177A	2			
			CACGAAGCT	C5 (89%)	N175H	3			
					G176E	4			
					T177A	2			
			CATGAAGCC	Toledo (100%)	N175H	3			
					G176E	4			
					T177A	2			
532	ECD2	AATGGACAA (99%)	AATGATATCAAT	C5 (89%)	G179D	3	5	4	4.7
					Q180N	2			
					-180I	5			
			GAT—GAA	Towne (66%)	N178D	1			
					G179-	5			
					Q180E	2			
			AACGATACCAAT	AD169 (95%)	G179D	3			
					Q180N	2			
					-180T	5			
			GAT—GAG	Towne (30%)	N178D	1			
					G179-	5			
					Q180E	2			
			AATGATACCACT	Toledo (92%)	G179D	3			
					-180T	5			
					Q180T	2			
541	ECD2	AGTAGCAAT (99%)	ACTAATGCCACT	Toledo (99%)	S181N	2	5	4	4.7
					S182A	1			
					N183T	3			
			AATACCAAC	C5 (100%)	S181N	2			
					S182T	2			
			AATACCAAT	Towne (100%)	S181N	2			
					S182T	2			
			AATACTAAT	AD169 (98%)	S181N	2			
					S182T	2			
586	ECD2	GAGGTGTAC (100%) / Towne (98%)	GAGGTGCAC	Toledo (97%)	Y198H	4	5	4	4.3
			GAGGTCTCT	C5 (78%)	Y198S	4			
			GAAGTGCAC	AD169 (100%)	Y198H	4			
829	ECD3	GACTGCGAA (100%) / Towne (100%)	AAATGTGAA	Toledo (26%)	D277K	3	5	4	4.0
			GACTGCAGA	C5 (78%)	E279R	3			
			GAATGTAAA	Toledo (74%)	D277E	2			
					E279K	2			
			CAGTGTGAA	AD169 (100%)	D277Q	3			

The columns **NUCLEOTIDE POSITION** and **DOMAIN** indicate nucleotide k-mers/variations starting position and protein domain where they are located. The columns **REFERENCE k-MER** and **VARIATION** indicate the form of the k-mer in the reference sequence and in the other classes of sequences. The column **AMINO ACID CHANGE** lists the amino acid changes implied by the variations compared to the reference k-mers. The column **CLASS** shows the classes and their percentages associated with the different variations. The columns **DISCRIMINATIVE SCORE**, **MUTATIONAL SCORE**, **PROTEIN SCORE**, and **k-MER SIGNIFICANCE SCORE** indicate the different importance scores calculated by the KSS.

The distribution of the most relevant k-mers/variations highlighted by KSS (Figure 7.4) reveals that the majority (10) are located in the N-terminal domain (NTD) and extracellular domains (ECD2/ECD3), while two are located in the transmembrane domain (TM4). Among these, five k-mers/variations are located in NTD and involve non-conservative mutations (with mutation scores between 3 and 5) : N10I (polar to hydrophobic), I20D (hydrophobic to negatively charged), E26G (negatively charged to very small non-polar), S29F (polar to aromatic hydrophobic), and deletions at positions N10 and D14. These structural alterations could potentially impact UL33 glycosylation, constitutive receptor activity, and binding of yet-unidentified ligands van Senten *et al.* (2020).

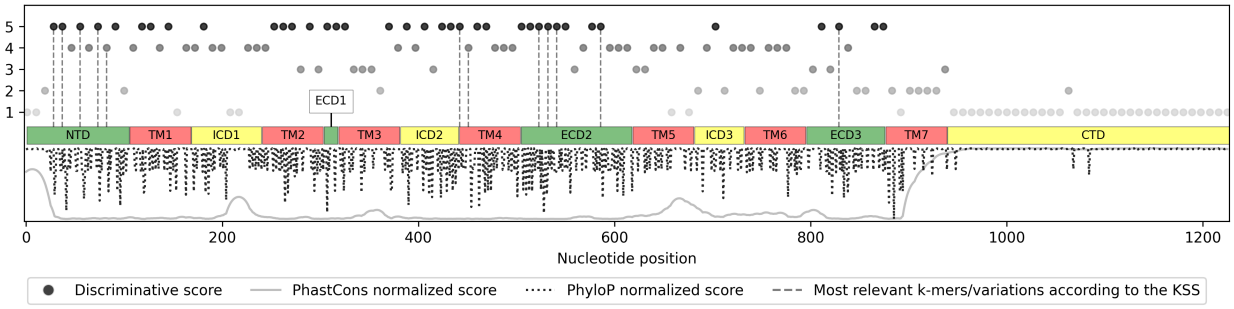


Figure 7.4 – Distribution of discriminative and variable nucleotide regions across UL33 domains.

Different protein regions are highlighted : extracellular domains and loops (green), intracellular domains and loops (yellow), and transmembrane domains (red).

In ECD2, several k-mers/variations implying significant mutations were identified, including G176E (neutral to negatively charged), G179D (non-polar to negatively charged tiny amino acid), along with insertions and deletions at positions 179 and 180. The transmembrane domain TM4 harbors two notable changes : I149T and L152T, both involving the substitution of hydrophobic residues with polar ones. In ECD3, two variants of position 198 were observed : Y198H (neutral aromatic to positively charged) and Y198S (aromatic to neutral polar).

Comparison of PhastCons and PhyloP conservation scores with KSS (Figure 7.4) revealed that the most relevant k-mers/variations strongly correlate with highly variable regions, while no significant k-mers/variations were identified in highly conserved regions such as the C-terminal domain. This extensive variability in key extracellular regions suggests potential functional differences in UL33 among strains van Senten *et al.* (2020).

To investigate potential immunological implications of this variability, we used NetMHCpan 4.1 Reynisson *et al.* (2020) to predict UL33-derived peptides across five MHC-I alleles : HLA-A*01 :01, HLA-A*02 :01, HLA-C*04 :01, HLA-C*06 :02, and HLA-C*07 :01. Notably, each k-mer predicted by the KSS framework was found within UL33-derived epitopes presented by MHC-I molecules. The analysis revealed that among the top 25 predicted epitopes (for each variant and HLA-I), several of them showed significant variability between the different variants/genotypes. This overlap between variable regions and predicted MHC-I epitopes suggests that UL33 variation could modulate immune recognition through altered HLA/epitope affinity, potentially enabling immune evasion through modified T-cell epitope presentation.

These variations in extracellular domains might not only affect immune recognition but could also impact receptor function. While most UL33 functional studies have focused on laboratory strains AD169/Merlin Krishna *et al.* (2021); van Senten *et al.* (2020), our identification of distinct genotypes with significant structural variations suggests the importance of examining other variants, such as Toledo, particularly for ligand identification studies. This is especially relevant given that no ligand has been identified to date qualifying UL33 as orphan receptor. Bio-ID experimentation based on the fused of BirA enzyme on UL33 could be a way for the identification of specific ligands Li *et al.* (2019).

7.6 Conclusion

In this study, our comparative analysis of HCMV GPCRs confirmed that UL33 exhibits notable hypervariability, registering the highest diversity metrics among all viral GPCRs. By examining mother-infant transmission pairs, we found that dominant maternal UL33 variants were predominantly transmitted through breastfeeding, suggesting their potential role in facilitating successful viral transmission.

Using the KSS framework coupled with hierarchical clustering, we identified five distinct UL33 genotypes defined by unique mutational profiles. Notably, we uncovered twelve sets of discriminative k-mers/variations linked to significant amino acid alterations, primarily situated in the extracellular (hilly variable) regions. These mutations, especially those localized in the N-terminal and extracellular domains, may influence receptor glycosylation, constitutive activity, and ligand binding, thereby aiding HCMV adaptability during host infection and helping infected cells evade immune detection. Although no antibody-mediated (humoral) response to UL33 has been described to date, several studies have documented cellular immune responses (T lymphocyte-based) against this receptor in HCMV-infected patients Lübke *et al.* (2019); Dhanwani *et al.* (2021). These responses depend on recognizing viral peptides presented on infected cells by MHC class I

molecules. The discovery of multiple well-defined genotypes, including a transitional C5 cluster, supports the idea that UL33 evolution is subject to functional constraints while retaining enough diversity to modulate receptor function across diverse host environments. Because current knowledge of UL33 primarily derives from the laboratory strains AD169/Merlin, our findings emphasize the need to investigate variants from other genotypes, particularly Toledo, in future work aimed at identifying UL33 ligands and clarifying its role in viral pathogenesis.

Second, applying the KSS framework combined with hierarchical clustering identified five distinct UL33 genotypes (Toledo, AD169, Merlin, Towne, and C5), each characterized by specific mutational profiles. The identification of an intermediate C5 cluster suggests ongoing evolutionary processes and possible recombination events in UL33. Our analysis highlighted twelve sets of discriminative k-mers/variations (KSS ≥ 4) primarily located in extracellular domains. The associated mutations, particularly concentrated in N-terminal and extracellular regions, could influence receptor glycosylation, constitutive activity, and ligand binding properties despite the orphan nature of UL33.

Using NetMHCpan 4.1, we predicted the top 25 UL33-derived peptides for each selected MHC-I allele of each variant. These epitopes, preferentially displayed by the respective MHC-I molecules, showed substantial variability/mutations. Crucially, each k-mer identified by the KSS framework was found within UL33-derived epitopes presented by MHC-I molecules. The variations tied to these k-mers and the epitopes they define may influence HLA/epitope affinity, thereby contributing to the immune evasion of infected cells by altering or disrupting the presentation of known UL33-derived T-cell epitopes.

Overall, this work provides new insights into UL33's genetic diversity and its putative role in viral transmission, highlighting new avenues for therapeutic interventions targeting this viral GPCR. Future investigations should explore the functional repercussions of genotype-specific variations to further elucidate UL33's contribution to HCMV pathogenesis and evaluate its suitability as an antiviral target. The results, along with the code to reproduce the experiments, the dataset, and supplementary materials, are available in the following GitHub repository : `https://github.com/bioinfoUQAM/UL33`

CONCLUSION

Dans cette thèse de doctorat, nous avons développé un ensemble d'approches méthodologiques pour l'identification et l'analyse des signatures génomiques dans le contexte de la classification des séquences biologiques virales. Ces approches, validées par des évaluations exhaustives sur des jeux de données simulés et réels, ont donné lieu à cinq contributions scientifiques publiées dans des conférences et journaux internationaux, ou actuellement en préparation. Ces contributions, allant des aspects théoriques aux applications pratiques, démontrent l'efficacité de nos approches sur des virus d'importance médicale majeure comme le SARS-CoV-2, le HIV-1 et le HCMV.

Tout d'abord, nous avons présenté KEVOLVE (Chapitre 3), une méthode combinant algorithmes génétiques et apprentissage automatique pour identifier des ensembles minimaux de caractéristiques génomiques basées sur les *k*-mers. Ces caractéristiques sont utilisées pour construire des modèles de prédiction ensemblistes basés sur des SVM. Les évaluations comparatives avec les outils de référence pour la prédiction des séquences du HIV-1 (REGA, COMET et KAMERIS) ont démontré la supériorité de KEVOLVE, notamment dans la classification des sous-types minoritaires comme les recombinants, y compris lors de l'introduction de bruit par mutation aléatoire. Nos analyses ont également souligné que les sous-types minoritaires du HIV-1 et d'autres virus sont souvent négligés dans les études ou masqués par l'utilisation exclusive de métriques de performance pondérées.

Deuxièmement, nous avons développé KANALYZER (Chapitre 4), un algorithme basé sur l'alignement par paires et le calcul parallèle pour identifier les variations des *k*-mers discriminatifs et leurs informations associées dans les séquences nucléotidiques. Il caractérise notamment les classes virales liées aux variations, leur fréquence d'occurrence, leur localisation génomique et leur impact sur les séquences d'acides aminés dans les régions codantes. L'évaluation de KANALYZER a porté sur plus de 150 000 variations de *k*-mers discriminatifs dans des jeux de données synthétiques, en variant des paramètres comme la longueur des séquences, la taille des *k*-mers et le taux de mutations aléatoires. L'algorithme a également été validé sur des ensembles de séquences virales hétérogènes, couvrant plus de 5 000 variations à identifier à travers différents groupes viraux, classifications taxonomiques et types de génomes. KANALYZER a identifié plus de 96% des variations dans l'ensemble des évaluations, bien qu'une limitation ait été observée dans l'identification des variations impliquant plus de trois substitutions par rapport aux *k*-mers discriminatifs originaux.

Troisièmement, nous avons développé un système de notation (*KmerSignificance Score*, KSS - Chapitre 6) pour évaluer l'importance des *k*-mers discriminants et de leurs variations. Ce score intègre deux composantes complémentaires : (1) un score discriminatif utilisant un cadre d'apprentissage supervisé pour évaluer leur capacité à distinguer les différentes sous-espèces virales, et (2) un score de signification biologique basé sur l'importance fonctionnelle des régions codantes et l'impact des substitutions d'acides aminés sur la dynamique protéique, évalué via des différences de propriétés biophysiques dérivées des matrices de substitution. Cette approche a démontré une forte cohérence avec les régions discriminantes associées aux variants du HIV-1 et du SARS-CoV-2, pour lesquels des données de référence extensives sont disponibles. Son application aux gènes *UL55*, *UL73* et *US28* du HCMV a non seulement validé les critères mutationnels établis sur des échantillons historiquement limités, mais a également permis de rendre disponibles des jeux de données annotés plus conséquents. Les résultats ont notamment montré que les *k*-mers et variations présentant les scores KSS les plus élevés permettent d'identifier des sous-ensembles de séquences pertinents tant pour la discrimination taxonomique que pour la caractérisation de régions de mutations potentiellement impactantes sur les fonctions protéiques. Ces travaux ont également conduit au développement de MIYATA_EVO, une version optimisée de la matrice MIYATA offrant une meilleure corrélation avec les distances des propriétés biophysiques des acides aminés.

En parallèle au développement de ces trois approches méthodologiques, nous avons validé leur utilisation à travers deux cas d'étude d'importance majeure en santé publique. Le premier portait sur l'identification et la caractérisation des variants du SARS-CoV-2. Le second concernait le cytomégalovirus humain (HCMV) dans le cadre d'une collaboration avec le Centre de recherche de l'Hôpital Sainte-Justine, avec le soutien de l'Institut de recherche en santé du Canada. Cette étude sur le HCMV visait spécifiquement à caractériser les souches virales dans le contexte de la transmission mère-enfant et à établir une classification des génotypes, particulièrement pour le gène *UL33*.

Dans l'étude portant sur le SARS-CoV-2 (Chapitre 5), nous avons évalué la performance de KEVOLVE en comparaison avec des outils statistiques spécialisés dans l'identification de motifs discriminants pour la classification des variants. Les modèles basés sur KEVOLVE ont surpassé les méthodes statistiques de référence tout en utilisant moins de motifs, démontrant notamment leur capacité à gérer efficacement la nature multiclasse du problème, contrairement aux approches limitées à la classification binaire. L'analyse de 334 956 séquences, couvrant les 10 principales classes de variants définies par l'OMS, a été réalisée avec KANALYZER pour identifier les variations des motifs discriminants. Les résultats montrent une concentra-

tion significative de ces motifs dans les protéines structurales, notamment la protéine S, cohérente avec les régions variables et mutations clés décrites dans la littérature.

Concernant l'application sur le HCMV (Chapitre 7), nous avons étudié le gène *UL33*, un récepteur couplé aux protéines G (GPCR). Notre analyse comparative des GPCRs du HCMV a confirmé l'hypervariabilité d'*UL33*, présentant les métriques de diversité les plus élevées parmi l'ensemble de ces récepteurs. L'étude des paires mère-enfant d'une cohorte kényane a révélé que les variants *UL33* maternels dominants étaient principalement transmis par l'allaitement. En combinant le KSS avec une approche de clustering hiérarchique sur des données publiques élargies, nous avons identifié cinq génotypes distincts d'*UL33* (Toledo, AD169, Merlin, Towne et C5), chacun caractérisé par des profils mutationnels uniques. L'identification de douze ensembles de *k*-mers et variations présentant les scores KSS les plus élevés, principalement localisés dans les domaines N-terminal et extracellulaires, suggère un rôle potentiel dans la glycosylation du récepteur, son activité constitutive et sa liaison aux ligands. De plus, ces variations coïncident avec des épitopes prédits de classe I du CMH, suggérant une possible implication dans l'échappement immunitaire via la modulation de la présentation des épitopes aux lymphocytes T.

Les méthodes développées dans cette thèse ont démontré leur utilité pour l'identification et l'analyse des signatures génomiques dans plusieurs contextes virologiques. Cependant, leur application reste actuellement limitée à un cadre de classification supervisée nécessitant des jeux de données d'apprentissage bien annotés. Un travail conséquent reste à réaliser pour étendre ces approches à la classification globale des espèces virales définies par l'ICTV et, plus particulièrement, à la détection automatique d'espèces virales encore inconnues. L'utilisation du KSS pour établir des profils mutationnels combinée à des méthodes de clustering, telle que réalisée dans le chapitre 7, peut fournir des pistes d'intérêt à explorer. Également, une bonne compréhension des données actuelles et de leurs seuils de différenciation génétique peut fournir un support solide pour concevoir ce type d'approches. De plus, bien que nos approches apportent des avancées méthodologiques significatives, leur accessibilité reste limitée pour les chercheurs sans expertise en bioinformatique. Pour répondre à ce défi, nous développons actuellement une plateforme bioinformatique spécialisée dans la classification et l'analyse des séquences virales. Cette plateforme intégrera nos méthodes et rendra accessibles nos jeux de données annotés, facilitant ainsi leur utilisation par la communauté scientifique.

BIBLIOGRAPHIE

Abecasis, A. B., Wang, Y., Libin, P., Imbrechts, S., de Oliveira, T., Camacho, R. J. et Vandamme, A.-M. (2010). Comparative performance of the rega subtyping tool version 2 versus version 1. *Infection, genetics and evolution*, *10*(3), 380–385.

Abergel, C., Legendre, M. et Claverie, J.-M. (2015). The rapidly expanding universe of giant viruses : Mimivirus, pandoravirus, pithovirus and mollivirus. *FEMS microbiology reviews*, *39*(6), 779–796.

Adriaenssens, E. M., Roux, S., Brister, J. R., Karsch-Mizrachi, I., Kuhn, J. H., Varsani, A., Yigang, T., Reyes, A., Lood, C., Lefkowitz, E. J. *et al.* (2023). Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification. *Nature biotechnology*, *41*(7), 898–902.

Ahmed, I. et Jeon, G. (2022). Enabling artificial intelligence for genome sequence analysis of covid-19 and alike viruses. *Interdisciplinary sciences : computational life sciences*, *14*(2), 504–519.

Aksamentov, I., Roemer, C., Hodcroft, E. B. et Neher, R. A. (2021). Nextclade : clade assignment, mutation calling and quality control for viral genomes. *Journal of open source software*, *6*(67), 3773.

Alam, M. N. U. et Chowdhury, U. F. (2020). Short k-mer abundance profiles yield robust machine learning features and accurate classifiers for rna viruses. *PloS one*, *15*(9), e0239381.

Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L. *et al.* (2023). The gene ontology knowledgebase in 2023. *Genetics*, *224*(1), iyad031.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402.

Alyazidi, R., Murthy, S., Slyker, J. A. et Gantt, S. (2018). The potential harm of cytomegalovirus infection in immunocompetent critically ill children. *Frontiers in pediatrics*, *6*, 96.

Arav-Boger, R., Willoughby, R. E., Pass, R. F., Zong, J.-C., Jang, W.-J., Alcendor, D. et Hayward, G. S. (2002). Polymorphisms of the cytomegalovirus (cmv)–encoded tumor necrosis factor–α and β-chemokine receptors in congenital cmv disease. *The Journal of infectious diseases*, *186*(8), 1057–1064.

Avila Cartes, J., Anand, S., Ciccolella, S., Bonizzoni, P. et Della Vedova, G. (2023). Accurate and fast clade assignment via deep learning and frequency chaos game representation. *GigaScience*, *12*, giac119.

Bailey, S. F., Alonso Morales, L. A. et Kassen, R. (2021). Effects of synonymous mutations beyond codon bias : the evidence for adaptive synonymous substitutions from microbial evolution experiments. *Genome biology and evolution*, *13*(9), evab141.

Bailey, T. L. (2021). Streme : accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834–2840.

Bailey, T. L., Bodén, M., Whitington, T. et Machanick, P. (2010). The value of position-specific priors in motif discovery using meme. *BMC bioinformatics*, *11*, 1–14.

Bailey, T. L., Elkan, C. *et al.* (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bailey, T. L., Johnson, J., Grant, C. E. et Noble, W. S. (2015). The meme suite. *Nucleic acids research*, *43*(W1), W39–W49.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological reviews*, *35*(3), 235–241.

Bao, Y., Chetvernin, V. et Tatusova, T. (2014). Improvements to pairwise sequence comparison (pasc) : a genome-based web tool for virus classification. *Archives of virology*, *159*, 3293–3304.

Barton, M. I., MacGowan, S. A., Kutuzov, M. A., Dushek, O., Barton, G. J. et Van Der Merwe, P. A. (2021). Effects of common mutations in the sars-cov-2 spike rbd and its ligand, the human ace2 receptor on binding affinity and kinetics. *Elife*, *10*, e70658.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E., Bye-A-Jee, H., Carpentier, D., Denny, P. *et al.* (2024). Uniprot : the universal protein knowledgebase in 2025. *Nucleic acids research*.

Bauer, D. C., Tay, A. P., Wilson, L. O., Reti, D., Hosking, C., McAuley, A. J., Pharo, E., Todd, S., Stevens, V., Neave, M. J. *et al.* (2020). Supporting pandemic response using genomics and bioinformatics : A case study on the emergent sars-cov-2 outbreak. *Transboundary and emerging diseases*, *67*(4), 1453–1462.

Bennet, S., Cohen, M. A. et Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering, Design and Selection*, *7*(11), 1323–1332.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D. et Sayers, E. W. (2018). Genbank. *Nucleic acids research*, *46*(D1), D41–D47.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, *456*(7218), 53–59.

Bernard, G., Chan, C. X., Chan, Y.-b., Chua, X.-Y., Cong, Y., Hogan, J. M., Maetschke, S. R. et Ragan, M. A. (2019). Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics*, *20*(2), 426–435.

Betts, M. J. et Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists*, 289–316.

Bleul, C. C., Wu, L., Hoxie, J. A., Springer, T. A. et Mackay, C. R. (1997). The hiv coreceptors cxcr4 and ccr5 are differentially expressed and regulated on human t lymphocytes. *Proceedings of the National Academy of Sciences*, *94*(5), 1925–1930.

Bonham-Carter, O., Steele, J. et Bastola, D. (2014). Alignment-free genetic sequence comparisons : a review of recent approaches by word analysis. *Briefings in bioinformatics*, *15*(6), 890–905.

Boppana, S. B., Ross, S. A. et Fowler, K. B. (2013). Congenital cytomegalovirus infection : clinical outcome. *Clinical infectious diseases*, *57*(suppl_4), S178–S181.

Boser, B. E., Guyon, I. M. et Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Dans *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Brehm, J. H., Koontz, D., Meteer, J. D., Pathak, V., Sluis-Cremer, N. et Mellors, J. W. (2007). Selection of mutations in the connection and rnase h domains of human immunodeficiency virus type 1 reverse transcriptase that increase resistance to 3'-azido-3'-dideoxythymidine. *Journal of virology*, *81*(15), 7852–7859.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Buchfink, B., Reuter, K. et Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, *18*(4), 366–368.

Buchfink, B., Xie, C. et Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, *12*(1), 59–60.

Bui, V.-K. et Wei, C. (2020). Cdkam : a taxonomic classification tool using discriminative k-mers and approximate matching strategies. *BMC bioinformatics*, *21*(1), 1–13.

Burke, H. G. et Heldwein, E. E. (2015). Crystal structure of the human cytomegalovirus glycoprotein b. *PLoS pathogens*, *11*(10), e1005227.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., Craig, P. A., Crichlow, G. V., Dalenberg, K., Duarte, J. M. *et al.* (2023). Rcsb protein data bank (rcsb. org) : delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, *51*(D1), D488–D508.

Caetano-Anollés, G., Claverie, J.-M. et Nasir, A. (2023). A critical analysis of the current state of virus taxonomy. *Frontiers in Microbiology*, *14*, 1240993.

Chaffey, N. (2003). Alberts, b., johnson, a., lewis, j., raff, m., roberts, k. and walter, p. molecular biology of the cell. 4th edn.

Chan, C. X. et Ragan, M. A. (2013). Next-generation phylogenomics. *Biology direct*, *8*(1), 1–6.

Chiu, T.-P., Xin, B., Markarian, N., Wang, Y. et Rohs, R. (2020). Tfbsshape : an expanded motif database for dna shape features of transcription factor binding sites. *Nucleic acids research*, *48*(D1), D246–D255.

Chou, S. et Dennison, K. M. (1991). Analysis of interstrain variation in cytomegalovirus glycoprotein b sequences encoding neutralization-related epitopes. *Journal of Infectious Diseases*, *163*(6), 1229–1234.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009). Biopython : freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Collier, D., Monit, C. et Gupta, R. (2019). The impact of hiv-1 drug escape on the global treatment landscape. *Cell host & microbe*, *26*(1), 48–60.

Compeau, P. E., Pevzner, P. A. et Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, *29*(11), 987–991.

Cortes, C. (1995). Support-vector networks. *Machine Learning*.

Couty, J.-P. et Gershengorn, M. C. (2005). G-protein-coupled receptors encoded by human herpesviruses. *Trends in pharmacological sciences*, *26*(8), 405–411.

Crawford, H., Prado, J. G., Leslie, A., Hué, S., Honeyborne, I., Reddy, S., van der Stok, M., Mncube, Z., Brander, C., Rousseau, C. *et al*. (2007). Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant hla-b* 5703-restricted gag epitope in chronic human immunodeficiency virus type 1 infection. *Journal of virology*, *81*(15), 8346–8351.

Crawford, L. B., Caposio, P., Kreklywich, C., Pham, A. H., Hancock, M. H., Jones, T. A., Smith, P. P., Yurochko, A. D., Nelson, J. A. et Streblow, D. N. (2019). Human cytomegalovirus us28 ligand binding activity is required for latency in cd34+ hematopoietic progenitor cells and humanized nsg mice. *MBio*, *10*(4), 10–1128.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. *et al*. (2021). Twelve years of samtools and bcftools. *Gigascience*, *10*(2), giab008.

Das, J. K. et Roy, S. (2021). A study on non-synonymous mutational patterns in structural proteins of sars-cov-2. *Genome*, *64*(7), 665–678.

Dash, M. et Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(3), 131–156.

Dayhoff, M. (1978). Atlas of protein sequence and structure. *(No Title)*, p. 345.

Dayhoff, M., Schwartz, R. et Orcutt, B. (1978). 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure*, *5*, 345–352.

De Castro, E., Hulo, C., Masson, P., Auchincloss, A., Bridge, A. et Le Mercier, P. (2024). Viralzone 2024 provides higher-resolution images and advanced virus-specific resources. *Nucleic Acids Research*, *52*(D1), D817–D821.

De la Fuente, R., Díaz-Villanueva, W., Arnau, V. et Moya, A. (2023). Genomic signature in evolutionary biology : A review. *Biology*, *12*(2), 322.

de Munnik, S. M., Smit, M. J., Leurs, R. et Vischer, H. F. (2015). Modulation of cellular signaling by herpesvirus-encoded g protein-coupled receptors. *Frontiers in pharmacology*, *6*, 40.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., Van Rensburg, E. J., Wensing, A. M., Van De Vijver, D. A. *et al*. (2005). An automated genotyping system for analysis of hiv-1 and other microbial sequences. *Bioinformatics*, *21*(19), 3797–3800.

de Vries, J. J., Vossen, A. C., Kroes, A. C. et van der Zeijst, B. A. (2011). Implementing neonatal screening for congenital cytomegalovirus : addressing the deafness of policy makers. *Reviews in medical virology*, *21*(1), 54–61.

Deckers, M., Hofmann, J., Kreuzer, K.-A., Reinhard, H., Edubio, A., Hengel, H., Voigt, S. et Ehlers, B. (2009). High genotypic diversity and a novel variant of human cytomegalovirus revealed by combined ul33/ul55 genotyping with broad-range pcr. *Virology journal*, *6*, 1–12.

Delviks-Frankenberry, K. A., Nikolenko, G. N., Barr, R. et Pathak, V. K. (2007). Mutations in human immunodeficiency virus type 1 rnase h primer grip enhance 3'-azido-3'-deoxythymidine resistance. *Journal of Virology*, *81*(13), 6837–6845.

Delviks-Frankenberry, K. A., Nikolenko, G. N., Boyer, P. L., Hughes, S. H., Coffin, J. M., Jere, A. et Pathak, V. K. (2008). Hiv-1 reverse transcriptase connection subdomain mutations reduce template rna degradation and enhance azt excision. *Proceedings of the National Academy of Sciences*, *105*(31), 10943–10948.

Desingu, P. A., Nagarajan, K. et Dhama, K. (2022). Emergence of omicron third lineage ba. 3 and its importance. *Journal of medical virology*, *94*(5), 1808–1810.

Dhanwani, R., Dhanda, S. K., Pham, J., Williams, G. P., Sidney, J., Grifoni, A., Picarda, G., Lindestam Arlehamn, C. S., Sette, A. et Benedict, C. A. (2021). Profiling human cytomegalovirus-specific t cell responses reveals novel immunogenic open reading frames. *Journal of Virology*, *95*(21), 10–1128.

Do, C. B., Mahabhashyam, M. S., Brudno, M. et Batzoglou, S. (2005). Probcons : Probabilistic consistency-based multiple sequence alignment. *Genome research*, *15*(2), 330–340.

Dong, J., Jiang, M., Hu, L. et He, Z. (2023). Hamming encoder : Mining discriminative k-mers for discrete sequence classification. *arXiv preprint arXiv :2310.10321*.

Duffy, S. (2018). Why are rna virus mutation rates so damn high ? *PLoS biology*, *16*(8), e3000003.

Duffy, S., Shackelton, L. A. et Holmes, E. C. (2008). Rates of evolutionary change in viruses : patterns and determinants. *Nature Reviews Genetics*, *9*(4), 267–276.

Durbin, R., Eddy, S. R., Krogh, A. et Mitchison, G. (1998). *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eddy, S. R. (2004). What is dynamic programming ? *Nature biotechnology*, *22*(7), 909–910.

Edgar, R. C. (2004). Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, *26*(19), 2460–2461.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al*. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Eschke, K., Trimpert, J., Osterrieder, N. et Kunec, D. (2018). Attenuation of a very virulent marek's disease herpesvirus (mdv) by codon pair bias deoptimization. *PLoS pathogens*, *14*(1), e1006857.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al*. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Dans *kdd*, volume 96, 226–231.

Fan, L.-q., Hu, X.-y., Chen, Y.-y., Peng, X.-l., Fu, Y.-h., Zheng, Y.-p., Yu, J.-m. et He, J.-s. (2021). Biological significance of the genomic variation and structural dynamics of sars-cov-2 b. 1.617. *Frontiers in Microbiology*, *12*, 750725.

Farrell, H. E., Abraham, A. M., Cardin, R. D., Sparre-Ulrich, A. H., Rosenkilde, M. M., Spiess, K., Jensen, T. H., Kledal, T. N. et Davis-Poynter, N. (2011). Partial functional complementation between human and mouse cytomegalovirus chemokine receptor homologues. *Journal of virology*, *85*(12), 6091–6095.

FAUCHÈRE, J.-L., Charton, M., Kier, L. B., Verloop, A. et Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research*, *32*(4), 269–278.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters : finding maximum likelihood estimates. *Evolution*, 1229–1242.

Fischer, S., Freuling, C. M., Müller, T., Pfaff, F., Bodenhofer, U., Höper, D., Fischer, M., Marston, D. A., Fooks, A. R., Mettenleiter, T. C. *et al.* (2018). Defining objective clusters for rabies virus sequences using affinity propagation clustering. *PLoS Neglected Tropical Diseases*, *12*(1), e0006182.

Fiscon, G., Weitschek, E., Cella, E., Lo Presti, A., Giovanetti, M., Babakir-Mina, M., Ciotti, M., Ciccozzi, M., Pierangeli, A., Bertolazzi, P. *et al.* (2016). Missel : a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification. *BioData mining*, *9*, 1–24.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, *19*(2), 99–113.

Fitch, W. M. (1971). Toward defining the course of evolution : minimum change for a specific tree topology. *Systematic Biology*, *20*(4), 406–416.

Fix, E. (1985). *Discriminatory analysis : nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine.

Flowers, P., Theopold, K. H., Langley, R. et Robinson, W. R. (2019). *Chemistry 2e*. OpenStax Houston, TX.

Fraile-Ramos, A., Kohout, T. A., Waldhoer, M. et Marsh, M. (2003). Endocytosis of the viral chemokine receptor us28 does not require beta-arrestins but is dependent on the clathrin-mediated pathway. *Traffic*, *4*(4), 243–253.

Fulkerson, H. L., Chesnokova, L. S., Kim, J. H., Mahmud, J., Frazier, L. E., Chan, G. C. et Yurochko, A. D. (2020). Hcmv-induced signaling through gb–egfr engagement is required for viral trafficking and nuclear translocation in primary human monocytes. *Proceedings of the National Academy of Sciences*, *117*(32), 19507–19516.

Gantt, S., Dionne, F., Kozak, F. K., Goshen, O., Goldfarb, D. M., Park, A. H., Boppana, S. B. et Fowler, K. (2016). Cost-effectiveness of universal and targeted newborn screening for congenital cytomegalovirus infection. *JAMA pediatrics*, *170*(12), 1173–1180.

Gao, Z., Liu, Y., Wang, X., Song, J., Chen, S., Ragupathy, S., Han, J. et Newmaster, S. G. (2017). Derivative technology of dna barcoding (nucleotide signature and snp double peak methods) detects adulterants and substitution in chinese patent medicines. *Scientific reports*, *7*(1), 1–11.

García-Gonzalo, E., Fernández-Muñiz, Z., Garcia Nieto, P. J., Bernardo Sánchez, A. et Menéndez Fernández, M. (2016). Hard-rock stability analysis for span design in entry-type excavations with learning classifiers. *Materials*, *9*(7), 531.

Gerdol, M., Dishnica, K. et Giorgetti, A. (2022). Emergence of a recurrent insertion in the n-terminal domain of the sars-cov-2 spike glycoprotein. *Virus research*, *310*, 198674.

Giagulli, C., D'Ursi, P., He, W., Zorzan, S., Caccuri, F., Varney, K., Orro, A., Marsico, S., Otjacques, B., Laudanna, C. *et al.* (2017). A single amino acid substitution confers b-cell clonogenic activity to the hiv-1 matrix protein p17. *Scientific reports*, *7*(1), 6555.

Gifford, R., De Oliveira, T., Rambaut, A., Myers, R. E., Gale, C. V., Dunn, D., Shafer, R., Vandamme, A.-M., Kellam, P., Pillay, D. *et al.* (2006). Assessment of automated genotyping protocols as tools for surveillance of hiv-1 genetic diversity. *Aids*, *20*(11), 1521–1529.

Gompels, U., Larke, N., Sanz-Ramos, M., Bates, M., Musonda, K., Manno, D., Siame, J., Monze, M., Filteau, S. et Group, C. S. (2012). Human cytomegalovirus infant infection adversely affects growth and development in maternally hiv-exposed and unexposed infants in zambia. *Clinical Infectious Diseases*, *54*(3), 434–442.

Gong, X. et Padhi, A. (2011). Evidence for positive selection in the extracellular domain of human cytomegalovirus encoded g protein-coupled receptor us28. *Journal of medical virology*, *83*(7), 1255–1261.

Goodwin, S., McPherson, J. D. et McCombie, W. R. (2016). Coming of age : ten years of next-generation sequencing technologies. *Nature reviews genetics*, *17*(6), 333–351.

Gorbalenya, A., Baker, S., Baric, R. S., de Groot, R., Drosten, C., Gulyaeva, A. A., Haagmans, B., Lauber, C., Leontovich, A. M., Neuman, B. W. *et al.* (2020). The species severe acute respiratory syndrome-related coronavirus : classifying 2019-ncov and naming it sars-cov-2. *Nature microbiology*, *5*(4), 536–544.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L. *et al.* (2020). A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, *583*(7816), 459–468.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *science*, *185*(4154), 862–864.

Greco, S. et Gerdol, M. (2022). Independent acquisition of short insertions at the rir1 site in the spike n-terminal domain of the sars-cov-2 ba. 2 lineage. *Transboundary and Emerging Diseases*, *69*(5), e3408–e3415.

Griffiths, P. et Reeves, M. (2021). Pathogenesis of human cytomegalovirus in the immunocompromised host. *Nature Reviews Microbiology*, *19*(12), 759–773.

Gusfield, D. (1997). Algorithms on stings, trees, and sequences : Computer science and computational biology. *Acm Sigact News*, *28*(4), 41–60.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J. *et al.* (2021). Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), 409–424.

Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., Schäffer, A. A. et Brister, J. R. (2017). Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research*, *45*(D1), D482–D490.

Hayer, J., Jadeau, F., Deleage, G., Kay, A., Zoulim, F. et Combet, C. (2013). Hbvdb : a knowledge database for hepatitis b virus. *Nucleic acids research*, *41*(D1), D566–D570.

He, X., Hong, W., Pan, X., Lu, G. et Wei, X. (2021). Sars-cov-2 omicron variant : characteristics and prevention. *MedComm*, *2*(4), 838–845.

Hemelaar, J. (2013). Implications of hiv diversity for the hiv-1 pandemic. *Journal of Infection*, *66*(5), 391–400.

Hemelaar, J., Gouws, E., Ghys, P. D. et Osmanov, S. (2006). Global and regional distribution of hiv-1 genetic subtypes and recombinants in 2004. *Aids*, *20*(16), W13–W23.

Henikoff, S. et Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Hirano, A. et Kameda, T. (2021). Aromaphilicity index of amino acids : Molecular dynamics simulations of the protein binding affinity for carbon nanomaterials. *ACS Applied Nano Materials*, *4*(3), 2486–2495.

Hoffmann, M., Hofmann-Winkler, H., Krüger, N., Kempf, A., Nehlmeier, I., Graichen, L., Arora, P., Sidarovich, A., Moldenhauer, A.-S., Winkler, M. S. *et al.* (2021). Sars-cov-2 variant b. 1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. *Cell reports*, *36*(3).

Holland, J. H. (1992). *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Howley, P. M., Knipe, D. M. et Enquist, L. W. (2023). *Fields virology : fundamentals*. Lippincott Williams & Wilkins.

Hraber, P., Kuiken, C., Waugh, M., Geer, S., Bruno, W. J. et Leitner, T. (2008). Classification of hepatitis c virus and human immunodeficiency virus-1 sequences with the branching index. *Journal of General Virology*, *89*(9), 2098–2107.

Hu, X., Wang, H.-Y., Otero, C. E., Jenks, J. A. et Permar, S. R. (2022). Lessons from acquired natural immunity and clinical trials to inform next-generation human cytomegalovirus vaccine development. *Annual review of virology*, *9*(1), 491–520.

Huang, Y., Yang, C., Xu, X.-f., Xu, W. et Liu, S.-w. (2020). Structural and functional properties of sars-cov-2 spike protein : potential antivirus drug development for covid-19. *Acta Pharmacologica Sinica*, *41*(9), 1141–1149.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. et Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, *294*(5550), 2310–2314.

Huggins, P., Zhong, S., Shiff, I., Beckerman, R., Laptenko, O., Prives, C., Schulz, M. H., Simon, I. et Bar-Joseph, Z. (2011). Decod : fast and accurate discriminative dna motif finding. *Bioinformatics*, *27*(17), 2361–2367.

Jain, M., Olsen, H. E., Paten, B. et Akeson, M. (2016). The oxford nanopore minion : delivery of nanopore sequencing to the genomics community. *Genome biology*, *17*(1), 239.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R. *et al.* (2020). The reactome pathway knowledgebase. *Nucleic acids research*, *48*(D1), D498–D503.

Javanmardi, K., Segall-Shapiro, T. H., Chou, C.-W., Boutz, D. R., Olsen, R. J., Xie, X., Xia, H., Shi, P.-Y., Johnson, C. D., Annapareddy, A. *et al.* (2022). Antibody escape and cryptic cross-domain stabilization in the sars-cov-2 omicron spike protein. *Cell host & microbe*, *30*(9), 1242–1254.

Johnson, B. A., Zhou, Y., Lokugamage, K. G., Vu, M. N., Bopp, N., Crocquet-Valdes, P. A., Kalveram, B., Schindewolf, C., Liu, Y., Scharton, D. *et al.* (2022). Nucleocapsid mutations in sars-cov-2 augment replication and pathogenesis. *PLoS pathogens*, *18*(6), e1010627.

Jukes, T. et Cantor, C. (1969). Evolution of protein molecules. pages 21–132inmammalian protein metabolism (hn munro, ed.).

Juneja, S., Dhankhar, A., Juneja, A. et Bali, S. (2022). An approach to dna sequence classification through machine learning : Dna sequencing, k mer counting, thresholding, sequence analysis. *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, *11*(2), 1–15.

Kandeel, M., Ibrahim, A., Fayez, M. et Al-Nazawi, M. (2020). From sars and mers covs to sars-cov-2 : Moving toward more biased codon usage in viral structural and nonstructural genes. *Journal of medical virology*, *92*(6), 660–666.

Kannan, S. R., Spratt, A. N., Sharma, K., Chand, H. S., Byrareddy, S. N. et Singh, K. (2022). Omicron sars-cov-2 variant : Unique features and their impact on pre-existing antibodies. *Journal of autoimmunity*, *126*, 102779.

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A. et Flouri, T. (2017). Multi-rate poisson tree processes for single-locus species delimitation under maximum likelihood and markov chain monte carlo. *Bioinformatics*, *33*(11), 1630–1638.

Kariin, S. et Burge, C. (1995). Dinucleotide relative abundance extremes : a genomic signature. *Trends in genetics*, *11*(7), 283–290.

Kasianowicz, J. J., Brandin, E., Branton, D. et Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, *93*(24), 13770–13773.

Katoh, K., Misawa, K., Kuma, K.-i. et Miyata, T. (2002). Mafft : a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, *30*(14), 3059–3066.

Katoh, K., Rozewicki, J. et Yamada, K. D. (2019). Mafft online service : multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, *20*(4), 1160–1166.

Katoh, K. et Standley, D. M. (2013). Mafft multiple sequence alignment software version 7 : improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780.

Kawashima, S., Ogata, H. et Kanehisa, M. (1999). Aaindex : amino acid index database. *Nucleic acids research*, *27*(1), 368–369.

Kimura, I., Kosugi, Y., Wu, J., Zahradnik, J., Yamasoba, D., Butlertanaka, E. P., Tanaka, Y. L., Uriu, K., Liu, Y., Morizako, N. *et al.* (2022). The sars-cov-2 lambda variant exhibits enhanced infectivity and immune resistance. *Cell reports*, *38*(2), 110218.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*, 111–120.

King, A. M., Adams, M. J., Carstens, E. B. et Lefkowitz, E. J. (2012). *Virus taxonomy : classification and nomenclature of viruses*.

King, A. M., Lefkowitz, E., Adams, M. J. et Carstens, E. B. (2011). *Virus taxonomy : ninth report of the International Committee on Taxonomy of Viruses*, volume 9. Elsevier.

Kira, K. et Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992* 249–256. Elsevier.

Kledal, T. N., Rosenkilde, M. M. et Schwartz, T. W. (1998). Selective recognition of the membrane-bound cx3c chemokine, fractalkine, by the human cytomegalovirus-encoded broad-spectrum receptor us28. *FEBS letters*, *441*(2), 209–214.

Koehler, M., Ray, A., Moreira, R. A., Juniku, B., Poma, A. B. et Alsteens, D. (2021). Molecular insights into receptor binding energetics and neutralization of sars-cov-2 variants. *Nature communications*, *12*(1), 6977.

Koonin, E. V., Dolja, V. V., Krupovic, M. et Kuhn, J. H. (2021). Viruses defined by the position of the virosphere within the replicator space. *Microbiology and Molecular Biology Reviews*, *85*(4), e00193–20.

Kosakovsky Pond, S. L., Posada, D., Stawiski, E., Chappey, C., Poon, A. F., Hughes, G., Fearnhill, E., Gravenor, M. B., Leigh Brown, A. J. et Frost, S. D. (2009). An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. *PLoS computational biology*, *5*(11), e1000581.

Koyama, T., Weeraratne, D., Snowdon, J. L. et Parida, L. (2020). Emergence of drift variants that may affect covid-19 vaccine development and antibody treatment. *Pathogens*, *9*(5), 324.

Krigbaum, W. et Komoriya, A. (1979). Local interactions as a structure determinant for protein molecules : Ii. *Biochimica et biophysica acta*, *576*(1), 204–248.

Krishna, B. A., Poole, E. L., Jackson, S. E., Smit, M. J., Wills, M. R. et Sinclair, J. H. (2017). Latency-associated expression of human cytomegalovirus us28 attenuates cell signaling pathways to maintain latent infection. *MBio*, *8*(6), 10–1128.

Krishna, B. A., Wass, A. B., Dooley, A. L. et O'Connor, C. M. (2021). Cmv-encoded gpcr pul33 activates creb and facilitates its recruitment to the mie locus for efficient viral reactivation. *Journal of cell science*, *134*(5), jcs254268.

Krupovic, M., Kuhn, J. H., Wang, F., Baquero, D. P., Dolja, V. V., Egelman, E. H., Prangishvili, D. et Koonin, E. V. (2021). Adnaviria : a new realm for archaeal filamentous viruses with linear a-form double-stranded dna genomes. *Journal of Virology*, *95*(15), 10–1128.

Kschonsak, M., Johnson, M. C., Schelling, R., Green, E. M., Rougé, L., Ho, H., Patel, N., Kilic, C., Kraft, E., Arthur, C. P. *et al.* (2022). Structural basis for hcmv pentamer receptor recognition and antibody neutralization. *Science advances*, *8*(10), eabm2536.

Kschonsak, M., Rougé, L., Arthur, C. P., Hoangdung, H., Patel, N., Kim, I., Johnson, M. C., Kraft, E., Rohou, A. L., Gill, A. *et al.* (2021). Structures of hcmv trimer reveal the basis for receptor recognition and cell entry. *Cell*, *184*(5), 1232–1244.

Kullback, S. et Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Lamers, M. M. et Haagmans, B. L. (2022). Sars-cov-2 pathogenesis. *Nature reviews microbiology*, *20*(5), 270–284.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L. *et al.* (2020). Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *nature*, *581*(7807), 215–220.

Lange, K. (2002). *Mathematical and statistical methods for genetic analysis*, volume 488. Springer.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R. *et al.* (2007). Clustal w and clustal x version 2.0. *bioinformatics*, *23*(21), 2947–2948.

Lassalle, F., Depledge, D. P., Reeves, M. B., Brown, A. C., Christiansen, M. T., Tutill, H. J., Williams, R. J., Einer-Jensen, K., Holdstock, J., Atkinson, C. *et al.* (2016). Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus evolution*, *2*(1), vew017.

Lau, B. T., Pavlichin, D., Hooker, A. C., Almeda, A., Shin, G., Chen, J., Sahoo, M. K., Huang, C. H., Pinsky, B. A., Lee, H. J. *et al.* (2021). Profiling sars-cov-2 mutation fingerprints that range from the viral pangenome to individual infection quasispecies. *Genome medicine*, *13*(1), 1–23.

Lauber, C. et Gorbalenya, A. E. (2012). Partitioning the genetic diversity of a virus family : approach and evaluation through a case study of picornaviruses. *Journal of virology*, *86*(7), 3890–3904.

Lebatteux, D. et Diallo, A. B. (2021). Combining a genetic algorithm and ensemble method to improve the classification of viruses. Dans *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 688–693. IEEE.

Lebatteux, D., Remita, A. M. et Diallo, A. B. (2019). Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *Journal of Computational Biology*, *26*(6), 519–535.

Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S. et Diallo, A. B. (2022). Kanalyzer : a method to identify variations of discriminative k-mers in genomic sequences. Dans *2022 IEEE International Conference On Bioinformatics And Biomedicine (BIBM)*, 757–762. IEEE.

Lebatteux, D., Soudeyns, H., Boucoiran, I., Gantt, S. et Diallo, A. B. (2024). Machine learning-based approach kevolve efficiently identifies sars-cov-2 variant-specific genomic signatures. *Plos one*, *19*(1), e0296627.

Lee, E. Y., Ng, M.-Y. et Khong, P.-L. (2020). Covid-19 pneumonia : what has ct taught us ? *The Lancet Infectious Diseases*, *20*(4), 384–385.

Lee, S., Chung, Y. H. et Lee, C. (2017). Us28, a virally-encoded gpcr as an antiviral target for human cytomegalovirus infection. *Biomolecules & Therapeutics*, *25*(1), 69.

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G. et Smith, D. B. (2018). Virus taxonomy : the database of the international committee on taxonomy of viruses (ictv). *Nucleic acids research*, *46*(D1), D708–D717.

Lengruber, R. B., Delviks-Frankenberry, K. A., Nikolenko, G. N., Baumann, J., Santos, A. F., Pathak, V. K. et Soares, M. A. (2011). Phenotypic characterization of drug resistance-associated mutations in hiv-1 rt connection and rnase h domains and their correlation with thymidine analogue mutations. *Journal of antimicrobial chemotherapy*, *66*(4), 702–708.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. et Liu, H. (2018). Feature selection : A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 94.

Li, M., Vitányi, P. *et al.* (2008). *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer.

Li, P., Meng, Y., Wang, L. et Di, L.-j. (2019). Bioid : a proximity-dependent labeling approach in proteomics study. *Functional Proteomics : Methods and Protocols*, 143–151.

Libbrecht, M. W. et Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332.

Liu, Y., Ghaffari, M. H., Ma, T. et Tu, Y. (2024). Impact of database choice and confidence score on the performance of taxonomic classification using kraken2. *aBIOTECH*, *5*(4), 465–475.

Liu, Y., Heim, K. P., Che, Y., Chi, X., Qiu, X., Han, S., Dormitzer, P. R. et Yang, X. (2021). Prefusion structure of human cytomegalovirus glycoprotein b and structural basis for membrane fusion. *Science Advances*, *7*(10), eabf3178.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Mulders, D. G., Molenkamp, R., Perez-Romero, C. A., Claassen, E., Garssen, J. et Kraneveld, A. D. (2021). Classification and specific primer design for accurate detection of sars-cov-2 using deep learning. *Scientific reports*, *11*(1), 1–11.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N. *et al.* (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus : implications for virus origins and receptor binding. *The lancet*, *395*(10224), 565–574.

Lu, Y. Y., Noble, W. S. et Keich, U. (2024). A blast from the past : revisiting blastp's e-value. *Bioinformatics*, *40*(12), btae729.

Lübke, M., Spalt, S., Kowalewski, D. J., Zimmermann, C., Bauersfeld, L., Nelde, A., Bichmann, L., Marcu, A., Peper, J. K., Kohlbacher, O. *et al.* (2019). Identification of hcmv-derived t cell epitopes in seropositive individuals through viral deletion models. *Journal of Experimental Medicine*, *217*(3), e20191164.

Mahmood, T. B., Hossan, M. I., Mahmud, S., Shimu, M. S. S., Alam, M. J., Bhuyan, M. M. R. et Emran, T. B. (2022). Missense mutations in spike protein of sars-cov-2 delta variant contribute to the alteration in viral structure and interaction with hace2 receptor. *Immunity, Inflammation and Disease*, *10*(9), e683.

Marie, V. et Gordon, M. L. (2022). The hiv-1 gag protein displays extensive functional and structural roles in virus replication and infectivity. *International Journal of Molecular Sciences*, *23*(14), 7569.

Matsen, F. A., Kodner, R. B. et Armbrust, E. V. (2010). pplacer : linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, *11*, 1–16.

Matthews, R. E. F. (2018). *A critical appraisal of viral taxonomy*. CRC Press.

Mayer, B. T., Krantz, E. M., Swan, D., Ferrenberg, J., Simmons, K., Selke, S., Huang, M.-L., Casper, C., Corey, L., Wald, A. *et al.* (2017). Transient oral human cytomegalovirus infections indicate inefficient viral spread from very few initially infected cells. *Journal of virology*, *91*(12), 10–1128.

Mayne, R. M., Aiewsakun, P., Turner, D., Adriaenssens, E. M. et Simmonds, P. (2024). Gravity-v2 : a grounded viral taxonomy application. *bioRxiv*, 2024–07.

McCallum, A., Nigam, K. *et al.* (1998). A comparison of event models for naive bayes text classification. Dans *AAAI-98 workshop on learning for text categorization*, volume 752, 41–48. Madison, WI.

McCallum, M., De Marco, A., Lempp, F. A., Tortorici, M. A., Pinto, D., Walls, A. C., Beltramello, M., Chen, A., Liu, Z., Zatta, F. *et al.* (2021). N-terminal domain antigenic mapping reveals a site of vulnerability for sars-cov-2. *Cell*, *184*(9), 2332–2347.

McMillen, T., Jani, K., Robilotti, E. V., Kamboj, M. et Babady, N. E. (2022). The spike gene target failure (sgtf) genomic signature is highly accurate for the identification of alpha and omicron sars-cov-2 variants. *Scientific Reports*, *12*(1), 1–8.

Meier-Kolthoff, J. P. et Göker, M. (2017). Victor : genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*, *33*(21), 3396–3404.

Miller, S., Janin, J., Lesk, A. M. et Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of molecular biology*, *196*(3), 641–656.

Mirjalili, S. (2019). Genetic algorithm. In *Evolutionary algorithms and neural networks* 43–55. Springer.

Mitchell, T. M. et Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Miyata, T., Miyazawa, S. et Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of molecular evolution*, *12*, 219–236.

Modha, S., Thanki, A. S., Cotmore, S. F., Davison, A. J. et Hughes, J. (2018). Victree : an automated framework for taxonomic classification from protein sequences. *Bioinformatics*, *34*(13), 2195–2200.

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A. et Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*.

Moraru, C., Varsani, A. et Kropinski, A. M. (2020). Viridic—a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses*, *12*(11), 1268.

Morgan, A. A. et Rubenstein, E. (2013). Proline : the distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. *PloS one*, *8*(1), e53785.

Motozono, C., Toyoda, M., Zahradnik, J., Saito, A., Nasser, H., Tan, T. S., Ngare, I., Kimura, I., Uriu, K., Kosugi, Y. *et al.* (2021). Sars-cov-2 spike l452r variant evades cellular immunity and increases infectivity. *Cell host & microbe*, *29*(7), 1124–1136.

Muhire, B. M., Varsani, A. et Martin, D. P. (2014). Sdt : a virus classification tool based on pairwise sequence alignment and identity calculation. *PloS one*, *9*(9), e108277.

Müller, A. C. et Guido, S. (2016). *Introduction to machine learning with Python : a guide for data scientists*. " O'Reilly Media, Inc.".

Muttineni, R., Putty, K., Marapakala, K., KP, S., Panyam, J., Vemula, A., Singh, S. M., Balachandran, S., ST, V. R. et Kondapi, A. K. (2022). Sars-cov-2 variants and spike mutations involved in second wave of covid-19 pandemic in india. *Transboundary and Emerging Diseases*, *69*(5), e1721–e1733.

Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., Atif, S. M., Hariprasad, G., Hasan, G. M. et Hassan, M. I. (2020). Insights into sars-cov-2 genome, structure, evolution, pathogenesis and therapies : Structural genomics approach. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1866*(10), 165878.

Narlikar, L., Gordân, R. et Hartemink, A. J. (2007). Nucleosome occupancy information improves de novo motif discovery. Dans *Annual International Conference on Research in Computational Molecular Biology*, 107–121. Springer.

Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.

Nelson, G., Buzko, O., Spilman, P., Niazi, K., Rabizadeh, S. et Soon-Shiong, P. (2021). Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant. *BioRxiv*.

Noriega, V. M., Gardner, T. J., Redmann, V., Bongers, G., Lira, S. A. et Tortorella, D. (2014). Human cytomegalovirus us28 facilitates cell-to-cell viral dissemination. *Viruses*, *6*(3), 1202–1218.

Notredame, C., Higgins, D. G. et Heringa, J. (2000). T-coffee : A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), 205–217.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N. *et al.* (2014). The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, *42*(D1), D358–D363.

Orozco-Arias, S., Candamil-Cortés, M. S., Jaimes, P. A., Piña, J. S., Tabares-Soto, R., Guyot, R. et Isaza, G. (2021). K-mer-based machine learning method to classify ltr-retrotransposons in plant genomes. *PeerJ*, *9*, e11456.

Otu, H. H. et Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, *19*(16), 2122–2130.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F. *et al.* (2021). The biogrid database : A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, *30*(1), 187–200.

Ounit, R., Wanamaker, S., Close, T. J. et Lonardi, S. (2015). Clark : fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, *16*, 1–13.

Pascarella, S. et Negro, F. (2011). Hepatitis d virus : an update. *Liver International*, *31*(1), 7–21.

Payne, R., Branch, S., Kløverpris, H., Matthews, P., Koofhethile, C., Strong, T., Adland, E., Leitman, E., Frater, J., Ndung'u, T. *et al.* (2014). Differential escape patterns within the dominant hla-b* 57 : 03-restricted hiv gag epitope reflect distinct clade-specific functional constraints. *Journal of virology*, *88*(9), 4668–4678.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L. *et al.* (2023). Interpro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427.

Pearson, K. (1895). Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, *58*(347-352), 240–242.

Pegg, C. L., Modhiran, N., Parry, R. H., Liang, B., Amarilla, A. A., Khromykh, A. A., Burr, L., Young, P. R., Chappell, K., Schulz, B. L. *et al.* (2024). The role of n-glycosylation in spike antigenicity for the sars-cov-2 gamma variant. *Glycobiology*, *34*(2), cwad097.

Peng, Q., Zhou, R., Liu, N., Wang, H., Xu, H., Zhao, M., Yang, D., Au, K.-K., Huang, H., Liu, L. *et al.* (2022). Naturally occurring spike mutations influence the infectivity and immunogenicity of sars-cov-2. *Cellular & Molecular Immunology*, *19*(11), 1302–1310.

Peng, S. (2024). Hiv-1 m group subtype classification using deep learning approach. *Computers in Biology and Medicine*, *183*, 109218.

Perez-Bercoff, L., Valentini, D., Gaseitsiwe, S., Mahdavifar, S., Schutkowski, M., Poiret, T., Pérez-Bercoff, Å., Ljungman, P. et Maeurer, M. J. (2014). Whole cmv proteome pattern recognition analysis after hsct identifies unique epitope targets associated with the cmv status. *PLoS One*, *9*(4), e89648.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z. *et al.* (2012). Vipr : an open bioinformatics database and analysis resource for virology research. *Nucleic acids research*, *40*(D1), D593–D598.

Pignatelli, S., Dal Monte, P. et Landini, M. (2001). gpul73 (gn) genomic variants of human cytomegalovirus isolates are clustered into four distinct genotypes. *Journal of General Virology*, *82*(11), 2777–2784.

Pignatelli, S., Lazzarotto, T., Gatto, M. R., Dal Monte, P., Landini, M. P., Faldella, G. et Lanari, M. (2010). Cytomegalovirus gn genotypes distribution among congenitally infected newborns and their relationship with symptoms at birth and sequelae. *Clinical infectious diseases*, *51*(1), 33–41.

Pineda-Peña, A.-C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., Gómez-López, A., Camacho, R. J., De Oliveira, T. et Vandamme, A.-M. (2013). Automated subtyping of hiv-1 genetic sequences for clinical and surveillance purposes : performance evaluation of the new rega version 3 and seven other tools. *Infection, genetics and evolution*, *19*, 337–348.

Platt, J. *et al.* (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61–74.

Pötzsch, S., Spindler, N., Wiegers, A.-K., Fisch, T., Rücker, P., Sticht, H., Grieb, N., Baroti, T., Weisel, F., Stamminger, T. *et al.* (2011). B cell repertoire analysis identifies new antigenic domains on glycoprotein b of human cytomegalovirus which are target of neutralizing antibodies. *PLoS pathogens*, *7*(8), e1002172.

Qin, L., Kufareva, I., Holden, L. G., Wang, C., Zheng, Y., Zhao, C., Fenalti, G., Wu, H., Han, G. W., Cherezov, V. *et al.* (2015). Crystal structure of the chemokine receptor cxcr4 in complex with a viral chemokine. *Science*, *347*(6226), 1117–1122.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Raju, R. S., Al Nahid, A., Dev, P. C. et Islam, R. (2022). Virustaxo : Taxonomic classification of viruses from the genome sequence using k-mer enrichment. *Genomics*, *114*(4), 110414.

Ramani, R., Krumholz, K., Huang, Y.-F. et Siepel, A. (2019). Phastweb : a web interface for evolutionary conservation scoring of multiple sequence alignments using phastcons and phylop. *Bioinformatics*, *35*(13), 2320–2322.

Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A. et Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens : Covid-19 case study. *Plos one*, *15*(4), e0232391.

Remita, A. M. et Diallo, A. B. (2019). Statistical linear models in virus genomic alignment-free classification : application to hepatitis c viruses. Dans *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 474–481. IEEE.

Remita, M. A., Halioui, A., Malick Diouara, A. A., Daigle, B., Kiani, G. et Diallo, A. B. (2017). A machine learning approach for viral genome classification. *BMC bioinformatics*, *18*, 1–11.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. et Sun, F. (2017). Virfinder : a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*, 1–20.

Reynisson, B., Alvarez, B., Paul, S., Peters, B. et Nielsen, M. (2020). Netmhcpan-4.1 and netmhciipan-4.0 : improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, *48*(W1), W449–W454.

Rokach, L. et Maimon, O. (2005). Clustering methods. *Data mining and knowledge discovery handbook*, 321–352.

Ronquist, F. et Huelsenbeck, J. P. (2003). Mrbayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. et Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, *229*(4716), 834–838.

Rosenbaum, D. M., Rasmussen, S. G. et Kobilka, B. K. (2009). The structure and function of g-protein-coupled receptors. *Nature*, *459*(7245), 356–363.

Rozanov, M., Plikat, U., Chappey, C., Kochergin, A. et Tatusova, T. (2004). A web-based genotyping resource for viral sequences. *Nucleic acids research*, *32*(suppl_2), W654–W659.

Rudnicki, W. R., Mroczek, T. et Cudek, P. (2014). Amino acid properties conserved in molecular evolution. *PLoS One*, *9*(6), e98983.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Saifi, S., Ravi, V., Sharma, S., Swaminathan, A., Chauhan, N. S. et Pandey, R. (2022). Sars-cov-2 vocs, mutational diversity and clinical outcome : Are they modulating drug efficacy by altered binding strength ? *Genomics*, *114*(5), 110466.

Saito, A., Irie, T., Suzuki, R., Maemura, T., Nasser, H., Uriu, K., Kosugi, Y., Shirakawa, K., Sadamasu, K., Kimura, I. *et al.* (2022). Enhanced fusogenicity and pathogenicity of sars-cov-2 delta p681r mutation. *Nature*, *602*(7896), 300–306.

Saitou, N. et Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406–425.

Sarafianos, S. G., Das, K., Tantillo, C., ClarkJr, A. D., Ding, J., Whitcomb, J. M., Boyer, P. L., Hughes, S. H. et Arnold, E. (2001). Crystal structure of hiv-1 reverse transcriptase in complex with a polypurine tract rna : Dna. *The EMBO journal*.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S. *et al.* (2022). Database resources of the national center for biotechnology information. *Nucleic acids research*, *50*(D1), D20–D26.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature methods*, *5*(1), 16–18.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shen, L., Bard, J. D., Triche, T. J., Judkins, A. R., Biegel, J. A. et Gai, X. (2021). Emerging variants of concern in sars-cov-2 membrane protein : a highly conserved target with potential pathological and therapeutic implications. *Emerging microbes & infections*, *10*(1), 885–893.

Siddell, S. G., Smith, D. B., Adriaenssens, E., Alfenas-Zerbini, P., Dutilh, B. E., Garcia, M. L., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J. *et al.* (2023). Virus taxonomy and the role of the international committee on taxonomy of viruses (ictv). *Journal of General Virology*, *104*(5), 001840.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034–1050.

Sijmons, S., Thys, K., Mbong Ngwese, M., Van Damme, E., Dvorak, J., Van Loock, M., Li, G., Tachezy, R., Busson, L., Aerssens, J. *et al.* (2015). High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. *Journal of virology*, *89*(15), 7673–7695.

Simmonds, P. (2024). A critique of the use of species and below-species taxonomic terms for viruses–time for change ? *Virus Evolution*, p. veae096.

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B. *et al.* (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168.

Simmonds, P. et Aiewsakun, P. (2018). Virus classification–where do you draw the line ? *Archives of virology*, *163*, 2037–2046.

Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C., Maertens, G. *et al.* (2005). Consensus proposals for a unified system of nomenclature of hepatitis c virus genotypes. *Hepatology*, *42*(4), 962–973.

Singh, P., Sharma, K., Singh, P., Bhargava, A., Negi, S. S., Sharma, P., Bhise, M., Tripathi, M. K., Jindal, A. et Nagarkar, N. M. (2022). Genomic characterization unravelling the causative role of sars-cov-2 delta variant of lineage b. 1.617. 2 in 2nd wave of covid-19 pandemic in chhattisgarh, india. *Microbial Pathogenesis*, *164*, 105404.

Slezak, T., Hart, B. et Jaing, C. (2020). Design of genomic signatures for pathogen identification and characterization. In *Microbial Forensics* 299–312. Elsevier.

Smith, T. F., Waterman, M. S. *et al.* (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195–197.

Sneath, P. (1966). Relations between chemical structure and biological activity in peptides. *Journal of theoretical biology*, *12*(2), 157–195.

Sokal, R. R. et Michener, C. D. (1958). A statistical method for evaluating systematic relationships.

Solis-Reyes, S., Avino, M., Poon, A. et Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes. *PloS one*, *13*(11), e0206409.

Spearman, C. (1961). The proof and measurement of association between two things.

Stagno, S., Pass, R. F., Dworsky, M. E. et Alford, C. A. (1982). Maternal cytomegalovirus infection and perinatal transmission. *Clinical obstetrics and gynecology*, *25*(3), 563–576.

Stamatakis, A. (2006). Raxml-vi-hpc : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690.

Stangherlin, L. M., de Paula, F. N., Icimoto, M. Y., Ruiz, L. G., Nogueira, M. L., Braz, A. S., Juliano, L. et da Silva, M. C. (2017). Positively selected sites at hcmv gb furin processing region and their effects in cleavage efficiency. *Frontiers in microbiology*, *8*, 934.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C. *et al.* (2020). Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell*, *182*(5), 1295–1310.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society : Series B (Methodological)*, *36*(2), 111–133.

Storato, D. et Comin, M. (2020). Improving metagenomic classification using discriminative k-mers from sequencing data. *BioRxiv*.

Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C. et Bercoff, D. P. (2014). Comet : adaptive context-based modeling for ultrafast hiv-1 subtype identification. *Nucleic acids research*, *42*(18), e144–e144.

Syed, A. M., Taha, T. Y., Tabata, T., Chen, I. P., Ciling, A., Khalid, M. M., Sreekumar, B., Chen, P.-Y., Hayashi, J. M., Soczek, K. M. *et al.* (2021). Rapid assessment of sars-cov-2–evolved variants using virus-like particles. *Science*, *374*(6575), 1626–1632.

Tadagaki, K., Tudor, D., Gbahou, F., Tschische, P., Waldhoer, M., Bomsel, M., Jockers, R. et Kamal, M. (2012). Human cytomegalovirus-encoded ul33 and ul78 heteromerize with host ccr5 and cxcr4 impairing their hiv coreceptor activity. *Blood, The Journal of the American Society of Hematology*, *119*(21), 4908–4918.

Tamanaha, E., Zhang, Y. et Tanner, N. A. (2022). Profiling rt-lamp tolerance of sequence variation for sars-cov-2 rna detection. *PLoS One*, *17*(3), e0259610.

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E. et Hammer, S. M. (2008). The challenge of hiv-1 subtype diversity. *New England Journal of Medicine*, *358*(15), 1590–1602.

Thakur, S., Sasi, S., Pillai, S. G., Nag, A., Shukla, D., Singhal, R., Phalke, S. et Velu, G. (2022). Sars-cov-2 mutations and their impact on diagnostics, therapeutics and vaccines. *Frontiers in Medicine*, *9*.

Thomas, A., Barriere, S., Broseus, L., Brooke, J., Lorenzi, C., Villemin, J.-P., Beurier, G., Sabatier, R., Reynes, C., Mancheron, A. *et al.* (2019). Gecko is a genetic algorithm to classify and explore high throughput sequencing data. *Communications biology*, *2*(1), 1–8.

Thompson, J. D., Higgins, D. G. et Gibson, T. J. (1994). Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, *22*(22), 4673–4680.

Tomii, K. et Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Engineering, Design and Selection*, *9*(1), 27–36.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y. et Kiyotani, K. (2020). Sars-cov-2 genomic variations associated with mortality rate of covid-19. *Journal of human genetics*, *65*(12), 1075–1082.

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D. et Corbett-Detig, R. (2021). Ultrafast sample placement on existing trees (usher) enables real-time phylogenetics for the sars-cov-2 pandemic. *Nature genetics*, *53*(6), 809–816.

Uversky, V. et Longhi, S. (2012). *Flexible viruses : structural disorder in viral proteins*, volume 11. John Wiley & Sons.

van Senten, J. R., Bebelman, M. P., van Gasselt, P., Bergkamp, N. D., van den Bor, J., Siderius, M. et Smit, M. J. (2020). Human cytomegalovirus-encoded g protein-coupled receptor ul33 facilitates virus dissemination via the extracellular and cell-to-cell route. *Viruses*, *12*(6), 594.

Venkatakrishnan, A., Anand, P., Lenehan, P. J., Suratekar, R., Raghunathan, B., Niesen, M. J. et Soundararajan, V. (2022). On the origins of omicron's unique spike gene insertion. *Vaccines*, *10*(9), 1509.

Venturini, C., Pang, J., Tamuri, A. U., Roy, S., Atkinson, C., Griffiths, P., Breuer, J. et Goldstein, R. A. (2022). Haplotype assignment of longitudinal viral deep sequencing data using covariation of variant frequencies. *Virus Evolution*, *8*(2), veac093.

Vinga, S. (2014a). Alignment-free methods in computational biology.

Vinga, S. (2014b). Information theory applications for biological sequence analysis. *Briefings in bioinformatics*, *15*(3), 376–389.

Vinga, S. et Almeida, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics*, *19*(4), 513–523.

Voichek, Y. et Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5), 534–540.

Vovk, V. et Wang, R. (2020). Combining p-values via averaging. *Biometrika*, *107*(4), 791–808.

Wang, L. et Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of computational biology*, *1*(4), 337–348.

Wang, Z., Schmidt, F., Weisblum, Y., Muecksch, F., Barnes, C. O., Finkin, S., Schaefer-Babajew, D., Cipolla, M., Gaebler, C., Lieberman, J. A. *et al.* (2021). mrna vaccine-elicited antibodies to sars-cov-2 and circulating variants. *Nature*, *592*(7855), 616–622.

Waters, S., Agostino, M., Lee, S., Ariyanto, I., Kresoje, N., Leary, S., Munyard, K., Gaudieri, S., Gaff, J., Irish, A. *et al.* (2021). Sequencing directly from clinical specimens reveals genetic variations in hcmv-encoded chemokine receptor us28 that may influence antibody levels and interactions with human chemokines. *Microbiology spectrum*, *9*(2), e00020–21.

Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J. et Gustafsson, C. (2009). Design parameters to control synthetic gene expression in escherichia coli. *PloS one*, *4*(9), e7002.

Wilkinson, D. A., Joffrin, L., Lebarbenchon, C. et Mavingui, P. (2020). Analysis of partial sequences of the rna-dependent rna polymerase gene as a tool for genus and subgenus classification of coronaviruses. *Journal of General Virology*, *101*(12), 1261–1269.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z. *et al.* (2018). Drugbank 5.0 : a major update to the drugbank database for 2018. *Nucleic acids research*, *46*(D1), D1074–D1082.

Wong, K. M., Suchard, M. A. et Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, *319*(5862), 473–476.

Wood, D. E., Lu, J. et Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome biology*, *20*, 1–13.

Wood, D. E. et Salzberg, S. L. (2014). Kraken : ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, *15*, 1–12.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y. *et al.* (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, *579*(7798), 265–269.

Wu, H., Xing, N., Meng, K., Fu, B., Xue, W., Dong, P., Tang, W., Xiao, Y., Liu, G., Luo, H. *et al.* (2021). Nucleocapsid mutations r203k/g204r increase the infectivity, fitness, and virulence of sars-cov-2. *Cell host & microbe*, *29*(12), 1788–1801.

Wu, T.-F., Lin, C.-J. et Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*(Aug), 975–1005.

Xia, S., Zhu, Y., Liu, M., Lan, Q., Xu, W., Wu, Y., Ying, T., Liu, S., Shi, Z., Jiang, S. *et al.* (2020). Fusion mechanism of 2019-ncov and fusion inhibitors targeting hr1 domain in spike protein. *Cellular & molecular immunology*, *17*(7), 765–767.

Xia, X. et Xia, X. (2018). Protein substitution model and evolutionary distance. *Bioinformatics and the Cell : Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*, 315–326.

Xu, D. et Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193.

Yang, S., Yu, Y., Jian, F., Song, W., Yisimayi, A., Chen, X., Xu, Y., Wang, P., Wang, J., Yu, L. *et al.* (2023). Antigenicity and infectivity characterisation of sars-cov-2 ba. 2.86. *The Lancet Infectious Diseases*, *23*(11), e457–e459.

Yin, C. (2020). Genotyping coronavirus sars-cov-2 : methods and implications. *Genomics*, *112*(5), 3588–3596.

Yousef, M., Nigatu, D., Levy, D., Allmer, J. et Henkel, W. (2017). Categorization of species based on their micrornas employing sequence motifs, information-theoretic sequence feature extraction, and k-mers. *EURASIP Journal on Advances in Signal Processing*, *2017*, 1–10.

Yu, L. et Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, *5*, 1205–1224.

Yu, Q., Huo, H., Chen, X., Guo, H., Vitter, J. S. et Huan, J. (2015). An efficient algorithm for discovering motifs in large dna data sets. *IEEE transactions on nanobioscience*, *14*(5), 535–544.

Yu, Y., Li, M., Ji, T. et Wu, Q. (2022). Fault location approach for distribution network with dynamic environment. *International Transactions on Electrical Energy Systems*, *2022*(1), 3065602.

Zhang, J., Xiao, T., Cai, Y., Lavine, C. L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M. L. *et al.* (2021). Membrane fusion and immune evasion by the spike protein of sars-cov-2 delta variant. *Science*, *374*(6573), 1353–1360.

Zhang, Q., Jun, S.-R., Leuze, M., Ussery, D. et Nookaew, I. (2017). Viral phylogenomics using an alignment-free method : A three-step approach to determine optimal length of k-mer. *Scientific reports*, *7*(1), 40712.

Zhou, N., Luo, Z., Luo, J., Liu, D., Hall, J. W., Pomerantz, R. J. et Huang, Z. (2001). Structural and functional characterization of human cxcr4 as a chemokine receptor and hiv-1 co-receptor by mutagenesis and molecular modeling studies. *Journal of Biological Chemistry*, *276*(46), 42826–42833.

Zhu, C., He, G., Yin, Q., Zeng, L., Ye, X., Shi, Y. et Xu, W. (2021). Molecular biology of the sars-cov-2 spike protein : A review of current knowledge. *Journal of Medical Virology*, *93*(10), 5729–5741.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. *et al.* (2020). A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*.

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., Lau, A. K., Röhling, S., Choi, J. J., Waterman, M. S. *et al.* (2019). Benchmarking of alignment-free sequence comparison methods. *Genome biology*, *20*, 1–18.

Zielezinski, A., Vinga, S., Almeida, J. et Karlowski, W. M. (2017). Alignment-free sequence comparison : benefits, applications, and tools. *Genome biology*, *18*, 1–17.

Zimmerman, J., Eliezer, N. et Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology*, *21*(2), 170–201.

Zuckerman, N. S., Fleishon, S., Bucris, E., Bar-Ilan, D., Linial, M., Bar-Or, I., Indenbaum, V., Weil, M., Lustig, Y., Mendelson, E. *et al.* (2021). A unique sars-cov-2 spike protein p681h variant detected in israel. *Vaccines*, *9*(6), 616.

Zuhair, M., Smit, G. S. A., Wallis, G., Jabbar, F., Smith, C., Devleesschauwer, B. et Griffiths, P. (2019). Estimation of the worldwide seroprevalence of cytomegalovirus : a systematic review and meta-analysis. *Reviews in medical virology*, *29*(3), e2034.

Zvelebil, M. et Baum, J. O. (2007). *Understanding bioinformatics*. Garland Science.