Aix Marseille Université

NNT : 2025AIXM0034

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 11 février 2025, par

Vincent WILDE

Développement d'outils bioinformatiques pour la caractérisation et l'analyse de nouveaux génomes viraux

Discipline
Biologie-Santé

Spécialité
Bioinformatique et Génomique

École doctorale
ED62

Laboratory
Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, PolytechCase 925, 13009 Marseille, France.

Composition du jury

Julie THOMPSON HDR, CNRS UMR 7357 - Université de Strasbourg	Rapporteure
Philippe Le MERCIER HDR, Swiss-Prot	Rapporteur
Anna-Sophie FISTON-LAVIER Professeure, Université de Montpellier	Examinatrice
Pedro COUTINHO Professeur, AFMB - CNRS UMR 7257	Président du jury
François FERRON CR, AFMB - CNRS UMR 7257	Directeur de thèse

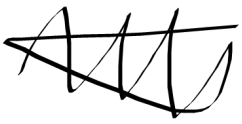
Affidavit

Je soussigné, Vincent Anthony Wilde, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de François Ferron, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 18 Juin 2024

Vincent WILDE



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Curriculum Vitae

TABLEAU 1. – **Liste des publications :**

Publications	Journal	Date de publication	Titre	Commentaire	liens
Article scientifique	Viruses	2023	Viral Instant Mutation Viewer (VIMVer) : a tool to speed up the identification and analysis of new SARS-CoV2 emerging variant and beyond	Présentation d'un outil bioinformatique for l'analyse des variants du SARS-CoV-2, donnant type et positions des mutations dans le génome et potéome.	Pmid : 37631971 PMCID : PMC10458308

TABLEAU 2. – **Participation aux conférences :**

Publications	Évènement	Date de publication	Titre	Commentaire	liens
poster	ViBioM2022	2022-03	Viral Instant Mutation Viewer (VIMVer) : a tool to speed up the identification and analysis of new SARS-CoV2 emerging variant	Exposant le concept du projet VIMVer	hal-04387599
poster	JFV	2022-06	Fast identification and Analysis of SARS-CoV-2 variants using : Viral Instant Mutation Viewer (VIMVer)	Lors du développement précose	doi :10.1684/vir.2022.0942
poster	ViBioM2024	2024-05	How did you annotate your new viral genome without it? VAZyMolO-2	Exposant le concept et le développement précose du porjet VAZyMolO-2	hal-04616074

Résumé et mots clés

La production massive de données bioinformatiques par le séquençage et la métagénomique ouvre des perspectives fascinantes sur le monde viral. L'émergence de variants génétiques impacte l'efficacité des vaccins et les options thérapeutiques, créant un besoin urgent d'outils d'analyse.

En réponse à cette demande croissante, nous présentons deux outils complémentaires :

— VIMVer (Viral Instant Mutation Viewer), un service web *user-friendly* permettant l'identification directe des mutations dans le génome et le protéome viral. Il permet une identification et une visualisation rapides des mutations aux niveaux nucléotidique et protéique, en comparaison avec une séquence de référence.

— VAZyMolO-2, une renaissance du projet original VAZyMolO, intègre des éléments modulaires couvrant la phylogénie, la génomique et la protéomique. Cette plateforme offre des outils analytiques pour visualiser les fonctions des protéines virales d'un point de vue phylogénétique, permettant une réflexion à l'échelle d'une famille virale entière.

Ces deux plateformes, avec leurs interfaces intuitives et leurs bases de données enrichies par notre équipe, comblent une lacune importante dans l'annotation des génomes viraux.

Mots clés : viromique, annotation, protéine, outil bioinformatique, python-django, service web.

Abstract and keywords

The massive production of bioinformatics data through sequencing and metagenomics opens fascinating perspectives into the viral world. The emergence of genetic variants impacts vaccine efficiency and therapeutic options, creating an urgent need for analytical tools.

In response to this growing demand, we present two complementary tools:

— VIMVer (Viral Instant Mutation Viewer), a user-friendly web service enabling direct identification of mutations in both viral genome and proteome. It allows rapid identification and visualization of mutations at both nucleotide and protein levels, in comparison with a reference sequence.

— VAZyMolO-2, a revival of the original VAZyMolO project, integrates modular elements covering phylogeny, genomics, and proteomics. This platform aimes to provide analytical tools to visualize viral protein functions from a phylogenetic perspective, enabling analysis at the scale of an entire viral family.

These two platforms, with their intuitive interfaces and databases enriched by our team, fill an important gap in viral genome annotation.

Keywords: viromics, annotation, protein, bioinformatic tool, python-django, webservice.

Remerciements

Je me lâche.

Je voudrais avant toute chose grandement remercier **François**, mon directeur/superviseur de thèse, un ami. Merci pour ton calme, ta bienveillance et ton optimisme sans faille. Ton savoir encyclopédique et ta pédagogie ont été maintes fois inspirants. J'appréciais et j'apprécierai toutes nos discussions.

Merci aux Directeurs **BOURNE Yves** et **Juan Javier REGUERA VIDAECHEA**, de m'avoir accuelli au sein du laboratoire, et de m'avoir offert la possibilité d'apprendre dans les meilleures conditions.

Bruno, je suis reconnaissant du cadre que tu as construit : favorisant l'entente, permettant des discussions variées et assurant un sérieux professionnel inspirant et inégalé. Je suis heureux d'avoir eu la chance d'arpenter la voie de la Recherche ici.

*

Je remercie ma **maman** pour son soutien indéfectible et absolu. C'est de toi que je dois mon cerveau déviant. Ton plus beau cadeau. Tu m'as aidé à construire un esprit observateur et réfléchi. Tu m'as également transmis ton courage, ta bienveillance et ta force de croire en l'impossible.

Laurent, il nous a toujours suffi d'un regard pour nous comprendre.

Je dois aussi remercier ma **grand-mère**, sans qui je ne serais pas là, en train d'écrire ma thèse. Tu m'as soutenu dans les pires moments. Et je garde en mémoire nos discussions et tes leçons, depuis la cuisine ou la chambre, à jamais.

Dad, I've have always enjoyed our time spent together online. Thank you for my blue eyes and my instictive love for british tea and culture.

Dany, mon frère, tu sais que je me suis toujours senti responsable de toi, étant gamin. La vérité est que j'avais plus besoin de toi que toi de moi. Je suis fier de toi.

Sache que je serrai toujours là pour toi.

Auré, par quoi commencer? Merci pour cette unique aventure que nous vivons ensemble.

**

I would like to thank our Research team members to be forthcoming, serious and playful. Special thanks to Camille, Pierre, Priscilla, Cécilia and Mika.

***

Il me tient à coeur d'aller plus loin dans ce chapitre de remerciements.

Je voudrais remercier le groupe de musique *Placebo*. Leur premier morceau, *Come Home*, est officiellement sorti en 1995, et reste l'un de mes morceaux préférés. ***Placebo*** est à part, non égalé et fascinant. Leurs morceaux m'ont accompagné dans mes pires comme meilleurs moments de ma vie.

Le code *Jedi* m'a suivi très longtemps. Ces 5 lignes m'ont apporté calme et sagesse. Cela pourrait paraître ridicule mais j'y crois vraiment : la Force est une énergie générée par tout organisme vivant. Elle nous entoure, nous traverse, maintient l'univers en équilibre.

There is no emotion, there is peace.
There is no ignorance, there is knowledge.
There is no passion, there is serenity.
There is no chaos, there is harmony.
There is no death, there is the Force.

"Some believe it is only GREAT POWER *that can hold evil in check. But that is not what I have found. I have found that it is the small everyday deeds of ordinary folk that keep the darkness at bay. Small acts of kindness and love."*

*-Gandalf (*J.R.R. TOLKIEN*)*

"Uuuuuuuuur Ahhhhhrrrrrrr

Uhrrr Ahhhhrrrrr

Aaaaaagh"

*-Chewbacca (*GEORGE LUCAS*)*

Table des matières

Affidavit 2

Curriculum Vitae 3

Résumé et mots clés 4

Abstract and keywords 5

Remerciements 6

Table des matières 9

Table des figures 12

Liste des tableaux 14

Liste des pseudo-codes 15

Liste des acronymes 16

Glossaire 18

Avant-propos 25

1. Introduction 27

1.1. La virologie 28

1.1.1. La classification des virus 29

1.1.1.1. La taxonomie 29

1.1.1.2. La classification de Baltimore 32

1.1.1.3. ICTV 32

1.1.2. Les virus émergents 34

1.2. La donnée biologique virale 37

1.2.1. La génomique 38

1.2.2. La séquence biologique 42
1.2.2.1. Le format GENBANK 44
1.2.2.2. Le format FASTA 45
1.2.3. La structure d'une protéine 46
1.2.3.1. La structure secondaire 48
1.2.3.2. La structure tertiaire 49
1.2.3.3. La structure quaternaire 50
1.2.4. Le domaine protéique fonctionnel 51
1.3. Enjeux et règles du jeu 56
1.3.1. Le mystère du protéome sombre. 56
1.3.2. Bio-donnée en abondance 57
1.3.3. L'annotation 59
1.3.4. La règle du jeu : Homologie 60
1.4. Quelle est la boîte à outils du bioinformaticien aujourd'hui? 63
1.4.1. La phylogénétique 63
1.4.1.1. Les principaux aligneurs 63
1.4.1.2. L'arbre phylogénétique 65
1.4.1.3. Distinction entre arbre taxonomique et arbre phylogénétique 65
1.4.2. La recherche de séquence 66
1.4.2.1. Le modèle "*Markov* caché" 67
1.4.2.2. Les prédicteurs de structure protéique 68
1.5. Quelle forme peut prendre un outil bioinformatique? 72
1.6. Buts et Objectifs 74

2. *Viral Instant Mutation Viewer* 79
2.1. Synthèse 80
2.2 Article 84
2.3. Discussion 94

3. VAZyMolO-2 96
3.1. Introduction 97
3.1.1. Origines et inspirations de VAZyMolO-2 97
3.1.2. La modularité dans les protéines virales 98
3.1.2.1. Le rôle de la polymérase dans le complexe de réplication 100
3.1.2.2. Les méthyltransférases des nidovirus 104

3.2. Matériels et Méthodes 106
3.2.1. Le cadre logiciel 106
3.2.1.1. Le déploiement 107
3.2.2. La structure de la base de donnée 108
3.2.2.1. La récupération de *vaz*1 109
3.2.2.2. Ressources Externes - Interopérabilité 112
3.2.3. La recherche de motifs 112
3.2.4. Homologie de séquence 113
3.2.5. Construction des profils de séquence 113
3.2.6. Taxonomie 115
3.2.7. La génération des figures 115
3.2.7.1. Représentation de l'objet *protein* 115
3.2.7.2. Représentation de l'objet *modulo* 117
3.2.7.3. Représentation de l'objet *profile* 118
3.2.8. La recherche par mots-clefs 118
3.3. Description de VAZyMolO-2 120
3.3.1. Le jeu de données test 121
3.3.2. Les utilisateurs 123
3.3.3. Les types d'entrées (input) 123
3.3.4. L'analytique et l'utilisation des données 124
3.4. Discussion 127

4. Conclusion **129**

Bibliographie **132**

ANNEXES **145**

Table des figures

1.1. Un clade : un groupe monophylétique 30
1.2. Déroulement du "chemin taxonomique" menant au sous-genre *Sarbecovirus*, passant par l'Ordre des *Nidovirales* 31
1.3. La classification de Baltimore 32
1.4. La composition hiérarchique de l'Ordre des *Nidovirales* 34
1.5. Taxonomie complète des *Nidovirales* 36
1.6. Les "-omiques" 37
1.7. Du génome à la fonction biologique 38
1.8. Le *frameshift* ribosomal 39
1.9. La Phylogenie des principaux nidovirus et organisation de la pp1ab . . 40
1.10. Les génotypes des principaux nidovirus 41
1.11. Le transcriptome du SARS-CoV-2 41
1.12. Les structures secondaires : hélice α et feuillet β 47
1.13. Les structures secondaires : pont disulfure 48
1.14. Qu'est-ce qu'un motif? 55
1.15. Le "*dark proteome.*" 56
1.16. Les différents types d'Homologie. 61
1.17. Représentation schématique d'un modèle de Markov caché (HMM) pour l'analyse de séquences biologiques. 68

3.1. La main droite 100
3.2. Le génome du SARS-CoV-2 : les protéines du RTC 101
3.3. L'architecture et structure de la nsp12 du SARS-CoV-2 102
3.4. Le complexe de réplication (RTC) minimal du SARS-CoV-2 103
3.5. Diagramme de classe : structure de *vazy2* 108
3.6. La définition de *Subseq* 109
3.7. Architecture of $vaz1$ dataset - UML 110
3.8. Diagramme de séquence : générer la figure représentant une protéine. 116
3.9. Diagramme de séquence : générer la figure représentant un domaine. . 118

3.10. Cycle de vie de la donnée dans VAZyMolO-2 120
3.11. Phylogénie du jeu de données test : les *Orthocoronavirinae* 122
3.12. Annotation du génome du SARS-CoV-2 126

** Chapitre 2 : figures de l'article non inclus*

Liste des tableaux

1. **Liste des publications :** 3
2. **Participation aux conférences :** 3

1.1. Les 20 acides aminés naturels 43

3.1. Certains champs liés à des sources externes dans VAZyMolO-2 112

Liste des pseudo-codes

1. Extraction de la méta-donnée des protéines 111
2. Extraction de la méta-donnée des domaines fonctionnels 111
3. Routine motif search 113
4. Traitement de la sortie de MMseqs2 pour la construction du profile de séquence 114
5. Génération de la figure représentant une protéine 117

Liste des acronymes

+ssRNA

ARN simple brin positif. 32

ADN

acide désoxyribonucléique. 60, 66, 71

AJAX

Asynchronous JavaScript and XML. 18

ARN

acide ribonucléique. 40, 42, 51, 71, 74

cryo-EM

Cryo-Microscopie Électronique. 50, 52

EC

Enzyme Commission (Commission des Enzymes). 52, 54

HIV

virus d'immunodéficience humaine (VIH). 25, 74

HTTPS

HyperText Transfer Protocol Secure. 73

ICTV

International Committee on Taxonomy of Viruses. 33

MERS-CoV

Coronavirus du Syndrome Respiratoire du Moyen-Orient. 28

mmCIF

Macromolecular Crystallographic Information File. 50

MSA

Multiple Sequence alignment. 64

ORF

open reading frame. 39, 41

PDB

Protein Data Bank. 50

RdRp

ARN dépendante ARN polymérase. 51

RMN

Résonance Magnétique Nucléaire. 50

RSV

virus Respiratoire Syncytia (VRS en français). 26, 74

RTC

Replication-Transcription Complex. 51

SARS-CoV-2

coronavirus 2 du syndrome respiratoire aigu et sévère (SRAS-Cov-2 en français). 25, 26, 28, 31, 58, 74, 75, 80, 82, 97

SRAS-CoV

coronavirus du syndrome respiratoire aigu et sévère. 28, 31, 97

Glossaire

Ajax

(*alias*) *Asynchronous JavaScript and XML* (AJAX) est une technique de développement web qui permet de charger et d'échanger des données avec un serveur de manière asynchrone, sans recharger la page entière. Cela permet de créer des applications web plus interactives et réactives, en mettant à jour dynamiquement le contenu de la page en fonction des actions de l'utilisateur. . 118

backend

Le ***backend*** désigne la partie d'une application web ou d'un site internet qui gère la logique, les bases de données et le traitement des données, mais qui n'est pas visible par les utilisateurs. Il inclut les serveurs, les applications et les bases de données, et est généralement développé avec des langages comme Python, Java, Ruby ou PHP. Le ***backend*** est responsable de la gestion des requêtes, de la sécurité et de la communication avec le ***frontend***. . 106, 115, 124

clade

Un **clade** est un groupe d'organismes qui comprend un ancêtre commun et tous ses descendants. En phylogénétique, un **clade** est utilisé pour représenter des relations évolutives, permettant de classer les espèces en fonction de leur histoire évolutive partagée.. 29, 30, 53, 54, 65

clustering

En bioinformatique, il s'agit de regrouper des séquences biologiques similaires.. 64

dark proteome

Le **protéome sombre** désigne les protéines dont la structure tridimensionnelle n'est pas définie et qui ne peuvent pas être analysées par modélisation homologue ou quantification analytique en raison de leur conformation moléculaire inconnue. Ces protéines, souvent appelées "protéines sombres", représentent une partie importante du protéome, notamment chez les eucaryotes et les virus, où elles constituent entre 44 et 54 % du protéome. . 56, 57, 62

docking in silico

Le **docking in silico** est une technique de modélisation informatique utilisée pour prédire la manière dont un ligand, souvent une petite molécule comme un médicament, interagit avec une protéine cible. Cette méthode simule les interactions moléculaires afin d'identifier la conformation la plus stable du complexe ligand-protéine, ce qui est crucial dans le processus de découverte de médicaments. . 71

fold

Un **fold** (repliement) en biologie et en bioinformatique désigne la structure tridimensionnelle qu'adopte une protéine lorsqu'elle se replie. Il représente l'arrangement spatial des différents segments d'une chaîne polypeptidique qui détermine sa fonction biologique. . 52

frameshift

Une mutation ***frameshift***, ou mutation avec décalage du cadre de lecture, est une mutation génétique causée par l'insertion ou la délétion d'un nombre de nucléotides qui n'est pas un multiple de trois dans une séquence d'ADN. Ce type de mutation a des conséquences importantes sur la synthèse des protéines. . 38, 39, 42

frontend

Le ***frontend*** désigne la partie d'une application web ou d'un site internet avec laquelle les utilisateurs interagissent directement. Cela inclut l'ensemble des éléments visuels, tels que la mise en page, les graphiques et les interfaces utilisateur, généralement développés à l'aide de langages comme HTML, CSS et JavaScript. Le ***frontend*** est responsable de l'expérience utilisateur et de l'esthétique de l'application. . 106, 124

in silico

In silico est un terme désignant des recherches ou expériences réalisées par simulation informatique. . 42, 53

input

En bioinformatique, il s'agit de la donnée en entrée d'un flux d'analyse.. 64

métagénomique

La **métagénomique** est une méthode d'étude du contenu génétique d'échantillons issus d'environnements complexes, comme l'intestin, l'océan, les sols ou l'air. Elle permet d'analyser l'ensemble des micro-organismes (bactéries, virus, champignons, levures) présents dans un milieu spécifique, appelé microbiome. . 57, 64, 66, 73, 76

motif structural

Un **motif structural** est un élément récurrent dans la structure tridimensionnelle des protéines qui peut avoir une signification fonctionnelle ou structurelle importante. Ces motifs sont souvent conservés au cours de l'évolution et peuvent être identifiés par l'analyse de séquences protéiques. . 54

open reading frame

Un cadre de lecture ouvert, ou phase ouverte de lecture (***open reading frame*** ou ORF en anglais), est une partie d'un cadre de lecture délimitée entre un codon START et un codon STOP, offrant la possibilité d'être transcrite en ARN, puis potentiellement traduite en protéine ou en peptide. . 38

orthologie

L'**orthologie** est un concept en biologie de l'évolution qui décrit la relation entre deux gènes présents chez deux espèces différentes, issus d'un ancêtre commun par spéciation. . 60

paralogie

Lien évolutif entre deux gènes issus d'un événement de duplication. Les gènes **paralogues** sont homologues et descendent d'un même gène ancestral, mais ont divergé suite à une duplication pour acquérir des fonctions différentes. . 60

parser

(*alias*) En bioinformatique, l'action de "**parser**", de l'anglais "to parse", renvoie au processus d'analyse et de désassemblage du format de fichier de données. On peut alors décortiquer et extraire les informations significatives contenues dans des séquences génomiques ou autres données biologiques. . 21

parsing

Renvoie à l'action de "parser".. 45

phylogénie

La **phylogénie** est l'étude des relations de parenté évolutive entre les êtres vivants. Elle vise à retracer l'histoire évolutive des espèces et à établir leur classification en fonction de leurs liens de parenté génétique. . 29, 30, 32, 33, 51, 59, 63, 66

prédicteur

En bioinformatique, les **prédicteurs** sont des outils ou des algorithmes utilisés pour estimer ou prédire certaines caractéristiques biologiques à partir de données génétiques ou protéiques. Ils jouent un rôle crucial dans des domaines tels que la prédiction de gènes, l'identification de motifs structuraux, et l'évaluation des interactions protéine-protéine. . 53, 68–71

profile

Un **profile** de séquences est une représentation statistique d'un alignement multiple de séquences biologiques. Il capture les caractéristiques conservées et variables d'une famille de séquences apparentées. . 15, 54, 64, 67, 70, 72, 99, 101, 104, 112–115, 125

regex

Une **regex**, ou expression régulière, est une séquence de caractères qui définit un motif de recherche. Utilisée dans le traitement de texte et la programmation, elle permet de rechercher, extraire, remplacer ou valider des chaînes de caractères selon des critères spécifiques. Les **regex** offrent une méthode puissante et flexible pour manipuler du texte en utilisant des symboles et des opérateurs spéciaux pour définir des patterns complexes. . 55

standalone

En informatique, le terme **standalone** désigne un logiciel ou un outil qui peut être exécuté de manière autonome sur un ordinateur, sans nécessiter de connexion à un serveur ou à d'autres ressources externes. . 72

taxon

Un **taxon** est une unité de classification qui regroupe des organismes vivants partageant des caractéristiques communes. En bioinformatique, le terme est souvent utilisé pour désigner des groupes d'organismes à différents niveaux hiérarchiques, tels que les espèces, les genres ou les familles, dans le cadre de la taxonomie systématique. Les taxons permettent d'organiser la biodiversité et de faciliter l'étude des relations évolutives entre les différentes formes de vie. . 30, 33

taxonomie

La **taxonomie**, ou taxinomie, est la science de la classification des organismes vivants. Elle vise à regrouper les êtres vivants en catégories appelées **taxons**, en fonction de leurs caractéristiques communes. . 29, 30, 32–34, 66

template

Un **template**, ou modèle, est un fichier ou une structure préétablie qui sert de base pour créer des documents, des pages web ou des applications. Il contient des éléments de mise en page, des styles et parfois des contenus prédéfinis, permettant de standardiser et de simplifier le processus de création tout en assurant une cohérence visuelle.
Dans Django, un **template** est un fichier HTML qui contient des balises et des variables spécifiques à Django. Il permet de séparer la présentation des données de la logique de l'application. Les **templates** définissent la structure et le design des pages web, tout en permettant l'insertion dynamique de contenu provenant des views. Ils facilitent la création de pages web dynamiques en combinant du HTML statique avec des éléments de syntaxe Django pour l'affichage de données variables.. 69

topologie

En bioinformatique, la **topologie** fait référence à l'arrangement spatial et aux relations structurelles des éléments au sein d'une molécule biologique, notamment les protéines ou les acides nucléiques. Elle décrit la façon dont ces

éléments sont connectés et organisés dans l'espace tridimensionnel, ce qui est crucial pour comprendre leur fonction et leurs interactions. . 63, 65

tree-of-life

Le **tree of life**, ou arbre de vie, est un concept fondamental qui représente les relations évolutives entre toutes les formes de vie sur Terre. Il illustre l'idée que tous les organismes descendent d'un ancêtre commun, et les divergences entre les espèces sont souvent représentées sous forme de branches sur un arbre. En biologie, cet arbre est utilisé pour visualiser l'évolution et la diversité des espèces à travers le temps, en utilisant des données génétiques et morphologiques pour construire des phylogénies. . 29

virion

Un **virion** est la particule virale qui est généralement composée du matériel génétique (ADN ou ARN) protégé par une couche protéique (appelée la capside). Le **virion** sert, à la fois, à la protection du génome, et de véhicule pour délivrer ce matériel génétique viral dans les cellules hôtes.. 25, 51

zoonose

En infectiologie, une **zoonose** est une maladie infectieuse qui est naturellement transmissible à l'humain depuis l'animal. Exemple : la rage, la maladie de Lyme ou encore la malaria.. 25

Avant-propos

L'épidémie du COVID-19 (SOHRABI et al. 2020) a profondément bouleversé la recherche mondiale sur les maladies infectieuses, en particulier celle portant sur les virus émergents. Le coupable de cette crise sanitaire globale est le SARS-CoV-2, plus connu sous le nom de coronavirus 2 du syndrome respiratoire aigu sévère (CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES et al. 2020). Ce virus tristement célèbre est l'exemple même de ce que l'on appelle un virus émergent.

Un virus émergent est un virus apparaissant soudainement au sein d'une population qui n'a jamais été exposée à la souche en circulation. Cette émergence est généralement marquée par une augmentation soudaine de son incidence. Tracer l'historique de ces apparitions constitue un véritable défi scientifique, un travail à la fois essentiel et complexe (DOMINGO 2021).

Ces virus émergents sont fascinants par leur capacité à "sauter" d'hôte en hôte, s'adaptant à chaque nouvel environnement cellulaire. Leurs transitions évolutives sont souvent causées par des zoonoses ou par une série de mutations dans leur génome. Dans tous les cas, leur survie et leur adaptation reposent principalement sur deux composantes fonctionnelles : le complexe réplicatif (PATHANIA et al. 2022 ; KOKIC et al. 2021) et les protéines de surface du virion (LAN et al. 2020).

Une équipe de recherche spécialisée dans l'étude approfondie des complexes de réplication des virus à ARN, dirigée par le Dr Bruno Canard, joue un rôle central dans ces investigations. Connue sous le nom de "Réplicases Virales : Structure, Mécanisme et *Drug-Design*", cette équipe mène des travaux allant de la recherche fondamentale – à travers l'élucidation de structures protéiques virales et de leurs mécanismes fonctionnels – jusqu'à la recherche appliquée en *drug-design* (SHANNON, LE et al. 2020 ; BARDIOT et al. 2022 ; SHANNON, SELISKO et al. 2020). Leur champ d'action couvre de nombreux virus, tels que le virus de l'immunodéficience humaine (HIV) (DECOMBE et al. 2024), le virus de la dengue (FERACCI et al. 2023), le virus respiratoire syncytial

(RSV) (SUTTO-ORTIZ et al. 2021), ou encore le SARS-CoV-2 (SHANNON, FATTORINI et al. 2022 ; SHANNON, CHAZOT et al. 2024 ; GRANDA et al. 2023). Ce groupe de chercheurs se positionne ainsi comme un acteur majeur dans la lutte contre les menaces virales.

Au cours de ces trois années de doctorat, j'ai eu l'opportunité d'évoluer dans cet environnement enrichissant, entouré d'experts en biologie structurale et de virologues. Cet encadrement de haut niveau m'a permis de mûrir mes propres objectifs, qui se sont naturellement alignés sur la volonté de contribuer à leurs efforts. Avec le soutien de mon directeur de thèse, le Dr François FERRON, j'ai pu jouer le rôle de pionnier en bioinformatique au sein de cette équipe, afin de développer des solutions innovantes adaptées à leurs besoins.

Ma thèse s'inscrit donc dans cette démarche, avec pour sujet le "**développement d'outils bioinformatiques pour la caractérisation et l'analyse de nouveaux génomes viraux**".

Ce manuscrit se structure en quatre chapitres. Il débute par une introduction aux concepts clés, destinée à faciliter la compréhension des lecteurs. Le deuxième chapitre est consacré à VIMVER, un outil bioinformatique qui détecte les mutations ponctuelles sur une séquence donnée en comparaison avec le génome de référence du SARS-CoV-2. Le troisième chapitre présente un projet plus ambitieux, actuellement en cours de développement : VAZYMOLO-2. Cet outil est conçu pour identifier les domaines fonctionnels sur des séquences protéiques virales à partir de données expérimentales. Les objectifs sont doubles : proposer une plateforme d'analyse et d'annotation de protéines virales, ainsi qu'un support visuel des connaissances accumulées, afin de guider la stratégie de recherche de l'équipe. Enfin, le dernier chapitre met en perspective l'ensemble du projet, en abordant les défis rencontrés et les enseignements tirés de cette recherche.

1. Introduction

Dans ce premier chapitre, je présenterai d'abord le contexte biologique du projet de thèse, suivi des principaux enjeux qui l'entourent. Ensuite, j'aborderai les connaissances bioinformatiques nécessaires à une compréhension approfondie. Enfin, le chapitre se terminera par une mise en perspective des objectifs et de la vision de cette thèse.

Sommaire

1.1. La virologie 28
1.1.1. La classification des virus 29
1.1.2. Les virus émergents 34
1.2. La donnée biologique virale 37
1.2.1. La génomique 38
1.2.2. La séquence biologique 42
1.2.3. La structure d'une protéine 46
1.2.4. Le domaine protéique fonctionnel 51
1.3. Enjeux et règles du jeu 56
1.3.1. Le mystère du protéome sombre. 56
1.3.2. Bio-donnée en abondance 57
1.3.3. L'annotation 59
1.3.4. La règle du jeu : Homologie 60
1.4. Quelle est la boîte à outils du bioinformaticien aujourd'hui? 63
1.4.1. La phylogénétique 63
1.4.2. La recherche de séquence 66
1.5. Quelle forme peut prendre un outil bioinformatique? 72
1.6. Buts et Objectifs 74

1.1. La virologie

Un virus est un parasite obligatoire qui alterne entre deux phases distinctes dans son cycle de vie :

1. Le virion : particule virale mobile considérée comme non vivante, capable de survivre dans divers environnements (sols, plantes, bactéries) ;
2. La phase d'infection : période durant laquelle le virus détourne les processus cellulaires de l'hôte pour sa propre réplication. Durant cette phase, le virus peut être considéré comme vivant car il manifeste des caractéristiques du vivant.

Les virus sont omniprésents dans la biosphère et peuvent infecter tous les types d'organismes. Ils ne peuvent se reproduire qu'en utilisant la machinerie cellulaire de leur hôte. Cette dépendance absolue envers l'hôte les classe parmi les parasites obligatoires, incapables de se développer sur un milieu nutritif artificiel, comme l'on peut le faire avec des bactéries.

L'étude des virus

Depuis toujours, les virus se sont imposés comme des ennemis invisibles de l'humanité, évoluant silencieusement dans l'ombre, prêts à émerger à tout moment.

Le SRAS-CoV, agent du syndrome respiratoire aigu sévère (SRAS), a provoqué une épidémie mondiale entre 2003 et 2004, devenant ainsi le premier coronavirus émergent du 21e siècle. L'émergence du MERS-CoV, identifiée en Arabie Saoudite en 2012, continue à sévir. Plus récemment encore, le SARS-CoV-2 est à l'origine d'une pandémie responsable d'une crise sanitaire, sociale et économique majeure à l'échelle mondiale depuis 2019. Ces épidémies causées par des coronavirus émergents illustrent le potentiel des *Nidovirales* à franchir la barrière des espèces et à déclencher des crises sanitaires d'ampleur internationale. Elles mettent en lumière l'importance de la surveillance et de la recherche sur ces virus à fort potentiel zoonotique. D'autres *Nidovirales*, comme les *Arteriviridae*, ont également causé des épidémies significatives chez les animaux d'élevage, avec un impact économique considérable, bien qu'aucune transmission à l'homme n'ait été documentée à ce jour.

Dans ce chapitre et au long de la thèse, nous utiliserons les coronavirus, représentants emblématiques de l'ordre des *Nidovirales*. Leur capacité à déclencher des crises sanitaires majeures et leurs spécificités génomiques en fait un modèle d'étude

particulièrement pertinent pour illustrer les mécanismes des virus émergents.

1.1.1. La classification des virus

La phylogénie

Comment peut-on exprimer la diversité des virus appartenant à l'ordre des *Nidovirales*?
La phylogénie est un domaine d'étude qui apporte une réponse directe à cette question. Elle permet de retracer l'histoire évolutive des espèces en caractérisant les liens de parenté qui les unissent. Ce principe repose sur la théorie de l'évolution moderne, qui s'appuie sur les travaux fondateurs de Darwin, selon laquelle tous les êtres vivants partagent un ancêtre commun hypothétique. Cette théorie nous permet de visualiser l'arbre du vivant, communément appelé tree-of-life, qui illustre les relations évolutives entre les espèces.

Cependant, les virus ne figurent pas encore pleinement dans le tree-of-life, bien qu'ils ne soient pas étrangers à la notion de phylogénie. Contrairement aux organismes vivants, les virus ne peuvent être classés sur la base de caractéristiques fossiles ; seule leur diversité génomique permet de les classer. Dans ce chapitre, nous explorerons les différentes stratégies qui ont permis de progresser dans leur classification.

1.1.1.1. La taxonomie

Pour y parvenir, la classification des virus passe par une étape cruciale de structuration des groupes phylogénétiques : la taxonomie. Celle-ci repose sur l'utilisation de mots-clefs spécifiques, permettant d'établir une hiérarchie claire entre les différents groupes d'espèces.

Un clade est un terme générique désignant tout groupe d'espèces reliées par un ancêtre commun, comme illustré par le schéma 1.1.

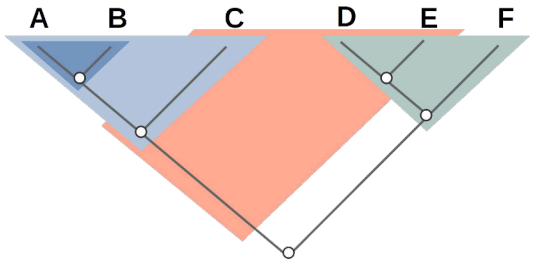
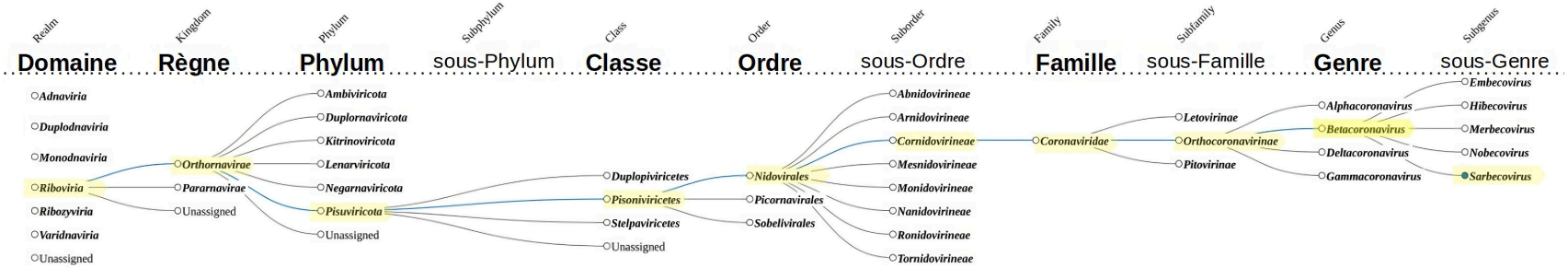


FIGURE 1.1. – Un clade : un groupe monophylétique
Les groupes AB (bleu foncé), ABC (bleu clair), DEF (vert clair) sont monophylétiques, donc aussi appelés clades. Au contraire le groupe CDE (rouge-orange) n'est pas monophylétique, car ses membres ne représentent pas tous les descendants du même ancêtre commun.

Un taxon est une entité conceptuelle regroupant tous les organismes vivants partageant des caractères taxonomiques ou diagnostiques bien définis. Ce terme est utilisé en classification biologique pour désigner des groupes d'organismes à différents niveaux hiérarchiques (figure 1.2), tels que l'espèce, le genre, la famille, ou encore l'ordre.

La taxonomie est l'art de nommer et de classer les taxons dans une hiérarchie composée de rangs et de sous-rangs. Contrairement à cette structure stable en apparence, la phylogénie est en constante évolution. De nouveaux virus sont régulièrement découverts, et il devient nécessaire de les classer correctement. Je vous propose donc de poursuivre cette discussion en explorant deux stratégies de classification qui ont tenté d'apporter de l'ordre dans ce domaine.

FIGURE 1.2. – Déroulement du "chemin taxonomique" menant au sous-genre *Sarbecovirus*, passant par l'Ordre des *Nidovirales*



Tout chemin commence par ***Viruses***, suivi du **Domaine**, **Règne**, ***Phylym*** (ou embranchement), **Classe**, **Famille**, **Genre** et **espèce**. Comme l'illustre la figure, certains chemins peuvent se complexifier en introduisant des sous-niveaux entre chaque rang, indiqués par le préfixe "sous-". Pour compléter l'exemple de la figure, le déroulement se poursuit sur le rang **espèce**, parmi eux : SRAS-CoV et SARS-CoV-2 (*Cladogramme raciné*).
La figure est tirée d'une fonctionnalité d' ICTV - version 2023(MSL-39).

1.1.1.2. La classification de Baltimore

La classification de Baltimore, mise au point par David Baltimore en 1971 (BALTIMORE 1971), est un système qui regroupe les virus selon la nature de leur génome et leur mécanisme de réplication. Elle comprend sept groupes distincts, numérotés de I à VII, comme illustré dans la figure 1.3.

L'ordre des *Nidovirales*, appartenant au groupe IV de la classification de Baltimore, regroupe des virus à ARN simple brin positif (+ssRNA). Leur génome peut directement servir de modèle pour la formation d'un brin complémentaire, appelé "négatif" ou "non-sens", par opposition au brin positif. Ce brin négatif va alors servir de matrice pour la réplication du génome viral ainsi que pour la production des ARN messagers (ARNm ou mRNA), qui seront pris en charge par les ribosomes de la cellule hôte afin de lancer la traduction des protéines virales (VRIES et al. 1997).

Ce système, bien que largement utilisé, présente des limites. Des avancées récentes dans la classification et l'étude de la phylogénie virale ont remis en question la définition de certains groupes (KOONIN, DOLJA et al. 2020). C'est pourquoi il est souvent utilisé en complément de la taxonomie virale officielle.

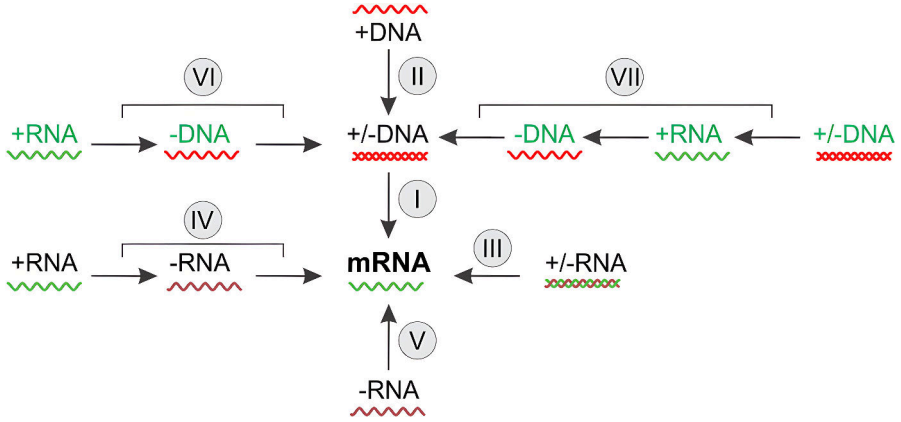


FIGURE 1.3. – La classification de Baltimore
Chaque groupe de virus correspondent à un nombre d'étapes qui séparent la génome à la production d'ARN messager (mRNA). (KOONIN, KRUPOVIC et al. 2021)

1.1.1.3. ICTV

Le Comité international de taxonomie des virus, ou *International Committee on Taxonomy of Viruses* (ICTV), est un regroupement de virologues chargé de la classifica-

tion et de la nomenclature des virus. Son rôle est d'établir et de maintenir la taxonomie officielle (LEFKOWITZ et al. 2018a). L'ICTV publie régulièrement des rapports sur l'état des phylogénies, comme celui portant sur les *Coronaviridae* (WOO et al. 2023). Ce comité applique la structure hiérarchique décrite précédemment pour classer les virus.
L'ICTV utilise une combinaison de critères pour la classification virale, incluant des caractéristiques phénotypiques, biologiques et moléculaires. Bien que l'ICTV intègre de plus en plus d'informations génomiques et phylogénétiques dans ses décisions taxonomiques (LEFKOWITZ et al. 2018b), il est important de noter que la classification de l'ICTV n'est pas exclusivement basée sur la phylogénie (KING et al. 2018 ; WALKER, SIDDELL, LEFKOWITZ, MUSHEGIAN, DEMPSEY et al. 2019).
Historiquement, l'ICTV s'est appuyé sur divers critères, et ce n'est que récemment qu'il a commencé à incorporer plus systématiquement des analyses phylogénétiques. Cette évolution vers une approche plus phylogénétique est en cours, mais la classification de l'ICTV reste un système qui prend en compte multiple facteurs, pas uniquement les séquences génomiques.

L'intégralité de la taxonomie officielle est disponible sur leur site, accessible via `https://ictv.global/`. Cette base de données (*ICTVdb*) tend vers l'exhaustivité du monde viral, mais elle présente certaines limites. En premier lieu, la taille colossale de l'*ICTVdb* implique une maintenance progressive, avec des mises à jour publiées sous forme de rapports. De plus, cette classification se concentre principalement sur les rangs supérieurs à l'espèce, offrant ainsi peu d'informations sur les sous-espèces, souches ou variants (SIDDELL et al. 2023). Enfin, la classification avance moins rapidement que les découvertes de nouvelles espèces, car le processus rigoureux d'acceptation nécessite d'éviter toute ambiguïté ou redondance entre les taxons déjà établis.

Ces limitations reflètent le défi de maintenir une taxonomie virale complète et à jour face à la diversité et à la rapidité des découvertes dans le domaine de la virologie. L'ICTV doit ainsi équilibrer la rigueur scientifique avec la nécessité de rester pertinent dans un domaine en constante évolution.

La classification des *Nidovirales*

En 2019, l'ICTV a publié un nouveau rapport annonçant un remaniement de la phylogénie des *Nidovirales* (KING et al. 2018). Cet ordre, qui était auparavant divisé

en quatre familles : *Coronaviridae, Arteriviridae, Roniviridae* et *Mesoviridae,* a été réorganisé. Depuis cette révision, les *Nidovirales* comprennent désormais huit sous-ordres, dont les anciennes familles, ont été subdivisées en un total de 14 familles. Le décompte des familles est illustré dans la figure 1.4. Le cladogramme raciné (figure 1.5), plus bas, présente l'intégralité de la taxonomie des *Nidovirales,* telle que définie par l'ICTV (KING et al. 2018 ; WALKER, SIDDELL, LEFKOWITZ, MUSHEGIAN, DEMPSEY et al. 2019 ; WOO et al. 2023 ; WALKER, SIDDELL, LEFKOWITZ, MUSHEGIAN, ADRIAENSSENS et al. 2022 ; BRINTON et al. 2021).

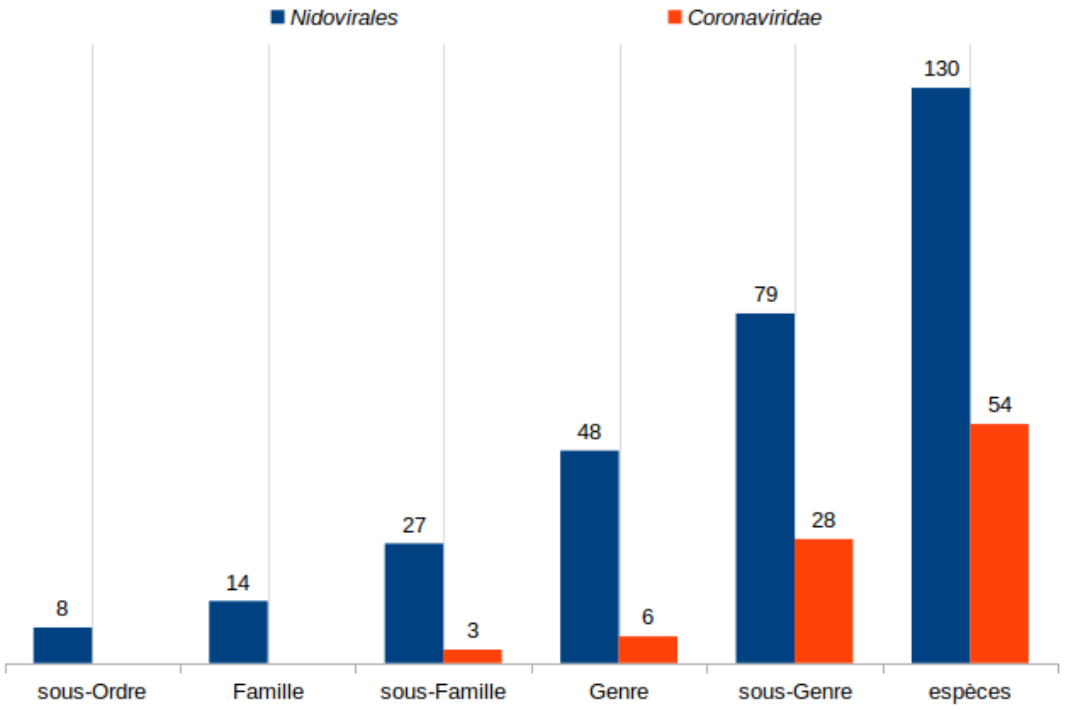


FIGURE 1.4. – La composition hiérarchique de l'Ordre des *Nidovirales* Chiffres officielles de la diversité en sein des *Nidovirales* (en bleu) et des *Coronaviridae* (en orange) - ICTV version 2024 (MSL 39) v.3.

1.1.2. Les virus émergents

Le principal obstacle à la taxonomie virale officielle est la constante découverte de virus émergents. Un virus émergent est un virus qui apparaît nouvellement dans une population ou dont l'incidence augmente rapidement. Il peut s'agir d'un virus totalement nouveau ou d'une nouvelle variante d'un virus existant. Son émergence est souvent liée à des changements écologiques ou démographiques et il représente une

menace pour la santé publique, pouvant causer des épidémies ou des pandémies. Ces virus sont souvent d'origine zoonotique (transmis des animaux à l'homme) (RAHMAN et al. 2020) et leur évolution est favorisée par des mutations (FLORES-VEGA et al. 2022 ; DIMONTE et al. 2024) ou recombinaisons génétiques (KUMAR et al. 2018). Leur surveillance est cruciale pour anticiper les risques pour la santé mondiale (HOWARD et al. 2012).

* * * *

Les virus émergents sont connus pour évoluer rapidement, accumulant des mutations qui leur permettent de s'adapter et de franchir les barrières d'espèces, posant un véritable casse-tête pour les virologistes. Leur impact mondial sur la santé publique en fait des cibles prioritaires pour la recherche médicale. Le développement de nouveaux vaccins et de nouvelles molécules thérapeutiques dépend avant tout d'une connaissance approfondie de ces virus. Pour y parvenir, nous appliquons une méthode bien établie : analyser ces virus en profondeur, comprendre leur fonctionnement, et les comparer à ceux que nous avons déjà étudiés et maîtrisés. Dans ces moments, qu'il s'agisse d'être penché sur une paillasse ou absorbé devant un écran, chaque instant est une quête de découverte.

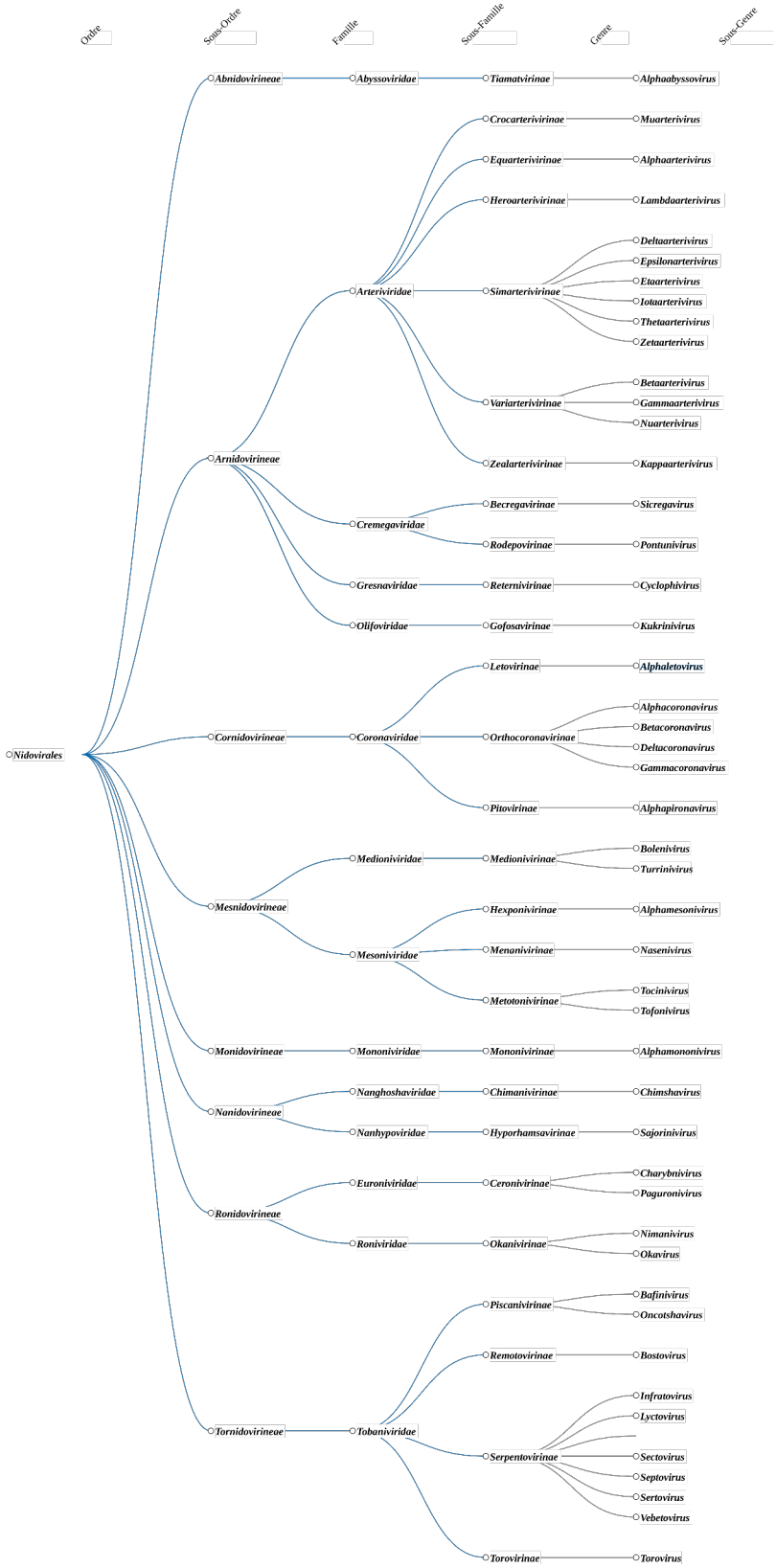


FIGURE 1.5. – Taxonomie complète des *Nidovirales*
La figure est tirée d'une fonctionnalité d' ICTV - version 2023(MSL-39).
(*Cladogramme raciné*)

1.2. La donnée biologique virale

Les "-omiques"

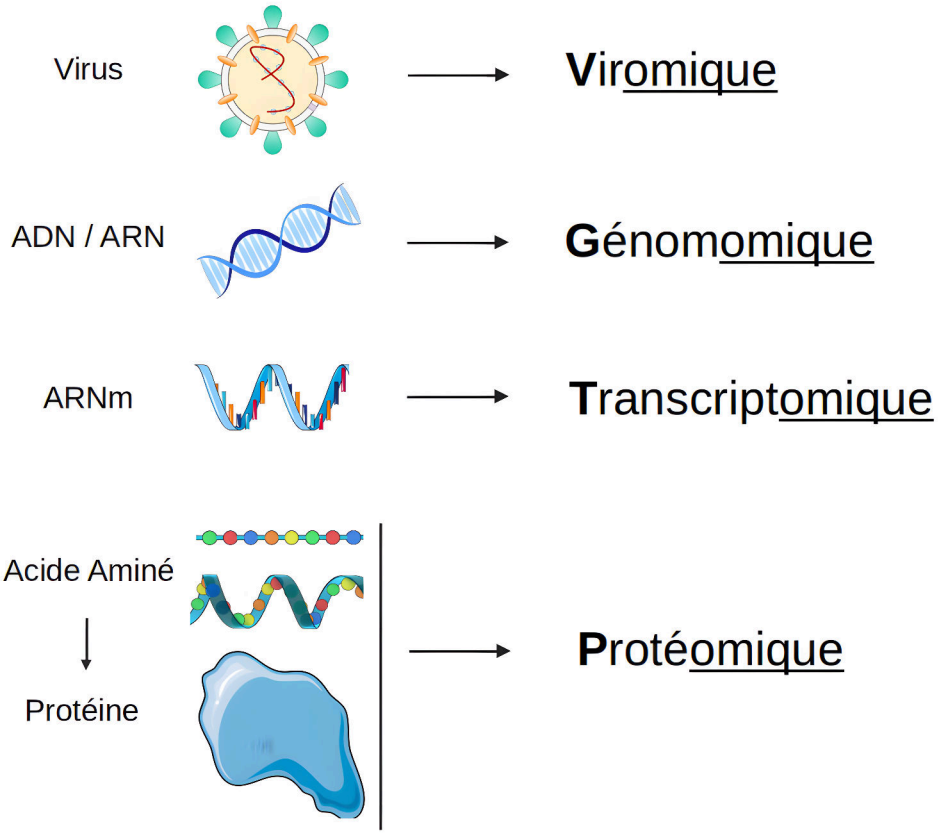


FIGURE 1.6. – Les "-omiques"
Quelques représentants des "-omiques" : disciplines bioinformatiques associées à leur type de donnée biologique de prédilection.

Qu'est-ce qu'une donnée biologique virale ? Il s'agit d'une information mesurable ou observable concernant un virus, sa structure, son fonctionnement, ou ses interactions avec un organisme hôte. Chaque type de donnée biologique est encadré par une discipline bioinformatique spécifique, désignée par le suffixe "-omique". Il existe ainsi plusieurs déclinaisons de ces disciplines, illustrées dans la figure 1.6.

Par exemple, le terme "génomique" est un mot-valise combinant "gène", "-ome" (du grec ancien "-$\omega\mu\alpha$", signifiant "ensemble" ou "totalité") , et "-ique", un suffixe indiquant la discipline. Les autres "-omiques" suivent une logique similaire.

Le virome fait référence à l'ensemble des données biologiques virales présentes dans un organisme ou un environnement donné, incluant tous les virus, qu'ils soient pathogènes ou non. La viromique est la discipline dédiée à l'étude de ces données.

Dans cette partie, nous adopterons une approche séquentielle, illustrée par la figure 1.7, pour passer en revue les différentes "-omiques" concernées par mon sujet de thèse.

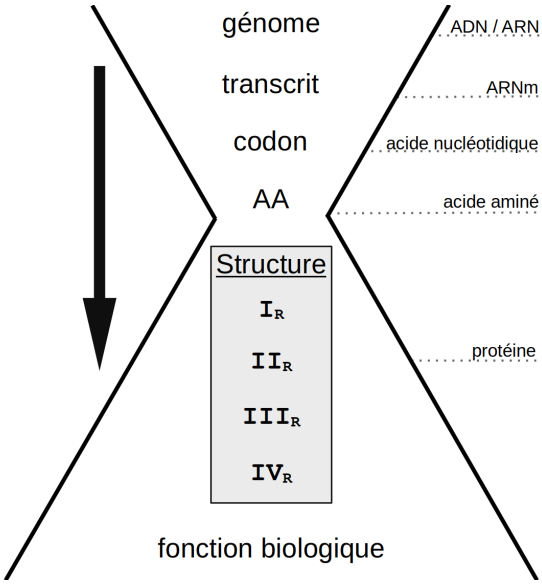


FIGURE 1.7. – Du génome à la fonction biologique
Représentation séquentielle des types de données biologiques, associées au type de molécules.

1.2.1. La génomique

Les virus sont des organismes parasitaires, cherchant à exploiter la machinerie de la cellule hôte pour subvenir à leurs propres besoins. Ce processus est appelé le "système viral". Le génome des *Nidovirales* est divisé en plusieurs ORF, open reading frame, (LAUBER et al. 2013), qui sont des cadres de lecture pris en charge par les ribosomes de l'hôte. Chaque ORF code pour une protéine (PASTERNAK et al. 2006), à l'exception du premier. Ce dernier produit une "polyprotéine", une longue chaîne peptidique qui est ensuite clivée par des protéases pour libérer les protéines virales fonctionnelles (ROE et al. 2021 ; FEHR et al. 2015 ; ZIEBUHR et al. 2000).

Une particularité de l'ORF1 est qu'il contient un mécanisme appelé "frameshift"

("-1PRF" dans la littérature), un glissement programmé du ribosome sur l'ARNm. Lorsque ce glissement n'a pas lieu, seule la première partie de l'ORF1 est exprimée, désignée ORF1a. En revanche, lorsque le frameshift se produit (voir figure 1.8), cela permet d'ignorer le codon-stop à la fin de l'ORF1a, et la deuxième partie, appelée ORF1b, est exprimée. Toutefois, l'ORF1b ne peut s'exprimer sans l'ORF1a, d'où l'utilisation du terme ORF1ab pour désigner l'ensemble complet (BEKAERT 2004 ; FIRTH 2014). Ce frameshift ribosomal a un rôle de régulation des protéines traduites par les ORF1ab, où les protéines de la partie 1b sont moins exprimées que les autres (GULYAEVA et al. 2021).

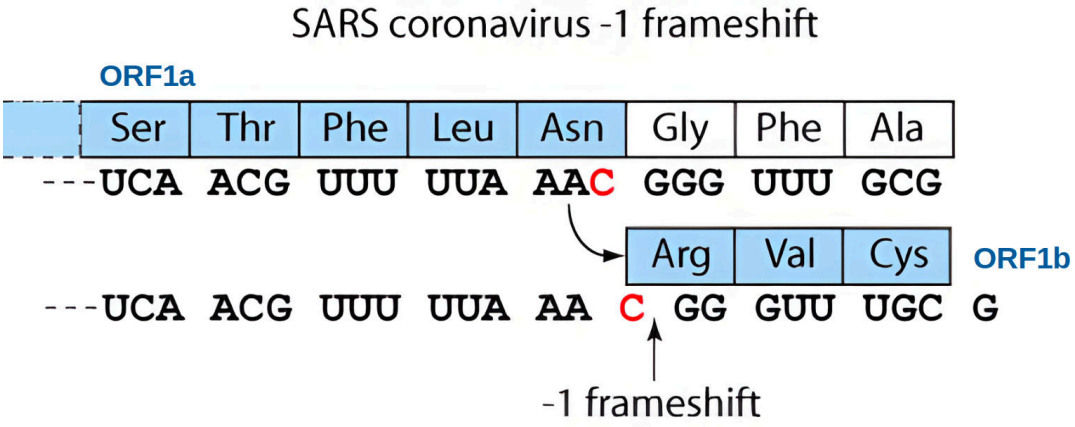


FIGURE 1.8. – Le *frameshift* ribosomal
Un évènement caractéristique pour tout nidovirus, entre l'ORF1a et l'ORF1b.
Figure tirée de viralzone.expasy.org.

Les protéines virales produites par les polyprotéines pp1a et pp1ab, correspondant respectivement à ORF1a et ORF1ab, sont appelées nsp (*non-structural proteins*) suivies d'un numéro. Ces nsp ont des fonctions principalement liées à la réplication et à la protection du génome viral. En revanche, les protéines produites par les autres ORF sont généralement des protéines de surface qui recouvrent le virion.

Les nidovirus, bien qu'ils appartiennent au même ordre, on peut cependant noter les différences dans leurs génomes, comme la taille et la structure : comme illustré dans les figures (1.9, 1.10, 1.11) suivantes. Avec des tailles allant de 20kb jusqu'à 50kb, ces génomes viraux cherchent un compromis entre la stabilité génétique, biophysique, biochimique et immunitaire dans leur cellule hôte (FERRON, SAMA et al. 2021). 0

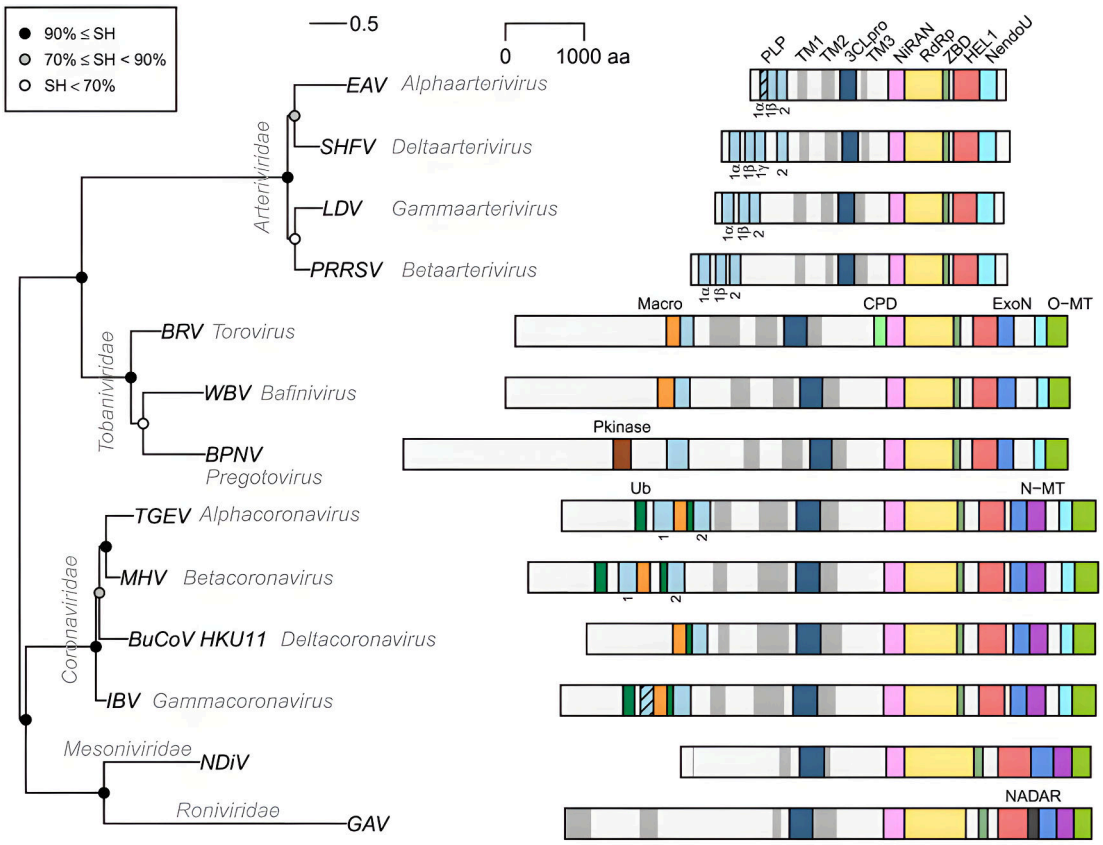


FIGURE 1.9. – La Phylogenie des principaux nidovirus et organisation de la pp1ab
Sont présents le principale famille des principaux sous-ordres : Coronaviridae, Mesoniviridae, Tobaniviridae, Arteriviridae ; les virus correspondants sont les suivants : virus de l'artérite équine (EAV), virus de la fièvre hémorragique simienne (SHFV), virus élévateur de la lactate déshydrogénase (LDV) virus du syndrome dysgénésique et respiratoire porcin (PRRSV), virus de Breda (BRV), virus de la brème blanche (WBV), nidovirus du python royal (BPNV), virus de la gastro-entérite transmissible porcine (TGEV), virus de l'hépatite murine (MHV), coronavirus HKU11 du bulbul (BuCoV_HKU11), virus de la bronchite infectieuse aviaire (IBV), virus de Nam Dinh (NDiV), virus associé aux branchies (GAV). Cette phylogénie est construite, (phylogramme non raciné) , avec Viralis (GORBALENYA et al. 2010a), sur le domaine polymerase (RdRp), le support de branche est calculé avec une méthode *SH-like* (Maximum de vraisemblance), utilisant *IQ-TREE* (NGUYEN et al. 2015).
Figure tirée de l'article suivant (GULYAEVA et al. 2021).

Les **ARN sous-génomiques** sont des molécules d'ARN plus courtes que le génome viral complet au cours de leur cycle de réplication.

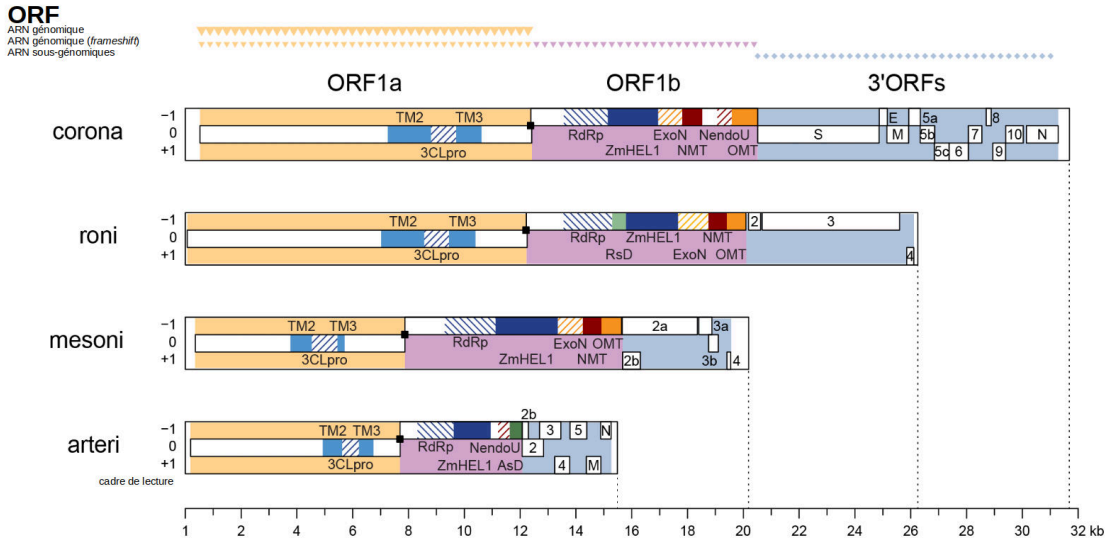


FIGURE 1.10. – Les génotypes des principaux nidovirus
Les régions codantes sont respectivement encadrées, l'ORF1a en orange, l'ORF1b en violet, les ORF suivants en bleu. Le **frameshift ribosomal** est indiqué par un carré noir. Certains domaines protéiques sont mise-en-évidence par des rectangles de couleurs, parmi eux : 3CLpro (nsp5), RdRp (nsp12), ZmHel1 (nsp13), ExoN (nsp14), NMT (nsp14), NendoU (nsp15) et OMT (nsp16). Les superpositions des cadres de lecture sont indiquées par la notation -1,0,+1.
Figure tirée et traduite du l'article suivant (LAUBER et al. 2013).

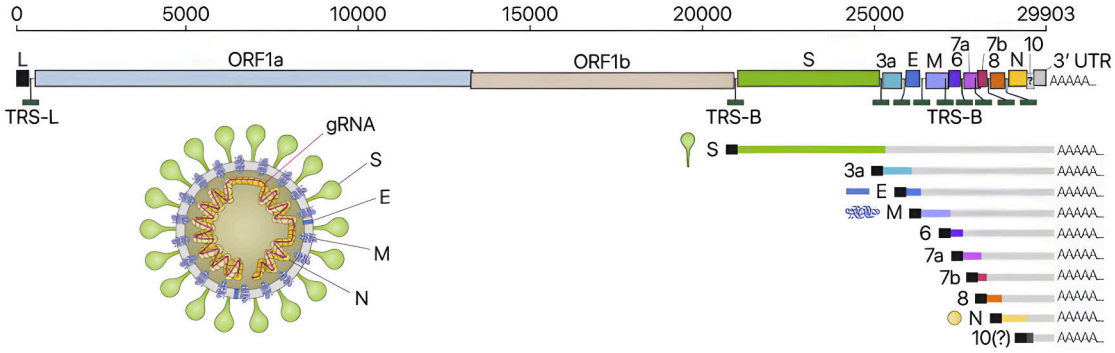


FIGURE 1.11. – Le transcriptome du SARS-CoV-2
La longueur entière du génome est de 29903 nucléotides (ou 29.9kb). Il y a plusieurs cadres de lecture (ORF), incluant l'ORF1a, l'ORF1ab et les ORF sous-génomiques. "gRNA" renvoie à l'ARN génomique ; "TRS" aux séquences régulatrices de la transcription.
Figure tirée du l'article suivant (GORDON et al. 2020).

* * * *

Cette organisation génomique complexe, avec le mécanisme de frameshift, permet aux *Nidovirales* de contrôler finement l'expression de leurs protéines et d'optimiser l'utilisation de leur matériel génétique limité. C'est un excellent exemple de l'adaptation évolutive des virus pour maximiser leur efficacité réplicative dans les cellules hôtes.

1.2.2. La séquence biologique

La donnée biologique virale doit être structurée pour devenir utile à la communauté scientifique. Parmi ces données, la séquence biologique est un type fondamental. Elle décrit l'enchaînement linéaire des éléments constitutifs d'une macromolécule biologique, généralement un acide nucléique (ADN ou ARN) pour le génome, ou un acide aminé pour les protéines.

— Pour le génome : il s'agit de la succession des nucléotides (A, T, G, C pour l'ADN ; A, C, G, U pour l'ARN). Cette séquence représente l'information génétique brute. Par convention, que le génome soit de l'ADN ou de l'ARN, la séquence est toujours représentée en ADN.

— Pour les protéines : il s'agit de la séquence d'acides aminés. Celle-ci peut également être obtenue par traduction in silico de la séquence génomique, en respectant le code génétique universel.

Une séquence est généralement représentée sous forme d'une chaîne de caractères (ou *string* pour les bioinformaticiens). On parle aussi de structure primaire (I_R). Ces séquences sont stockées dans des fichiers informatiques au format texte, dont les standards les plus courants sont les formats FASTA et GENBANK.

TABLEAU 1.1. – Les 20 acides aminés naturels

Nom complet	Code		Radical			
	3 lettres	1 lettre	taille	polaire	charge	spécificité
Glycine	Gly	G	petit			
Alanine	Ala	A	petit			
Valine	Val	V	petit			
Leucine	Leu	L				
Isoleucine	Ile	I				
Proline	Pro	P	petit			coudé
Methionine	Met	M	long			
Serine	Ser	S	petit	X		
Thréonine	Thr	T		X		
Cystéine	Cys	C	petit	X		souffre
Tyrosine	Tyr	Y		X		
Asparagine	Asn	N		X		
Glutamine	Gln	Q		X		
Aspartate	Asp	D		X	-	
Glutamate	Glu	E		X	-	
Lysine	Lys	K	long	X	+	
Arginine	Arg	R	long	X	+	
Tryptophane	Trp	W	aromatique			
Phenyalanine	Phe	F	aromatique			
Histidine	His	H	aromatique	X	+	

Chaque acide aminé est représenté sous son code à 3 et 1 lettre, puis décrit par les propriétés de la chaîne radicale (ce qui différencie les aa les un aux autres).

1.2.2.1. Le format GENBANK

GenBank a été officiellement fondée en 1982 par le NIH aux États-Unis comme banque de gènes publique. Entre 1989 et 1992, sa gestion a été transférée au NCBI (National Center for Biotechnology Information) (SMITH 2013). Son développement a été accéléré par l'adoption de systèmes de soumission informatisés à la fin des années 1980, ce qui a conduit à la standardisation du format de fichier.

Le format de fichier GENBANK (extension ".gb" ou ".genbank") est en effet devenu un standard incontournable en bioinformatique pour stocker les séquences génomiques et protéiques ainsi que leurs annotations, que l'on appellera "métadonnées". Chaque information est associée à un mot-clef, en suivant le principe du couple "clef valeur". La structure du fichier est composée de quatre sections. Voici un aperçu de la structure d'un fichier GENBANK, qui illustre bien l'idée que ce format constitue un compromis entre la lisibilité humaine et l'utilisabilité informatique :

— La section d'en-tête (ou HEADER) : Elle fournit des informations sur la séquence, telles que le nom, l'origine, les dates de soumission, et les références à des articles scientifiques.
  LOCUS : position de la séquence, sa longueur, type de macromolécule, date de la dernière modification
  DEFINITION : Brève description de la séquence
  ACCESSION : Numéro d'accession unique assigné par la base de données
  VERSION : Numéro d'accession suivi d'un numéro de la version
  KEYWORDS : Mots-clés pertinents (peut être vide)
  SOURCE : Organisme source de la séquence
  ORGANISM : Classification taxonomique de l'organisme source (sous forme de chemin taxonomique)
— La section des références (ou REFERENCE) est une liste dont chaque élément est composé du trio suivant :
  AUTHORS : Auteurs de la publication associée
  TITLE : Titre de la publication
  JOURNAL : Détails du journal où l'étude a été publiée
— La section des composants caractéristiques (ou FEATURES) : Elle décrit en détail chaque élément annoté (*gene*, *CDS*, etc.), souvent avec des positions spécifiques dans la séquence. Les qualificateurs (*/gene*, */product*, etc.) fournissent des infor-

mations supplémentaires.
— La section de la séquence (ou ORIGIN) : Enfin, la séquence d'ADN ou de protéine en elle-même est présentée sous forme de lettres, groupées par blocs.
— La fin du fichier est toujours marqué par "//".

Ce format texte est à la fois suffisamment simple pour être interprété par un humain, et structuré de manière cohérente pour les logiciels bioinformatiques. Il est ainsi devenu un standard incontournable, avec une grande flexibilité dans la représentation des informations génétiques et des annotations biologiques associées.

Le format GBSON est une alternative qui offre une meilleure structure et facilite de parsing.

1.2.2.2. Le format FASTA

Le format FASTA (extension ".fa", ".fasta" ou ".fst") est un format plus simple et très largement utilisé en bioinformatique pour stocker les séquences nucléotidiques ou protéiques. Il est composé de deux parties :

1. L'en-tête (header) : La première ligne commence par le symbole '>' suivi d'un identifiant et, éventuellement, d'une description ou d'autres informations. Cette ligne est unique pour chaque séquence.
2. La séquence : La deuxième ligne contient la séquence (nucléotidique ou protéique). Si la séquence est longue, elle est généralement divisée en plusieurs lignes de longueur fixe (souvent 60 ou 80 caractères).

Le format MULTIFASTA permet de stocker plusieurs séquences dans un seul fichier, en répétant ce schéma pour chaque séquence, séparées par un retour à la ligne. Cette simplicité et compacité font du FASTA un choix idéal lorsque les annotations détaillées ne sont pas nécessaires, et que l'on souhaite simplement stocker ou transférer des séquences.

* * * *

Le choix du format dépend des besoins spécifiques en termes de stockage et d'analyse des données de séquences.

1.2.3. La structure d'une protéine

Les protéines, véritables piliers du monde vivant, remplissent des fonctions aussi bien structurales que fonctionnelles. Leur description repose sur quatre niveaux ou dimensions de structure. Ces quatre niveaux structuraux sont étroitement interconnectés et essentiels à la fonction biologique d'une protéine. La structure primaire influence les niveaux supérieurs de repliement, lesquels définissent la forme tridimensionnelle nécessaire à l'activité spécifique de la protéine, qu'elle soit structurelle ou enzymatique.

La **structure primaire** : C'est la séquence linéaire des acides aminés (aa) qui compose la chaîne polypeptidique. Déterminée par le code génétique, elle représente le niveau fondamental de la structure protéique.

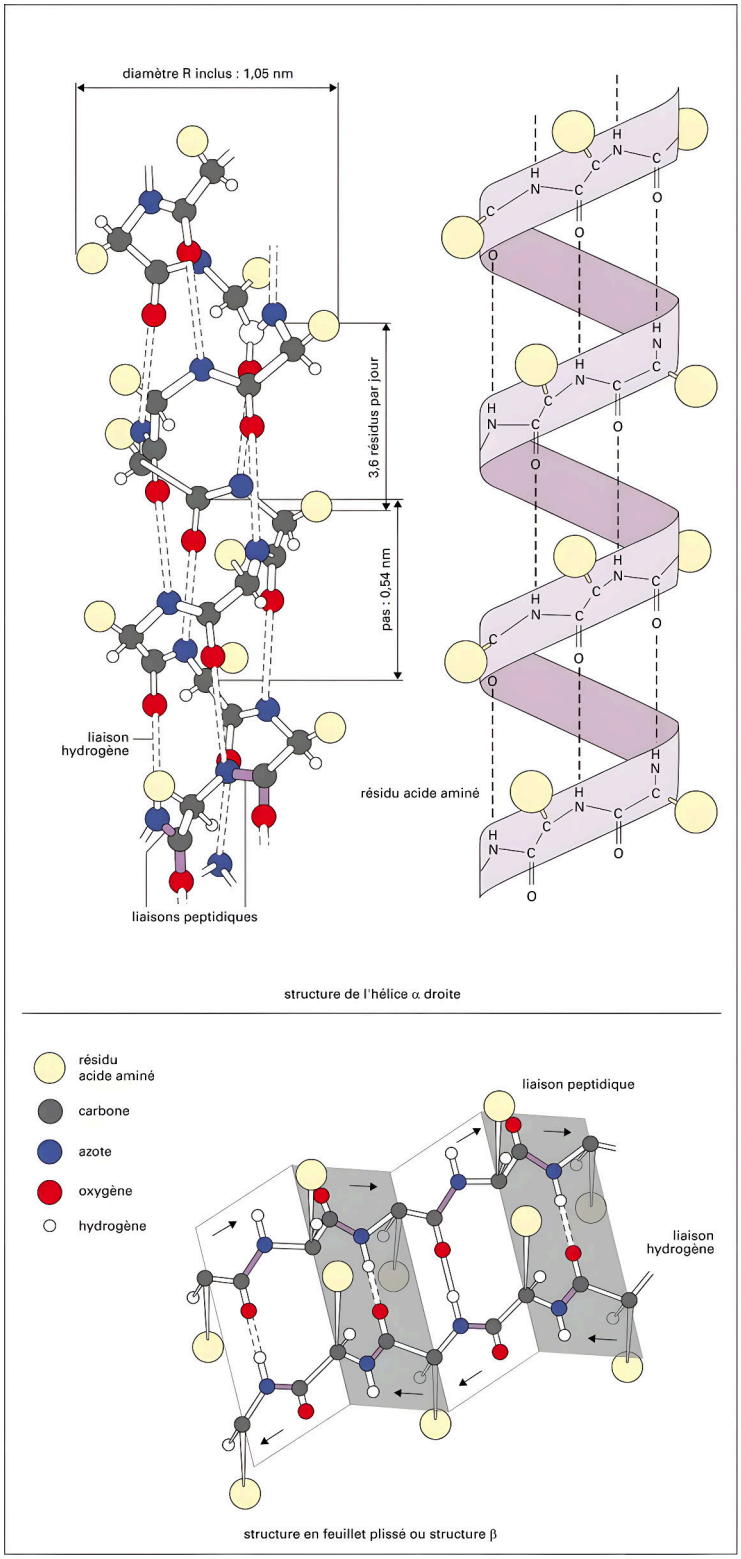


FIGURE 1.12. – Les structures secondaires : hélice α et feuillet β
Source : (ENCYCLOPÆDIA UNIVERSALIS 2024)

1.2.3.1. La structure secondaire

La structure secondaire des protéines correspond aux motifs réguliers qui résultent du repliement local de la chaîne polypeptidique, principalement sous la forme d'hélices α et les brins β. Ces structures sont stabilisées par des liaisons hydrogènes entre les groupements du squelette peptidique.

Les **hélices** α sont caractérisées par un enroulement hélicoïdal de la chaîne polypeptidique autour d'un axe central, où des liaisons hydrogènes entre le groupement C=O d'un résidu i et le groupement N-H du résidu i+4 (voir figure 1.12).

Les **brins** β peuvent être parallèles ou antiparallèles. Dans le cas parallèle, les brins adjacents sont orientés dans le même sens, tandis que dans le cas antiparallèle, ils sont orientés en sens opposés. Les liaisons hydrogènes se forment entre les groupements C=O et N-H de brins adjacents (voir figure 1.12). L'assemblage de 3 brins et plus forment un feuillet.

Outre ces motifs principaux, d'autres éléments, comme les **coudes** β, jouent un rôle dans la structure secondaire. Ces coudes relient différents éléments de structure et sont souvent induits par la présence de résidus proline, qui agissent comme des "casseurs d'hélices" en imposant des contraintes conformationnelles. Les **boucles** ou *coils*, quant à elles, offrent plus de flexibilité. Elles se trouvent souvent à la surface des protéines, interagissant avec le solvant et exposant leurs résidus polaires, tels que la glycine ou la proline, qui facilitent ces interactions.

Les cystéines stabilisent la structure des protéines en formant des **ponts disulfures** (voir figure 1.13). Ces liaisons covalentes unissent deux résidus cystéines distants, renforçant la structure tridimensionnelle et la résistance de la protéine, surtout dans les environnements extracellulaires.

Source : (*pont disulfure — Wiktionnaire, le dictionnaire libre* 2024)

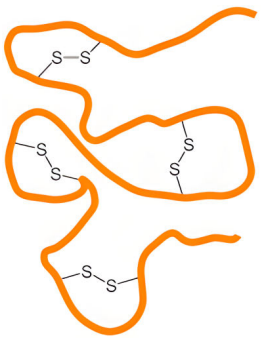


FIGURE 1.13. – Les structures secondaires : pont disulfure

Ces structures secondaires sont thermodynamiquement favorisées, car elles permettent de réduire l'énergie potentielle en formant des liaisons hydrogènes. En exposant les groupes polaires vers le solvant, elles contribuent à la solubilité de la protéine, pour tendre vers une forme globulaire. En revanche, les chaînes latérales non polaires favorisent des interactions hydrophobes, contribuant à l'agencement global de la protéine.

1.2.3.2. La structure tertiaire

Le processus de repliement au troisième niveau aboutit à la structure tertiaire d'une protéine, où les éléments de structure secondaire se combinent pour former des régions tridimensionnelles spécifiques :

- région globulaire : Ces régions organisent les résidus polaires à la surface pour interagir avec l'environnement cellulaire (solvant ou partenaires). Elles abritent souvent l'activité catalytique des enzymes. Le cœur hydrophobe stabilise la structure tandis que les résidus en surface forment des cavités qui permettent d'accéder au site actif.
- régions de basse complexité :
  - région transmembranaire : Séquences riches en résidus non polaires permettant à une protéine de s'ancrer dans la membrane. Elles sont surtout impliquées dans le positionnement plutôt que dans l'activité enzymatique.
  - région intrinsèquement désordonnée (UVERSKY 2013) : Séquences sans structure tridimensionnelle définie, flexibles et souvent riches en résidus polaires ou chargés. Elles jouent des rôles essentiels comme l'interaction avec d'autres protéines, la facilitation de la mobilité, et la régulation de l'activité protéique, ou piéger l'immunité innée cellulaire de l'hôte (ZHANG et al. 2021). Leur flexibilité leur permet d'adopter différentes conformations pour interagir avec divers partenaires fonctionnels

La structure tertiaire est stabilisée par des ponts disulfures, des interactions hydrophobes, des liaisons hydrogènes, et des interactions ioniques.

Crée en 1971 par le *Protein Data Bank* (PDB), le format de fichier PDB (extension ".pdb") est un format texte, lisible par l'humain, couramment utilisé pour stocker la structure tridimensionnelle des protéines et autres macromolécules. Ce format contient notamment :

— Les coordonnées atomiques de chaque atome dans l'espace.
— La séquence protéique, qui peut différer légèrement de la séquence de référence.
— La position des éléments de structure secondaire (hélices α, brins β, etc.).
— Des informations sur les conditions expérimentales (comme la cristallographie aux rayons X, la cryo-EM, ou la RMN).
— Des métadonnées (auteurs, date de dépôt, références)

Le format PDB est essentiel en biologie structurale, fournissant une représentation standardisée et détaillée des macromolécules pour une analyse et une visualisation précises de leurs structures tridimensionnelles.

Le format MMCIF a été introduit pour pallier les limitations du format PDB, notamment sa restriction à 80 caractères par ligne. Développé dans les années 1990 à partir du format CIF utilisé en cristallographie, mmCIF offre une plus grande flexibilité, permettant de gérer un nombre illimité d'atomes, de résidus, et de chaînes, ainsi qu'une meilleure organisation des données. En 2014, il devient le format standard de la Protein Data Bank (PDB), avec une adoption complète en 2019 par la *worldwide* PDB (wwPDB). Bien que le format PDB soit encore supporté, certaines structures complexes ne sont disponibles qu'en mmCIF. Des outils de conversion et de mise à jour des logiciels ont facilité cette transition.

1.2.3.3. La structure quaternaire

Les protéines peuvent être décrites à différents niveaux hiérarchiques, mais la règle traditionnelle est que le repliement dans l'espace d'une séquence en acides aminés produit une structure qui porte la fonction enzymatique. Cette relation entre la structure et la fonction est souvent représentée par l'équation suivante :

protéine - structure - fonction

Cependant, il existe des exceptions à ce paradigme. Certaines protéines ne développent leur fonction que lorsqu'elles sont en complexe avec d'autres molécules (protéiques ou non). Ces complexes sont considérés comme des structures quaternaires, qui décrit le lien dynamique entre les sous-unités protéiques et la fonction biologique réelle. Il s'agit de la quatrième dimension d'une protéine.

L'exemple des "super-structures" illustre cette notion. Ainsi, l'actine forme un réseau au sein du cytosquelette cellulaire ; tandis que la polymerase (ou RdRp) des nidovirus

est une protéine qui forme le cœur du complexe de réplication (ou RTC), nécessaire pour assurer la réplication du génome viral (MALONE et al. 2022).

1.2.4. Le domaine protéique fonctionnel

Un domaine est de conception modulaire, c'est-à-dire qu'une protéine peut porter plusieurs fonctions individualisées en domaines distincts. Un domaine protéique fonctionnel est une fonction que l'on annote à une protéine. Il peut être décrit selon 4 approches : par sa fonction, sa structure, son approche analytique et sa phylogénie.

L'approche fonctionnel

On peut distinguer plusieurs types de domaines fonctionnels dans les protéines virales, notamment :

— Les domaines structuraux ont un rôle de soutien et de maintien de la structure globale du virion.
Exemples : E (Enveloppe), N (Nucléocapside), M (Membrane) des coronavirus.

— Les domaines de surface ont un rôle crucial dans l'entrée du virus dans les cellules hôtes. Leur fonction est d'interagir avec les récepteurs cellulaires (principe "clef-serrure").
Exemple : Domaine S (communément appelé "Spike") des coronavirus.

— Les domaines enzymatiques liés à la réplication du génome viral sont essentiels pour la multiplication du virus.
Exemple : Domaine ARN polymérase ARN dépendente ("RdRp" pour *RNA-dependent RNA polymerase*).

— D'autres domaines fonctionnels participent au système viral : les protéases virales pour le clivage des polyprotéines virales (FEHR et al. 2015) ; l'hélicase pour le déroulement et l'adressage de l'ARN viral à la RdRp (MALONE et al. 2022) ; les méthyltransférases pour la modification des coiffes des ARN viraux pour mimer les ARNm de l'hôte (FERRON, DEBAT et al. 2020).

Il est important de noter que certaines protéines virales peuvent combiner plusieurs de ces fonctions dans différents domaines. Par exemple, la protéine Spike des coronavirus a à la fois un rôle structural (en tant que composant du virion) et un rôle fonctionnel crucial dans l'entrée du virus dans la cellule hôte (WALLS et al. 2020).

La classification des enzymes est un moyen de faciliter la communication scientifique et de permettre une classification systématique. Le code EC (*Enzyme Commission*) est un système standardisé de nomenclature qui décrit les réactions chimiques catalysées par chaque enzyme, mais il ne reflète pas la structure ou l'évolution de celle-ci (McDonald et al. 2023). Bien que le code EC soit utile pour classer les enzymes, il peut avoir des limitations pour représenter toutes les activités enzymatiques, notamment celles des virus. Les virus peuvent présenter des enzymes avec des activités uniques ou peu communes qui ne correspondent pas facilement aux catégories existantes dans le système EC. De plus, certaines protéines virales peuvent exercer plusieurs fonctions enzymatiques distinctes (Khan et al. 2021), ce qui les rend difficiles à classer dans le système EC.

L'approche structurale

Un domaine protéique est un ensemble particulier de régions d'une protéine qui adopte une forme globulaire indépendamment du reste de sa séquence. Cette structure peut être simplifiée pour faciliter sa classification et on parle alors de repliement ou de "fold". La **topographie** est une projection en deux dimensions de l'organisation spatiale des éléments secondaires dans une protéine. Il est important de noter que le même fold peut se retrouver dans différents domaines ne supportant pas les mêmes fonctions. En effet, une fonction peut être exprimée par plusieurs folds différentes.

Le **squelette** (ou *core*) d'une protéine est une représentation simplifiée de l'espace qui conserve les éléments de structure importants et conservés dans l'évolution.

La résolution expérimentale de la structure des protéines sert de référence, et chaque méthode apporte ses propres avantages et limitations :

— La cristallographie aux rayons X fournit une structure tridimensionnelle à haute résolution, idéale pour les protéines globulaires. Elle nécessite d'obtenir des cristaux de protéines, une manipe courante en biologie moléculaire pratique dans notre laboratoire.

— La Cryo-microscopie électronique (cryo-EM) permet d'observer les protéines dans un état proche de leur environnement naturel, figées à l'échelle cellulaire. Elle est particulièrement utile pour les grands complexes protéiques et les interactions protéine-protéine. Ces dernières années, cette méthode a connu des améliorations significatives en résolution.

— La Résonance Magnétique Nucléaire (RMN) quant à elle, renseigne plutôt des informations sur la structure et la dynamique des protéines en solution. Elle est très utile pour caractériser les protéines et régions intrinsèquement désordonnées. Elle est aussi appliquée dans l'étude les interactions protéine-ligand. Elle est limitée par la taille des protéines qu'elle peut analyser efficacement.

Malgré ces limitations, ces méthodes expérimentales restent supérieures aux prédicteurs in silico en termes de fiabilité et de précision. Les approches computationnelles, bien qu'en constante amélioration, sont généralement utilisées en complément des données expérimentales plutôt que comme substitut pour formuler des hypothèses.

En utilisant ces notions, il est possible de retracer l'histoire évolutive des structures en les classant par famille, c'est-à-dire en regroupant les protéines qui partagent une topographie ou un *core* similaire, où le squelette représente le consensus des parties communes.

L'approche phylogénique

Un domaine protéique fonctionnel, d'un point de vue phylogénétique, est une unité structurale et fonctionnelle conservée au cours de l'évolution au sein d'un groupe phylogénétique.

Le domaine persiste à travers le temps évolutif, témoignant de son importance fonctionnelle. Au sein d'un même clade viral, le domaine existe en plusieurs versions, résultat de l'accumulation indépendante de mutations dans le génome de chaque virus. Malgré la diversité génétique, ces mutations ne perturbent pas le repliement global de la protéine, maintenant ainsi la structure tridimensionnelle essentielle du domaine. La fonction du domaine est conservée malgré les variations dans la séquence, indiquant une forte pression de sélection sur les résidus critiques pour l'activité. Les mutations bénéfiques ou neutres sont maintenues, tandis que les mutations délétères sont éliminées, reflétant l'équilibre entre la diversification génétique et la conservation fonctionnelle (O'TOOLE et al. 2022).

L'étude phylogénétique de ces domaines permet d'établir des liens évolutifs entre différentes protéines et organismes.

L'approche analytique

Un domaine se décrit au-delà de sa fonction, sa structure ou de son histoire évolutive.

Un profile est un regroupement de séquences protéiques, aux quelles ont été annotées un même domaine, appartenant au même clade ou famille virale. Les séquences ont a priori un repliement similaire, donc comparables. Elles peuvent être représentées sous la forme d'un multifasta.

Un motif structural est un marqueur constitué de résidus conservés proches dans l'espace, qui se représente à la fois en 1D (sur la séquence) comme en 3D (voir figure 1.14). Les motifs peuvent se déduire de l'analyse d'un profile de séquences, ensemble ils constituent une véritable "signature" servant d'identification plus précise qu'un code EC, dans le cas des virus.

* * * *

Le domaine est un objet puissant pour décrire la transdisciplinarité d'une protéine dans son ensemble. Il est le point de départ du modèle bioinformatique que j'ai développé durant mes travaux.

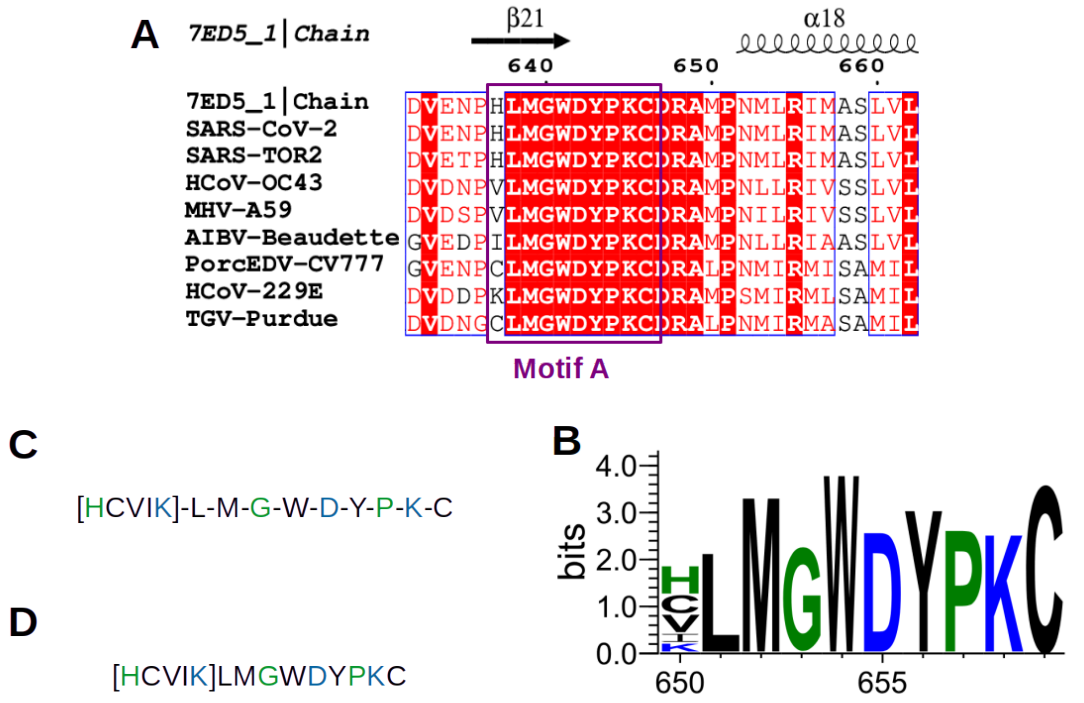


FIGURE 1.14. – Qu'est-ce qu'un motif?
(**A**) Le motif A du domaine RdRp des coronavirus, mis en évidence par ces résidus conservés, avec ESPript (ROBERT et al. 2014). Les séquences sont alignées avec Muscle (EDGAR 2004). (**B**) Le logo de séquence est la représentation la plus courant d'un motif, avec WebLogo (CROOKS et al. 2004). (**C**) Il s'agit du pattern de la séquence du motif, que je nomme "la formule". (**D**) Ici, j'exprime cette "formule" convertie en pattern pour un usage avec regex.
Structure PDB de la polymérase du SARS-CoV-2 (7ED5), coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2), coronavirus du syndrome respiratoire aigu sévère isolat Tor2 (SARS-TOR2), coronavirus humain isolat OC43 (HCoV-OC43), virus de l'hépatite murine (MHV-A59), virus de la bronchite infectieuse aviaire (AIBV-Beaudette), virus de la gastro-entérite transmissible porcine (PorcEDV-CV777), coronavirus humain isolat 229e (HCoV-229E), virus de la gastro-entérite transmissible isolat Purdue P115 (TGV-Purdue).

1.3. Enjeux et règles du jeu

1.3.1. Le mystère du protéome sombre.

"Je sais que je ne sais rien" - Socrate

La virologie est confrontée à un défi majeur : comprendre la fonction et la structure tridimensionnelle du protéome "sombre" ou "profond" (dark proteome), qui représente plus de 50% des protéines virales. Bien que les génomes viraux soient séquencés, une grande partie des protéines reste encore mal caractérisée. Cette compréhension est pourtant cruciale pour décrypter la biologie virale et développer des stratégies antivirales efficaces.

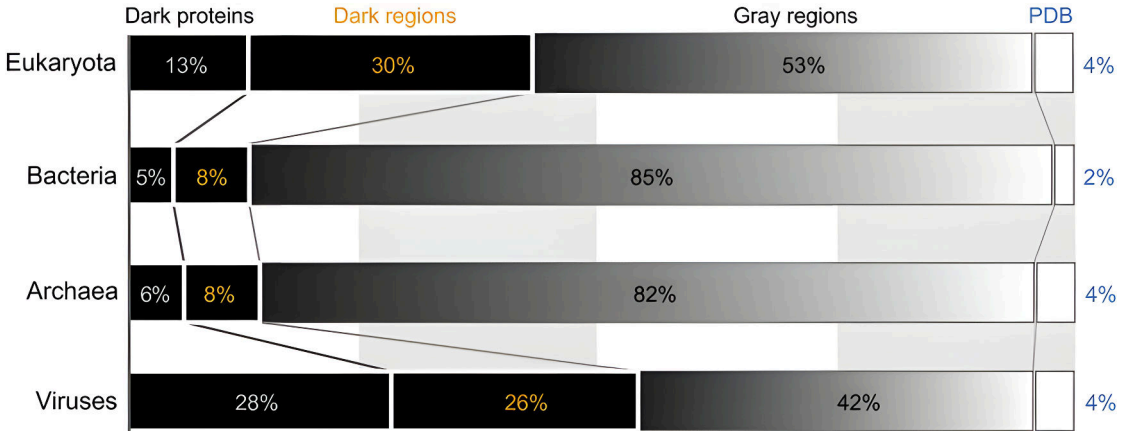


FIGURE 1.15. – Le "*dark proteome.*"
Figure tirée du l'article suivant (PERDIGÃO, HEINRICH et al. 2015).

Le dark proteome est un terrain inexploré qui recèle un potentiel immense pour la découverte de nouvelles fonctions biologiques. Il représente non seulement un défi pour les chercheurs, mais aussi une opportunité de faire des avancées fondamentales. Dans une étude récente, un échantillonnage de 67715 de nouvellement prédites protéines sur un ensemble de 4463 d'espèces de virus eucaryotes, on observe que 62% des protéines virales sont structurellement distinctes et présentent peu d'homologie dans la base de données de AlphaFold (NOMBURG et al. 2024; VARADI et al. 2022; INIGO BARRIO-HERNANDEZ et al. 2023).

Les études sur le dark proteome ont mis en évidence l'importance des régions intrinsèquement désordonnées (IDR) (PERDIGÃO, HEINRICH et al. 2015). Ces segments protéiques n'adoptent pas de structure tridimensionnelle définie et ne forment pas de

repliement globulaire stable, offrant ainsi une grande flexibilité structurelle.

Ces régions désordonnées jouent souvent un rôle essentiel dans les interactions protéiques complexes, s'adaptant à différents environnements et abritant fréquemment des sites de modifications post-traductionnelles. Cependant, leur flexibilité rend leur caractérisation structurelle plus difficile (BHOWMICK et al. 2016).

Contrairement aux attentes, une partie importante des protéines dites "sombres" présente une structure bien ordonnée et globulaire. Cela montre que notre compréhension du protéome viral reste incomplète. De plus, de nombreuses protéines sombres diffèrent considérablement des protéines déjà étudiées, posant des défis supplémentaires pour leur caractérisation. Ces protéines uniques, qui sortent de l'ordinaire, peuvent révéler de nouvelles fonctions biologiques.

Par opposition, on appelle *white proteome*, la partie que l'on considère connue. La zone grise (soit *twilight zone*), à l'entre-deux, n'est ni complètement définie ni complètement inconnue (PERDIGÃO, ROSA et al. 2017).

* * * *

Le dark proteome viral demeure une "masse sombre" fascinante dans la recherche virologique. Sa caractérisation est une priorité pour plusieurs raisons :

— Développement de nouvelles stratégies antivirales : Une meilleure compréhension des protéines sombres pourrait révéler de nouvelles cibles thérapeutiques ou vaccinales.

— Exploration de la biologie virale : L'étude des régions désordonnées et des protéines uniques ouvre des perspectives inédites pour comprendre la diversité fonctionnelle des virus.

1.3.2. Bio-donnée en abondance

La métagénomique transforme radicalement la bioinformatique en générant une quantité massive de données de séquençage. Avec des techniques en constante évolution, le volume et la complexité des données augmentent, posant un défi méthodologique majeur pour leur caractérisation et analyse (RU et al. 2023).

Ce défi technique pousse au développement d'outils adaptés au traitement de grandes quantités de données. La bioinformatique joue ici un rôle crucial en permet-

tant de filtrer, organiser et extraire des informations à partir de ces données brutes. Cela souligne l'importance de l'optimisation des algorithmes et des infrastructures de calcul pour répondre aux besoins croissants en traitement des données.

Les bases de données sont essentielles pour centraliser et structurer ces informations, facilitant ainsi leur accès et leur analyse par la communauté scientifique. On distingue notamment :

- Bases de données généralisées :
  - NCBI (National Center for Biotechnology Information) : une référence mondiale dans le domaine des données biologiques (SMITH 2013)
- Bases de données spécialisées
  - *GISAID* : dédiée aux virus influenza et au SARS-CoV-2 (KHARE et al. 2021)
    Histoire : créée pour le partage rapide des données sur les virus émergents
    Potentiel : favorise la collaboration internationale et la surveillance des épidémies.
  - *UniProt* : pour les séquences et les fonctions des protéines (THE UNIPROT CONSORTIUM 2023)
  - PDB (*Protein Data Bank*) : pour les structures tridimensionnelles des protéines (ZARDECKI et al. 2022)
  - *eggNOG* : spécialisée dans la recherche de groupes d'orthologues annotés (HUERTA-CEPAS et al. 2018)
  - *Dark Proteome Database* : répertorie le protéome profond, incluant les protéines mal caractérisées du plus "sombre" au plus "clair" (PERDIGÃO, ROSA et al. 2017),
  - *Polymerase Structure Alignments* : superpose toutes les polymérase virales (PEERSEN 2019).

* * * *

Malgré l'acquisition d'une somme considérable de connaissances et de développement une logique analytique pour extraire des informations à partir de cette abondance de données, les outils pour donner du sens à cette masse d'information manquent encore. Ce que nous pourrions découvrir pourrait remettre en question certaines certitudes et nous mener vers une compréhension plus approfondie du monde qui nous entoure.

Nous entrons dans une nouvelle ère où la simple récolte de données ne suffit plus. Il est désormais essentiel de penser à la visualisation des connaissances. Cartographier ces données les rendrait non seulement plus accessibles et compréhensibles, mais aiderait aussi à nourrir la réflexion scientifique.

1.3.3. L'annotation

L'annotation protéomique consiste à attribuer des significations biologiques à des séquences ou des structures protéiques. Ce processus est fondamental pour :

— assigner des fonctions biologiques aux éléments identifiés,

— relier les protéines à des processus biologiques et des voies métaboliques, interactions avec leur environnement,

— caractériser l'aspect enzymatique d'une protéine (emploie du code EC),

— mettre en évidence l'aspect phylogénie, avec la caractérisation des domaines protéiques conservés,

— identifier les motifs structuraux et les sites actifs.

En bioinformatique, l'annotation est un enjeu majeur, notamment face à l'explosion des données protéomiques. Elle repose sur des approches automatisées, comme l'utilisation d'outils de similarité tels que BLAST (ALTSCHUL, GISH et al. 1990), et manuelles, nécessitant l'intervention humaine pour affiner et valider les résultats. La mise à jour constante des annotations est également essentielle avec l'évolution des techniques et des connaissances. Des méthodes plus avancées, comme les modèles de Markov et l'apprentissage automatique, permettent une meilleure sensibilité dans l'analyse.

Enfin, l'annotation des protéines "sombres", dont les fonctions et structures sont encore largement inconnues, reste une priorité importante pour comprendre des aspects inexplorés du protéome.

* * * *

L'annotation en bioinformatique est un processus dynamique qui évolue constamment avec les avancées en séquençage et en analyse informatique. Elle agit comme un système d'"étiquetage", essentiel pour interpréter et donner du sens aux données issues des génomes, notamment viraux. Ce processus est crucial pour notre compréhension des systèmes biologiques, permettant de relier des séquences à des fonctions

biologiques spécifiques et de mieux appréhender la complexité des organismes. L'annotation contribue ainsi à éclairer des aspects encore inconnus et à structurer les informations nécessaires pour de futures recherches et applications.

1.3.4. La règle du jeu : Homologie

Qu'entend-on par séquences "homologues" ?

En biologie moléculaire et en évolution, l'homologie désigne une relation évolutive entre deux séquences (ADN, ARN ou protéines) ayant un ancêtre commun. Deux séquences sont dites "homologues" si elles proviennent d'une origine ancestrale partagée, ce qui implique qu'elles ont divergé au cours de l'évolution. Il est important de noter que l'homologie est une propriété binaire : deux séquences sont soit homologues, soit non homologues. Contrairement à certaines idées reçues, il n'existe pas de "degré" ou de "pourcentage" d'homologie, mais on peut parler de similitude ou d'identité de séquence pour mesurer leur ressemblance.

On peut distinguer deux types d'homologie, illustrés par la figure ci-dessous (1.16) :

— orthologie : homologie entre gènes de différentes espèces issus d'un événement de spéciation,

— paralogie : homologie entre gènes issus d'une duplication au sein d'un même génome.

La similarité de séquence, souvent exprimée en pourcentage d'identité, mesure combien de positions entre deux séquences sont identiques. Une forte similarité entre deux séquences peut indiquer une homologie, mais ce n'est pas une preuve définitive. En effet, deux séquences peuvent sembler similaires sans avoir un ancêtre commun, en raison de pressions évolutives similaires qui ont conduit à des adaptations comparables, phénomène connu sous le nom de convergence évolutive.

Un exemple classique en biologie pour illustrer cela est celui des ailes des oiseaux et des chauves-souris. Bien que les deux possèdent des ailes pour voler, ces structures ne sont pas homologues puisqu'elles ne partagent pas un ancêtre commun ayant des ailes. Les ailes sont un exemple de convergence évolutive, car elles ont évolué indépendamment dans ces deux groupes pour remplir une fonction similaire, le vol.

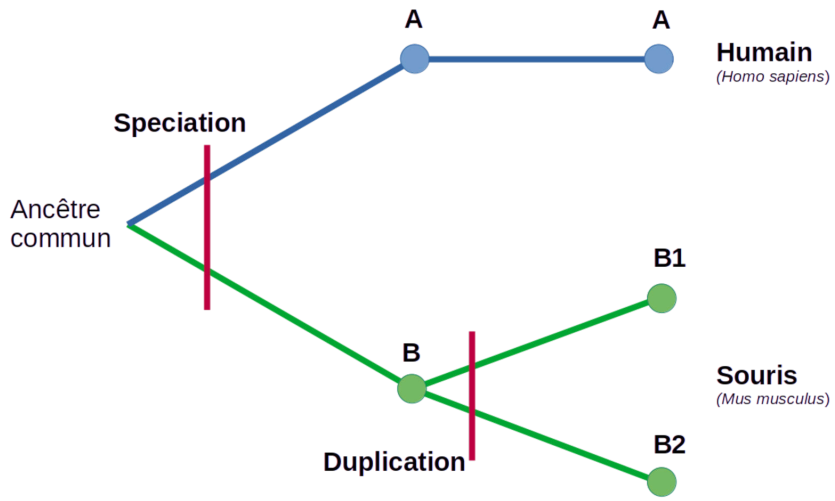


FIGURE 1.16. – Les différents types d'Homologie.
Spéciation : Événement qui marque la différenciation entre deux espèces ; les séquences **A** et **B** sont **orthologues**. **Duplication** : Processus où un gène au sein d'un même génome est dupliqué, entraînant l'évolution indépendante de ces deux copies ; les séquences **B1** et **B2** sont **paralogues**.

Conservation structurale

Les structures tridimensionnelles des protéines homologues tendent à être mieux conservées que leurs séquences (ILLERGÅRD et al. 2009). Cela signifie que même si deux protéines homologues présentent des différences significatives au niveau de leur séquence d'acides aminés, leur structure 3D reste souvent très similaire. Cette conservation de la structure est essentielle pour préserver la fonction biologique des protéines, car la forme tridimensionnelle d'une protéine est étroitement liée à sa fonction.

L'homologie est un concept clef en biologie évolutive, permettant de regrouper des séquences partageant une origine commune. Identifier des groupes de séquences homologues est crucial pour plusieurs raisons :

— Caractérisation et analyse des séquences : L'homologie aide à mieux comprendre les relations entre différentes protéines ou gènes et leurs fonctions potentielles.

— Prédiction fonctionnelle pour les gènes non caractérisés : Notamment dans le cadre du dark proteome, l'homologie permet d'inférer des fonctions à partir de gènes connus.

— Modélisation par homologie des structures protéiques : Utilisé pour prédire la structure de protéines en se basant sur celles déjà connues, cette méthode est largement appliquée pour les protéines « blanches » (connues).

— Études évolutives et phylogénétiques : L'homologie est essentielle pour retracer les liens évolutifs entre espèces.

* * * *

En bioinformatique, l'homologie est la pierre angulaire de nombreuses analyses, en particulier pour l'annotation fonctionnelle des génomes, reliant les séquences à des fonctions connues, et facilitant des découvertes sur les fonctions et structures protéiques non résolues.

1.4. Quelle est la boîte à outils du bioinformaticien aujourd'hui ?

La bioinformatique n'est pas simplement une discipline, mais une super-compétence qui s'applique à de nombreux domaines de pointe et projets interdisciplinaires. Aujourd'hui, elle dispose d'une gamme d'outils variés et puissants.

1.4.1. La phylogénétique

Nous avons vu un peu plus tôt la phylogénie et ses principes. Ici, nous abordons la phylogénétique qui est son aspect analytique et méthodologique appliqué à la bioinformatique. La phylogénétique s'appuie sur différents piliers comme :

1. l'alignement de séquences : étape essentielle pour comparer les séquences.
2. la construction d'arbres phylogénétiques : représentation graphique des relations évolutives.
3. les méthodes de construction d'arbre.

* * * *

L'analyse phylogénétique joue un rôle crucial dans la compréhension de l'évolution et des relations entre les génomes viraux. Les méthodes de construction d'arbres phylogénétiques, telles que le maximum de vraisemblance, l'inférence bayésienne et le maximum de parcimonie, utilisent des données d'alignement de séquences pour proposer la meilleure topologie possible. Bien que ces méthodes soient importantes pour l'étude des virus, notre travail se concentre sur le développement d'outils bioinformatiques spécifiques à l'annotation des protéines virales, un aspect complémentaire mais distinct de l'analyse phylogénétique.

1.4.1.1. Les principaux aligneurs

Un aligneur est un algorithme qui permet de comparer des séquences entre elles, c'est-à-dire mettre en évidence leurs similarités comme leurs différences. Il existe plusieurs stratégies pour aligner :

- La stratégie "*greedy*" est représentée par BLAST (ALTSCHUL, GISH et al. 1990) et ses variations. Cette gamme d'outils est idéale pour comparer deux à deux une séquence contre chaque entité d'une large base de donnée. BLAST n'est plus utilisé au profit des autres.

  PSI-BLAST : une version itérative améliorant grandement sa sensibilité, plus spécifiquement pour détecter des homologues distants de protéines (ALTSCHUL, MADDEN et al. 1997 ; ODA et al. 2017).

  DIAMOND : une version plus rapide que BLAST, adaptée pour être confrontée à de plus grands ensembles de données métagénomiques (BUCHFINK et al. 2021).

- MMSEQS2 est un outil de recherche de séquence qui se démarque de la stratégie de BLAST. Il se veut plus rapide, et aussi précis que PSI-BLAST ou DIAMOND (STEINEGGER et al. 2017). Cet outil permet aussi de regrouper des séquences similaires, on parlera alors de *clustering*. MMSEQS2 propose d'aligner plusieurs séquences contre une autre multitude. Cette fonctionnalité le rend particulièrement efficace pour la recherche de séquences homologues dans une large base de données. Il peut également faire une recherche à partir d'un profile de séquences, augmentant ainsi sa sensibilité. Le *clustering* peut être généralement utilisé pour générer des profiles à partir de séquences similaires.

- L'alignement multiple de séquence (MSA) est une autre stratégie, celle-ci permet d'aligner plus de trois séquences en même temps. Le but est de mettre en évidence les régions conservées, les motifs communs et les variations sur un ensemble de séquences. Le MSA vise à révéler les relations évolutives et fonctionnelles que partagent ces séquences. Voici quelques représentants : MAFFT (KURAKU et al. 2013 ; KATOH et al. 2019), CLUSTALΩ (SIEVERS et al. 2011), MUSCLE (EDGAR 2004). Ces outils peuvent aligner jusqu'à un volume de l'ordre de la centaine au millier de séquences, de longueur du même ordre. Les vitesses et précisions des *runtimes* sont relativement similaires et dépendent surtout des données d'entrée (input) (KOSTENKO et al. 2022).

* * * *

Un choix peut être difficile parmi ces aligneurs. Chacun apporte des nuances aux méthodes employées. La différence se joue surtout sur le paramétrage, l'*input*, les préférences de chacun et les ressources matérielles à disposition.

1.4.1.2. L'arbre phylogénétique

Un arbre, en bioinformatique, est décrit comme n'importe quel arbre :

— La **racine** est le début de l'arbre. Dans notre contexte, il s'agit du plus ancien ancêtre commun.

— Les **feuilles** sont les extrémités de l'arbre, et représentent les espèces actuelles.

— Les **nœuds** intermédiaires sont tous les ancêtres communs hypothétiques.

— Les **branches** connectent deux nœuds par lien de parenté évolutive. Elles peuvent aussi porter des informations supplémentaires comme les "longueurs de branches", pour représenter le temps évolutif.

— L'arrangement d'un arbre est appelée topologie. Il existe plusieurs topologies possibles pour exprimer les mêmes relations entre les feuilles.

Un arbre est un objet graphique pour représenter un clade étudié. Il peut prendre plusieurs formes selon les convenances :

— *cladogramme* : seul les relations de parenté importent, les longueurs de branches ne sont pas significatives.

— *phylogramme* : à l'inverse, les branches fournissent des informations sur les taux évolutif.

— *circulaire* : une représentation à 360 degrés pour un grand nombre d'espèces.

— *raciné* : un point d'origine (racine) donne la direction évolutive.

— *non raciné* : ici, la racine n'est pas définie avec certitude.

— *newick* : un format texte universel.

1.4.1.3. Distinction entre arbre taxonomique et arbre phylogénétique

Dans cette thèse, il est important de souligner la différence fondamentale entre un arbre taxonomique et un arbre phylogénétique. Bien que ces deux types d'arbres puissent avoir une apparence similaire, ils diffèrent significativement dans leur conception et leur méthodologie de construction.

Arbre taxonomique

Un arbre taxonomique est une représentation visuelle de la classification actuelle des organismes. Il se base sur la taxonomie établie et ne nécessite pas de méthode de construction phylogénétique. Cet arbre reflète simplement la hiérarchie des catégories taxonomiques (comme espèce, genre, famille, etc.) telles qu'elles sont actuellement définies.

Arbre phylogénétique

En revanche, un arbre phylogénétique est le résultat d'une analyse spécifique visant à reconstruire les relations évolutives entre les organismes. Sa construction nécessite :

1. Un alignement de séquences (généralement ADN ou protéines) comme point de départ.
2. L'application d'une méthode de reconstruction phylogénétique particulière.

Par conséquent, un arbre phylogénétique représente une hypothèse ou une proposition d'une phylogénie possible. Il est important de noter que différents alignements ou méthodes de construction peuvent conduire à des arbres phylogénétiques différents pour le même groupe d'organismes. Cette distinction est cruciale pour comprendre la nature et les limites des informations fournies par chaque type d'arbre dans le contexte de cette thèse.

1.4.2. La recherche de séquence

Il est également essentiel d'aborder les principes bioinformatiques fondamentaux, ci-dessous, qui étendent les capacités de la recherche de séquences. Typiquement, le point de départ est une séquence, qu'elle soit d'origine expérimentale ou de référence, appliquée dans un contexte de métagénomique ou d'annotation fonctionnelle. La recherche de séquences permet alors d'identifier des séquences annotées et homologues à notre séquence initiale, enrichissant ainsi notre compréhension et nos analyses.

1.4.2.1. Le modèle "*Markov* caché"

Le modèle de Markov caché (HMM) est un outil statistique puissant largement utilisé en bioinformatique pour l'analyse et la prédiction de séquences de données. Son principe repose sur la conversion d'un alignement de séquences en un ensemble fini d'états, associés à une matrice de probabilités de transition. Son fonctionnement est illustré dans la figure (1.17) ci-dessous. L'utilisation des HMM s'est considérablement répandue dans le domaine de la bioinformatique, devenant incontournable dans de nombreux outils d'analyse. On les retrouve notamment dans :

— La prédiction de structures secondaires et tertiaires des protéines, avec des outils comme HHpred (HILDEBRAND et al. 2009)

— La classification des protéines en familles, grâce à des programmes tels que HHblits (REMMERT et al. 2011)

— La création d'alignements multiples de séquences, où ils modélisent efficacement les insertions, délétions et substitutions (comme dans MUSCLE (EDGAR 2004), MAFFT (KATOH et al. 2019), ClustalΩ (SIEVERS et al. 2011) et d'autres).

L'un des principaux avantages des HMM réside dans leur sensibilité accrue, permettant d'approfondir l'étude des relations évolutives entre espèces par l'analyse fine des variations de séquences. De plus, ils excellent dans la représentation de profiles de séquences homologues, offrant ainsi une vision plus complète des similitudes et différences entre protéines apparentées.

* * * *

Cette approche probabiliste a donc révolutionné de nombreux aspects de l'analyse bioinformatique, en apportant une précision et une sensibilité accrues dans la détection et la caractérisation des motifs et structures biologiques. De plus, étant un modèle statistique, le HMM requiert un grand volume de séquences pour être véritablement efficace.

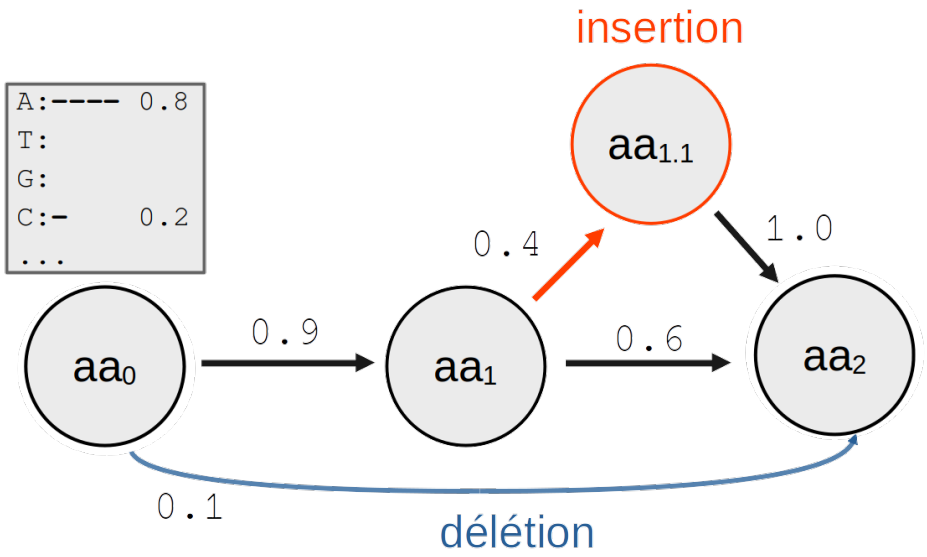


FIGURE 1.17. – Représentation schématique d'un modèle de Markov caché (HMM) pour l'analyse de séquences biologiques.
Le schéma illustre les composants clefs d'un HMM appliqué aux séquences biologiques :

— État initial (aa0) : Point de départ du modèle, associé à une matrice de probabilités décrivant sa composition. La taille de cette matrice varie selon le type de séquence analysée : 4 pour l'ADN et 20 pour les protéines, correspondant au nombre de nucléotides ou d'acides aminés possibles.
— Transitions entre états : Représentées par des flèches, elles indiquent les passages possibles d'un état à l'autre. Trois types de transitions sont mis en évidence :

1. Correspondance/Discordance : transition horizontale.
2. Délétion : Illustrée par la transition directe de aa0 à aa2, sautant aa1.
3. Insertion : Montrée par la transition de aa1 à aa1.1, introduisant un état supplémentaire.

Cette représentation permet de visualiser comment le HMM modélise la structure et les variations des séquences biologiques, en tenant compte des événements d'insertion, de délétion et de substitution couramment observés dans l'évolution des molécules.

1.4.2.2. Les prédicteurs de structure protéique

Les prédicteurs de structure protéique jouent un rôle crucial dans l'annotation des protéines virales. Ces outils bioinformatiques permettent d'anticiper la conformation

tridimensionnelle d'une protéine à partir de sa séquence d'acides aminés.

Processus général de prédiction

La plupart des prédicteurs suivent une succession d'étapes similaires :

1. ***Input*** : Entrée de la, ou des, séquence protéique à analyser.
2. ***Pooling*** : Regroupement des informations pertinentes, telles que les propriétés physico-chimiques des acides aminés ou les données évolutives. Cette étape comprend aussi un alignement multiple de séquence, comparaison de la séquence cible avec des séquences homologues pour extraire des informations évolutives.
3. **Intégration des annotations** : Fusion des annotations provenant de différentes sources, comme les bases de données de structures connues et les résultats d'alignements multiples, pour enrichir l'analyse.
4. ***Modeling*** : Modélisation de la structure tridimensionnelle de la protéine, incluant des étapes supplémentaires comme :
   - **Recherche de templates** : Identification de structures similaires connues.
   - Intégration des éléments de structures secondaires.
   - **Prédiction des contacts résiduels** : Estimation des interactions entre acides aminés distants.
   - **Optimisation énergétique** : Affinage du modèle pour minimiser l'énergie.
   - **Évaluation de la qualité du modèle** : Utilisation de métriques pour estimer la fiabilité.
   - **Raffinement itératif** : Amélioration progressive du modèle par des cycles répétés de prédiction et d'évaluation.

Évolution des prédicteurs

On peut distinguer deux générations de prédicteurs :

- **Prédicteurs "traditionnels"** : Ces prédicteurs, basés sur des méthodes statistiques et des algorithmes classiques d'apprentissage automatique, sont représentés par des outils comme PHYRE2 (KELLEY et al. 2015).

  PHYRE2 utilise le *threading*, une méthode qui "enfile" les acides aminés de la

séquence cible sur une structure *template*, puis évalue la qualité de cet ajustement. Cette technique est particulièrement efficace pour les séquences ayant une faible homologie (<30% d'identité). PHYRE2 optimise ce processus en combinant *threading* et *multi-threading*, améliorant ainsi le temps d'exécution et l'utilisation de la mémoire.

La phase de *modeling* de PHYRE2 illustre la complexité de l'approche :

1. Construction d'un modèle de squelette grossier : PHYRE2 superpose les acides aminés correspondants, créant un *core* qui sert de point d'ancrage. Cette étape exploite le fait que les structures protéiques sont généralement plus conservées que les séquences, surtout dans les régions du cœur de la protéine. L'utilisation de profiles de séquences et de modèles de Markov cachés (HMM) améliore la détection d'homologues distants.
2. Modélisation des boucles : Cette étape corrige les défauts dus aux insertions et délétions, reflétant la variabilité évolutive.
3. Ajout des chaînes latérales : PHYRE2 utilise une base de données de structures courantes et intègre des prédictions de structure secondaire et de régions désordonnées pour affiner la structure finale.

L'efficacité des prédicteurs "traditionnels" s'est accrue grâce à l'augmentation du nombre de structures protéiques annotées, permettant une recherche d'homologie plus performante.

— **Nouvelle génération** : Utilise des modèles avancés de langage, notamment des réseaux de neurones profonds, pour améliorer la précision des prédictions.

Les modèles de langage en intelligence artificielle sont des algorithmes capables de comprendre et de générer des séquences protéiques comme s'il s'agissait d'un langage. Cette approche novatrice a considérablement amélioré la précision des prédictions de structure protéique, comme le démontre le succès d'ALPHAFOLD développé par DEEPMIND. ALPHAFOLD2 (JUMPER et al. 2021), introduit en 2020, a marqué une avancée significative dans ce domaine.

Cette nouvelle génération se démarque de ses prédécesseurs par l'intégration d'informations évolutives plus complexes, notamment les relations entre résidus distants dans la structure et la covariance. ROSETTAFOLD (MARCHAL 2024) a amélioré cette approche en maîtrisant la notion de covariance, tandis

qu'ESMFOLD (LIN et al. 2023) propose une modélisation *de novo*, particulièrement efficace pour les séquences orphelines et de taille moyenne.

Les modèles de langage continuent d'évoluer et s'étendent à d'autres domaines, comme le docking in silico. La récente introduction d'ALPHAFOLD3 (ABRAMSON et al. 2024) en 2024 illustre cette progression, avec sa capacité à prédire non seulement les structures protéiques, mais aussi les interactions avec l'ADN, l'ARN, et certains ligands. Cette version utilise une nouvelle architecture appelée "*Pairformer*" et un modèle de diffusion pour générer des prédictions encore plus précises.

Ces avancées ouvrent de nouvelles perspectives dans la compréhension des mécanismes biologiques et accélèrent potentiellement la découverte de médicaments et la recherche biomédicale.

* * * *

Les prédicteurs de nouvelle génération, basés sur l'apprentissage profond et les modèles de langage, permettent une annotation plus précise des protéines virales, améliorant notre compréhension de leur fonction et évolution. Ces outils offrent des prédictions fiables même en l'absence d'homologie, surpassant les méthodes traditionnelles (TERWILLIGER, LIEBSCHNER, CROLL, WILLIAMS, MCCOY, POON, AFONINE, OEFFNER, RICHARDSON et al. 2024 ; TERWILLIGER, LIEBSCHNER, CROLL, WILLIAMS, MCCOY, POON, AFONINE, OEFFNER, SCHLICKSUP et al. 2023).

Ces avancées ont révolutionné la prédiction de structures protéiques, intégrant des étapes complexes qui reflètent l'évolution des méthodes dans ce domaine. L'intelligence artificielle transforme divers domaines, de la médecine à la recherche scientifique. À l'intersection de la technologie et de la biologie, ces outils puissants soulèvent la question : Quelles seront les conséquences de l'intégration des IA dans le quotidien du chercheur ou du bioinformaticien ?

1.5. Quelle forme peut prendre un outil bioinformatique ?

Un outil bioinformatique peut prendre plusieurs formes, adaptées au profile des utilisateurs ciblés et à la portée visée. Les principales formes sont :

Application locale :

— **Profil** : Utilisateurs ayant des compétences techniques suffisantes pour l'installation.
— **Portée** : Généralement limitée, car nécessite une installation sur chaque poste.
— **Accessibilité** : Installation complexe, mais utilisation rapide une fois installée. Souvent standalone.
— **Interface** : Aspect "logiciel".
— **Sécurité et confidentialité des données** : Offre un meilleur contrôle sur la sécurité des données, mais nécessite une gestion locale des mises à jour de sécurité.
— **Protection du code source** : Le code peut être obfusqué ou compilé pour limiter l'accès.
— **Contraintes matérielles** : Dépend des ressources de la machine de l'utilisateur.
— **Exemples** : Editeurs de séquences biologiques (*Seaview* (GOUY et al. 2010), *Jalview* (WATERHOUSE et al. 2009)), outils en biologie structurale (*chimeraX* (MENG et al. 2023), *COOT* (EMSLEY et al. 2004)).

Paquet (*package*) :

— **Profil** : Bioinformaticiens confirmés, capables d'utiliser des langages de programmation ou des lignes de commande.
— **Portée** : Moyenne à large, facilement intégrable dans des pipelines.
— **Accessibilité** : Nécessite des connaissances en programmation, offre une grande flexibilité. Peut dépendre d'autres paquets.
— **Interface** : Généralement non interfacé, lignes de commande.
— **Sécurité et confidentialité des données** : La sécurité dépend de l'environnement d'exécution et des pratiques de l'utilisateur.
— **Protection du code source** : Souvent open-source, mais peut utiliser des licences restrictives.

— **Contraintes matérielles** : Peut nécessiter des ressources importantes selon la complexité des calculs.
— **Exemples** : *Biopython* (paquet *python*), pipelines d'analyse métagénomiques.

Service web :
— **Profil** : Large éventail d'utilisateurs, des débutants aux experts.
— **Portée** : Potentiellement très large, accessible via internet.
— **Accessibilité** : Facile d'accès, ne nécessite qu'un navigateur web.
— **Interface** : Forme "formulaire" ou interface "dynamique".
— **Sécurité et confidentialité des données** : Nécessite une attention particulière à la sécurité des transmissions (HTTPS) et au stockage des données. Doit souvent se conformer à des réglementations strictes (RGPD en Europe).
— **Protection du code source** : Le code *backend* reste généralement privé et protégé.
— **Contraintes matérielles** : Les ressources sont gérées côté serveur, mais peuvent limiter le nombre d'utilisateurs simultanés, la taille des données traitées ou la complexité des tâches.
— **Exemples** : *Bioinformatic toolkit* (GABLER et al. 2020), *Nextstrain* (HADFIELD et al. 2018).

* * * *

Le choix de la forme d'un outil bioinformatique dépend des objectifs du développeur, des besoins des utilisateurs et des contraintes spécifiques du projet. Un service web favorise une large diffusion et une facilité d'utilisation, tandis qu'un *package* convient mieux aux utilisateurs techniques pour des workflows complexes. Les considérations de sécurité, de confidentialité des données et de ressources de calcul influencent également ce choix, assurant ainsi la viabilité et l'acceptabilité de l'outil dans divers contextes d'utilisation.

1.6. Buts et Objectifs

Le cadre de la recherche

La pandémie de COVID-19 a radicalement changé le paysage de la recherche en virologie, notamment en ce qui concerne les virus émergents comme le SARS-CoV-2 (CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES et al. 2020). Ces virus, capables de surgir soudainement dans des populations naïves et de se propager rapidement, posent des défis scientifiques majeurs, en particulier pour suivre leur évolution génétique et comprendre leur adaptation. Ces enjeux sont exacerbés par la capacité de ces virus à passer d'un hôte à l'autre, rendant l'étude de leur histoire et de leur dynamique particulièrement complexe (DOMINGO 2021).

Mon travail de thèse s'inscrit au sein d'une équipe de recherche dirigée par le Dr Bruno Canard, spécialisée dans l'étude des complexes réplicatifs des virus à ARN. L'équipe "Réplicases Virales : Structure, Mécanisme et Drug-Design" joue un rôle clef dans la recherche fondamentale et appliquée sur les virus émergents, tels que le (HIV) (DECOMBE et al. 2024) , le virus de la dengue (FERACCI et al. 2023), le RSV (SUTTO-ORTIZ et al. 2021) et le SARS-CoV-2 (SHANNON, FATTORINI et al. 2022 ; SHANNON, CHAZOT et al. 2024 ; GRANDA et al. 2023). Cette équipe s'attache à élucider les structures protéiques des virus et à concevoir des médicaments antiviraux (SHANNON, LE et al. 2020 ; BARDIOT et al. 2022 ; SHANNON, SELISKO et al. 2020), dans un contexte où la lutte contre les menaces virales devient de plus en plus urgente.

La pandémie de COVID-19 a agi comme un catalyseur pour des recherches fondamentales et appliquées sur les virus émergents (SOHRABI et al. 2020). En particulier, l'analyse des variants du SARS-CoV-2 (TORBATI et al. 2021) et la mise en place de stratégies de séquençage ont mis en évidence la nécessité de développer des outils bioinformatiques pour traiter les données biologiques complexes. Cette crise sanitaire a permis à notre laboratoire de jouer un rôle central, en particulier durant les périodes de confinement, où la collecte de données et l'étude des variantes ont été accélérées.

Le laboratoire dirigé par le Dr Bruno Canard est activement impliqué dans des projets européens visant à répondre aux futures pandémies, intégrant ainsi ce projet de thèse dans un cadre international de recherche. Ces projets soutiennent les efforts

de recherche en promouvant des collaborations internationales sur les virus émergents. Une initiative soutenue par l'ANRS, dirigée par le Dr Canard, se concentre sur l'intégration et le partage des données, un aspect crucial pour l'étude des variants du SARS-CoV-2 et d'autres virus émergents. Cette approche vise à renforcer la capacité de réponse face aux menaces virales futures en facilitant une coopération nationale interdisciplinaire.

En 2020, alors que je terminais mon master, j'ai eu l'occasion de travailler avec l'équipe sur la caractérisation et l'annotation du génome du SARS-CoV-2. Cette première expérience m'a permis de m'intégrer de manière fluide au projet de thèse en 2021, où j'ai continué à développer des outils bioinformatiques pour l'annotation des génomes viraux. Mon implication dans ce projet est naturelle, car mon parcours en bioinformatique a directement répondu aux besoins de l'équipe dans ce domaine crucial.

Pour répondre aux défis actuels, le laboratoire développe plusieurs solutions bioinformatiques, dont l'outil VAZyMOLO (FERRON, RANCUREL et al. 2005), ainsi que d'autres alternatives pour améliorer l'annotation des génomes viraux et identifier les mutations pertinentes. Ces outils sont essentiels pour fournir des informations exploitables dans la lutte contre les virus émergents.

Le projet thèse

Face à l'abondance de données générées par les techniques modernes de séquençage, de la métagénomique et les campagnes de séquençage liées à la pandémie, les scientifiques se trouvent souvent débordés et frustrés par des flux constants de données désorganisées. Traditionnellement, il est courant de passer beaucoup de temps à rechercher des séquences homologues ou des modèles viraux adaptés aux questions posées.

Mon projet de thèse vise à répondre à ces défis en proposant un système d'organisation du savoir qui rend ce dernier accessible et facilite le travail quotidien des biologistes.

1. **Développement d'outils pour la caractérisation et l'analyse des génomes viraux** : Ces outils sont spécifiquement conçus pour être appliqués dans le contexte de la pandémie de COVID-19.
2. **Développement d'outils pour organiser et cartographier le savoir** : L'objectif est de structurer les connaissances maîtrisées par notre équipe et d'en permettre la diffusion au sein de la communauté scientifique.

Ce projet ambitionne ainsi de transformer la manière dont les données virales sont gérées et partagées, offrant une solution pratique aux défis actuels rencontrés par les chercheurs.

Résultats

Première Partie

Chapitre 2

Développement d'outils pour la caractérisation
et l'analyse des génomes viraux.

2. *Viral Instant Mutation Viewer* :
A Tool to Speed Up the Identification and Analysis of New SARS-CoV-2 Emerging Variants and Beyond

article publié (WILDE et al. 2023)
dans *Viruses special issue* : **Virus Bioinformatics 2023**, soumis en juin 2023.

Dans ce chapitre, je souhaite proposer un premier élément de réponse à la thèse avec un premier outil intitulé VIMVER (pour *Viral Instant Mutation Viewer*). Sa création est le fruit de la rencontre entre le potentiel bioinformatique, croissant au sein de l'équipe, et mon engagement à développer un outil afin de soutenir leur travail de recherche.

Sommaire

2.1. Synthèse 80
2.2 Article 84
2.3. Discussion 94

2.1. Synthèse

Le contexte

Les coronavirus (CoVs), appartenant à l'ordre des *Nidovirales*, sont des virus à ARN simple brin de grande taille. Le SARS-CoV-2, à l'origine de la pandémie de COVID-19, reste un défi mondial malgré les campagnes de vaccination. La surveillance continue de l'évolution de son génome est essentielle pour :

— Identifier les lignées circulantes et émergentes.

— Étudier les variants préoccupants : *Variants of Concern* (TORBATI et al. 2021).

— Évaluer l'impact des mutations sur l'entrée virale, la pathogénèse et la réponse immunitaire (HOWARD et al. 2012).

Les variantes génétiques impactent l'efficacité des vaccins et des traitements (TORBATI et al. 2021 ; FLORES-VEGA et al. 2022), rendant essentielle la cartographie des mutations au niveau des génomes et des protéomes viraux. Pour répondre à ce besoin, nous avons développé VIMVer (Viral Instant Mutation Viewer), un outil web permettant l'identification directe des mutations dans le génome viral ainsi que leurs équivalents au niveau protéique.

Avec plus de 15 millions de séquences génomiques disponibles via des plateformes comme GISAID (KHARE et al. 2021), l'analyse des mutations est devenue cruciale. Les méthodes de séquençage de nouvelle génération (NGS) sont largement utilisées dans les hôpitaux et les laboratoires de recherche. Cependant, l'identification rapide des séquences mutantes et leur impact protéique demeure complexe.

VIMVer simplifie ce processus en permettant :

— La localisation des régions critiques de mutation sur le génome.

— L'attribution instantanée des lignées.

— Une visualisation intuitive des impacts aux niveaux du génomes.

La base de données

Actuellement, VIMVer se concentre sur le génome du SARS-CoV-2. La base de données comprend le génome complet du SARS-CoV-2 (*Wuhan-1* : NC_045512.2 ou

YP_009725297.1 de NCBI), comprenant les 26 protéines en séquences nucléotidiques et en acides aminés, organisées en deux fichiers fasta distincts.

La base de données de VIMVer intègre des séquences virales de référence sous forme de séquences génomiques rétro-traduites à partir des séquences protéiques correspondantes. Le génome du SARS-CoV-2 est divisé en sous-séquences fonctionnelles, chacune correspondant à une protéine spécifique. Cette méthode préserve les cadres de lecture des codons et gère les décalages ribosomiques typiques des coronavirus, notamment l'ORF1 qui conduit à la synthèse des polyprotéines pp1a et pp1ab (BHATT et al. 2021). Cette méthode assure une annotation précise et une représentation fidèle de la structure génomique complexe du virus.

Ce décalage affecte la traduction de la protéine nsp12, dont l'extrémité N-terminale est codée à la fin de l'ORF1a, chevauchant nsp11, tandis que le reste de la séquence codante subit un décalage de +1. Dans la base de données locale, nsp11 a été fusionné au début de nsp12 pour s'aligner sur le cadre de lecture codant, entraînant la duplication du *n*ième nucléotide de la séquence rétro-traduite correspondante.

De plus, Django utilise la base de données *MySQL* pour stocker tous les éléments des modèles créés, y compris les résultats des requêtes utilisateurs, accessibles pendant 24 heures. VIMVer génère des fichiers fasta contenant les alignements résultants.

Méthodes d'alignement et de recherche

VIMVer combine l'usage de BlastN (ALTSCHUL, GISH et al. 1990) et MUSCLE (EDGAR 2004) pour la génération des alignements de séquences et la gestion des requêtes à la base de données. Toutes les comparaisons sont effectuées par rapport aux séquences nucléotidiques de référence, servant d'ancrage pour la co-traduction codon par codon ultérieure. Le résultat se présente sous forme d'une liste de correspondances, chacune affichant deux panneaux interactifs : les alignements protéiques à gauche et nucléotidiques à droite. Les mutations sont mises en évidence par des astérisques. Les résultats sont téléchargeables au format multi-fasta.

Déploiement

Viral Instant Mutation Viewer (VIMVer) est déployé sur un serveur *Apache2* de notre laboratoire, avec un protocole SSL fourni par *Let's Encrypt*. Le code source est disponible gratuitement pour une installation locale, mais nécessite une configuration spécifique pour fonctionner. Un fichier de paramètres (settings.py) contenant une clé de chiffrement doit être présent, généré par l'administrateur lors de la création du projet dans Django. Ce fichier définit également les paramètres du serveur avant le déploiement. Des instructions détaillées pour l'installation sont disponibles sur le dépôt GitHub du projet.

Accessibilité

VIMVer est disponible publiquement en ligne : https://vimver.afmb.univ-mrs.fr/. Bien qu'initialement développé pour le SARS-CoV-2, cet outil peut être adapté à d'autres pathogènes grâce à son architecture flexible.

Article

viruses

MDPI

Article

Viral Instant Mutation Viewer: A Tool to Speed Up the Identification and Analysis of New SARS-CoV-2 Emerging Variants and Beyond

Vincent Wilde [1,2], Bruno Canard [1,2,3] and François Ferron [1,2,3,*]

[1] Architecture et Fonction des Macromolécules Biologiques, CNRS-UMR 7257, Polytech Case 925, 13009 Marseille, France; vincent.wilde@univ-amu.fr (V.W.); bruno.canard@univ-amu.fr (B.C.)
[2] Marseille and Laboratoire Architecture et Fonction des Macromolécules Biologiques (AFMB), Aix-Marseille Université, AFMB UMR 7257, 13288 Marseille, France
[3] European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany
* Correspondence: francois.ferron@univ-amu.fr; Tel.: +33-(0)491828628

Abstract: The appearance of genetic variants impacts vaccination efficiency and therapeutic options, generating a need to map and relate mutations observed in the proteome and the genome. We develop an user-friendly web service software (Viral Instant Mutation Viewer or VIMVer) which allows a direct identification of mutations in the genome and its counterpart in the viral proteome. Since its emergence in 2019, the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), responsible for the COVID-19 pandemic, has generated an overwhelming amount of data while becoming one of the most studied viruses of the Nidovirales order. We originally developed this tool during the COVID pandemic; thus, for any SARS-CoV-2 nucleotide sequence, the web service gives a fast identification, mapping, and display of new mutations simultaneously at the nucleotide and amino acid level in comparison to a reference sequence (Wuhan-1). Furthermore, the lineage or the relative position to the known lineage of the variant of interest is available on the link to Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN COVID-19). The workflow presented here is available online. The source code is released under public license and can be easily adapted for further development to other viruses.

Keywords: genomics; proteomics; Django; mutation; variants; virus bioinformatics; software; SARS-CoV-2

check for updates

Citation: Wilde, V.; Canard, B.; Ferron, F. Viral Instant Mutation Viewer: A Tool to Speed Up the Identification and Analysis of new SARS-CoV-2 Emerging Variants and Beyond. *Viruses* **2023**, *15*, 1628. https://doi.org/10.3390/v15081628

Academic Editors: Manja Marz, Franziska Hufsky, Alba Pérez-Cataluña and Fernando González-Candelas

Received: 30 June 2023
Revised: 19 July 2023
Accepted: 20 July 2023
Published: 26 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Coronaviruses (CoVs) are large-genome, positive-strand RNA viruses of the Nidovirales order that have recently attracted global attention due to the ongoing COVID-19 pandemic. Despite significant efforts to control its spread through worldwide vaccination campaign strategies, SARS-CoV-2 is still causing substantial health and economic burden, emphasizing the need for a continuous monitoring of its genome evolution to identify the circulating and emerging SARS-CoV-2 lineages [1]. Thus, the scale of the COVID-19 pandemic has led to unprecedented efforts by the research community to rapidly identify variant sequences, to test therapeutics and vaccines, and to understand the molecular basis of SARS-CoV-2 entry, pathogenesis, and immune targeting in light of these emerging mutations. This effort resulted in the availability of more than 15 million genomic sequences available through the GISAID [2] data portal: "https://www.gisaid.org/ (accessed on 30 June 2023)". Whole-genome sequencing allows us to discriminate one lineage from the other and spot variants of concern (VOCs). Next-generation sequencing (NGS) methods are widely used, from hospitals to academic research labs, and yet the process to rapidly identify mutant sequences and its impact at the protein level is still tedious, although maybe less so since the development of tools like nextclade [3]. In this work, we present the Viral

Instant Mutation Viewer (VIMVer) workflow, which allows the user to easily identify the mutation hot spots on the genome, to assign its lineage, and to instantly visualize its potential impact at the protein level. This resource can help practitioners and researchers quickly discriminate the new emergence of mutations at the genome level and its potential impact at the protein level. VIMVer is publicly available at: "https://vimver.afmb.univ-mrs.fr/" (accessed on 21 September 2022).

2. Materials and Methods

Our goal is to propose a tool to quickly analyze viral genomes and to rapidly detect mutations and their consequences on the protein structure relative to a reference strain present in the database. The VIMVer workflow is developed under Ubuntu 22.04 LTS and is an early-stage web service developed in Python 3.8+ using the Django framework. It is intended to be friendly to the virology community. The Django development framework has a heavy and rigorous but adaptable architecture which simplifies any integration of new functions during development.

The workflow is presented in Figure 1 and can be summarized as follows. A nucleic sequence is submitted as an input to a search engine, currently a blastn [4], to query the VIMVer reference database. The multiple sub-alignments generated by blastn are optimized through a realign procedure using Muscle version 3.8 [5] to proof the alignment. All alignments are carried out using the nucleotide sequences and are dynamically translated to generate the corresponding protein alignment. The protein alignment reflects the genomic variability and is independent of amino acid composition. The final step is the dynamic visualization of both alignments (nucleotide/amino acids), which allows us to spot mutations, whether silent or not. The user can submit the resulting output to the PANGOLIN webserver to assess lineage.

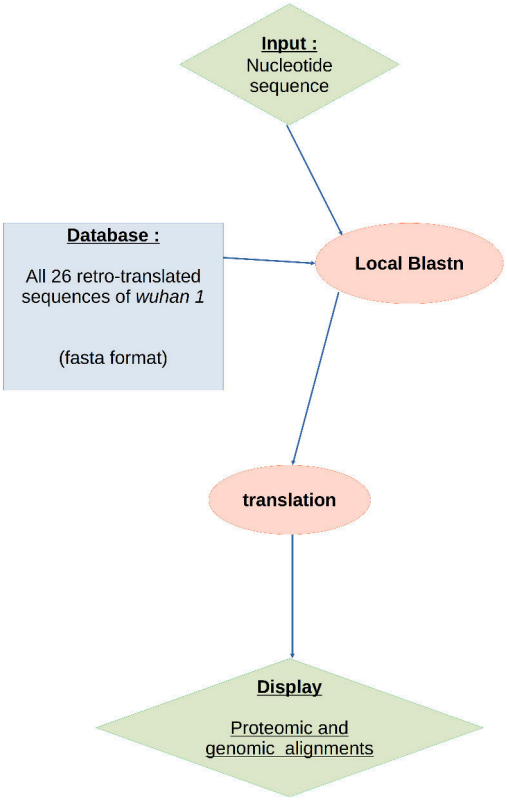


Figure 1. The workflow scheme of VIMVer.

2.1. Database

Any viral reference sequence could be implemented in the database and constituted of the retro-translated genomic sequence(s) from the corresponding viral protein sequences. In this context, viral genome sequences are treated as modular sequences representing their translated states. Therefore, genomes coding for polyproteins, which are subsequently processed into multiple proteins, are pre-annotated at the genomic level. Currently, the development focus of VIMVer lies in the SARS-CoV-2 genome. Consequently, the database currently consists of the complete SARS-CoV-2 genome (Wuhan-1: NC_045512.2 or YP_009725297.1 from NCBI), encompassing 26 proteins in both nucleotide and amino acid sequences. Each sequence type is organized into separate files in fasta format. To accomplish this, the entire genome was divided into sub-sequences to match each functional protein. These nucleotide sub-sequences were meticulously retro-translated to ensure the preservation of codon-wise reading frames and to address ribosomal frameshifts, a characteristic feature of CoV genomes. Notably, the first open reading frame (ORF) of CoV involves a ribosomal frameshift, leading to the synthesis of two polyproteins, pp1a and pp1ab [6]. This frameshift has implications for the translation of the landmark protein nsp12, as its N-terminal is encoded at the end of ORF1a, overlapping nsp11, while the remaining coding sequence is shifted by +1 frameshift. In our local database, we have merged nsp11 at the beginning of nsp12 to align with the coding reading frame, resulting in the duplication of the corresponding retro-translated sequence's nth nucleotide. Therefore, both genomic and protein nsp12 sequences are accurately assigned and numbered in our database.

Furthermore, Django utilizes the MySQL database to store all items from the created models, including user request results, which are only accessible for 24 h. VIMVer generates fasta files containing the resulting alignments.

2.2. Search Engine and Alignments

As shown in Figure 1, we are using blastn [4] mainly to confront the query to our database. Blastn parameters are left by default. In the event that the inquiry is too distant from the reference or the inquiry's sequence has too many 'N's (unspecified nucleotide(s)), blastn produces multiple sub-alignments for a hit. Muscle [5] (v3.8) is used to correct the alignment [7]. All alignments are carried out with respect to the reference nucleotide sequences to serve as anchors for the following codon-wise co-translation. Here blastn was selected for its ability to deploy and parameterize, considering the size of the database.

The output is a list of as many hits as blastn found for the maximum sequences of the prepared dataset. Each hit opens a two-frame window that shows on the left panel the protein alignments and on the right panel the nucleotide alignments. Sequences are interactive, so one can follow its position on the sequence using the mouse cursor on the corresponding sequence. Mutations are highlighted by an asterisk under each alignment. Outputs are downloadable in multi-fasta format.

2.3. Accessibility

Each job runs under few seconds and is identified by a key made of 6 printable characters, generated when a query is launched. This job name allows the users to retrieve their run within 24 h. Additional information is available on the project's GitHub.

2.4. Deployment

Viral Instant Mutation Viewer is deployed on a apache2 server hosted from our laboratory server. The ssl protocol was obtained with Encrypt's services. For users or developers who want to install VIMVer locally, the source code is freely available; however, it will not be functional straight away after download. Indeed the settings and configurations file has to be protected. A setting file (setting.py) including an encryption key used to run Django functionality needs to be present in the path. This setting file is generated once by the administrator when creating the VIMVer project within Django. Moreover, it is also within

the setting file that the administrator will define its server setting prior to deployment. Details for installation are provided in the GitHub repository.

3. Results

3.1. Web Interface and Practical Example

VIMVer is a bioinformatics tool that quickly shows mutations on the user's query with the proper numbering both in genome and protein. Our goal is to propose a tool to quickly analyze viral genomes, and for any SARS-CoV-2 sequence of interest, VIMVer rapidly detects mutations and their consequences on the protein structure relative to the Wuhan-1 reference strain. This is particularly important in the case of the frameshift, which obscures the recognition of the 1b fragment. Our web tool is composed of three pages named query, finder, and viewer. VIMVer is accessible through the internet using any web browser.

The user can paste his query in fasta format on the homepage, shown in Figure 2. The inquiry must be a nucleotide sequence possibly from a sequencing result in fasta format, while the length can range from a random recommended minimum of 30 nucleotides, or a single coding sequence, to a whole viral genome. In this version, queries have to be SARS-CoV- or SARS-CoV-2-related to retrieve hits from blastn. When the query is submitted, VIMVer generates a job-key, which is only shown once. On the same page, there is a link to the PANGOLIN [8] web service available though: "https://pangolin.cog-uk.io/" (accessed on 3 May 2023), to retrieve information to determine the lineage of the query sequence.

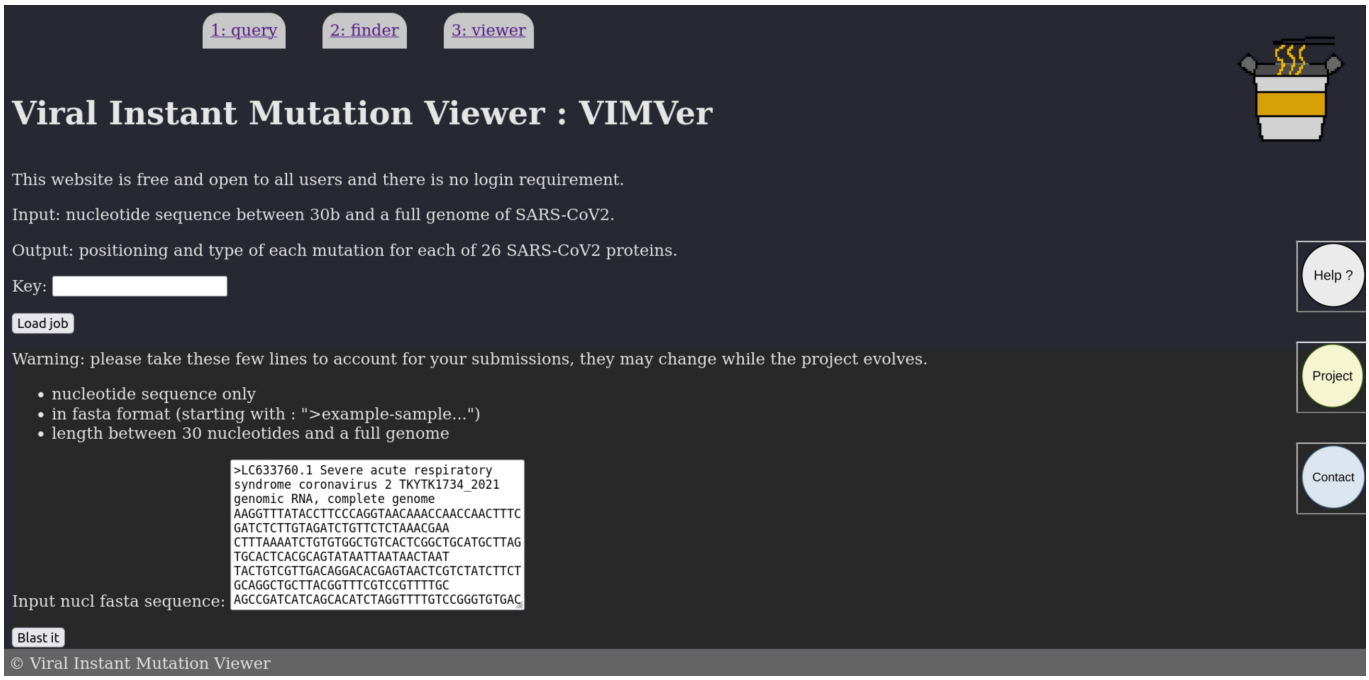


Figure 2. Snapshot of the input form page of VIMVer. A fasta sequence with the header can be pasted into the form section as query.

The result, shown in Figure 3, is returned as a table for which each entry corresponds to a "blastn" hit, described by the score, length, and identity percentage. Considering that our database is small, the E-value is not relevant and therefore not shown, albeit calculated; instead, the score is shown, as it represents the number of aligned residues, which allows us to infer the identity percentage. If the query nucleotide sequence is not a silent mutation, then the corresponding hit is marked as "Mutant" in red. At the end of the page, the user can download the blastn results in fasta format. To access the resulting alignment sampled in Figure 4, click on the 'pick' button on the same line.

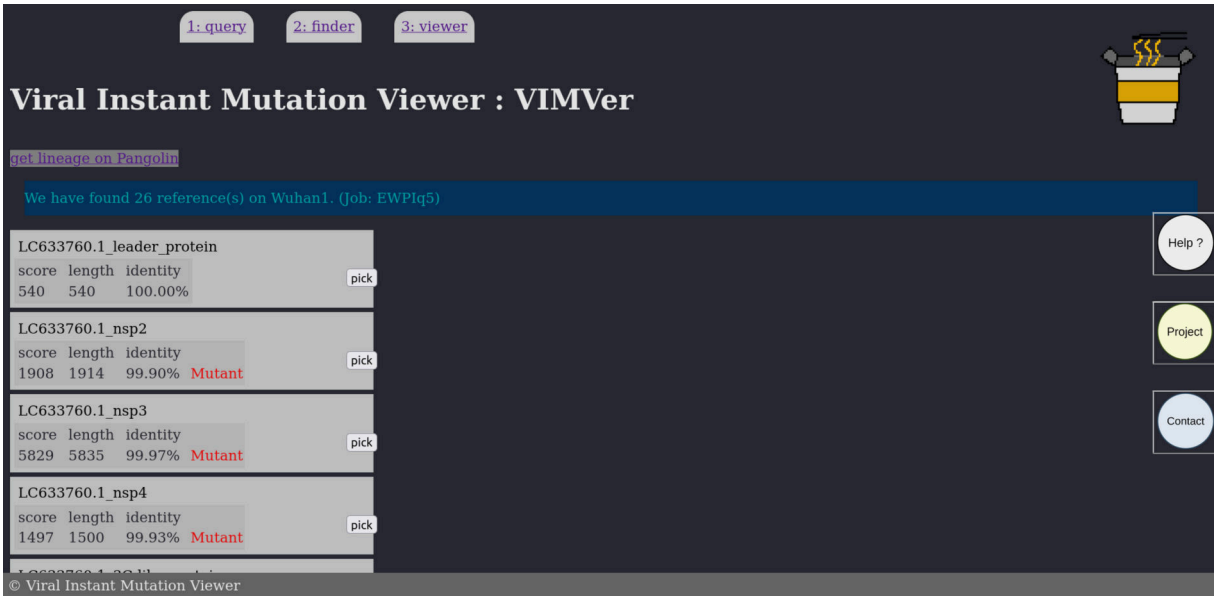


Figure 3. Snapshot of the resulting VIMVer search. The hits are presented as a list with an identification number corresponding to the identified domain. The score, the matching length, and the sequence identity to the reference sequence are shown. If non-silent mutations are detected, then a Mutant mention appears in red. A pick button allows us to access individual results. Response time around 1 s.



Figure 4. VIMVer result page presenting a dual alignment of the GISAID entry EPI_ISL_14769247 on nsp12 Wuhan reference protein (**left**) and genome (**right**). The alignment is partial due to space reasons and was shortened from positions 81 to 281. Shown are protein numbering 1 to 40, 321 to 360, and 921 to 932. Shown are nucleotide numbering 1 to 120, 961 to 1080, and 2761 to 2796. On the left we have the protein side (20 residues per line), on the right the nucleotide side (60 residues per line). Any mutations are marked by '*'. The mouse cursor is dynamic on both alignments, so the user can highlight the amino acid/codon pair. Response time around 2 s.

The VIMVer result page presents a dual alignment of the query against the reference sequence. The left alignment corresponds to the aligned proteins, while the genome alignment is on the right. Any mutations are marked by an asterisk. And all the codons or the amino acids of the same position are correlated when one is hovered over with the mouse cursor. Nucleic residues are colored by the following rule: A red, C green, G yellow, T blue, while proteic residues are colored by hydrophobicity logic as follows: hydrophobic (AILPVM) white, small polar (CSTG) green, positively charged (RHK) blue, negatively charged (DE) red, aromatic (FWY) pink, amphiphilic (NQ) purple.

3.2. Benchmark

As proof of concept, we tested our project in a real-life scenario. To do so, we looked for two well-known proteins: the domain RdRp and the spike protein S.

Indeed, the S protein is the third ORF of the SARS-CoV-2 genome, and the gene most prone to accumulating mutations that influence the infectivity of emerging variant [9]. If one wants to observe and analyze the spike nucleotide and protein sequence evolution straight out of the sequencing campaign or from database analysis, VIMVer makes the task easy and quick.

The nsp12 gene (the main replicative RdRp core) is also a point of interest for our project since it is often reported with incorrect numbering due to the Nidovirales frameshift [6,10].

As an example, a random entry in GISAID (EPI_ISL ID: 14769247) is selected and examined for mutations. We selectively focus on the RdRp domain for the frameshift management and the spike protein for its high mutation susceptibility. The goal is to process a whole genome, which is why the test sequence has to be a complete genome with a high coverage rate. On the query page of VIMVer, a single fasta sequence was pasted in the input text box.

In the alignments of the RdRp domain and the spike protein, shown in Figure 4 and Supplementary Figure S1, respectively, the numbering is respected and each mutation is easily identified as silent or otherwise. In the RdRp example, the frameshift is discretely managed and one non-synonymous mutation (P323L) is found in the RdRp domain. In the spike example, several non-synonymous mutations were found: T19I, L24-, P25-, P26-, A27S, H69-, V70-, G142D, V213G, G339D, S371F, S373P, S375F, T376A, D405N, R408S, K417N, N440K, L452R, S477N, T478K, E484A, F486V, Q498R, N501Y, Y505H, D614G, H655Y, N679K, P681H, N764K, D796Y, Q954H, N969K. All of this indicates that this spike is deviant from the original strain. A quick request to the PANGOLIN web service confirms the lineage: BF10, corresponding to 22B-like (Omicron variant). Indeed, according to nextstrain, the clade 22B includes the lineage BF10, as presented in Figure 5.

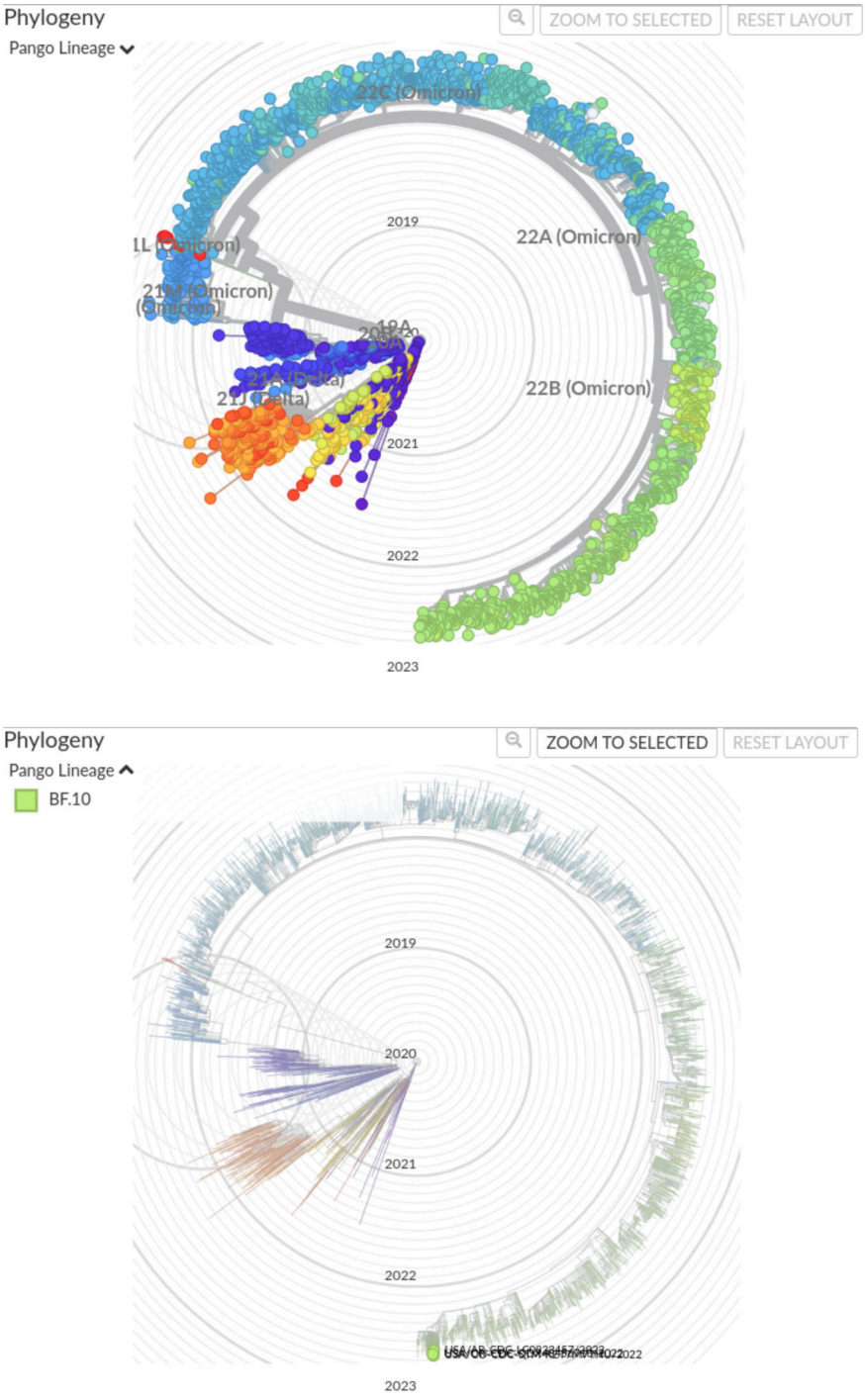


Figure 5. Position of lineage BF10 on the most recent SARS-COV-2 phylogeny (at the date of the writing of the paper): "Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months" (Dataset: ncov, open, global, 6 m; Data range: from 17 December 2019 to 4 September 2022; data form GenBank), carried out on nextstrain.org at "https://nextstrain.org/ncov/open/global/6m?f_pango_lineage=BF.10" (accessed on 9 September 2022). (**top**) All major clades; (**bottom**) focus on BF10 lineage.

4. Discussion

In its final stage, VIMVer is designed as a transversal tool. It should target a wide audience from research studies to clinical applications. It could be implemented as an automatized survey workflow synergized with wet-lab applications and resources, such as

the database on RNA viruses described in the VAZyMolO [11] database used in structural and functional studies [12–19]. The application of VIMVer in the research lab is quite straightforward, as it allows the technician, engineer, or researcher to analyze their data straight out of the sequencer. It provides fundamental insight that allows us to sort out the emergence of new data. This quick identification can provide rapid annotation for new genomes which can be exploited by the epidemiologist or medical expert. The double nucleic/proteic interface facilitates the interpretation of a given mutation for any biologist that is neither a biochemist nor bioinformatics expert. Further developments concerning the online VIMVer include extending the reference database to the annotated sequences of VAZyMolO. The latter is a highly curated database of soluble functional or structural protein domains. The combination of both would then allow us to not only spot an emerging mutation but also to correlate it with structural information and compare it within a defined domain family.

The connection of VIMVer to protein structure analysis software, such as AlphaFold [20] or other structural prediction and analysis tools, should facilitate the connection of genotype to phenotype for emerging virus gene products. The main spirit of VIMVer is to propose a deep analysis of each mutation found in the user's genomic sequence of interest against the corresponding completely annotated reference. Considering its specific conception, VIMVer diverges from existing web services, such as NextClade [3], which does not provide a specific annotated protein-related dataset.

5. Conclusions

Viral Instant Mutation Viewer (VIMVer) is a sequence analysis tool of potential use to the community deciphering genomic sequences and connecting them to the phenotype. Our workflow allows the instant visualization of SARS-CoV-2 mutations on both genomic and protein sequences with proper numbering. It delivers an interactive alignment and allows rapid lineage identification that could be easily used in genomic and epidemiological reports. VIMVer can be easily implemented in a larger infrastructure or adapted to all RNA viruses and is interoperable with already existing workflows, as it is able to deliver results in a few seconds, even using a modest computational infrastructure. At the time of writing, the pandemic is now over but still drives us to better prepare ourselves for the next episode. Thus, VIMVer contributes to existing services to better serve the scientific community and allow it to quickly analyze mutational changes detected in the enormous wealth of data generated daily.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v15081628/s1, Figure S1: Alignment of the GISAID entry EPI_ISL_14769247 on spike protein against Wuhan reference (NC_045512.2).

Author Contributions: Conceptualization, V.W., B.C. and F.F.; Funding acquisition, B.C. and F.F.; Methodology, V.W.; Supervision, B.C. and F.F.; Writing—original draft, V.W. and F.F.; Writing—review and editing, V.W., B.C. and F.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fondation de la Recherche Médicale (Aide aux équipes), and the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101005077 (IMI-CARE).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.

Data Availability Statement: VIMVer is available through a web service (https://vimver.afmb.univ-mrs.fr/). The code and workflow documentation is available at https://github.com/wildevince/VIMVer (accessed on 19 April 2022). The project is under a free software license CeCILL (see more: http://www.cecill.info).

Acknowledgments: The authors acknowledge F. Debarre from "Ecole Normale Supérieur" IESS, Paris for useful discussion.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VIMVer	Viral Instant Mutation Viewer
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
COVID-19	Coronavirus disease 2019
CoV	Coronavirus
NCBI	National center of biotechnology information
GISAID	Global Initiative on Sharing All Influenza Data
NGS	next-generation sequencing
VOCs	variants of concern
ORF#	open reading frame #number
pp#	polyprotein #number
nsp#	non-structural protein #number
ssl	Secure Sockets Layer certificate
RdRp	RNA-dependant RNA-polymerase
RNA	ribonucleic acid

References

1. Sohrabi, C.; Alsafi, Z.; O'Neill, N.; Khan, M.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* **2020**, *76*, 71–76. [CrossRef] [PubMed]
2. Fehr, A.R.; Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Coronaviruses* **2015**, *1282*, 1. [CrossRef]
3. Aksamentov, I.; Roemer, C.; Hodcroft, E.; Neher, R. Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **2021**, *6*, 3773. [CrossRef]
4. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef] [PubMed]
5. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef] [PubMed]
6. Bhatt, P.R.; Scaiola, A.; Loughran, G.; Leibundgut, M.; Kratzel, A.; Meurs, R.; Dreos, R.; O'Connor, K.M.; McMillan, A.; Bode, J.W.; et al. Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science* **2021**, *372*, 1306–1313. [CrossRef]
7. Nuin, P.A.S.; Wang, Z.; Tillier, E.R.M. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinf.* **2006**, *7*, 471. [CrossRef] [PubMed]
8. O'Toole, Á.; Pybus, O.G.; Abram, M.E.; Kelly, E.J.; Rambaut, A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genom.* **2022**, *23*, 121. [CrossRef] [PubMed]
9. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; COVID-19 Genomics UK (COG-UK) Consortium; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **2021**, *19*, 409. [CrossRef] [PubMed]
10. Plant, E.P.; Dinman, J.D. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front. Biosci. J. Virtual Libr.* **2023**, *13*, 4873. [CrossRef] [PubMed]
11. Ferron, F.; Rancurel, C.; Longhi, S.; Cambillau, C.; Henrissat, B.; Canard, B. VaZyMolO: A tool to define and classify modularity in viral proteins. *J. Gen. Virol.* **2005**, *86*, 743–749. [CrossRef] [PubMed]
12. Uversky, V.N.; Dunker, A.K. Understanding Protein Non-Folding. *Biochim. Biophys. Acta* **2010**, *1804*, 1231. [CrossRef]
13. Gorbalenya, A.E.; Lieutaud, P.; Harris, M.R.; Coutard, B.; Canard, B.; Kleywegt, G.J.; Kravchenko, A.A.; Samborskiy, D.V.; Sidorov, I.A.; Leontovich, A.M.; et al. Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antivir. Res.* **2010**, *87*, 95. [CrossRef]
14. Imbert, I.; Snijder, E.J.; Dimitrova, M.; Guillemot, J.C.; Lécine, P.; Canard, B. The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. *Virus Res.* **2008**, *133*, 136. [CrossRef] [PubMed]
15. Coutard, B.; Gorbalenya, A.E.; Snijder, E.J.; Leontovich, A.M.; Poupon, A.; De Lamballerie, X.; Charrel, R.; Gould, E.A.; Gunther, S.; Norder, H.; et al. The VIZIER project: Preparedness against pathogenic RNA viruses. *Antivir. Res.* **2008**, *78*, 37. [CrossRef] [PubMed]
16. John, S.P.; Wang, T.; Steffen, S.; Longhi, S.; Schmaljohn, C.S.; Jonsson, C.B. Ebola Virus VP30 Is an RNA Binding Protein. *J. Virol.* **2007**, *81*, 8967. [CrossRef] [PubMed]
17. Radivojac, P.; Iakoucheva, L.M.; Oldfield, C.J.; Obradovic, Z.; Uversky, V.N.; Dunker, A.K. Intrinsic Disorder and Functional Proteomics. *Biophys. J.* **2007**, *92*, 1439. [CrossRef]

18. Habchi, J.; Mamelli, L.; Darbon, H.; Longhi, S. Structural Disorder within Henipavirus Nucleoprotein and Phosphoprotein: From Predictions to Experimental Assessment. *PLoS ONE* **2010**, *5*, e11684. [CrossRef]
19. Ogola, E.O.; Kopp, A.; Bastos, A.D.S.; Slothouwer, I.; Marklewitz, M.; Omoga, D.; Rotich, G.; Getugi, C.; Sang, R.; Torto, B.; et al. Jingmen Tick Virus in Ticks from Kenya. *Viruses* **2022**, *14*, 1041. [CrossRef] [PubMed]
20. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

2.3. Discussion

Le projet VIMVer a été initié lors de mon stage de master M2, suivant la dynamique du laboratoire face à la pandémie de COVID-19. Il a conduit à la définition de la première problématique de ma thèse, visant à répondre au besoin d'identifier et de localiser simultanément les mutations sur les séquences génomiques et protéiques virales.

Ce projet m'a permis de développer des compétences en développement web, notamment avec Django, un framework Python pour les services web. Pour ma thèse, VIMVer apporte une solution :

— accessible via un navigateur,

— visuelle grâce à des figures générées à la volée,

— rapide et fournissant des résultats en moins de 3 minutes,

L'investissement nécessaire en recherches personnelles sur les divers outils de bioinformatique et de développement web. a résulté en une expérience enrichissante et formatrice.

Depuis son lancement et la soumission de l'article, VIMVer a évolué pour atteindre la version 1.0, avec une mise à jour vers la version 1.1 en préparation. Ces améliorations reflètent mon expérience croissante en programmation web, notamment en matière de sécurité et d'optimisation du *backend*. Le projet nécessite encore du temps pour atteindre son plein potentiel. Des retours d'utilisation, mesurés par la fréquence des visites, permettront d'ajuster la capacité serveur nécessaire pour optimiser ses performances.

En parallèle, une réflexion est en cours sur une potentielle fusion avec le projet VAzyMolo2, visant à créer un projet global à l'échelle de l'équipe. Cette initiative s'accompagne d'efforts de diffusion, d'assistance et de formation aux pratiques simples en bioinformatique au sein du laboratoire.

Deuxième Partie

Chapitre 3

Développement d'outils pour organiser
et cartographier le savoir.

3. *VAZyMolO-2* :

Une cartographie des génomes viraux à portée de main

Manuscrit en cours de rédaction.

Ce chapitre est dédié au projet **VAZyMolO-2**, qui constitue le cœur même de ce travail. Il répond pleinement à mon deuxième objectif : collecter, caractériser et cartographier des données expérimentales qui servent directement les études de notre équipe de recherche.

Sommaire

3.1. Introduction . . . 97
3.1.1. Origines et inspirations de VAZyMolO-2 . . . 97
3.1.2. La modularité dans les protéines virales . . . 98
3.2. Matériels et Méthodes . . . 106
3.2.1. Le cadre logiciel . . . 106
3.2.2. La structure de la base de donnée . . . 108
3.2.3. La recherche de motifs . . . 112
3.2.4. Homologie de séquence . . . 113
3.2.5. Construction des profils de séquence . . . 113
3.2.6. Taxonomie . . . 115
3.2.7. La génération des figures . . . 115
3.2.8. La recherche par mots-clefs . . . 118
3.3. Description de VAZyMolO-2 . . . 120
3.3.1. Le jeu de données test . . . 121
3.3.2. Les utilisateurs . . . 123
3.3.3. Les types d'entrées (input) . . . 123
3.3.4. L'analytique et l'utilisation des données . . . 124
3.4. Discussion . . . 127

3.1. Introduction

3.1.1. Origines et inspirations de VAZyMolO-2

La première moitié de la thèse a été consacrée au développement et à la diffusion du projet VIMVer, un outil web innovant pour l'identification et l'analyse des mutations virales, principalement axé sur le génome du SARS-CoV-2. Cette période a permis d'approfondir mes compétences en programmation et en biologie structurale appliquée à la virologie.

Au cours de cette phase, mes interactions avec divers projets de l'équipe ont révélé une frustration commune : la difficulté à extraire des données pertinentes d'un flux constant et désordonné d'informations. Cette observation a suscité ma curiosité et m'a conduit à définir un second objectif pour ma thèse : développer un outil pour organiser et cartographier le savoir maîtrisé par l'équipe, facilitant ainsi sa diffusion au sein de la communauté scientifique.

Ce nouvel objectif vise à répondre aux défis de gestion des connaissances dans un domaine en rapide évolution, comme l'illustre l'expérience acquise avec VIMVer. L'outil envisagé pourrait s'appuyer sur les mêmes principes, tels que l'accessibilité via une interface web et la visualisation intuitive des données, tout en les étendant à un spectre plus large de connaissances en virologie et en biologie structurale.

L'inspiration pour VAZyMolO-2 provient d'un projet antérieur développé par mon directeur, et superviseur, de thèse : VAZyMolO. Cet outil bioinformatique, en service de 2005 à 2011, proposait un service d'annotation de structure protéique pour une large variété de familles de virus à ARN.

VAZyMolO, l'original (FERRON, RANCUREL et al. 2005), visait à définir et classifier la modularité dans les protéines virales. Il permettait d'identifier des modules protéiques viraux susceptibles d'être exprimés sous une forme soluble et fonctionnellement active. Cette approche a contribué à la cartographie de génomes viraux, notamment celui du virus du Syndrome Respiratoire Aigu Sévère (SRAS-CoV), et a facilité la résolution structurale de certaines protéines virales (LIEUTAUD et al. 2008 ; COUTARD et al. 2010 ; GORBALENYA et al. 2010b).

L'outil combinait l'analyse de séquences, l'intégration de données structurales exis-

tantes et l'incorporation d'informations bibliographiques pour définir des modules fonctionnels et structuraux au sein des protéines virales. Cette approche multidimensionnelle a jeté les bases pour combler les lacunes dans l'annotation des génomes viraux.

VAZyMolO-2, inspiré par son prédécesseur, vise à étendre ces principes à un spectre plus large de connaissances en virologie et en biologie structurale, tout en intégrant les avancées technologiques récentes en matière d'accessibilité web et de visualisation intuitive des données.

3.1.2. La modularité dans les protéines virales

La validation et le test de la modularité sur le génome du SARS-CoV-2 constituaient une partie essentielle du travail de conception. Ce choix s'est avéré judicieux en raison de l'importance actuelle du virus et de l'abondance de données disponibles.

Cette démarche comprenait plusieurs aspects clefs :

— La délimitation précise des domaines fonctionnels sur les séquences protéiques

— L'identification des résidus structuraux conservés et la déduction des motifs sur la séquence

— La caractérisation des fonctions virales dans leur contexte cellulaire et leurs interactions avec l'hôte

— L'analyse de l'histoire évolutive des domaines fonctionnels

— La caractérisation et la source des données utilisées

— La maîtrise des méthodes analytiques pour exploiter ces données

— L'examen du code source du projet original

L'approche adoptée visait à intégrer les perspectives du structuraliste, du biologiste et du bioinformaticien, soulignant ainsi l'importance d'une vision multidisciplinaire pour comprendre la modularité virale.

Suite à cette approche multidisciplinaire, je me suis approprié le concept de "modulo". Cet objet représente l'association entre trois éléments clefs d'une protéine virale :

1. Sa structure tridimensionnelle
2. Son profile de séquences
3. Ses motifs caractéristiques

Cette définition du "modulo" permet d'intégrer les aspects structuraux, séquentiels et fonctionnels des protéines virales en une seule entité conceptuelle, facilitant ainsi l'analyse de la modularité dans les génomes viraux.

Pour illustrer le concept de modulo à travers deux exemples pertinents dans le contexte des nidovirus, nous pouvons examiner les méthyltransférases et le rôle de la polymérase dans le complexe de réplication.

3.1.2.1. Le rôle de la polymérase dans le complexe de réplication

La polymérase virale, également appelée ARN polymérase ARN-dépendante (RdRp), est un composant essentiel du complexe de réplication des virus à ARN, y compris les nidovirus. Cette enzyme joue un rôle crucial dans la réplication du génome viral et présente une structure tridimensionnelle caractéristique en forme de main (voir la figure 3.1 ci-dessous et la figure 3.3 plus bas).

La structure de la RdRp comprend trois sous-domaines principaux :

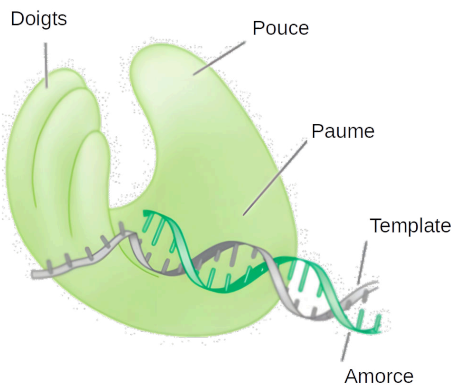


FIGURE 3.1. – La main droite

1. Le sous-domaine "paume" : C'est le cœur catalytique de l'enzyme, contenant les résidus essentiels pour la synthèse de l'ARN.
2. Le sous-domaine "doigts" : Il interagit avec l'ARN matrice et les nucléotides entrants.
3. Le sous-domaine "pouce" : Il joue un rôle dans le positionnement de l'ARN matrice et la processivité de l'enzyme.

Cette structure en forme de main est hautement conservée parmi les virus à ARN, soulignant son importance fonctionnelle. La conservation de cette structure tertiaire est cruciale pour maintenir l'activité enzymatique de la polymérase (PEERSEN 2019).

Dans le contexte du complexe de réplication, la RdRp interagit avec d'autres protéines non structurales (nsp) pour former le complexe de réplication-transcription (RTC). Par exemple, chez les coronavirus, la RdRp (nsp12) interagit avec nsp7 et nsp8 pour former un complexe plus efficace pour la synthèse d'ARN : le RTC minimal (voir la figure 3.2) ci-dessous et la figure 3.4 plus bas).

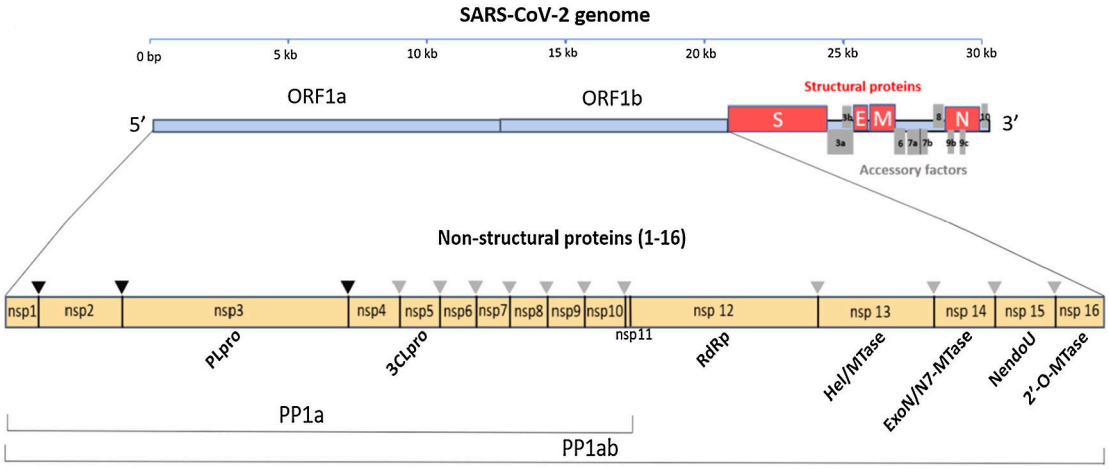


FIGURE 3.2. – Le génome du SARS-CoV-2 : les protéines du RTC
Figure adaptée de l'article suivant (ROMANO et al. 2020).

Chez les nidovirus, le domaine fonctionnel de la RdRp (ARN polymérase ARN-dépendante) est effectivement bien conservé dans sa séquence. Ce clade des *Nidovirales*, qui comprend notamment les *Coronaviridae*, les *Mesoniviridae*, les *Tobaniviridae* et les *Arteriviridae*, présente une forte conservation de la RdRp. Toutes les séquences de polymérase au sein de ce clade constituent un profile de séquences conséquent, permettant une analyse approfondie de la structure et de la fonction de cette enzyme. L'analyse structurale de l'alignement de ces séquences, combinée aux données de la littérature (SHANNON, LE et al. 2020), permet d'identifier des motifs conservés essentiels à la fonction de la RdRp.

Ainsi l'association des informations sur la structure, le profile de séquence et la détermination des motifs résume la somme du savoir sur un domaine fonctionnel : autrement dit son modulo.

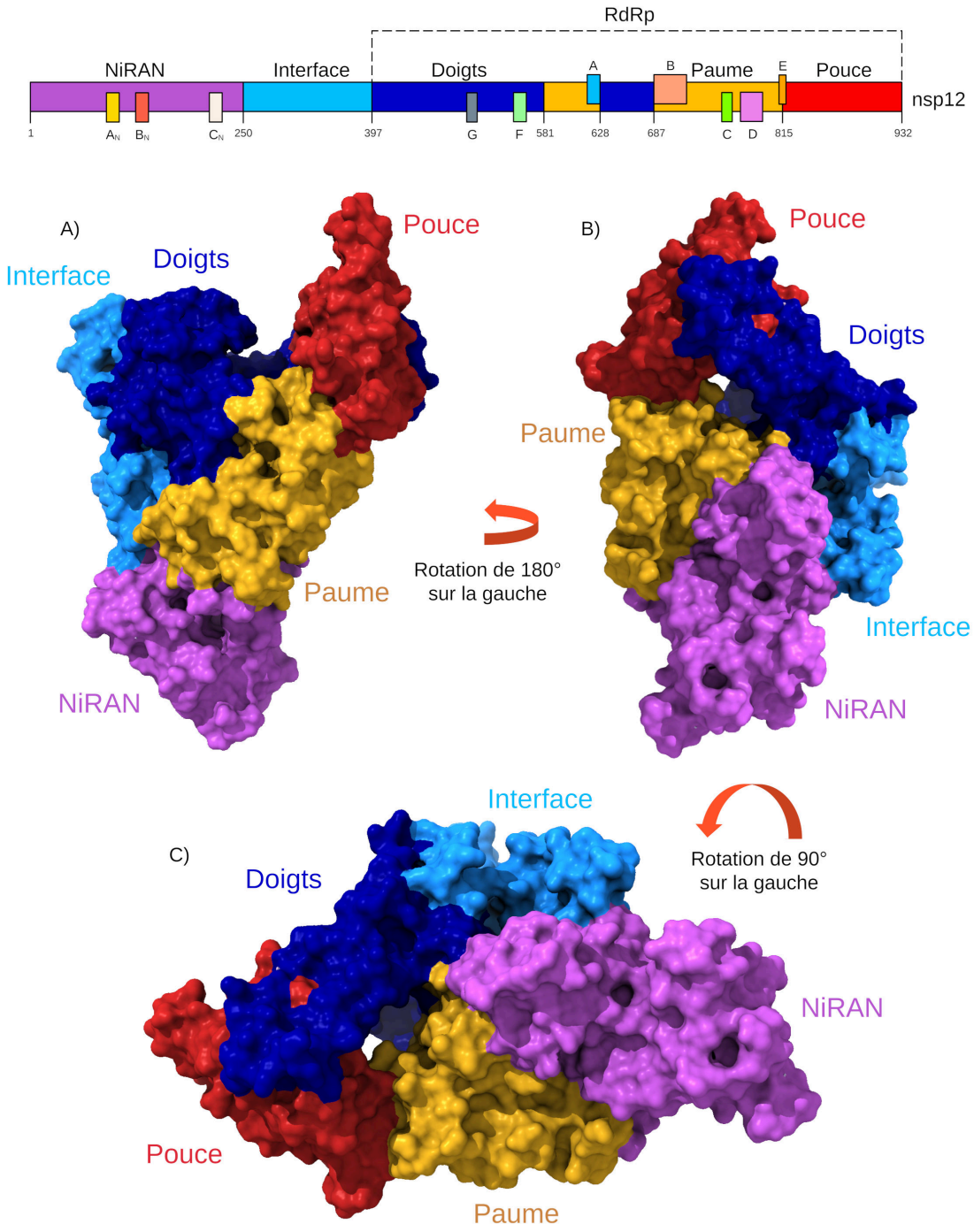


FIGURE 3.3. – L'architecture et structure de la nsp12 du SARS-CoV-2
Le schéma au-dessus représente les domaines : NiRAN en violet, polymérase (RdRp) et ses sous-domaines. La structure peut être observée sur 3 angles, elle a la forme d'une main droite.

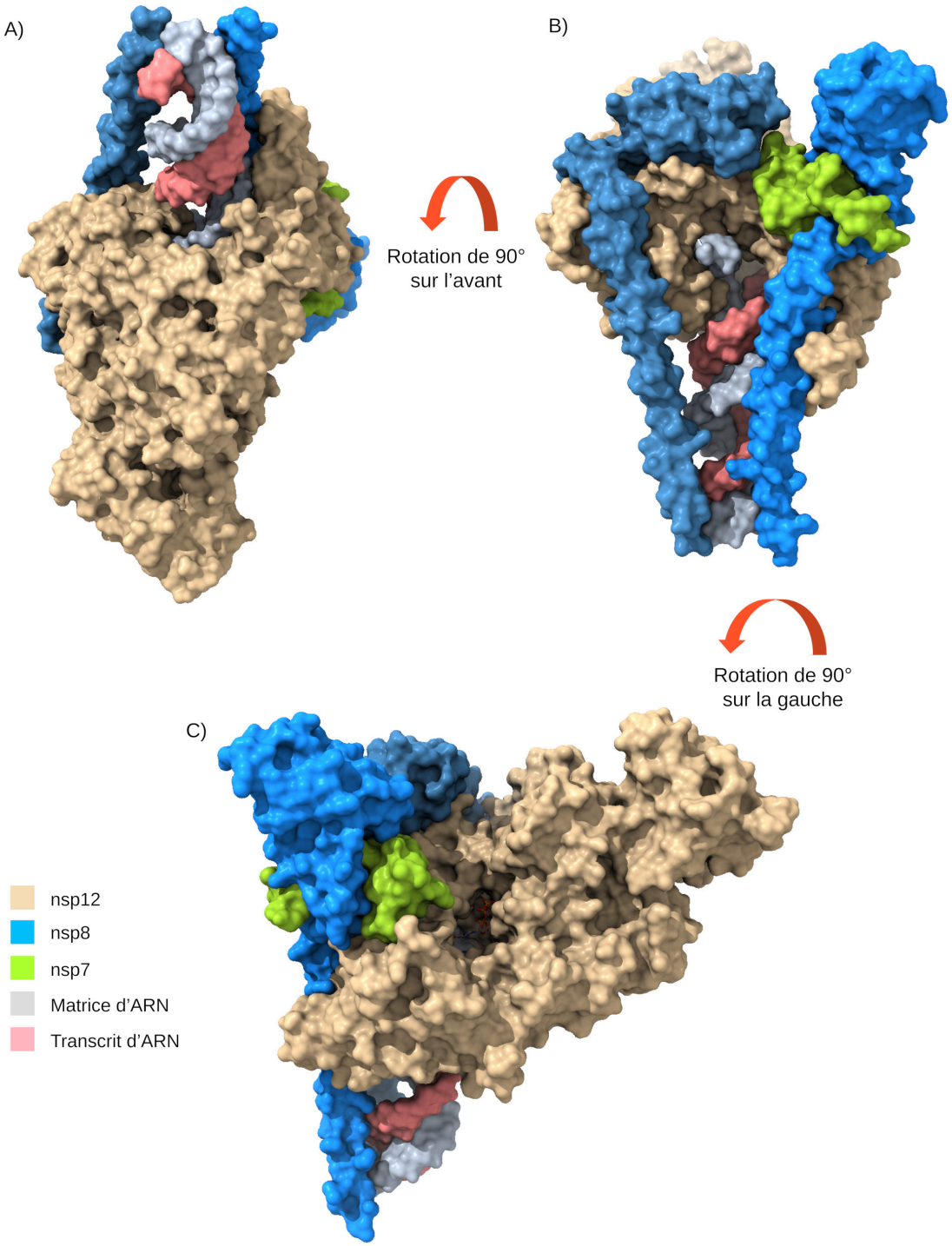


FIGURE 3.4. – Le complexe de réplication (RTC) minimal du SARS-CoV-2
Il s'agit de l'assiaciation de la nsp12, nsp7, deux nsp8 et deux brins d'ARN.

3.1.2.2. Les méthyltransférases des nidovirus

Les méthyltransférases (MTases) des nidovirus jouent un rôle crucial dans la stabilité de l'ARNm viral et l'échappement immunitaire chez les nidovirus. Elles sont impliquées dans la formation de la coiffe des ARN viraux, un processus essentiel pour la protection et l'utilisation des extrémités des ARN viraux (DECROLY et al. 2011).

Chez les nidovirus, on distingue principalement deux types de MTases :

1. N7-méthyltransférase (N7-MTases) :
   — Locus : nsp14, dans l'ORF1b, des *Coronaviridae*
   — Activité : méthylation de la coiffe en position N7 du résidu
   — Phylogénie : *Coronaviridae* et *Mesoniviridae*
2. 2'-O-méthyltransférase (2'-O-MTases) :
   — Locus : nsp16, dans l'ORF1b, des *Coronaviridae*
   — Activité : méthylation de la coiffe en position 2'O du résidu
   — Phylogénie : *Tobaniviridae*, *Coronaviridae* et *Mesoniviridae*
   — Structure : Repliement de type *Rossmann-like fold* (SCHUBERT et al. 2003 ; SCHUBERT et al. 2003)

Récemment, une nouvelle MTase potentielle a été découverte dans le génome des *Tobaniviridae* (FERRON, DEBAT et al. 2020) :

3. putative N7-Mtase :
   — Locus : dans l'ORF1a des *Tobaniviridae*
   — Validation expérimental : SHANNON, SAMA et al. 2022,
   — Activité : méthylation de la coiffe en position N7 du résidu
   — Structure : Repliement de type *Rossmann-like fold*, inhabituel pour une N7-MTase virale (SCHUBERT et al. 2003 ; SCHUBERT et al. 2003),
   — profile de séquence : Distinct des N7-MTases, comme des 2'-O-MTases, connues chez les nidovirus
   — Motifs caractéristiques : Activité N7-MTase spécifique, localisation génomique unique

La modularité peut apporter une réponse intéressante concernant cette nouvelle N7-MTase des *Tobaniviridae*. En effet, son repliement de type *Rossmann*, caractéristique des 2'-O-MTases mais inhabituel pour une N7-MTase virale, suggère une évolution convergente ou une adaptation fonctionnelle unique. Sa localisation dans l'ORF1a,

différente des autres N7-MTases virales, indique une organisation génomique distincte qui pourrait être liée à des contraintes évolutives spécifiques aux *Tobaniviridae*.

Cette découverte illustre la diversité structurale et fonctionnelle des MTases chez les nidovirus, soulignant l'importance d'une approche modulaire pour comprendre leur évolution et leur fonction. La caractérisation de cette nouvelle N7-MTase pourrait non seulement clarifier la classification au sein de l'ordre des *Nidovirales*, mais aussi aider à l'identification et à l'attribution de MTases dans d'autres domaines de la vie.

3.2. Matériels et Méthodes

Ce projet de thèse a évolué à partir du projet original VAZyMolO, en intégrant ses concepts clés ainsi que les données accumulées au fil du temps. L'objectif principal est de refondre et de mettre à jour l'ensemble de l'architecture, tout en respectant les principes fondamentaux établis.

VAZyMolO-2 est conçu pour allier la bioinformatique à la biologie structurale appliquée à la virologie. Il vise à fournir une aide et un soutien fonctionnel tant aux bioinformaticiens qu'aux biologistes.

Étant actuellement en cours de développement, cette section de la thèse présentera tous les composants du projet, qu'ils soient complétés ou non. Cette approche offre une vue d'ensemble de l'état actuel du projet ainsi que des orientations à court terme.

3.2.1. Le cadre logiciel

VAZyMolO-2 est un service web en phase précoce de développement, conçu sous Ubuntu 24.04 LTS. Actuellement, il est hébergé sur un serveur local. Cette application est développée avec Python (3.10.12) et Django (4.2), auxquels sont associés plusieurs autres paquets : Biopython (1.81), Django Extensions (3.2.3), Django MySQL (4.9.0), Matplotlib (3.7.1), NumPy (1.24.2), Pandas (2.0.2), Pillow (9.0.1) et Whitenoise (6.4.0).

Le frontend et le backend sont tous deux gérés par des scripts que j'ai conçus en Python, intégrés à l'architecture de Django.

Pour gérer les requêtes courantes et les sessions utilisateur, VAZyMolO-2 s'interface nativement avec une base de données locale sous MySQL, qui sert à la gestion des données temporaires. Cette application utilise des cookies, essentiels à son bon fonctionnement.

Il n'y a pas d'identification de session utilisateur afin d'améliorer l'ergonomie. À la place, les sessions sont reconnues automatiquement grâce à des cookies temporaires. En effet, un identifiant est généré et envoyé côté client via un unique cookie à courte durée de vie. Ce petit fichier temporaire est stocké et accessible uniquement par le navigateur client, avec une recommandation particulière pour **Firefox**.

En parallèle du cookie de session, un nouvel objet **session** est créé spécifiquement

pour répondre à cette clé de session. Ainsi, cet objet **session** est conçu pour suivre uniquement la progression de la session utilisateur et se met à jour à chaque interaction. Sa durée de vie est limitée à 24 heures consécutives. Il est donc conseillé de ne pas supprimer les cookies du navigateur pendant une session, car cela pourrait entraîner des erreurs non encore traitées et compromettre la session en cours. Toutes les traces de session seront détruites 24 heures après leur création, même si l'onglet du navigateur reste ouvert côté client, sans message ni alerte indiquant la fin d'une session.

L'intégralité du code source a été rendue publiquement accessible pendant la durée de l'examen sur GitHub à l'adresse suivante : wildevince/Testymolo.

3.2.1.1. Le déploiement

VAZyMolO-2 adopte une stratégie de déploiement similaire à celle de VIMVer. L'application sera mise en ligne sur un serveur géré par le laboratoire, mais uniquement après avoir atteint un certain seuil critique de développement. Cette approche garantit que le système est suffisamment mature et stable avant d'être rendu accessible sur l'infrastructure du laboratoire.

3.2.2. La structure de la base de donnée

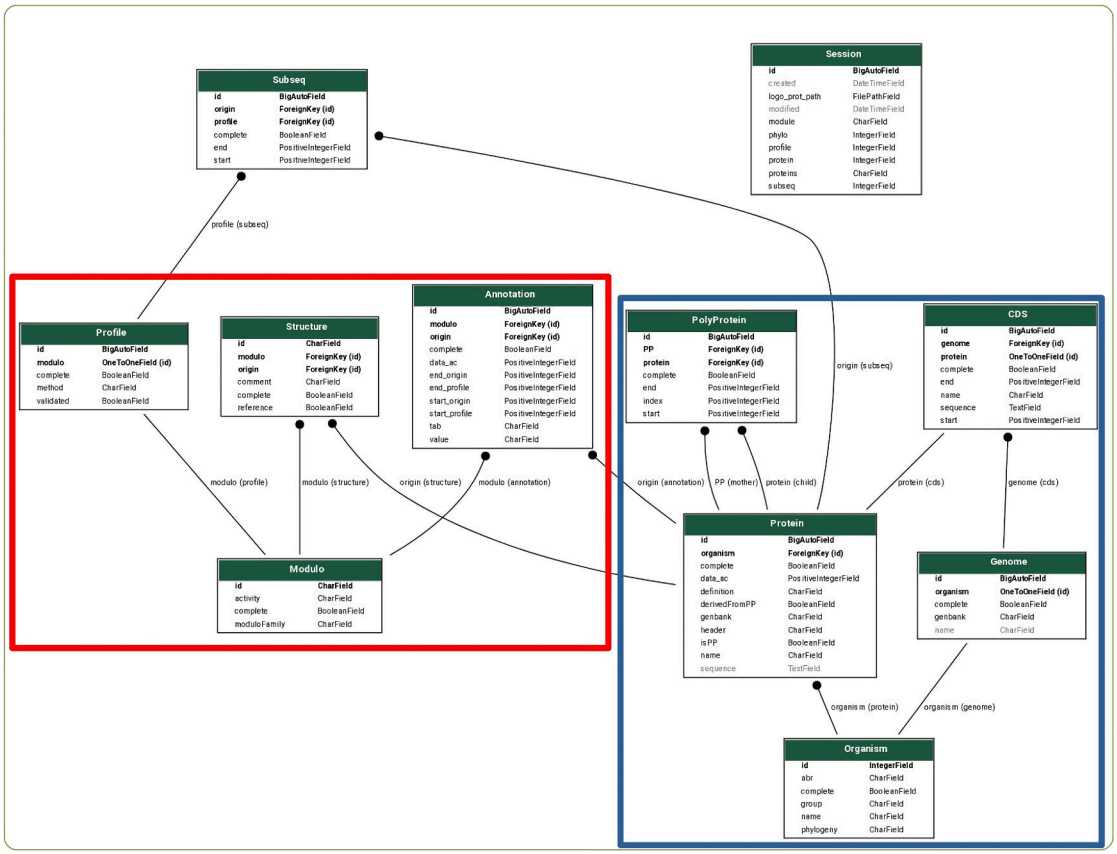


FIGURE 3.5. – Diagramme de classe : structure de *vazy2*
Le diagramme est généré avec *graphviz* depuis le fichier "models.py" qui structure *vazy2*. Le cadre bleu décrit la méta-donnée d'une séquence. La classe MODULO est défini par le cadre rouge, tel que :

$$modulo = structure + profile + motifs$$

La structure de la base de données a été conçue et adaptée pour le service MySQL, qui est natif dans le projet Django. Ainsi, toutes les relations entre les classes SQL sont établies par le biais de clés étrangères simples, d'agrégations ou de compositions. La figure 3.5 ci-dessus illustre ces relations logiques.

En parallèle des sauvegardes du fichier en MySQL, les données sensibles sont également enregistrées au format JSON, avec un fichier JSON par classe d'objet. Au sein de chacune de ces classes, des méthodes de sérialisation assurent les conversions de formats. Cela facilite la gestion des modifications, ajouts et mises à jour pendant le

développement de l'outil et l'implémentation progressive de la base de données.

Au centre, la classe **subseq** fait le lien entre ces deux parties. Comme montré dans la figure 3.6, un objet **subseq** représente une sous-séquence portant un domaine fonctionnel. Cet objet détaille la position sur la protéine et fait référence au modulo correspondant.

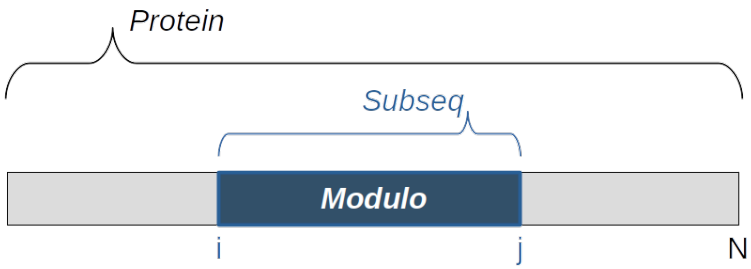


FIGURE 3.6. – La définition de *Subseq*
Soit une séquence d'une protéine (*Protein*) de longueur N, qui porte un domaine fonctionnel (*Modulo*), alors *Subseq* définit la sous-séquence de i à j positionnant le domaine sur cette protéine. Sachant que : 0 <= i, j <= N.

3.2.2.1. La récupération de $vaz1$

Un effort considérable a été fait pour extraire, caractériser et mettre à jour les données essentielles de l'ancien VAZyMolO.

Une fois le code source original en main, deux fichiers SQL ont été extraits. L'un de ces fichiers contenait l'architecture de la base de données originale, c'est-à-dire la description de chacune des 22 classes. L'autre fichier était une liste de toutes les entrées qui constituaient la base de données depuis sa création entre 2003 et 2004. DBEAVER a été utilisé pour lire et visualiser cette structure SQL.

Par la suite, une longue période a été consacrée à comprendre, qualifier et restaurer toute information manquante. L'architecture de VAZyMolO-2 a été rapidement finalisée après cela.

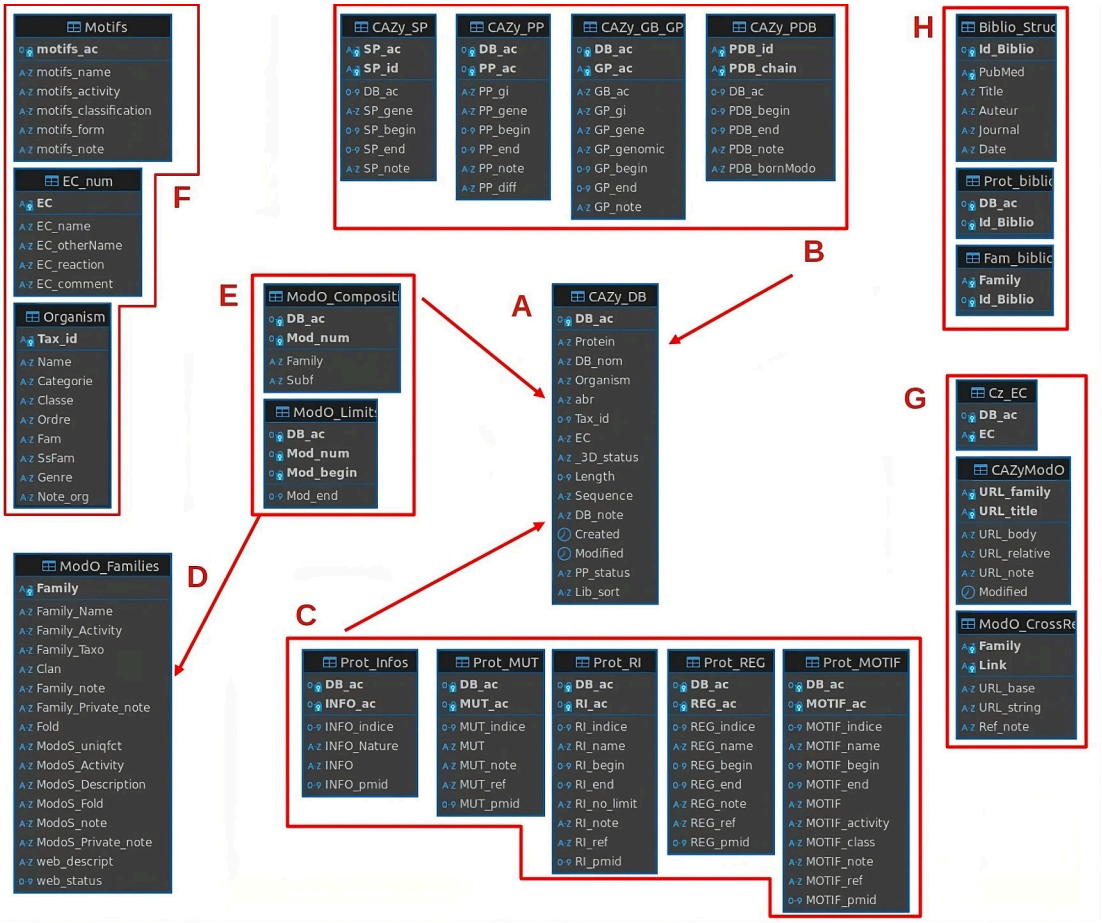


FIGURE 3.7. – Architecture of *vaz*1 dataset - UML

A : la séquence protéique
B : la référence genbank de la séquence
C : l'annotation de séquence (mutation, région désordonné, motif, région d'interaction, ou autre commentaire)
D : le domaine
E : l'équivalent de *Subseq*
F : la description des motifs, la classification des activités virales (code EC), la taxonomie des virus
G : des références à des bases de données extérieures, telles que EC, l'ancien CAZy ou autre.
H : des références à la littérature.
Le diagramme est généré avec DBeaver. L'image est transformée avec *Upscale*

Jusqu'à présent, la récupération des données s'est concentrée sur l'ordre des *Nidovirales*, avec une exploration en cours au sein de la sous-famille des *Orthocoronavirinae*. Pour ce faire, une série de scripts Python a été développée au format *Jupyter Notebook*. La procédure de récupération a été structurée en trois étapes distinctes.

En tant qu'exigence préalable, les deux fichiers SQL originaux ont été convertis en un ensemble de 22 dictionnaires imbriqués, représentant l'architecture de VAZyMolO. Ces "super-dictionnaires" ont ensuite été transformés en tables CSV, considérées comme le jeu de données de VAZyMolO.

Pseudo-code 1 Extraction de la méta-donnée des protéines

Require: all 22 tables of $vaz1$ ▷ as nested dictionaries from previous csv parsing
1: $metadata$: nested dictionaries
2: **for each** protein entry **in** *DB-ac* table **do**
3: get protein id as $prot_ID$
4: **while** parsing all other tables **do**
5: **if** related to $prot_ID$ **then**
6: saving corresponding $metadata$
7: **end if**
8: **end while**
9: **end for**
10: **Output :** protein $metadata$ in $json$ format

La première étape, décrite dans l'algorithme 1, consiste à extraire les métadonnées du côté des séquences protéiques (classe **DB-ac** dans VAZyMolO).

Ensuite, la deuxième partie, présentée dans l'algorithme 2, vise à extraire les méta-données du côté des domaines fonctionnels (classe **mod-family** dans VAZyMolO).

Pseudo-code 2 Extraction de la méta-donnée des domaines fonctionnels

Require: all 22 tables of $vaz1$ ▷ as nested dictionaries from previous csv parsing
1: $metadata$: nested dictionaries
2: **for each** $modulo$ **in** *mod-family* table **do**
3: get $modulo$ id as mod_ID ▷ $modulo$ as functional domain
4: **while** parsing all other tables **do**
5: **if** related to mod_ID **then**
6: saving corresponding $metadata$
7: **end if**
8: **end while**
9: **end for**
10: **Output :** $modulo\ metadata$ in $json$ format

TABLEAU 3.1. – Certains champs liés à des sources externes dans VAZyMolO-2

Classe	Champs	Clef primaire	Description	lien externe
Organism	id	oui	numéro taxonomique des espèces	NCBI *taxonomy*
Modulo	activity	non	classification EC	ExplorEnz
Protein	genbank	non	numéro d'accession genbank	NCBI *protein*
Structure	id	oui	numéro d'accession PDB	PDB

Dans la dernière étape, les deux sorties précédentes sont analysées ensemble afin de construire un ensemble de séquences pour chaque **modulo**, comme référencé dans **Mod-limits** dans VAZyMolO, qui répertorie toutes les occurrences de chaque **modulo** dans n'importe quelle séquence protéique. Ces séquences protéiques sont ensuite alignées par ensembles pour constituer un profile préétabli pour chaque **modulo**.

L'objectif de ce pipeline est de reformater les données sensibles de VAZyMolO afin qu'elles correspondent à l'architecture de VAZyMolO-2. Toute addition provenant de VAZyMolO dans VAZyMolO-2 doit également être mise à jour directement à partir de ressources externes, comme décrit dans le tableau 3.1. Je reviendrai sur la mise à jour des taxons plus tard, en 3.2.6.

3.2.2.2. Ressources Externes - Interopérabilité

Pour garantir l'interopérabilité, VAZyMolO-2 a intégré certains champs liés à des ressources externes, comme détaillé dans le tableau 3.1.

3.2.3. La recherche de motifs

Les motifs sont déterminés lors de la création ou de la révision de l'objet modulo associé. Ils résultent d'un consensus entre une caractérisation expérimentale et les motifs identifiés dans VAZyMolO. Ce travail est encore en cours.

Pour l'instant, la position et la signification des motifs sont établies par des recherches menées par l'équipe et des études approfondies. Une fois que des preuves solides concernant le motif sont collectées dans la littérature, la formule regex du motif est extraite du profile de séquence correspondant (voir l'exemple du motif A pour le modulo F189 dans la figure 1.14).

La recherche de motifs, effectuée dans le cadre de la routine analytique, est réalisée

par un script Python décrit dans l'algorithme 3.

Pseudo-code 3 Routine motif search

Require: re, matplotlib
Input : a $profile$ or a alignment ; Motifs
$hasMotifs$: nested list of boolean
for each sequence **in** $profile$ **do**
    append empty list to $hasMotifs$
    for each motif **in** Motifs **do**
        append the boolean (motif is found in sequence) ? to the latest list of $hasMotifs$ ▷ using $re.find$
    end for
end for
Generate $heatMap$ from $hasMotifs$
Output : $heatMap$

3.2.4. Homologie de séquence

La recherche d'homologie de séquence fait partie du pipeline analytique de routine. Cette recherche est effectuée à l'aide de l'outil MMseqs2, plus précisément via la fonctionnalité *easySearch*. La séquence d'intérêt est comparée à l'ensemble des séquences présentes dans la base de données JSON. Cet ensemble comprend à la fois les sous-séquences définies dans **Subseq** et les séquences encore non caractérisées ou en cours de caractérisation. En sortie, cette recherche retourne uniquement les premiers résultats parmi **Subseq**. Les résultats complets sont consignés dans des logs, accessibles uniquement côté serveur.

3.2.5. Construction des profils de séquence

Comme indiqué dans la figure 3.5, chaque **modulo** est associé à un profile de sous-séquences provenant de **Subseq**. Pour l'instant, les profils contenus dans VAZyMolO-2 sont issus de la récupération de VAZyMolO, comme exposé dans la sous-section 3.2.2.1. En effet, le point de départ pour la construction d'un profile est le proto-profile extrait de VAZyMolO.

Par la suite, une étape d'enrichissement du profile en séquences homologues est effectuée. Cet enrichissement est similaire à la recherche d'homologie des séquences

décrite en 3.2.4. De plus, pour distinguer l'orthologie de la paralogie ou de la divergence, une contrainte taxonomique est ajoutée. Il a été décidé de restreindre un profile à un certain clade taxonomique afin de limiter ces risques.

Pseudo-code 4 Traitement de la sortie de MMseqs2 pour la construction du profile de séquence

```
Input : return from MMseqs2 easySearch
pairwise processing :
if identité = 100% then
    if identical taxa then
        Redundancy : keep only one occurrence
    else
        Call taxonKit
        if the taxa are distinct then
            Are the sequences relevant ?
        else
            Redundancy : keep only one occurrence                ▷ taxonomy
        end if
    end if
else
    if identical taxa then
        What is the origin of the sequence ?                     ▷ mutant or error ?
    else
        retain sequences above the 70% identity threshold
    end if
end if
Output : a list of sequences to be added to the profile
```

Il est également possible de rencontrer des redondances parmi les séquences, par exemple dans les cas où l'identité est de 100% dans le retour de MMseqs2. Le pseudo-code 4 décrit le traitement de ces situations. Le seuil minimal d'identité est fixé à 70%. Dans le cas particulier où deux séquences proviennent du même taxon mais diffèrent, il est nécessaire de déterminer la cause de cette différence. La base de données d'origine de la séquence peut fournir un élément de réponse.

Jusqu'à présent, l'efficacité d'un profile est jugé par le nombre de séquence qui le compose. Chaque séquence qui compose un profile doit être pertinente, c'est-à-dire n'être ni redondante ni trop divergente. Il a été arbitrairement défini qu'un profile soit opérationnel à partir de 30 séquences. Au delà de ce nombre, il a été prévu de faire appel au modèle de Markov caché (introduit en 1.4.2.1) pour grandement améliorer l'usage analytique des profils.

3.2.6. Taxonomie

Dans VAZyMolO-2, chaque séquence protéique est associée à un taxon afin de conserver son origine taxonomique. Un numéro d'accès taxonomique (alias **taxId**) est attribué à chaque taxon, comme indiqué dans le tableau 3.1. La mise à jour du **taxId** se fait lors de l'ajout d'un nouveau taxon dans la table **Organism** de VAZyMolO-2, qu'il provienne de VAZyMolO ou d'un enrichissement de profile. Des scripts auxiliaires ont été développés pour encadrer et semi-automatiser ces étapes.

Dans un premier temps, *bio.Entrez* est utilisé pour corriger l'obsolescence du **taxId** et vérifier que le taxon représente bien un clade. Ensuite, dans un second temps, *taxonkit* permet d'identifier les taxons feuilles à partir d'un **taxId**.

3.2.7. La génération des figures

Les figures sont générées par des fonctions en Python intégrées au projet. Elles peuvent être appelées par une vue, en fonction des actions de l'utilisateur. Trois types de figures sont possibles : une représentation d'une protéine (voir 3.2.7.1 ci-dessous), une représentation d'un domaine (voir 3.2.7.2), et une représentation d'un profile de séquence (voir 3.2.7.3).

3.2.7.1. Représentation de l'objet *protein*

Ce diagramme, présenté dans la figure 3.8, illustre les interactions entre toutes les parties impliquées, depuis la demande d'accès à la page d'accueil jusqu'à l'affichage de la figure. Une fois la page web entièrement chargée, un script JavaScript lance une requête asynchrone (Ajax) qui est traitée par Django au niveau du backend.

Après que la requête a été relayée au backend, une vue appelle une fonction qui génère la figure, comme le décrit le pseudo-code 5. Le résultat de cette fonction est ensuite intégré au *template* pour finaliser la génération de la figure. La vue retourne alors ce contenu SVG à l'Ajax sous forme de simple réponse HTML, qui est ensuite redirigée vers la balise Django/HTML correspondante.

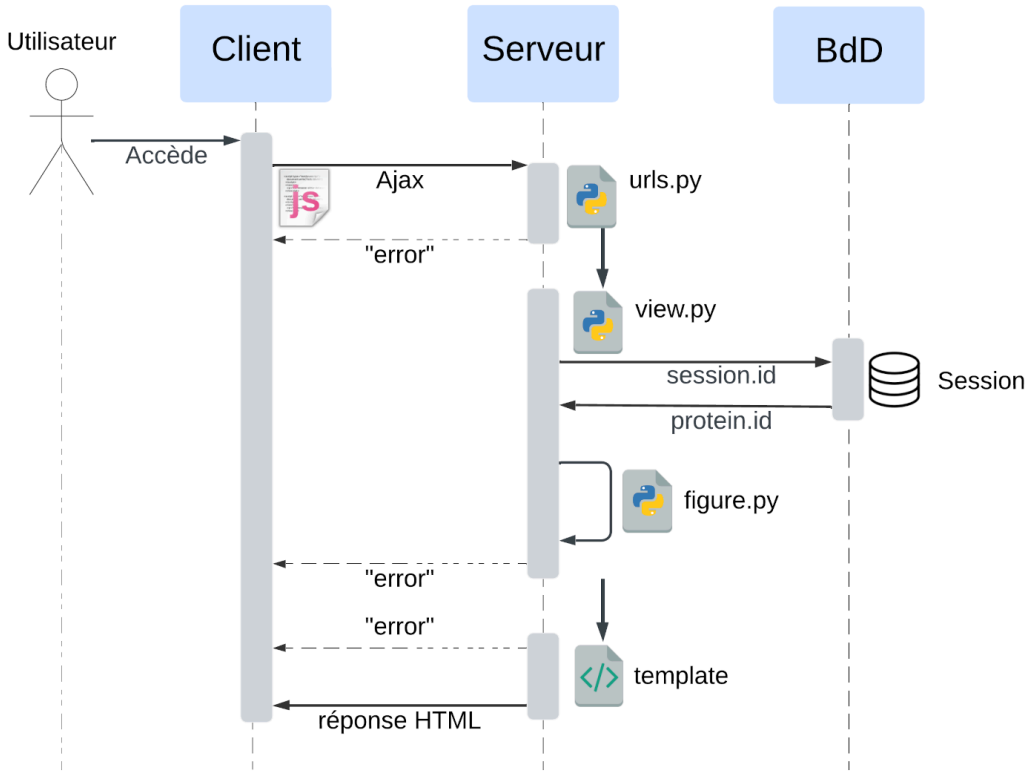


FIGURE 3.8. – Diagramme de séquence : générer la figure représentant une protéine. Le "Client" représente le navigateur côté client (autrement dit le *frontend*). Le "serveur" représente à l'inverse le *backend*, géré par django. La base de donnée ("BdD" dans la figure), sous sa forme *mysql*, est uniquement accessible depuis le *backend* de django. Les messages d'erreurs ("error") sont à titre indicatif et natifs au service que propose django.

Pseudo-code 5 Génération de la figure représentant une protéine

```
1: Input : relevant serialization data related to protein P
2: HTML − data : a dictionary to prepare the SVG
3: calculate N : number of lines to display          ▷ If P is a polyprotein then N > 1
4: adjust the height of the figure based on N
5: prepare the legend for protein P
6: prepare the legend for the organism
7: for all protein np linked to protein P do                ▷ spacing between proteins
8:     calculate the relative position of protein np
9:     if P is a polyprotein then
10:        prepare a dashed separator line
11:    end if
12:    prepare the legend for protein np
13:    for all subseq linked to protein np do
14:        calculate the relative position of subseq
15:        prepare the legend for the modulo related to subseq
16:        calculate the final separator line position
17:    end for
18: end for
19: return HTML − data
```

3.2.7.2. Représentation de l'objet *modulo*

Le déroulement est similaire au diagramme précédent (3.8). Seule l'action d'initiation change.

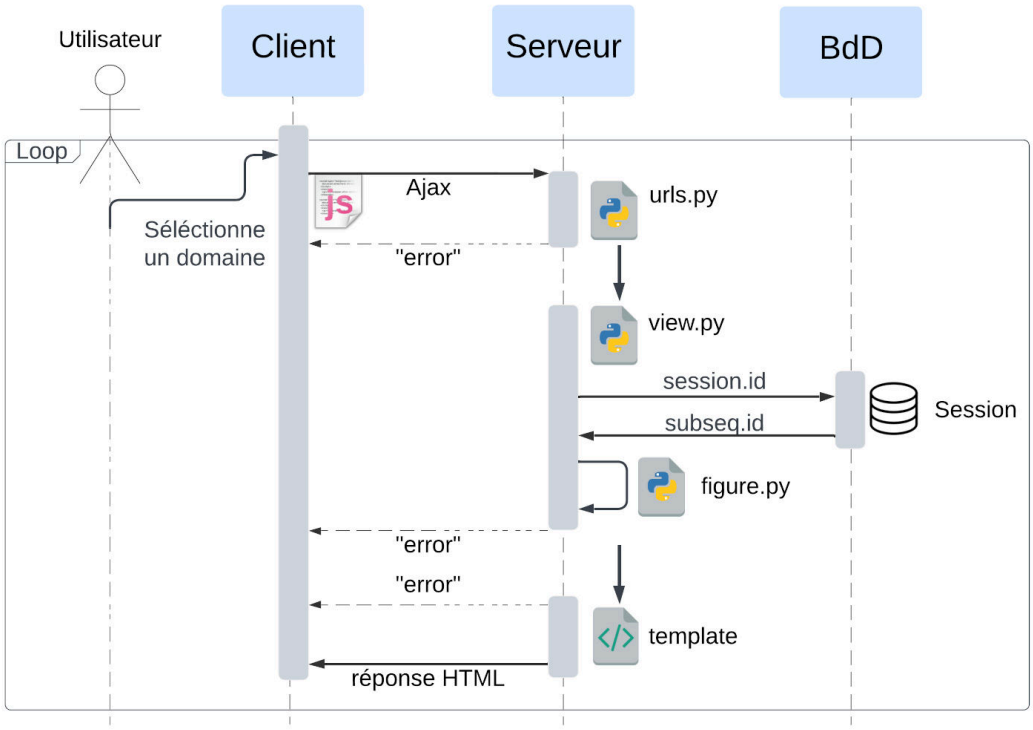


FIGURE 3.9. – Diagramme de séquence : générer la figure représentant un domaine. Ici une boucle ("loop") signale la possibilité à l'utilisateur que l'action peut être répétée.

3.2.7.3. Représentation de l'objet *profile*

Une requête Ajax est déclenchée par une interaction d'un utilisateur sur la figure générée par la séquence décrite dans le diagramme 3.9. A l'image des figures précédentes, une fonction python dans le fichier "figure.py" prépare les éléments nécessaires à la génération de la figure SVG. Dans ce cas, il s'agit d'un ensemble de $subseq$ pour un $modulo$ donné.

3.2.8. La recherche par mots-clefs

Pour le moment, ce travail est en cours : voici la méthode envisagée.

Je propose une première solution naïve qui sera amenée à évoluer le moment venu. Un dictionnaire servira de lexique pour contenir les écritures alternatives des termes considérés comme mots-clés. L'entretien de ce lexique se fait déjà progressivement.

Chaque requête de recherche par mots-clés est décomposée en une série de motifs (patterns). Un motif est le résultat d'une conversion d'un mot-clé ou d'un terme, en masquant sa casse et ses déclinaisons. Cette série de motifs est ensuite confrontée à la base de données sous sa forme JSON, à l'aide de regex. Le retour sera pondéré en fonction du champ, afin de proposer un score de pertinence pour chaque recherche. Cependant, je m'attends à rencontrer de nombreux faux négatifs, mais peu de faux positifs.

À terme, lorsque la base de données s'agrandira, je souhaite mettre en place une gestion automatique des GO terms.

3.3. Description de VAZyMolO-2

Cette section présente les résultats du développement de VAZyMolO-2 en suivant séquentiellement les différentes étapes du cycle de vie de la donnée, comme illustré par la figure 3.10 ci-dessous. Le projet a atteint une étape clé de son développement : la finalisation de la phase α. La démonstration qui suit reflétera l'état de transition actuel entre cette phase α achevée et la phase β en cours.

Le développement de VAZyMolO-2 a été structuré en trois phases principales :

1. Phase α (finalisée) : Mise en place de l'architecture de base et intégration des données héritées de VAZyMolO-1.
2. Phase β (en cours) : Optimisation des fonctionnalités et amélioration de l'interface utilisateur.
3. Phase γ (à venir) : Finalisation du projet avec des fonctionnalités avancées et préparation au déploiement.

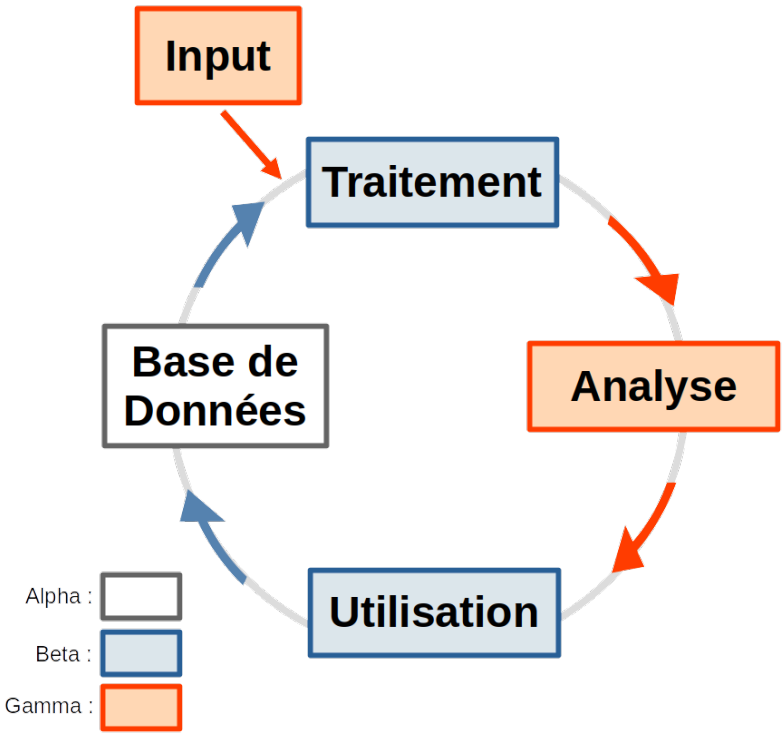


FIGURE 3.10. – Cycle de vie de la donnée dans VAZyMolO-2

TABLEAU 3.2. – Le jeu de données test

taxID	abr	Nom complet	Groupe
11137	HCoV-229E	Human coronavirus 229E	Alphacoronavirus
11149	TGV-Purdue	Transmissible gastroenteritis virus Purdue	Alphacoronavirus
28295	PEDV-CV777	Porcine epidemic diarrhea virus CV777	Alphacoronavirus
11128	BCoV	Bovine coronavirus	Betacoronavirus
11138	MHV-A59	Murine hepatitis virus	Betacoronavirus
31631	HCoV-OC43	Human coronavirus OC43	Betacoronavirus
227984	SARS-TOR2	Severe acute respiratory syndrome-related coronavirus TOR2	Betacoronavirus
233262	BCoV-ENT	Bovine enteric coronavirus	Betacoronavirus
694007	BatCoV-HK4	Tylonycteris bat coronavirus HKU4	Betacoronavirus
694008	BatCoV-HK5	Pipistrellus bat coronavirus HKU5	Betacoronavirus
1335626	MERS	Middle East respiratory syndrome-related coronavirus	Betacoronavirus
2697049	SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2	Betacoronavirus
11120	AIBV-Beaudette	Infectious bronchitis virus Beaudette	Gammacoronavirus

Pour chaque étape du cycle de vie de la donnée présentée, nous préciserons à quelle phase de développement elle appartient. Cette approche permettra de mettre en lumière les progrès réalisés et les développements en cours, tout en offrant un aperçu des fonctionnalités futures prévues pour VAZyMolO-2.

3.3.1. Le jeu de données test

L'objectif principal du projet est d'intégrer les données de la base de données du projet original VAZyMolO et d'incorporer les études en cours selon la dynamique du laboratoire.

Phase Alpha

Toutefois, il était nécessaire de disposer d'un jeu de données (*dataset*) préliminaire pour tester l'infrastructure du projet et valider l'affichage des résultats potentiels.

Phase Beta

Le développement a débuté avec l'étude des *Orthocoronavirinae* (voir figure 3.11 plus bas), pour s'étendre aux *Nidovirales*. Cette expansion progressive permettra l'enrichissement des profils en séquences homologues (voir section 3.2.5). La base de données (tableau 3.2 ci-dessus) comprends actuellement 13 génomes viraux, dont 5 sont en cours d'annotation ; cela revient à un total de 252 séquences protéiques.



FIGURE 3.11. – Phylogénie du jeu de données test : les *Orthocoronavirinae*
Alignement réalisé avec MUSCLE (EDGAR 2004), sur le domaine RdRP (nsp12 entière), des Coronavirus : en rose, le représentant de la sous-famille des *Letovirinae*; en gris, le représentant de la sous-famille des *Pitovirinae*; les autres taxons appartiennent aux principales sous-familles des *Orthocoronavirinae*. L'arbre construit par une méthode aLRT (SH-like), depuis SeaView (GOUY et al. 2010).

3.3.2. Les utilisateurs

Phase *Alpha*

Actuellement, un seul type d'utilisateur est implémenté, disposant de tous les droits d'accès. Un système d'authentification n'est pas nécessaire à ce stade, étant donné que le projet évolue dans un environnement fermé, hébergé localement (localhost)

Phase *Beta*

Deux types d'utilisateurs sont prévus pour cette phase :

— **Administrateur** (admin) : Réservé à la gestion administrative du projet, avec un accès protégé par mot de passe.

— **Utilisateur Public** (p.u.) : Destiné à un usage général, tant interne qu'externe au laboratoire. La méthode d'authentification pour ce type d'utilisateur est détaillée dans la section 3.2.1.

Cette distinction permettra une gestion plus fine des droits d'accès et une meilleure sécurisation des fonctionnalités administratives

3.3.3. Les types d'entrées (input)

Phase Alpha

La gestion des entrées reste actuellement simple. Les données sont directement accessibles via leur identifiant dans la base de données MySQL. Sur l'interface web, la barre de recherche est associée à un AJAX qui retourne l'objet demandé. Cette méthode est suffisante pour explorer le jeu de données test.

Phase Beta

Deux types d'entrées sont prévus :

1. Séquence protéique
   — Format FASTA
   — Traitement d'une séquence à la fois

— Voir section 3.2.4

2. Mots-clefs
   — Recherche par termes phylogénétiques ou fonctionnels
   — Conversion des mots-clefs en motifs pour expression régulière
   — Voir section 3.2.8

Cette évolution permettra une recherche plus flexible et adaptée aux besoins des utilisateurs. La conception de cette fonctionnalité est réalisée en phase *beta*, cependant son implémentation est prévue en phase *gamma*.

3.3.4. L'analytique et l'utilisation des données

Phase Alpha

La première phase du projet a permis de mettre en place les fonctionnalités fondamentales :

— Visualisation des informations liées à un modulo via une interface graphique (figure 3.12)
— Implémentation d'une méthode d'annotation manuelle (voir section 3.1.2)
— Développement des méthodes de sérialisation/désérialisation au format JSON pour la gestion des séquences annotées dans la base de données
— Validation initiale avec la polymérase des *Orthocoronaviridae* comme point de départ pour la base de données

Phase Beta

Cette phase se concentre sur la consolidation du projet :

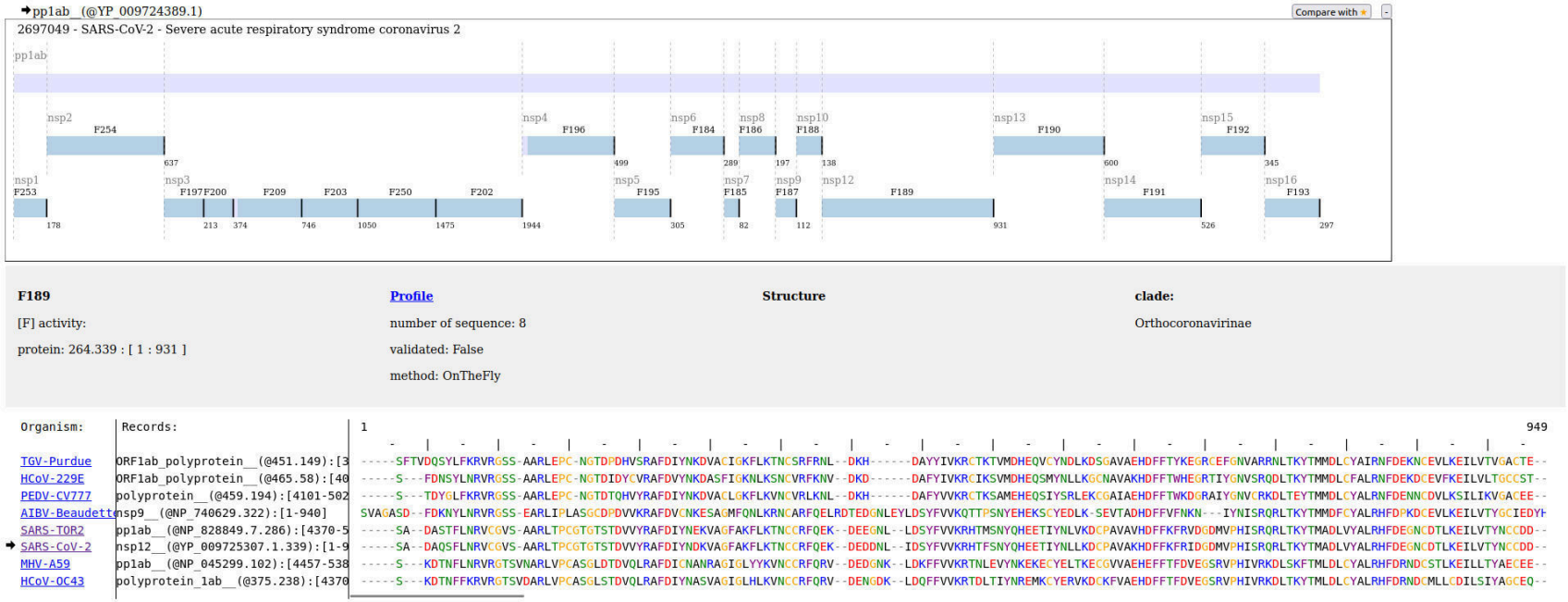
— Renforcement de la robustesse de l'architecture, tant en frontend qu'en backend
— Conception des méthodes automatisées d'annotation
— Acquisition des compétences nécessaires pour l'automatisation

Phase Gamma

Les objectifs de cette phase finale incluent :

— Implémentation d'un usage analytique rapide : identification de l'orthologue annoté le plus proche
— Mise en place d'une analyse en continu
— suivi de l'utilisation de l'outil, pour orienter le travail d'annotation selon l'usage
— Optimisation des paramètres clefs :
  — Définition de la taille minimale d'un profile pour l'application des HMM
  — Établissement des seuils d'identité pour l'intégration fiable des séquences dans un profile

FIGURE 3.12. – Annotation du génome du SARS-CoV-2



De haut en bas :
La représentation de la séquence entière de la polyprotéine (pp1ab) du coronavirus 2 du syndrome respiratoire aigu et sévère (SARS-CoV-2), appartenant au groupe des *Betacoronavirus*. Les domaines fonctionnels sont exposés et indiqués par leur identifiant, par exemple la polymérase a l'identifiant *F189*. En sélectionnant le **modulo** *F189*, on obtient un bandeau qui renseigne les informations relatives. En sélectionnant "Profile", on obtient alors le profil de séquence du **modulo**.

3.4. Discussion

La refonte de VAZyMolO en VAZyMolO-2 représente une potentielle avancée significative dans l'annotation des génomes viraux, comblant un manque important d'outils spécifiques dans ce domaine. Cette nouvelle version vise à simplifier et moderniser la structure originale tout en conservant sa richesse fonctionnelle. La réorganisation de l'architecture de la base de données initiale en un ensemble de modèles plus cohérent, associé à une meilleure connexion aux ressources externes, améliore considérablement l'exploitation des données.

Les défis rencontrés, notamment dans la gestion des séquences issues de différentes sources (PDB, NCBI) et leurs variations (polymorphisme, mutations structurales, erreurs de saisie), conduisent au développement d'une approche plus robuste pour la constitution des profils de séquences. L'utilisation flexible de formats JSON et MySQL offre une adaptabilité accrue pour les différents cas d'usage : accès, lecture et édition.

Avec un état précoce de développement : les perspectives d'évolution sont nombreuses, et constituent les objectifs pour la phase *gamma* :

— Extension de l'analyse à d'autres Nidovirus et au-delà (en citant les Flavivirus)
— Enrichissement des profils et évolution vers des modèles HMM, pour se projeter vers un usage analytique efficace
— Développement d'une interface multi-utilisateurs, tant externes qu'internes.
— Cartographie approfondie de la taxonomie virale, en employant des arbres interactifs

VAZyMolO-2 s'inscrit ainsi dans une démarche continue d'amélioration, visant à répondre aux besoins croissants de la communauté scientifique en matière d'annotation et d'analyse des génomes viraux.

Conclusion et perspectives

4. Conclusion

L'héritage de VAZyMolO

VAZyMolO se distingue par son approche modulaire de l'annotation des génomes viraux, s'inscrivant dans la lignée d'outils comme VADR (SCHÄFFER et al. 2020). Cette modularité représente un équilibre entre rigueur organisationnelle et flexibilité, permettant une modélisation efficace et évolutive du monde viral. Cette approche, bien qu'imparfaite, est conçue pour s'adapter aux défis complexes de la virologie en constante évolution. L'amélioration du modèle se fera progressivement, au fil des cas d'études rencontrés.

La revalidation des travaux antérieurs, bien que chronophage, a permis d'établir une base solide pour le projet avec trois objectifs principaux :

— Centraliser les connaissances du laboratoire
— Cartographier les forces et les zones d'incertitude dans notre compréhension
— Fournir un support pour l'analyse et la prise de décision

Si VAZyMolO-2 s'éloigne quelque peu de son prédécesseur, qui se concentrait davantage sur l'application expérimentale des données, cette évolution s'est faite sans perte d'information.

Perspectives d'évolution

Le développement futur du projet pourrait intégrer une couche de fonctionnalités basée sur les modèles de langage (INIGO BARRIO-HERNANDEZ et al. 2023). Cette évolution s'inscrit dans la tendance actuelle des interfaces en langage naturel, facilitant l'interaction avec les outils technologiques. L'enjeu sera de maintenir un équilibre entre accessibilité et précision scientifique, en combinant l'intuitivité des interfaces modernes avec la rigueur des méthodes d'annotation traditionnelles.

Ce projet ambitieux continuera au-delà de cette thèse, avec un engagement personnel dans son développement futur et la formation potentielle de nouveaux contributeurs.

Réflexion personnelle

Défis et apprentissages :

— Le développement en solitaire a nécessité un temps d'apprentissage conséquent

— L'adoption progressive des bonnes pratiques a ralenti le développement initial

— La réimplémentation de solutions existantes, bien que chronophage, a permis une compréhension approfondie des outils actuels

— La création d'un modèle abstrait pour l'organisation et l'analyse des connaissances

Compétences développées :

— Capacité d'auto-formation

— Double expertise : pharmacien et bioinformaticien

— Acquisition de compétences en :
  — Biologie structurale
  — Virologie
  — Bioinformatique appliquée
  — Rédaction scientifique
  — Montage de projets

Contributions à l'équipe :

— Support bioinformatique :

— Installation et configuration d'outils

— Formation aux commandes de base

— Participation aux réflexions sur des projets interdisciplinaires

— Interface entre biologie virale structurale et bioinformatique

Cette expérience a permis de développer non seulement un outil d'annotation viral, mais aussi des compétences transversales essentielles à la recherche moderne.

Bibliographie

[Abr+24] Josh ABRAMSON, Jonas ADLER, Jack DUNGER et al. « Accurate structure prediction of biomolecular interactions with AlphaFold 3 ». In : *Nature* 630.8016 (mai 2024), p. 493. DOI : 10.1038/s41586-024-07487-w. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC11168924 (cf. p. 71).

[Alt+90] S. F. ALTSCHUL, W. GISH, W. MILLER et al. « Basic local alignment search tool ». In : *J. Mol. Biol.* 215.3 (oct. 1990), p. 403-410. ISSN : 0022-2836. DOI : 10.1016/S0022-2836(05)80360-2. eprint : 2231712. URL : https://pubmed.ncbi.nlm.nih.gov/2231712/ (cf. p. 59, 64, 81).

[Alt+97] S. F. ALTSCHUL, T. L. MADDEN, A. A. SCHÄFFER et al. « Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. » In : *Nucleic Acids Res.* 25.17 (sept. 1997), p. 3389. DOI : 10.1093/nar/25.17.3389. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/ (cf. p. 64).

[Bal71] D. BALTIMORE. « Expression of animal virus genomes. » In : *Bacteriol. Rev.* 35.3 (sept. 1971), p. 235. DOI : 10.1128/br.35.3.235-241.1971 (cf. p. 32).

[Bar+22] Dorothée BARDIOT, Laura VANGEEL, Mohamed KOUKNI et al. « Synthesis, Structure–Activity Relationships, and Antiviral Profiling of 1-Heteroaryl-2-Alkoxyphenyl Analogs as Inhibitors of SARS-CoV-2 Replication ». In : *Molecules* 27.3 (fév. 2022). DOI : 10.3390/molecules27031052 (cf. p. 25, 74).

[Bek04] Michaël BEKAERT. « Étude du décalage de phase de lecture dans le génome de Saccharomyces cerevisiæ ». Thèse de doct. Paris, France : Université Pierre et Marie Curie - Paris VI, nov. 2004. URL : https://theses.hal.science/tel-00007928 (cf. p. 39).

[Bha+21] Pramod R. BHATT, Alain SCAIOLA, Gary LOUGHRAN et al. « Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome ». In : *Science* 372.6548 (mai 2021), p. 1306-1313. ISSN : 0036-8075. DOI : 10.1126/science.abf3546. URL : https://www.science.org/doi/10.1126/science.abf3546 (cf. p. 81).

[Bho+16] Asmit BHOWMICK, David H. BROOKES, Shane R. YOST et al. « Finding Our Way in the Dark Proteome ». In : *J. Am. Chem. Soc.* 138.31 (juill. 2016), p. 9730. DOI : 10.1021/jacs.6b06543. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC5051545 (cf. p. 57).

[Bri+21] Margo A. BRINTON, Anastasia A. GULYAEVA, Udeni B. R. BALASURIYA et al. « ICTV Virus Taxonomy Profile : Arteriviridae 2021 ». In : *J. Gen. Virol.* 102.8 (2021). DOI : 10.1099/jgv.0.001632. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8513641 (cf. p. 34).

[BRD21] Benjamin BUCHFINK, Klaus REUTER et Hajk-Georg DROST. « Sensitive protein alignments at tree-of-life scale using DIAMOND ». In : *Nat. Methods* 18.4 (2021), p. 366. DOI : 10.1038/s41592-021-01101-x. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8026399/ (cf. p. 64).

[Cor+20] CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES, Alexander E. GORBALENYA, Susan C. BAKER et al. « The species Severe acute respiratory syndrome-related coronavirus : classifying 2019-nCoV and naming it SARS-CoV-2 ». In : *Nat. Microbiol.* 5.4 (avr. 2020), p. 536-544. ISSN : 2058-5276. DOI : 10.1038/s41564-020-0695-z. eprint : 32123347. URL : https://pubmed.ncbi.nlm.nih.gov/32123347 (cf. p. 25, 74).

[CC10] Bruno COUTARD et Bruno CANARD. « The VIZIER project : overview ; expectations ; and achievements ». In : *Antiviral Res.* 87.2 (août 2010), p. 85-94. ISSN : 1872-9096. DOI : 10.1016/j.antiviral.2010.02.326. eprint : 20226212. URL : https://pubmed.ncbi.nlm.nih.gov/20226212 (cf. p. 97).

[Cro+04] Gavin E. CROOKS, Gary HON, John-Marc CHANDONIA et al. « WebLogo : a sequence logo generator ». In : *Genome Res.* 14.6 (juin 2004), p. 1188-1190. ISSN : 1088-9051. DOI : 10.1101/gr.849004. eprint : 15173120. URL : https://pubmed.ncbi.nlm.nih.gov/15173120 (cf. p. 55).

[Dec+24] Alice DECOMBE, Olve PEERSEN, Priscila SUTTO-ORTIZ et al. « Internal RNA 2′-O-methylation on the HIV-1 genome impairs reverse transcription ». In : *Nucleic Acids Res.* 52.3 (fév. 2024), p. 1359. DOI : 10.1093/nar/gkad1134 (cf. p. 25, 74).

[Dec+11] Etienne DECROLY, François FERRON, Julien LESCAR et al. « Conventional and unconventional mechanisms for capping viral mRNA ». In : *Nat. Rev. Microbiol.* 10.1 (déc. 2011), p. 51. DOI : 10.1038/nrmicro2675 (cf. p. 104).

[Dim+24] Salvatore DIMONTE, Muhammed BABAKIR-MINA, Taib HAMA-SOOR et al. « Genetic Variation and Evolution of the 2019 Novel Coronavirus ». In : *Public Health Genomics* (2024), p. 1. DOI : 10.1159/000513530 (cf. p. 35).

[Dom21] Jose L. DOMINGO. « What we know and what we need to know about the origin of SARS-CoV-2 ». In : *Environ. Res.* 200 (sept. 2021), p. 111785. DOI : 10.1016/j.envres.2021.111785 (cf. p. 25, 74).

[Edg04] Robert C. EDGAR. « MUSCLE : multiple sequence alignment with high accuracy and high throughput ». In : *Nucleic Acids Res.* 32.5 (2004), p. 1792. DOI : 10.1093/nar/gkh340. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337 (cf. p. 55, 64, 67, 81, 122).

[EC04] Paul EMSLEY et Kevin COWTAN. « Coot : model-building tools for molecular graphics ». In : *Acta Crystallogr., Sect. D : Biol. Crystallogr.* 60.Pt (déc. 2004), p. 12. ISSN : 0907-4449. DOI : 10.1107/S0907444904019158. eprint : 15572765. URL : https://pubmed.ncbi.nlm.nih.gov/15572765 (cf. p. 72).

[Enc24] ENCYCLOPÆDIA UNIVERSALIS. *Encyclopædia Universalis*. [Online ; accessed 19. Sep. 2024]. Sept. 2024. URL : https://www.universalis.fr/media/structures-fines-dhelice-alpha-et-feuillet-beta-de072721 (cf. p. 47).

[FP15] Anthony R. FEHR et Stanley PERLMAN. « Coronaviruses : An Overview of Their Replication and Pathogenesis ». In : *Coronaviruses* 1282 (2015), p. 1. DOI : 10.1007/978-1-4939-2438-7_1. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4369385 (cf. p. 38, 51).

[Fer+23] Mikael FERACCI, Cécilia EYDOUX, Véronique FATTORINI et al. « AT-752 targets multiple sites and activities on the Dengue virus replication enzyme NS5 ». In : *Antiviral Research* 212 (avr. 2023), p. 105574. ISSN : 1872-9096. DOI : 10.1016/j.antiviral.2023.105574 (cf. p. 25, 74).

[Fer+20] François FERRON, Humberto J. DEBAT, Ashleigh SHANNON et al. « A N7-guanine RNA cap methyltransferase signature-sequence as a genetic marker of large genome, non-mammalian Tobaniviridae ». In : *NAR Genomics Bioinf.* 2.1 (mars 2020). DOI : 10.1093/nargab/lqz022 (cf. p. 51, 104).

[Fer+05] François FERRON, Corinne RANCUREL, Sonia LONGHI et al. « VaZyMolO : a tool to define and classify modularity in viral proteins ». In : *J. Gen. Virol.* 86.Pt (mars 2005), p. 3. ISSN : 0022-1317. DOI : 10.1099/vir.0.80590-0. eprint : 15722535. URL : https://pubmed.ncbi.nlm.nih.gov/15722535 (cf. p. 75, 97).

[Fer+21] François FERRON, Bhawna SAMA, Etienne DECROLY et al. « The enzymes for genome size increase and maintenance of large (+)RNA viruses ». In : *Trends Biochem. Sci.* 46.11 (nov. 2021), p. 866-877. ISSN : 0968-0004. DOI : 10.1016/j.tibs.2021.05.006. eprint : 34172362. URL : https://pubmed.ncbi.nlm.nih.gov/34172362 (cf. p. 39).

[Fir14] Andrew E. FIRTH. « Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses ». In : *Nucleic Acids Res.* 42.20 (nov. 2014), p. 12425. DOI : 10.1093/nar/gku981. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4227794 (cf. p. 39).

[Flo+22] Verónica Roxana FLORES-VEGA, Jessica Viridiana MONROY-MOLINA, Luis Enrique JIMÉNEZ-HERNÁNDEZ et al. « SARS-CoV-2 : Evolution and Emergence of New Viral Variants ». In : *Viruses* 14.4 (mars 2022), p. 653. ISSN : 1999-4915. DOI : 10.3390/v14040653. eprint : 35458383 (cf. p. 35, 80).

[Gab+20] Felix GABLER, Seung-Zin NAM, Sebastian TILL et al. « Protein Sequence Analysis Using the MPI Bioinformatics Toolkit ». In : *Current Protocols in Bioinformatics* 72.1 (déc. 2020), e108. ISSN : 1934-3396. DOI : 10.1002/cpbi.108. URL : https://doi.org/10.1002/cpbi.108 (cf. p. 73).

[Gor+10a] Alexander E. GORBALENYA, Philippe LIEUTAUD, Mark R. HARRIS et al. « Practical application of bioinformatics by the multidisciplinary VIZIER consortium ». In : *Antiviral Res.* 87.2 (août 2010), p. 95. DOI : 10.1016/j.antiviral.2010.02.005. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7172516 (cf. p. 40).

[Gor+10b] Alexander E. GORBALENYA, Philippe LIEUTAUD, Mark R. HARRIS et al. « Practical application of bioinformatics by the multidisciplinary VIZIER consortium ». In : *Antiviral Res.* 87.2 (août 2010), p. 95-110. ISSN : 1872-9096. DOI : 10.1016/j.antiviral.2010.02.005. eprint : 20153379. URL : https://pubmed.ncbi.nlm.nih.gov/20153379 (cf. p. 97).

[Gor+20] David E. GORDON, Gwendolyn M. JANG, Mehdi BOUHADDOU et al. « A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug-Repurposing ». In : *Nature* 583.7816 (juill. 2020), p. 459. DOI : 10.1038/s41586-020-2286-9. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7431030 (cf. p. 41).

[GGG10] Manolo GOUY, Stéphane GUINDON et Olivier GASCUEL. « SeaView Version 4 : A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building ». In : *Mol. Biol. Evol.* 27.2 (fév. 2010), p. 221-224. ISSN : 0737-4038. DOI : 10.1093/molbev/msp259. URL : https://doi.org/10.1093/molbev/msp259 (cf. p. 72, 122).

[Gra+23] Oney Ortega GRANDA, Karine ALVAREZ, Maria J. MATE-PEREZ et al. « Macro1 domain residue F156 : A hallmark of SARS-CoV-2 de-MARylation specificity ». In : *Virology* 587 :109845. (Oct. 2023). ISSN : 1096-0341. DOI : 10.1016/j.virol.2023.109845. eprint : 37517331 (cf. p. 26, 74).

[GG21] Anastasia A. GULYAEVA et Alexander E. GORBALENYA. « A nidovirus perspective on SARS-CoV-2 ». In : *Biochem. Biophys. Res. Commun.* 538 :24-34. (Jan. 2021). ISSN : 1090-2104. DOI : 10.1016/j.bbrc.2020.11.015. eprint : 33413979 (cf. p. 39, 40).

[Had+18] James HADFIELD, Colin MEGILL, Sidney M. BELL et al. « Nextstrain : real-time tracking of pathogen evolution ». In : *Bioinformatics* 34.23 (déc. 2018), p. 4121-4123. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bty407. URL : https://doi.org/10.1093/bioinformatics/bty407 (cf. p. 73).

[Hil+09] Andrea HILDEBRAND, Michael REMMERT, Andreas BIEGERT et al. « Fast and accurate automatic structure prediction with HHpred ». In : *Proteins* 77.Suppl (2009), p. 9128-9132. ISSN : 1097-0134. DOI : 10.1002/prot.22499. eprint : 19626712. URL : https://pubmed.ncbi.nlm.nih.gov/19626712 (cf. p. 67).

[HF12] Colin R. HOWARD et Nicola F. FLETCHER. « Emerging virus diseases : can we ever expect the unexpected ? » In : *Emerging Microbes Infect.* 1.12 (déc. 2012), e46. DOI : 10.1038/emi.2012.47 (cf. p. 35, 80).

[Hue+18] Jaime HUERTA-CEPAS, Damian SZKLARCZYK, Davide HELLER et al. « eggNOG 5.0 : a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses ». In : *Nucleic Acids Res.* 47.Database (nov. 2018), p. D309. DOI : 10.1093/nar/gky1085. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC6324079 (cf. p. 58).

[IAE09] Kristoffer ILLERGÅRD, David H. ARDELL et Arne ELOFSSON. « Structure is three to ten times more conserved than sequence–a study of structural response in protein cores ». In : *Proteins* 77.3 (nov. 2009), p. 499-508. ISSN : 1097-0134. DOI : 10.1002/prot.22458. eprint : 19507241. URL : https://pubmed.ncbi.nlm.nih.gov/19507241 (cf. p. 61).

[Ini+23] # INIGO BARRIO-HERNANDEZ, # JINGI YEO, Jürgen JÄNES et al. « Clustering predicted structures at the scale of the known protein universe ». In : *Nature* 622.7983 (oct. 2023), p. 637-645. ISSN : 1476-4687. DOI : 10.1038/s41586-023-06510-w. eprint : 37704730. URL : https://pubmed.ncbi.nlm.nih.gov/37704730 (cf. p. 56, 129).

[Jum+21] John JUMPER, Richard EVANS, Alexander PRITZEL et al. « Highly accurate protein structure prediction with AlphaFold ». In : *Nature* 596 (août 2021), p. 583-589. ISSN : 1476-4687. DOI : 10.1038/s41586-021-03819-2. URL : https://www.nature.com/articles/s41586-021-03819-2 (cf. p. 70).

[KRY19] Kazutaka KATOH, John ROZEWICKI et Kazunori D. YAMADA. « MAFFT online service : multiple sequence alignment, interactive sequence choice and visualization ». In : *Briefings Bioinf.* 20.4 (juill. 2019), p. 1160. DOI : 10.1093/bib/bbx108. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6781576 (cf. p. 64, 67).

[Kel+15] Lawrence A. KELLEY, Stefans MEZULIS, Christopher M. YATES et al. « The Phyre2 web portal for protein modeling, prediction and analysis ». In : *Nat. Protoc.* 10.6 (juin 2015), p. 845-858. ISSN : 1750-2799. DOI : 10.1038/nprot.2015.053. eprint : 25950237. URL : https://pubmed.ncbi.nlm.nih.gov/25950237 (cf. p. 69).

[KMN21] Kinaan Aamir KHAN, Safyan Aman MEMON et Hammad NAVEED. « A hierarchical deep learning based approach for multi-functional enzyme classification ». In : *Protein Science : A Publication of the Protein Society* 30.9 (sept. 2021), p. 1935. DOI : 10.1002/pro.4146. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8376422 (cf. p. 52).

[Kha+21] Shruti KHARE, Céline GURRY, Lucas FREITAS et al. « GISAID's Role in Pandemic Response ». In : *China CDC Weekly* 3.49 (déc. 2021), p. 1049. DOI : 10.46234/ccdcw2021.255. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC8668406 (cf. p. 58, 80).

[Kin+18] Andrew M. Q. KING, Elliot J. LEFKOWITZ, Arcady R. MUSHEGIAN et al. « Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018) ». In : *Arch. Virol.* 163.9 (sept. 2018), p. 2601-2631. ISSN : 1432-8798. DOI : 10.1007/s00705-018-3847-1. eprint : 29754305 (cf. p. 33, 34).

[Kok+21] Goran KOKIC, Hauke S. HILLEN, Dimitry TEGUNOV et al. « Mechanism of SARS-CoV-2 polymerase stalling by remdesivir ». In : *Nat. Commun.* 12.1 (jan. 2021), p. 279. ISSN : 2041-1723. DOI : 10.1038/s41467-020-20542-0. eprint : 33436624 (cf. p. 25).

[Koo+20] Eugene V. KOONIN, Valerian V. DOLJA, Mart KRUPOVIC et al. « Global Organization and Proposed Megataxonomy of the Virus World ». In : *Microbiology and Molecular Biology Reviews : MMBR* 84.2 (juin 2020). DOI : 10.1128/MMBR.00061-19 (cf. p. 32).

[KKA21] Eugene V. KOONIN, Mart KRUPOVIC et Vadim I. AGOL. « The Baltimore Classification of Viruses 50 Years Later : How Does It Stand in the Light of Virus Evolution ? » In : *Microbiology and Molecular Biology Reviews : MMBR* 85.3 (sept. 2021). DOI : 10.1128/MMBR.00053-21 (cf. p. 32).

[KK22] Dimitrii O. KOSTENKO et Eugene V. KOROTKOV. « Application of the MAHDS Method for Multiple Alignment of Highly Diverged Amino Acid Sequences ». In : *Int. J. Mol. Sci.* 23.7 (avr. 2022). DOI : 10.3390/ijms23073764. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8998981 (cf. p. 64).

[Kum+18] Binod KUMAR, Kumari ASHA, Madhu KHANNA et al. « The emerging influenza virus threat : status and new prospects for its therapy and control ». In : *Arch. Virol.* 163.4 (avr. 2018), p. 831-844. ISSN : 1432-8798. DOI : 10.1007/s00705-018-3708-y. eprint : 29322273 (cf. p. 35).

[Kur+13] Shigehiro KURAKU, Christian M. ZMASEK, Osamu NISHIMURA et al. « aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity ». In : *Nucleic Acids Res.* 41.W1 (juill. 2013), W22-W28. ISSN : 0305-1048. DOI :

10.1093/nar/gkt389. URL : https://doi.org/10.1093/nar/gkt389 (cf. p. 64).

[Lan+20] Jun LAN, Jiwan GE, Jinfang YU et al. « Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor ». In : *Nature* 581.7807 (mai 2020), p. 215-220. ISSN : 1476-4687. DOI : 10.1038/s41586-020-2180-5. eprint : 32225176 (cf. p. 25).

[Lau+13] Chris LAUBER, Jelle J. GOEMAN, Maria del CARMEN PARQUET et al. « The Footprint of Genome Architecture in the Largest Genome Expansion in RNA Viruses ». In : *PLoS Pathog.* 9.7 (juill. 2013). DOI : 10.1371/journal.ppat.1003500. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3715407 (cf. p. 38, 41).

[Lef+18a] Elliot J. LEFKOWITZ, Donald M. DEMPSEY, Robert Curtis HENDRICKSON et al. « Virus taxonomy : the database of the International Committee on Taxonomy of Viruses (ICTV) ». In : *Nucleic Acids Res.* 46.Database (jan. 2018), p. D708. DOI : 10.1093/nar/gkx932 (cf. p. 33).

[Lef+18b] Elliot J. LEFKOWITZ, Donald M. DEMPSEY, Robert Curtis HENDRICKSON et al. « Virus taxonomy : the database of the International Committee on Taxonomy of Viruses (ICTV) ». In : *Nucleic Acids Res.* 46.D1 (jan. 2018), p. 708-717. ISSN : 1362-4962. DOI : 10.1093/nar/gkx932. eprint : 29040670. URL : https://pubmed.ncbi.nlm.nih.gov/29040670 (cf. p. 33).

[LCL08] Philippe LIEUTAUD, Bruno CANARD et Sonia LONGHI. « MeDor : a metaserver for predicting protein disorder ». In : *BMC Genomics* 9.Suppl (sept. 2008), p. 2. ISSN : 1471-2164. DOI : 10.1186/1471-2164-9-S2-S25. eprint : 18831791. URL : https://pubmed.ncbi.nlm.nih.gov/18831791 (cf. p. 97).

[Lin+23] Zeming LIN, Halil AKIN, Roshan RAO et al. « Evolutionary-scale prediction of atomic-level protein structure with a language model ». In : *Science* 379.6637 (mars 2023), p. 1123-1130. ISSN : 0036-8075. DOI : 10.1126/science.ade2574. URL : https://www.science.org/doi/abs/10.1126/science.ade2574 (cf. p. 71).

[Mal+22] Brandon MALONE, Nadya URAKOVA, Eric J. SNIJDER et al. « Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design ». In : *Nat. Rev. Mol. Cell Biol.* 23.1 (2022), p. 21. DOI : 10.1038/s41580-021-00432-z. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8613731 (cf. p. 51).

[Mar24] Iris MARCHAL. « RoseTTAFold expands to all-atom for biomolecular prediction and design ». In : *Nat. Biotechnol.* 42 (avr. 2024), p. 571. ISSN : 1546-1696. DOI : 10.1038/s41587-024-02211-5. URL : https://www.nature.com/articles/s41587-024-02211-5 (cf. p. 70).

[MT23] Andrew G. MCDONALD et Keith F. TIPTON. « Enzyme nomenclature and classification : the state of the art ». In : *FEBS J.* 290.9 (mai 2023), p. 2214-2231. ISSN : 1742-4658. DOI : 10.1111/febs.16274. eprint : 34773359. URL : https://pubmed.ncbi.nlm.nih.gov/34773359 (cf. p. 52).

[Men+23] Elaine C. MENG, Thomas D. GODDARD, Eric F. PETTERSEN et al. « UCSF ChimeraX : Tools for structure building and analysis ». In : *Protein Sci.* 32.11 (nov. 2023), e4792. ISSN : 0961-8368. DOI : 10.1002/pro.4792. URL : https://doi.org/10.1002/pro.4792 (cf. p. 72).

[Ngu+15] Lam-Tung NGUYEN, Heiko A. SCHMIDT, Arndt von HAESELER et al. « IQ-TREE : A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies ». In : *Mol. Biol. Evol.* 32.1 (jan. 2015), p. 268-274. ISSN : 0737-4038. DOI : 10.1093/molbev/msu300. URL : https://doi.org/10.1093/molbev/msu300 (cf. p. 40).

[Nom+24] Jason NOMBURG, Erin E. DOHERTY, Nathan PRICE et al. « Birth of protein folds and functions in the virome ». In : *Nature* 633.8030 (août 2024), p. 710. DOI : 10.1038/s41586-024-07809-y. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC11410667 (cf. p. 56).

[OTo+22] Áine O'TOOLE, Oliver G. PYBUS, Michael E. ABRAM et al. « Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences ». In : *BMC Genomics* 23 (2022). DOI : 10.1186/s12864-022-08358-2. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8832810 (cf. p. 53).

[OLT17] Toshiyuki ODA, Kyungtaek LIM et Kentaro TOMII. « Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance ». In : *BMC Bioinf.* 18 (2017). DOI : 10.1186/s12859-017-1686-9. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5455086/ (cf. p. 64).

[PSS06] Alexander O. PASTERNAK, Willy J. M. SPAAN et Eric J. SNIJDER. « Nidovirus transcription : how to make sense...? » In : *J. Gen. Virol.* 87.6 (juin 2006), p. 1403-1421. ISSN : 0022-1317. DOI : 10.1099/vir.0.81611-0. URL : https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.81611-0#tab2 (cf. p. 38).

[PRS22] Shelly PATHANIA, Ravindra K. RAWAL et Pankaj Kumar SINGH. « RdRp (RNA-dependent RNA polymerase) : A key target providing anti-virals for the management of various viral diseases ». In : *J. Mol. Struct.* 1250 (fév. 2022), p. 131756. DOI : 10.1016/j.molstruc.2021.131756 (cf. p. 25).

[Pee19] Olve B. PEERSEN. « A Comprehensive Superposition of Viral Polymerase Structures ». In : *Viruses* 11.8 (août 2019). DOI : 10.3390/v11080745 (cf. p. 58, 100).

[Per+15] Nelson PERDIGÃO, Julian HEINRICH, Christian STOLTE et al. « Unexpected features of the dark proteome ». In : *Proc. Natl. Acad. Sci. U.S.A.* 112.52 (déc. 2015), p. 15898. DOI : 10.1073/pnas.1508380112. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702990 (cf. p. 56).

[PRO17] Nelson PERDIGÃO, Agostinho C. ROSA et Seán I. O'DONOGHUE. « The Dark Proteome Database ». In : *BioData Mining* 10 (juill. 2017), p. 24. DOI : 10.1186/s13040-017-0144-6. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC5520327 (cf. p. 57, 58).

[24] *pont disulfure — Wiktionnaire, le dictionnaire libre*. [Online ; accessed 19. Sep. 2024]. Sept. 2024. URL : https://fr.wiktionary.org/wiki/pont_disulfure (cf. p. 48).

[Rah+20] Md Tanvir RAHMAN, Md Abdus SOBUR, Md Saiful ISLAM et al. « Zoonotic Diseases : Etiology, Impact, and Control ». In : *Microorganisms* 8.9 (sept. 2020), p. 1405. ISSN : 2076-2607. DOI : 10.3390/microorganisms8091405. eprint : 32932606 (cf. p. 35).

[Rem+11] Michael REMMERT, Andreas BIEGERT, Andreas HAUSER et al. « HHblits : lightning-fast iterative protein sequence searching by HMM-HMM alignment ». In : *Nat. Methods* 9.2 (déc. 2011), p. 173-175. ISSN : 1548-7105. DOI : 10.1038/nmeth.1818. eprint : 22198341. URL : https://pubmed.ncbi.nlm.nih.gov/22198341 (cf. p. 67).

[RG14] Xavier ROBERT et Patrice GOUET. « Deciphering key features in protein structures with the new ENDscript server ». In : *Nucleic Acids Res.* 42.W1 (juill. 2014), W320-W324. ISSN : 0305-1048. DOI : 10.1093/nar/gku316. URL : https://doi.org/10.1093/nar/gku316 (cf. p. 55).

[Roe+21] Molly K. ROE, Nathan A. JUNOD, Audrey R. YOUNG et al. « Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19 ». In : *J. Gen. Virol.* 102.3 (2021). DOI : 10.1099/jgv.0.001558. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515871 (cf. p. 38).

[Rom+20] Maria ROMANO, Alessia RUGGIERO, Flavia SQUEGLIA et al. « A Structural View of SARS-CoV-2 RNA Replication Machinery : RNA Synthesis, Proofreading and Final Capping ». In : *Cells* 9.5 (mai 2020), p. 1267. ISSN : 2073-4409. DOI : 10.3390/cells9051267. eprint : 32443810. URL : https://pubmed.ncbi.nlm.nih.gov/32443810 (cf. p. 101).

[Ru+23] Jinlong RU, Mohammadali Khan MIRZAEI, Jinling XUE et al. « ViroProfiler : a containerized bioinformatics pipeline for viral metagenomic data analysis ». In : *Gut Microbes* 15.1 (mars 2023), p. 2192522. DOI : 10.1080/19490976.2023.2192522. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC10072060 (cf. p. 57).

[Sch+20] Alejandro A. SCHÄFFER, Eneida L. HATCHER, Linda YANKIE et al. « VADR : validation and annotation of virus sequence submissions to GenBank ». In : *BMC Bioinf.* 21 (mai 2020), p. 211. DOI : 10.1186/s12859-020-3537-3. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC7245624 (cf. p. 129).

[SBC03] Heidi L. SCHUBERT, Robert M. BLUMENTHAL et Xiaodong CHENG. « Many paths to methyltransfer : a chronicle of convergence ». In : *Trends Biochem. Sci.* 28.6 (juin 2003), p. 329. DOI : 10.1016/S0968-0004(03)00090-2. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC2758044 (cf. p. 104).

[Sha+24] Ashleigh SHANNON, Aurélie CHAZOT, Mikael FERACCI et al. « An exonuclease-resistant chain-terminating nucleotide analogue targeting the SARS-CoV-2 replicase complex ». In : *Nucleic Acids Res.* 52.3 (fév. 2024), p. 1325-1340. ISSN : 1362-4962. DOI : 10.1093/nar/gkad1194. eprint : 38096103 (cf. p. 26, 74).

[Sha+22a] Ashleigh SHANNON, Véronique FATTORINI, Bhawna SAMA et al. « A dual mechanism of action of AT-527 against SARS-CoV-2 polymerase ». In : *Nat. Commun.* 13.1 (fév. 2022), p. 621. ISSN : 2041-1723. DOI : 10.1038/s41467-022-28113-1. eprint : 35110538 (cf. p. 26, 74).

[Sha+20a] Ashleigh SHANNON, Nhung Thi-Tuyet LE, Barbara SELISKO et al. « Remdesivir and SARS-CoV-2 : Structural requirements at both nsp12 RdRp and nsp14 Exonuclease active-sites ». In : *Antiviral Res.* 178 :104793. (Juin 2020). ISSN : 1872-9096. DOI : 10.1016/j.antiviral.2020.104793. eprint : 32283108. URL : https://pubmed.ncbi.nlm.nih.gov/32283108 (cf. p. 25, 74, 101).

[Sha+22b] Ashleigh SHANNON, Bhawna SAMA, Pierre GAUFFRE et al. « A second type of N7-guanine RNA cap methyltransferase in an unusual locus of a large RNA virus genome ». In : *Nucleic Acids Res.* 50.19 (oct. 2022), p. 11186-11198. ISSN : 0305-1048. DOI : 10.1093/nar/gkac876. URL : https://doi.org/10.1093/nar/gkac876 (cf. p. 104).

[Sha+20b] Ashleigh SHANNON, Barbara SELISKO, Nhung-Thi-Tuyet LE et al. « Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis ». In : *Nat. Commun.* 11 (2020). DOI : 10.1038/s41467-020-18463-z. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7499305 (cf. p. 25, 74).

[Sid+23] Stuart G. SIDDELL, Donald B. SMITH, Evelien ADRIAENSSENS et al. « Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV) ». In : *J. Gen. Virol.* 104.5 (2023). DOI : 10.1099/jgv.0.001840 (cf. p. 33).

[Sie+11] Fabian SIEVERS, Andreas WILM, David DINEEN et al. « Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega ». In : *Molecular Systems Biology* 7 (11 oct. 2011), p. 539. ISSN : 1744-4292. DOI : 10.1038/msb.2011.75 (cf. p. 64, 67).

[Smi13] Kent SMITH. « A Brief History of NCBI's Formation and Growth ». In : *The NCBI Handbook [Internet]. 2nd edition*. National Center for Biotechnology Information (US), 2013. URL : https://www.ncbi.nlm.nih.gov/books/NBK148949 (cf. p. 44, 58).

[Soh+20] Catrin SOHRABI, Zaid ALSAFI, Niamh O'NEILL et al. « World Health Organization declares global emergency : A review of the 2019 novel coronavirus (COVID-19) ». In : *International Journal of Surgery (London, England)* 76 (avr. 2020), p. 71. DOI : 10.1016/j.ijsu.2020.02.034 (cf. p. 25, 74).

[SS17] Martin STEINEGGER et Johannes SÖDING. « MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets ». In : *Nat. Biotechnol.* 35 (nov. 2017), p. 1026-1028. ISSN : 1546-1696. DOI : 10.1038/nbt.3988. URL : https://www.nature.com/articles/nbt.3988 (cf. p. 64).

[Sut+21] Priscila SUTTO-ORTIZ, Sergey TCHERNIUK, Nina YSEBAERT et al. « The methyltransferase domain of the Respiratory Syncytial Virus L protein catalyzes cap N7 and 2'-O-methylation ». In : *PLoS Pathog.* 17.5 (mai 2021). DOI : 10.1371/journal.ppat.1009562 (cf. p. 26, 74).

[Ter+24] Thomas C. TERWILLIGER, Dorothee LIEBSCHNER, Tristan I. CROLL, Christopher J. WILLIAMS, Airlie J. MCCOY, Billy K. POON, Pavel V. AFONINE, Robert D. OEFFNER, Jane S. RICHARDSON et al. « AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination ». In : *Nat. Methods* 21.1 (jan. 2024), p. 110-116. ISSN : 1548-7105. DOI : 10.1038/s41592-023-02087-4. eprint : 38036854. URL : https://pubmed.ncbi.nlm.nih.gov/38036854 (cf. p. 71).

[Ter+23] Thomas C. TERWILLIGER, Dorothee LIEBSCHNER, Tristan I. CROLL, Christopher J. WILLIAMS, Airlie J. MCCOY, Billy K. POON, Pavel V. AFONINE, Robert D. OEFFNER, Christopher J. SCHLICKSUP et al. « Alphafold changes everything (and nothing) ». In : *Acta Crystallographica Section A* 79.a2 (août 2023), p. C1. DOI : 10.1107/S2053273323096079. URL : https://doi.org/10.1107/S2053273323096079 (cf. p. 71).

[The23] THE UNIPROT CONSORTIUM. « UniProt : the Universal Protein Knowledgebase in 2023 ». In : *Nucleic Acids Res.* 51.D1 (jan. 2023), p. D523-D531. ISSN : 0305-1048. DOI : 10.1093/nar/gkac1052. URL : https://doi.org/10.1093/nar/gkac1052 (cf. p. 58).

[TKU21] Elham TORBATI, Kurt L. KRAUSE et James E. USSHER. « The Immune Response to SARS-CoV-2 and Variants of Concern ». In : *Viruses* 13.10 (sept. 2021), p. 1911. DOI : 10.3390/v13101911. URL : https://pmc.ncbi.nlm.nih.gov/articles/PMC8537260 (cf. p. 74, 80).

[Uve13] Vladimir N. UVERSKY. « Unusual biophysics of intrinsically disordered proteins ». In : *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834.5 (mai 2013), p. 932-951. ISSN : 1570-9639. DOI : 10.1016/j.bbapap.2012.12.008. URL : https://www.sciencedirect.com/science/article/pii/S1570963912002889 (cf. p. 49).

[Var+22] Mihaly VARADI, Stephen ANYANGO, Mandar DESHPANDE et al. « AlphaFold Protein Structure Database : massively expanding the structural coverage of protein-sequence space with high-accuracy models ». In : *Nucleic Acids Res.* 50.D1 (jan. 2022), p. 439-444. ISSN : 1362-4962. DOI : 10.1093/nar/gkab1061. eprint : 34791371. URL : https://pubmed.ncbi.nlm.nih.gov/34791371 (cf. p. 56).

[Vri+97] Antoine A. F. de VRIES, Marian C. HORZINEK, Peter J. M. ROTTIER et al. « The Genome Organization of the Nidovirales : Similarities and Differences between Arteri-, Toro-, and Coronaviruses ». In : *Semin. Virol.* 8.1 (fév. 1997), p. 33-47. ISSN : 1044-5773. DOI : 10.1006/smvy.1997.0104. eprint : 32288441 (cf. p. 32).

[Wal+22] Peter J. WALKER, Stuart G. SIDDELL, Elliot J. LEFKOWITZ, Arcady R. MUSHEGIAN, Evelien M. ADRIAENSSENS et al. « Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022) ». In : *Arch. Virol.* 167.11 (nov. 2022), p. 2429. DOI : 10.1007/s00705-022-05516-5. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10088433 (cf. p. 34).

[Wal+19] Peter J. WALKER, Stuart G. SIDDELL, Elliot J. LEFKOWITZ, Arcady R. MUSHEGIAN, Donald M. DEMPSEY et al. « Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019) ». In : *Arch. Virol.* 164.9 (sept. 2019), p. 2417-2429. ISSN : 1432-8798. DOI : 10.1007/s00705-019-04306-w. eprint : 31187277 (cf. p. 33, 34).

[Wal+20] Alexandra C. WALLS, Young-Jun PARK, M. Alejandra TORTORICI et al. « Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein ». In : *Cell* 181.2 (avr. 2020), p. 281-2926. ISSN : 1097-4172. DOI : 10.1016/j.cell.2020.02.058. eprint : 32155444. URL : https://pubmed.ncbi.nlm.nih.gov/32155444 (cf. p. 51).

[Wat+09] Andrew M. WATERHOUSE, James B. PROCTER, David M. A. MARTIN et al. « Jalview Version 2–a multiple sequence alignment editor and analysis workbench ». In : *Bioinformatics* 25.9 (mai 2009), p. 1189-1191. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btp033. eprint : 19151095. URL : https://pubmed.ncbi.nlm.nih.gov/19151095 (cf. p. 72).

[WCF23] Vincent WILDE, Bruno CANARD et François FERRON. « Viral Instant Mutation Viewer : A Tool to Speed Up the Identification and Analysis of New SARS-CoV-2 Emerging Variants and Beyond ». In : *Viruses* 15.8 (juill. 2023), p. 1628. ISSN : 1999-4915. DOI : 10.3390/v15081628 (cf. p. 79).

[Woo+23] Patrick C. Y. WOO, Raoul J. de GROOT, Bart HAAGMANS et al. « ICTV Virus Taxonomy Profile : Coronaviridae 2023 ». In : *J. Gen. Virol.* 104.4 (avr. 2023), p. . ISSN : 1465-2099. DOI : 10.1099/jgv.0.001843. eprint : 37097842 (cf. p. 33, 34).

[Zar+22] Christine ZARDECKI, Shuchismita DUTTA, David S. GOODSELL et al. « PDB-101 : Educational resources supporting molecular explorations through biology and medicine ». In : *Protein Sci.* 31.1 (jan. 2022), p. 129-140. ISSN : 1469-896X. DOI : 10.1002/pro.4200. eprint : 34601771. URL : https://pubmed.ncbi.nlm.nih.gov/34601771 (cf. p. 58).

[Zha+21] Yabin ZHANG, Hao SUN, Rongjuan PEI et al. « The SARS-CoV-2 protein ORF3a inhibits fusion of autophagosomes with lysosomes ». In : *Cell Discovery* 7.1 (mai 2021), p. 31. ISSN : 2056-5968. DOI : 10.1038/s41421-021-00268-z. eprint : 33947832 (cf. p. 49).

[ZSG00] J. ZIEBUHR, E. J. SNIJDER et A. E. GORBALENYA. « Virus-encoded proteinases and proteolytic processing in the Nidovirales ». In : *J. Gen. Virol.* 81.Pt (avr. 2000), p. 4. ISSN : 0022-1317. DOI : 10.1099/0022-1317-81-4-853. eprint : 10725411. URL : https://pubmed.ncbi.nlm.nih.gov/10725411 (cf. p. 38).

ANNEXES

A. Financements

— Du 01/10/2021 au 31/05/2023 Aix-Marseille Université ANR - IMI CARE

— Du 01/06/2023 au 31/05/2024 Aix-Marseille Université ANR

— Du 01/06/2024 au 30/11/2024 Aix-Marseille Université Laboratoire