UNIVERSITÉ Clermont Auvergne

MASTER 1 Bio-informatique

Parcours analyse et modélisation des données

RAPPORT DE STAGE PRÉSENTÉ PAR :

Youlen IGLESIAS

Titre

Exploitation des modèles de langage protéique et de la projection UMAP pour améliorer la prédiction de gènes codants chez les phages : application aux *Microviridae*

Titre court

Prédiction de gènes de *Microviridae* par PLM

Title

Leveraging Protein Language Models and UMAP Projection to Enhance Gene Prediction in Bacteriophages: A Case Study on *Microviridae*

Short title

Gene Prediction in Microviridae via PLMs

Responsable du stage : **François Enault**

Adresse complète du laboratoire : 1 Impasse Amélie Murat, 63170 Aubière

05/07-2025

Résumé

La prédiction des gènes codants chez les phages, en particulier les microvirus, demeure complexe du fait des limitations actuelles des méthodes classiques, notamment pour identifier les gènes de petite taille, variables ou surimprimés. Dans cette étude exploratoire, nous avons évalué si l'utilisation d'un modèle de langage protéique (ProtT5), transformant des séquences protéiques en vecteurs de haute dimension (1 024) à partir de millions de paramètres inconnus, couplée à une projection en deux dimensions (UMAP), pouvait discriminer efficacement les séquences codantes des non codantes. L'idée centrale était d'exploiter l'intuition acquise implicitement par le modèle durant son entraînement afin de révéler une discrimination pertinente au sein des projections réalisées. À partir de 40 génomes sélectionnés parmi : (i) 8 143 génomes de microvirus assemblés à partir de métagénomes intestinaux, et (ii) 14 génomes références, nous avons généré des embeddings protéiques pour 2 275 ORFs détectés par Getorf ainsi que 287 protéines prédites par Prodigal. Les résultats obtenus montrent une séparation pertinente des grandes catégories fonctionnelles connues (capside, réplication, protéines Pilot) et révèlent plusieurs groupes prometteurs, comprenant potentiellement des gènes non détectés par les méthodes standards. Malgré certaines limites, notamment la tendance au regroupement par affiliation taxonomique plutôt que fonctionnelle, ces résultats indiquent que la méthode proposée complète efficacement les approches classiques, sous réserve d'ajustements méthodologiques tels qu'une sélection préalable et diversifiée des séquences soumises. En perspective, l'adaptation spécifique des modèles protéiques aux séquences virales ainsi que l'intégration d'analyses complémentaires (évolutives, expérimentales) pourraient encore renforcer l'efficacité de cette approche pour la prédiction des gènes de phages.

Abstract

Predicting protein-coding genes in bacteriophages, particularly microviruses, remains challenging due to limitations of current computational methods, especially regarding the detection of small, variable, or overlapping genes. In this exploratory study, we assessed whether employing a protein language model (ProtT5), which transforms protein sequences into high-dimensional vectors (embeddings of 1,024 dimensions) based on millions of unknown parameters, combined with dimensionality reduction (UMAP), could effectively discriminate coding from non-coding sequences. The core idea was to leverage the implicit knowledge acquired by the model during its training to reveal meaningful patterns through embedding projections. Using 40 genomes selected from (i) 8,143 microvirus genomes assembled from intestinal metagenomes, and (ii) 14 reference genomes, we generated protein embeddings for 2,275 ORFs identified by Getorf and 287 proteins predicted by Prodigal. The results demonstrated effective separation of known functional categories (capsid, replication, Pilot proteins), and uncovered several promising clusters potentially containing genes undetected by standard approaches. Despite certain limitations, such as some of the clustering driven more by taxonomic affiliation than functional properties, these findings suggest that our proposed method effectively complements conventional approaches, provided methodological adjustments such as a prior selection of sequences and greater genomic diversity are implemented. Future work involving fine-tuning protein models specifically on viral sequences, alongside complementary analyses (evolutionary, experimental), could further enhance the efficacy of this approach for bacteriophage gene prediction.

Remerciements

Je remercie Monsieur Françoi Enault qui m'a offert l'opportunité de réaliser ce stage en me proposant exploratoire original et qui s'est rendu disponible malgré ses autres obligations.

Je remercie Madame Anne-Catherine Lehours, pour m'avoir grandement aidé dans la réflexion scientifique et dans la rédaction, elle m'a permis de prendre du recul et de mieux me situer dans le contexte biologique.

Je remercie Celtill Dumont pour m'avoir rassuré quand parfois je perdais confiance quant à la pertinence de mes avancées.

Je remercie également Clovis Galiez et Paul C. Kirchberger qui m'ont fourni des explications ainsi que des données qui m'ont permis d'avancer.

J'ai été sensible à l'accueil et l'atmosphère agréable de toute l'équipe ainsi que du personnel du laboratoire.

Sommaire

SYNTHÈSE BIBLIOGRAPHIQUE

Les Phages : importance écologique

Avec une population mondiale estimée à 10^{31} particules virales, les phages (*i.e*. virus infectant les bactéries) représentent les entités biologiques les plus abondantes et les plus diversifiées génétiquement au sein de la biosphère. Parasites obligatoires, ils peuvent influencer la mortalité et le métabolisme bactérien par le biais d'infections lytiques et lysogéniques **(Figure 1)**. Selon leur stratégie de réplication, l'impact des phages sur la structure et le fonctionnement du microbiome diffère. Les phages lytiques (ou virulents), qui induisent la lyse de la cellule hôte, constituent des agents de mortalité importants pour les communautés bactériennes. Ils sont ainsi généralement responsables de 10 à 50 % de la mortalité bactérienne dans la colonne d'eau océanique et jusqu'à 80 % dans les systèmes profonds. Les virus lysogéniques (ou tempérés), en s'intégrant sous forme de prophage dans le chromosome hôte, semblent augmenter la fitness de l'hôte infecté en l'immunisant contre de nouvelles infections et en lui apportant des fonctions nouvelles **[Chevallereau *et al.* 2022]**[1].

Les Phages : diversité morphologique et génomique

Les phages sont de minuscules particules dont la taille est généralement comprise entre 20 et 150 nm. Ils sont constitués d'un génome, dont la taille est comprise entre 1 680 et 2,5 millions de nucléotides, qui peut être soit de l'ADN ou de l'ARN simple ou double brin, protégé dans une coque de protéines (la capside) et parfois d'une enveloppe phospholipidique interne ou externe. La capside a une structure qui peut être polyédrique (*microviridae*, *corticoviridae*, *tectiviridae*, *leviviridae* et *cystoviridae*), filamenteuse (*inoviridae*), pléomorphe (plasmaviridae) ou reliée à une queue (*caudoviricetes*) **(Figure 2)**. Cependant, ces critères morphologiques ne permettent pas de classer les phages en groupes monophylétiques, ils sont classés principalement selon leur contenu génomique **[Dion et al. 2020]**[2]. Cette classification est réalisée au sein du comité international pour la taxonomie des virus (ICTV, https://ictv.global/).

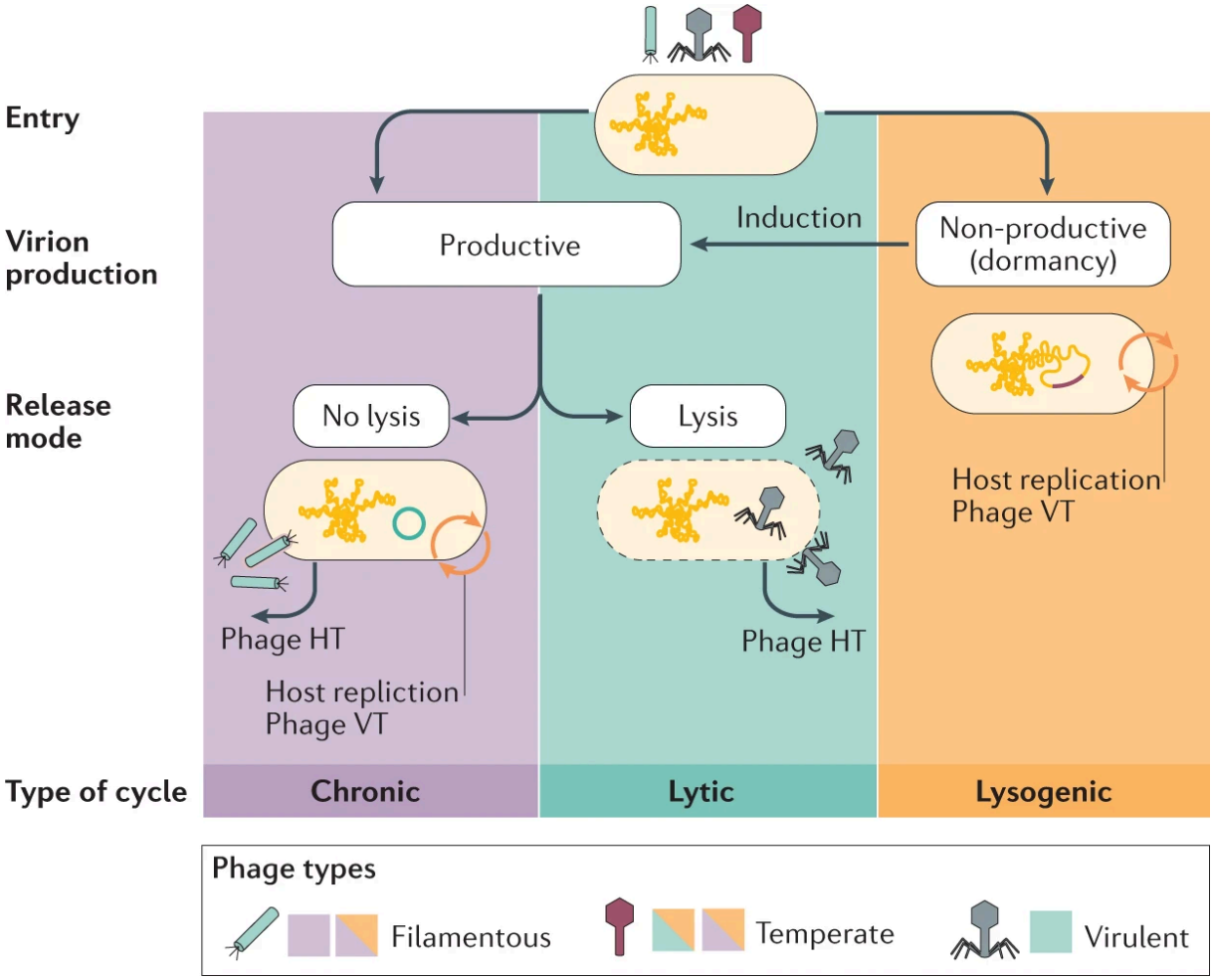


Figure 1: Les différents cycles d'infection des hôtes bactériens par les phages

Une fois qu'ils ont pénétré dans leur cellule hôte, les phages peuvent entrer dans un cycle de réplication productif qui aboutit à la libération de nouveaux virions, soit sans lyser l'hôte (cycle chronique), soit après la lyse de l'hôte (cycle lytique). Ils peuvent également suivre un cycle non productif, dans lequel leur génome s'intègre dans le chromosome de l'hôte et se réplique avec lui (cycle lysogène). Ils peuvent sortir de cet état dormant, soit spontanément, soit sous l'effet de stimuli exogènes, pour passer à l'un des cycles productifs.

Selon leur cycle d'infection, les phages peuvent être classés en trois types : les phages filamenteux suivent généralement un cycle chronique productif ; certains d'entre eux (mais pas tous) ont la capacité d'entrer dans un cycle lysogénique non productif ; les phages tempérés se caractérisent par leur capacité à être lysogéniques ; lors de l'induction, ils peuvent produire de nouveaux virions par le biais d'un cycle chronique ou lytique ; enfin, les phages virulents se répliquent uniquement par le biais d'un cycle lytique. HT : transmission horizontale ; VT : transmission verticale.

Source de la figure: Chevallereau *et al.* 2021

Prédire les gènes des phages : une tâche complexe

L'essor de la métagénomique virale a rendu accessible de grandes quantités de fragments ou de génomes complets de phages, mais pour mieux comprendre leur rôle dans les écosystèmes, il est nécessaire d'identifier les gènes encodés dans leurs génomes. En théorie, les techniques de protéomique sont capables de repérer ces gènes ; en pratique leur portée est limitée aux virus cultivables, à cela vient s'ajouter la difficulté de la distinction entre les gènes exprimés de l'hôte et du phage.

La prédiction *in silico* s'impose donc comme principal moyen de trouver les gènes encodés chez la majorité des souches. L'approche la plus répandue, pour laquelle le logiciel le plus utilisé est Prodigal **[Hyatt et *al*]**[3], consiste à **(i)** lister tous les cadres de lecture ouverts (Open Reading Frames ou ORF) puis **(ii)** à déterminer les ORFs ayant la plus forte probabilité de correspondre réellement à des gènes à partir de leurs caractéristiques (taille, composition en acides aminés, signaux en amont) ainsi que de la maximisation de la fraction codante du génome. Ces prédictions peuvent ensuite être confirmées par la recherche de séquences homologues dans d'autres génomes, mais **(i)** certaines séquences non codantes peuvent être conservées et donner lieu à des ORFs non codants eux-mêmes conservés dans différents génomes et **(ii)** l'absence de séquences similaires dans d'autres génomes avec une protéine prédite n'implique pas forcément que celle-ci soit une fausse prédiction, les similarités de séquences étant parfois difficiles à détecter pour les séquences virales évoluant rapidement. De plus, les logiciels basés sur ces approches peinent à détecter correctement les gènes de petites tailles (<150 nucléotides), chaque génome contenant une grande quantité d'ORF non codant de taille comparable. Les gènes surimprimés sont eux systématiquement non prédits, Prodigal évitant au maximum les chevauchements entre gènes **(Annexe 1)**. Pour contourner cette limite, des analyses dites évolutives se sont avérées utiles pour la détection *de novo* des régions codantes. Par exemple, le logiciel RNAcode[4] discrimine les régions codantes et non codantes à partir d'alignements multiples de gènes potentiels. Pour un alignement de séquences multiples donné et une matrice de similarité des protéines, un score de substitution est dérivé pour évaluer les mutations observées afin de détecter des preuves de sélection négative en les comparant aux scores de fond estimés à partir d'alignements aléatoires. Ce logiciel a été utilisé récemment pour identifier des milliers de nouveaux petits gènes (<150 nucléotides) qui ne sont généralement pas prédits dans les génomes de phages. Cette

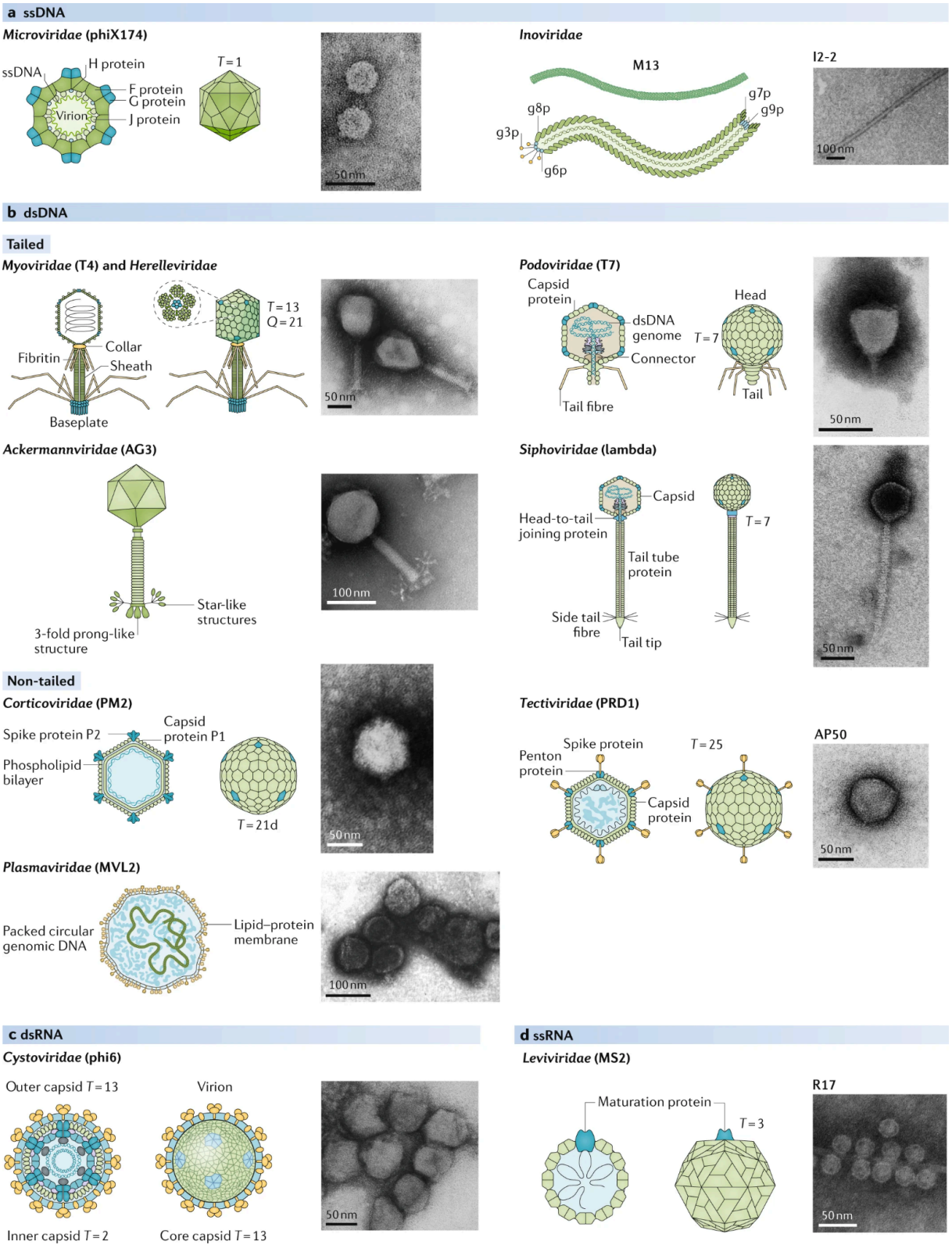


Figure 2: Diversité morphologique des phages

Une représentation schématique (SR) et une photographie en microscopie électronique à transmission (TEM) sont présentées pour chaque morphologie. a) Les *Microviridae* ont des capsides icosaédriques et de petits génomes circulaires à ADN simple brin (ADNss). Le génome des phages filamenteux de la famille des

approche donne des résultats intéressants mais nécessite d'avoir des séquences similaires à la séquence d'intérêt.

Objectifs du stage

Dans le cadre de ce stage, nous avons voulu savoir si l'utilisation d'une approche d'encodage des séquences pouvait aider à discriminer les séquences réellement codantes des séquences non codantes. Pour tester cette méthode, nous avons utilisé les génomes des virus de la famille des *Microviridae* **(Figure 2)** car : **(i)** il existe un important socle de connaissances sur la diversité génomique des microvirus dans l'équipe BioAdapt au sein de laquelle ce stage a été réalisé, (ii) les microvirus ont un rôle écologique important, ils sont ubiquitaires et font partie des phages les plus abondants dans la plupart des écosystèmes et (iii) pour plusieurs microvirus cultivés, les gènes codants pour des protéines ont été déterminés expérimentalement. En outre, les microvirus possèdent un petit génome (<9 000 nucléotides) dans lequel sont encodés une dizaine de gènes, tous sur le même brin **(Figure 3)**. Malgré leur apparente facilité d'analyse, ces génomes contiennent de nombreux petits gènes (pouvant aller jusqu'à 75 nucléotides) et également des gènes chevauchant voire surimprimés, tous ces gènes étant particulièrement difficiles à distinguer des ORFs non codants et étant tout simplement ignorés par les logiciels de prédiction de gènes comme Prodigal.

MATÉRIEL ET MÉTHODES

Ressources informatiques

Un ordinateur portable équipé d'un processeur 16 coeurs AMD Ryzen 7 7745HX, 32 Go de mémoire vive, d'une carte graphique laptop-GPU RTX 4070 8 Gb et 1To de stockage ssd a été utilisé.

Sources des génomes

A partir de métagénomes viraux, issus de microbiotes intestinaux, 8 143 génomes ont été assemblés avec MEGAHIT (v1.2.9, paramètres : --presets meta-large) par Celtill Dumont lors

Inoviridae est composé d'une molécule d'ADN simple brin circulaire superenroulée, emballée dans un long filament (> 500 nm) composé de milliers de copies de la protéine majeure de la capside (MCP). b) La plupart des phages caractérisés sont dotés d'une queue et d'un génome à ADN double brin (ADNdb) et appartiennent à l'ordre des Caudovirales. Cinq familles ont été décrites pour cet ordre : les *Myoviridae* et le *Herelleviridae* , les *Podoviridae*, les *Ackermannviridae* et les *Siphoviridae*. Les *Corticoviridae* ont un génome à ADN double brin circulaire et une capside composée d'une membrane lipidique interne entourée de MCP . Les *Tectiviridae* ont une capside icosaédrique contenant un génome à ADN double brin linéaire et une membrane lipidique interne. Les virus appartenant à la famille des *Plasmaviridae* ont un génome à ADN double brin circulaire entouré d'une enveloppe lipidique et ne possèdent pas de capside. Les *Cystoviridae* possèdent un génome à ARN double brin (ARNdb) contenu dans une capside sphérique. d) Les *Leviviridae* possèdent un génome à ARN simple brin codant pour seulement quatre protéines (MCP, réplicase, protéines de maturation et de lyse) et une capside de forme icosaédrique et sphérique.

Source de la figure: Dion *et al.* 2020

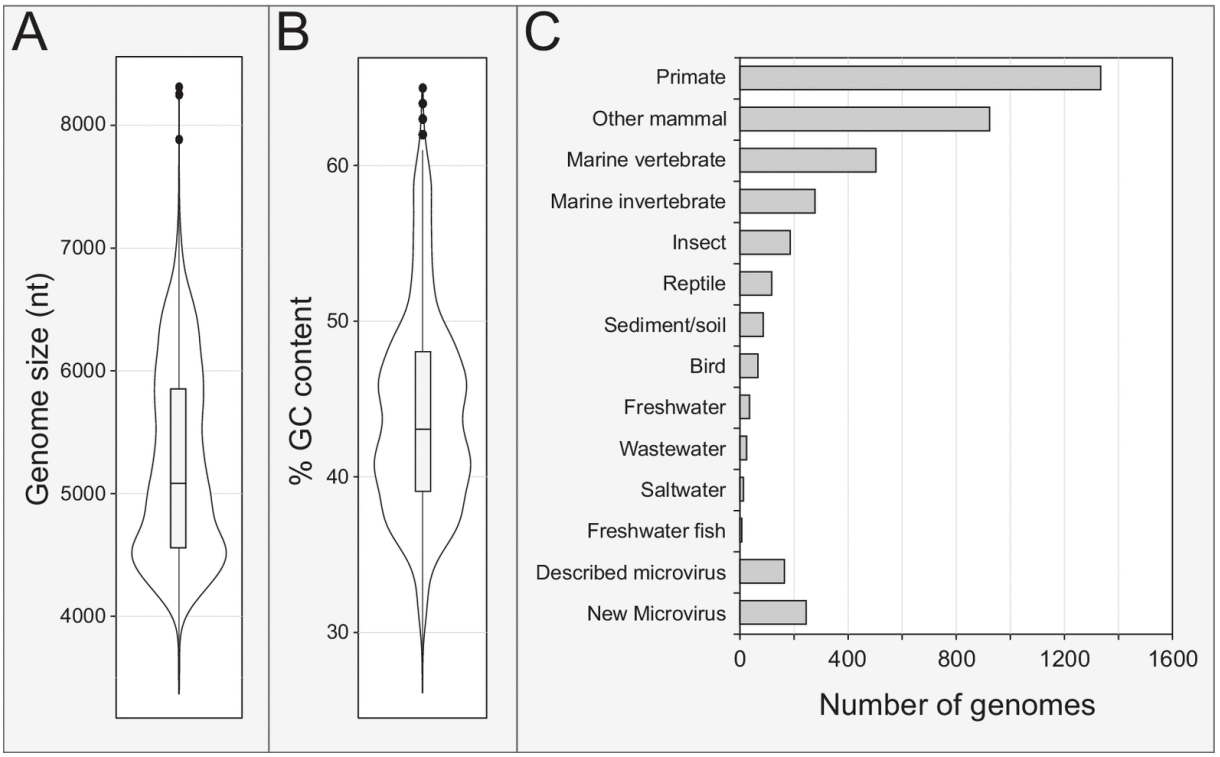


Figure 3: Propriétés et origines des échantillons des génomes de *Microviridae.*

(A) Diagrammes en violon et diagrammes en boîte à moustaches représentant la distribution de la taille des génomes de microvirus. (B) Diagrammes en violon et diagrammes en boîte à moustaches représentant la teneur en GC des génomes de microvirus (même ensemble de données que dans le panneau A). Les diagrammes en boîte et à moustaches indiquent les valeurs médianes, les 25e et 75e centiles, les intervalles interquartiles 1,5, ainsi que les points de données aberrants. (C) Histogramme montrant les sources des échantillons à partir desquels les génomes microviraux ont été détectés/isolés. La barre intitulée « Microvirus décrits » désigne les génomes pouvant être attribués à des genres de microvirus officiels, et la barre intitulée « Nouveaux microvirus » désigne les génomes pouvant être attribués à des hôtes bactériens soit en tant que prophages, soit via des réseaux CRISPR.

Source de la figure: Kirchberger *et al.* 2022

de ses premiers mois de stage de Master II en 2025. Ces génomes sont considérés comme complets car ils sont circulaires (le début et la fin des contigs sont identiques), ont une taille comprise entre deux et 8 kb et encodent tous les gènes codants pour la protéine de capside et de réplication. Ces génomes circulaires ont été linéarisés à partir du début de la protéine de capside. A ces 8 143 génomes, 14 génomes de référence correspondant à des microvirus cultivés ont été ajoutés et également linéarisés sur la protéine de capside.

Prédiction des potentielles régions codantes

Tout d'abord, tous les cadres ouverts de lecture, ou ORFs en anglais pour Open Reading Frame, ont été déterminés. Pour cela, le logiciel Getorf de la suite logicielle EMBOSS **[Rice *et al*]**[5], repère toutes les régions entre un codon start et un codon stop suffisamment longues. Ici, nous avons utilisé les codons start alternatifs, le fait que les génomes soient circulaires et une taille d'ORFs de minimum 25 acides aminés.

Les potentiels gènes codants pour des protéines ont été parallèlement identifiés par le logiciel de prédictions de gènes Prodigal. Prodigal a été exécuté en mode *single* (car appliqué sur des génomes entiers alors que le mode meta est plus adapté aux contigs courts et variables) et avec une longueur minimale de 25 résidus. Afin d'enlever les séquences redondantes, les ORFs correspondants à des gènes prédits par Prodigal ont été repérés en comparant ces séquences avec MMseqs2 **[Steinegger et Söding]**[6] (commande *easy-search* avec les paramètres --min-seq-id 0.95 -c 0.95) et ne seront pas considérés dans le reste des analyses.

Ces séquences ont été annotées fonctionnellement en les comparant aux protéines de microvirus (mmse) annotées lors d'un travail précédent menés dans l'équipe par Eric Olo Ndela avec MMseqs2 easy-search (-c 0.3, –max_seqs 1, identity par défaut).

Clusterisation des génomes et conservation des séquences

Afin de calculer un indice de la conservation des ORFs et des protéines prédites, les génomes ont tout d'abord été regroupés en comparant les protéines de capside de chaque génome avec MMseqs2 (commande *easy-cluster* avec les paramètres --min-seq-id 0.50 -c 0.75), pour chaque cluster, la plus longue protéine (capside) avec le plus grand nombre de connexions au sein du cluster a été désignée par MMseqs2 comme la séquence représentative du cluster. Ensuite, pour chaque ORF ou gène, la proportion des génomes du même cluster contenant une séquence similaire a été déterminée à partir de la comparaison entre toutes ces séquences (MMseqs2,easy -search --max-seqs 1000 -c 0.5 --cov-mode 0).

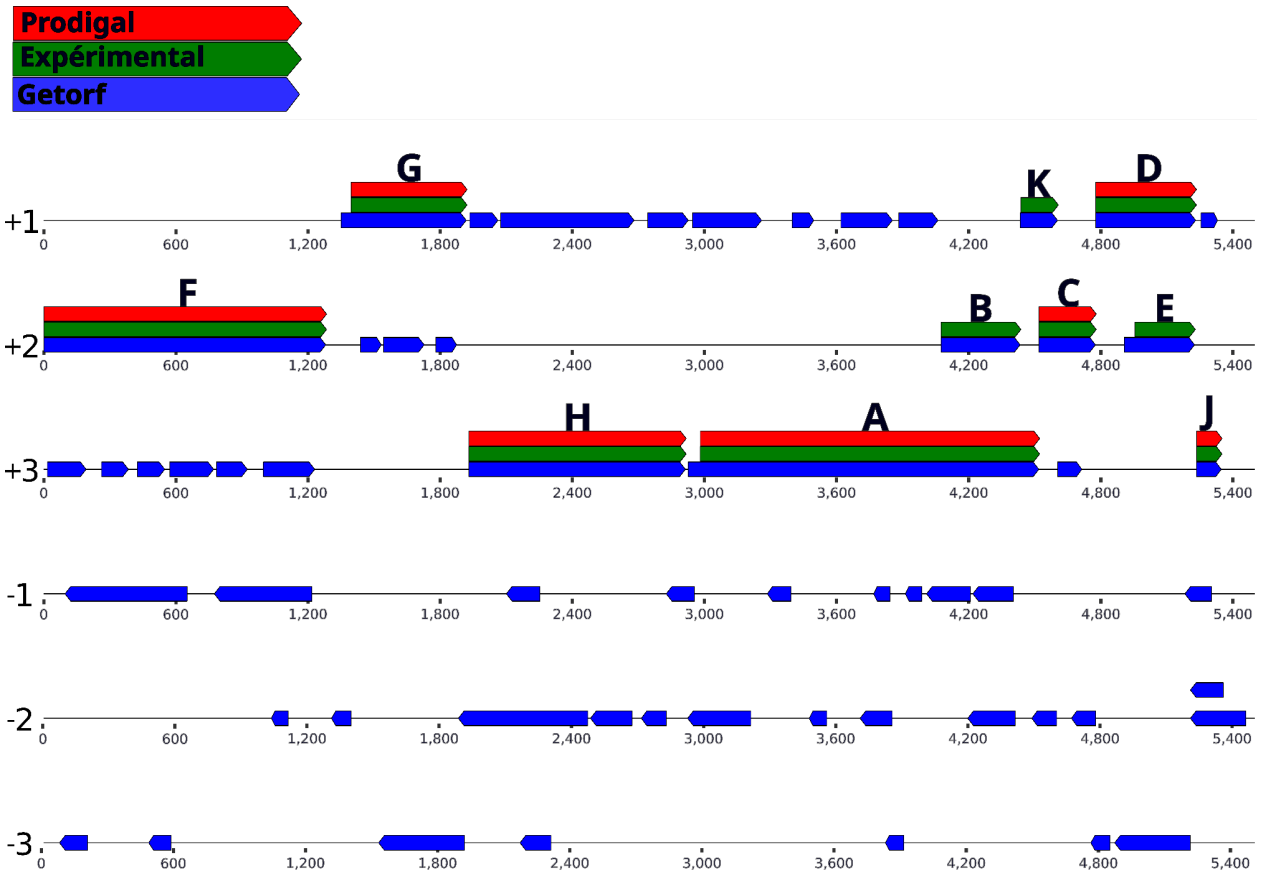


Figure 4: Carte génomique du phage ΦX174 (5 386 nt).

Représentation via DNAViewer: montrant les six cadres de lecture, trois par brin. Les segments rouges correspondent aux protéines prédites par Prodigal; les segments bleus aux ORFs donnés par Getorf ; les segments verts indiquent les protéines validées expérimentalement, considérées exhaustives au vu du nombre d'études consacrées à ce phage modèle. Chaque protéine est annotée par une lettre : dans l'ordre génomique: F (capside majeure), G (spike majeur), H (spike mineur / pilot), A (réplication), C (inhibiteur de synthèse dsDNA), D (scaffolding externe), J (liaison ADN). Trois protéines surimprimées: B (scaffolding interne), K (burst size modulation) et E (lysine) sont visibles en vert : elles disposent d'un ORF détecté mais ne sont pas reconnues par Prodigal, illustrant les limites de la prédiction classique face aux petits gènes chevauchants.

Embedding des séquences protéiques

Chaque ORF et gène (sous forme de séquences d'acides aminés) a ensuite été "encodé" ou "vectorisé" (en anglais, *embedded*) à l'aide d'un modèle de langage large (LLM) protéique, à savoir ProtT5-XL-UniRef50 **[Elnaggar *et al*]**[7]. Pour chaque séquence de départ, un vecteur de 1024 dimensions est obtenu. Ces calculs ont été effectués sur un GPU (Geforce RTX 4070 8Go) en vectorisant d'abord les résidus des séquences protéiques, puis en calculant la moyenne pour obtenir un vecteur par séquence protéique. Les séquences trop longues pour être vectorisées en une seule fois par GPU l'ont été en utilisant le mode CPU.

Réduction de dimension avec UMAP

Pour permettre la visualisation des vecteurs, l'algorithme UMAP (*Uniform Manifold Approximation and Projection*) a été appliqué. L'implémentation GPU de cuML-UMAP v25.02 **[Nolet *et al*]**[8] a été choisie, permettant d'utiliser le mode brute force KNN et de calculer les distances exactes même sur des jeux de données volumineux. La métrique cosinus a été choisie pour le calcul des distances car plus adaptée aux vecteurs à haute dimension que les métriques plus classiques comme la métrique euclidienne.

Dans le but de séparer les familles fonctionnelles, les deux hyper-paramètres principaux, n_neighbors et min_dist ont été ajustés de manière empirique selon le cas: un couple (n_neighbors = 15 / min_dist = 0.1) optimisé pour la vue globale et un réglage plus serré (n_neighbors = 10 / min_dist = 0.01) plus adapté pour la résolution locale dans le but de détecter de petits sous clusters ou gènes plus difficilement détectables et différenciables du bruit de fond.

Deux autres paramètres ont été fixés : n-epochs à 5000 qui consiste en le nombre d'itérations durant lesquels umap ajuste les coordonnées des points, une valeur élevée favorisant la convergence des résultats , et random-state (avec 42 comme seed) garantissant la reproductibilité exacte. Tous les autres paramètres ont été laissés par défaut.

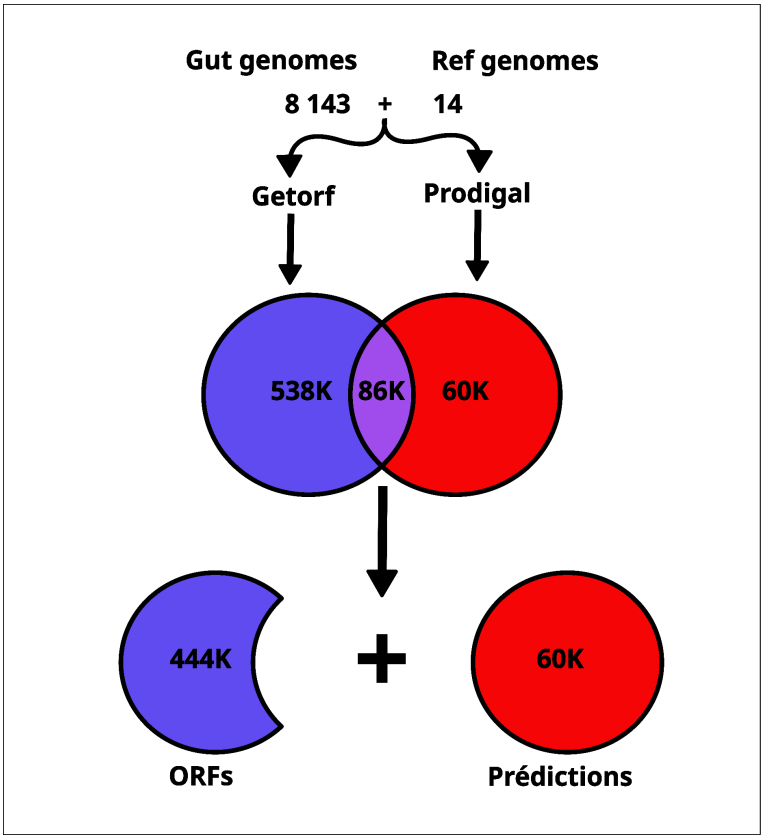


Figure 5: diagramme de Venn illustrant le recoupement entre ORFs et prédictions protéiques.

La recherche d'ORFs ainsi que la prédiction protéique ont été réalisés sur un ensemble de 8 157 génomes (8 143 provenant du Gut ainsi que 14 génomes de référence), aboutissant à approximativement 538 000 ORFs en bleu et 60 000 protéines prédites en rouge. Parmi ces ORFs, 86 000 en violet présentaient une correspondance avec au moins une des protéines prédites et ont donc été retirés de la liste des ORFs.

Protéines surimprimées vérifiées expérimentalement

Quatre ORFs non détectés par Prodigal ont été annotés manuellement sur la base de données expérimentales. Trois d'entre eux correspondent aux protéines surimprimées B, K et E du phage ΦX174 **(Figure 4)**. Le quatrième est la protéine K du microvirus *Ebor Gok_MT185428_EC6098*, appartenant aux Gokusho. L'existence de cette dernière nous a été confirmée par une communication personnelle avec J. Kirchberger, qui en a démontré l'expression et prévoit d'en publier les résultats prochainement.

Accessibilité des cripts et reproductibilité

L'ensemble des scripts est déposé dans un dépôt Git public, deux fichiers YAML contiennent les environnements Conda nécessaires : env_cpu.yml (analyses légères) et env_gpu.yml (embeddings ProtT5 et UMAP cuML): https://github.com/Tulgar-BioInformatic/phage-pppes Deux visualisations UMAP interactives (umap_global.html, umap_central.html) sont également déposées dans le dépôt Git et consultables directement via GitHub Pages avec le lien suivant : https://tulgar-bioinformatic.github.io/phage-pppes/ .

RESULTATS / DISCUSSION

Dans ce qui suit, « ORF » désigne exclusivement les cadres de lecture ouverts détectés par Getorf, « protéine » (ou « prédiction protéique ») les gènes codants appelés par Prodigal, et le terme générique « séquence » englobe l'un ou l'autre de ces deux ensembles.

Analyse de la projection globale

A partir des 8 143 génomes de microvirus issus de l'intestin humain et des 14 génomes de référence, 538 805 ORFs et 60 059 protéines ont été identifiées par Getorf et Prodigal respectivement et ainsi 86 493 ORFs correspondant à une des ces protéines prédites ont été supprimées (Figure 5). A partir des similarités de leur protéine de capside (>50 % d'identité), ces génomes ont été regroupés en 289 clusters, 26 d'entre eux ayant une taille supérieure à 60. Les génomes représentatifs de ces 26 clusters et 14 génomes de référence additionnels

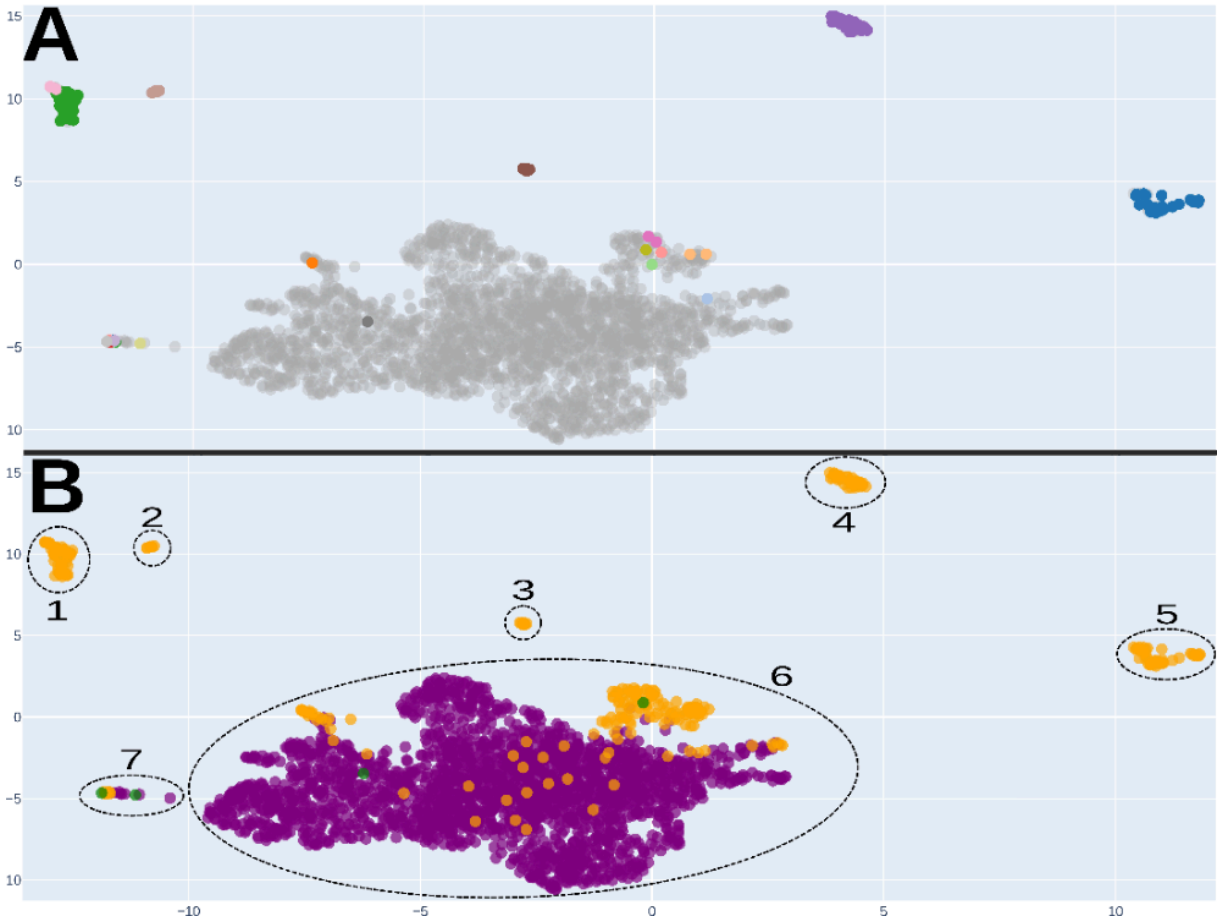


Figure 6: Projection UMAP des 2 562 séquences protéiques potentielles issues de 8 157 génomes de Microviridae.

Les 2 562 vecteurs ProtT5 (1 024 dimensions) ont été réduits en deux dimensions par UMAP (n_neighbors = 15, min_dist = 0,1) puis représentés deux fois pour mettre en évidence différents paramètres. Les axes représentent les deux dimensions issues de la projection, leurs valeurs sont sans unité et servent à visualiser la proximité relative des points dans l'espace d'origine. **(A)** Annotation fonctionnelle (MOG ID) : cinq familles majeures se détachent : Replication (vert), Pilot (mauve), Capsid (bleu), DNA Binding (brun), Internal Scaffolding (marron-clair), alors que les séquences non annotées (gris) occupent le noyau central. **(B)** Outil de prédiction et Synthèse : la majorité des points provient de Getorf (violet, n = 2 271), tandis que les gènes identifiés par Prodigal (orange, n = 287) ou validés expérimentalement (vert, n = 4) forment plusieurs îlots périphériques. Sept régions numérotées (pointillés noirs) sont distinguées ; les îlots 1–5 et 7 contiennent presque exclusivement des protéines Prodigal, dont le groupe 7 correspond aux séquences du phage ΦX174, tandis que le noyau 6 rassemble ~94 % des ORFs Getorf et quelques protéines périphériques.

contiennent 2 275 ORFs et 287 protéines prédites. Les embeddings de ces 2 562 séquences ont été soumis à une analyse UMAP dont le but est de réduire la répartition spatiale de ces vecteurs de 1024 valeurs en deux dimensions **(Figure 6)**.

Les séquences sont séparées en un grand groupe central (94% des séquences) et six petits groupes différents. Cinq de ces petits groupes ne sont constitués que de protéines prédites par Prodigal **(Figure 6-B)** qui sont fortement conservées **(Annexes 2 et 3)**, sont dominés par une ou deux annotations fonctionnelles **(Figure 6-A)** : **(i)** le groupe 1 contient des protéines de réplication (MOG2) et des peptidases (MOG9) et quelques protéines non annotées, **(ii)** le groupe 2 contient les protéines dites d'échafaudage aidant la formation de la particule virale ("Internal scaffolding", MOG5) et une protéine non annotée, **(iii)** le groupe 3 contient les protéines se fixant à l'ADN ("DNA binding", MOG4), **(iv)** le groupe 4 contient des protéines dites "Pilot", associées soit aux MOG3 ou aux MOG26 ainsi que cinq non annotées et **(v)** le groupe 5 contient les protéines de capsides et des protéines non annotées.

Ces résultats préliminaires nous montrent que les protéines de fonctions différentes sont pour la plupart projetées de manière distincte, et que la plupart de ces protéines sont réparties distinctement de la majorité des ORFs. Il est également intéressant de noter que toutes les protéines Pilot, dont dix copies sont embarquées dans le virion puis assemblées pour former un tube par lequel est injecté l'ADN dans la cellule hôte, ont des embeddings proches alors que pour certaines d'entre elles, aucune similarité de séquence n'est trouvée, même avec des outils de détection sensibles comme les HMMs.

Le dernier îlot (groupe 7)**(Figure 6-B et Annexe 4)** est composé de 14 séquences, toutes issues de phiX174, parmi lesquelles les six protéines prédites par Prodigal et huit ORFs, dont deux présents sur le brin négatif et cinq sur le brin positif dont deux protéines surimprimées vérifiées expérimentalement E ("endolysin") et B ("internal scaffolding"), ces deux protéines n'étant pas prédites par prodigal. Contrairement aux autres microvirus, les protéines de capside, de réplication, la "Pilot", ou encore les protéines d'échafaudage de phiX174 ne sont pas dans les groupes précédemment cités et sont ici regroupées, suggérant des caractéristiques particulières liées à sa distance phylogénétique notable avec les autres microvirus **[Rokyta et *al.* 2026]**[9].

Enfin, un grand groupe central regroupe la quasi-totalité des ORFs 2 269 ainsi que 149 protéines prédites. Parmi ces dernières, 26 sont au milieu du groupe et 123 sont à la périphérie. Les séquences du brin négatif sont réparties un peu différemment de celles du brin positif avec certaines "péninsules", composées uniquement de séquences provenant d'un seul brin **(Annexe 3)**.

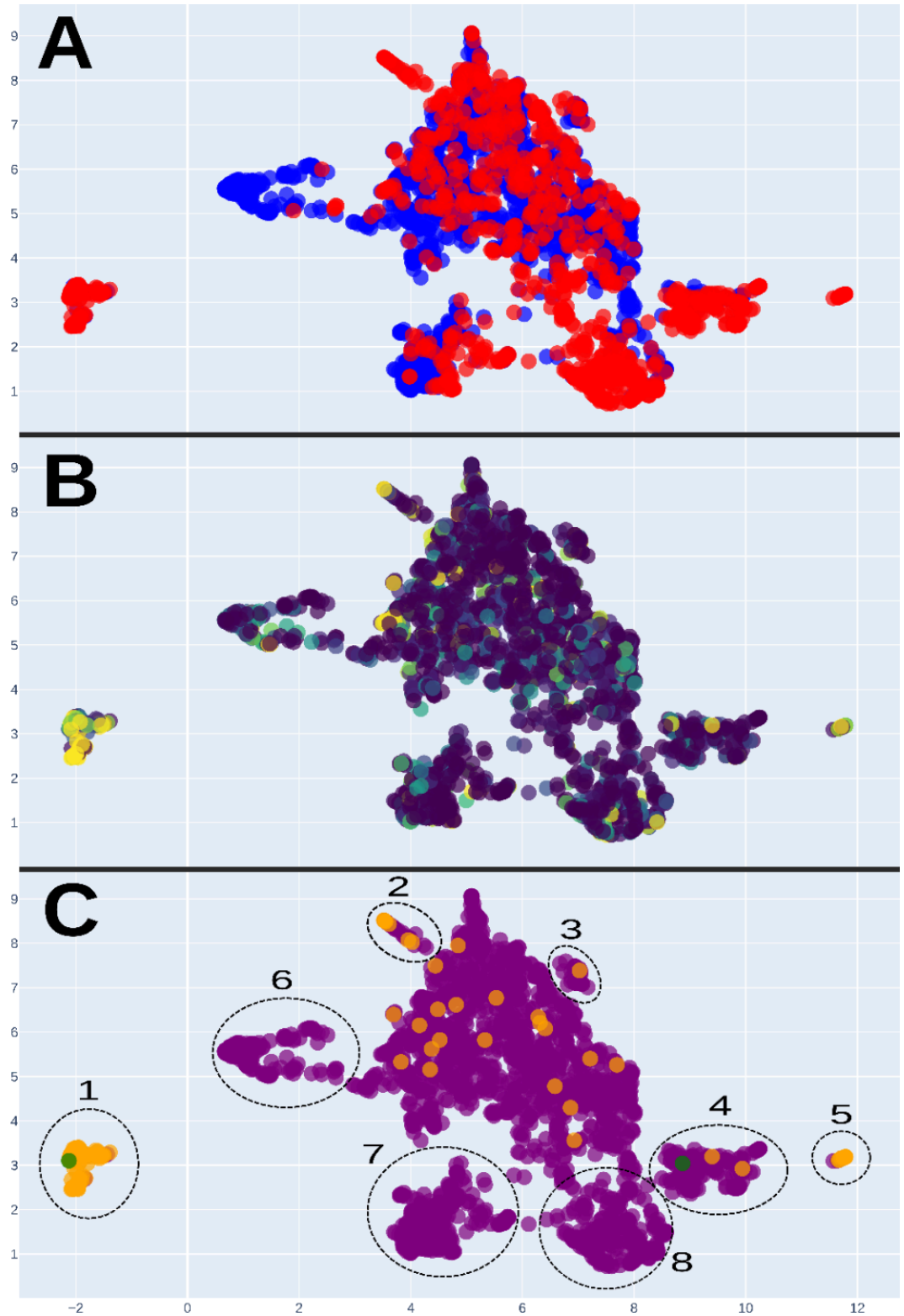


Figure 7: UMAP ciblant le « groupe central » (2 418 séquences) issu de la projection globale.

Les embeddings ProtT5 des ORFs/gènes situés dans la zone centrale de la Figure 4 ont été une nouvelle fois projetés par UMAP (n_neighbors = 10, min_dist = 0,01) pour accroître la résolution locale. Les axes représentent les deux dimensions issues de la projection, leurs valeurs sont sans unité et servent à visualiser la proximité relative des points dans l'espace d'origine. La même carte est montrée trois fois : **(A)** orientation du brin (rouge = brin +, bleu = brin –) ; **(B)** degré de conservation au sein de chaque cluster capsidique (indice de conservation, de mauve à jaune: 0 – 100 %) ; **(C)** origine de la prédiction (violet = ORFs Getorf, orange = gènes Prodigal, vert = gènes validés expérimentalement).

Analyse de la projection centrale

Afin d'affiner la discrimination au sein du groupe central, une seconde projection UMAP a été réalisée sur les 2 418 embeddings de ce groupe. Neuf groupes ressortent clairement avec un noyau central (85 % des séquences) et huit îlots ou « péninsules » distincts **(Figure 7)** :

(i) le groupe 1 **(Figure 8-A)** rassemble 109 séquences dont sept seulement appartenant au brin négatif, parmi ces dernières, quatre prédites par Prodigal. L'ensemble du groupe est constitué de protéines prédites avec huit ORFs. On observe une variabilité de catégories fonctionnelles avec une protéine surimprimée vérifiée expérimentalement (protéine K) ainsi que des protéines annotées "RHH", "Spike", "External Scaffolding" et "Turquoise", **(ii)** le groupe 2 **(Figure 8-B)** comptent 37 séquences dont deux seulement du brin négatif, 9 protéines prédites dont 7 se concentrant à l'une des extrémités du pic ou de la ligne formée par l'allure générale du groupe. On y retrouve une protéine annotée "ssDNA binding" de phix174, **(iii)** le groupe 3 avec une seule protéine prédite, sur le brin négatif et des ORFs principalement du brin négatif, **(iv)** le groupe 4 **(Figure 9-A)** avec deux protéines et 169 ORFs, majoritairement des ORFs du brin positif dont une protéine surimprimée vérifiée expérimentalement qui précédemment faisait partie du noyau central, **(v)** le groupe cinq **(Figure 9-B)** avec 23 séquences, toutes du brin positif et six ORFs s'ajoutant aux 17 prédictions, **(vi)** les groupes 6, 7 et 8, péninsules ne contenant que des ORFs, le groupe 6 composé quasi exclusivement d'ORFs du brin négatif, le groupe 8 du brin positif et le groupe 7 d'un mélange des deux brins, **(vii)** le reste des séquences formant le noyau central avec approximativement 400 ORFs contre 19 prédictions protéiques probablement non codantes.

Huit îlots ou « péninsules » numérotés 1-8 (cercles pointillés) se détachent d'un noyau central (1 351 points ; 55% du total).

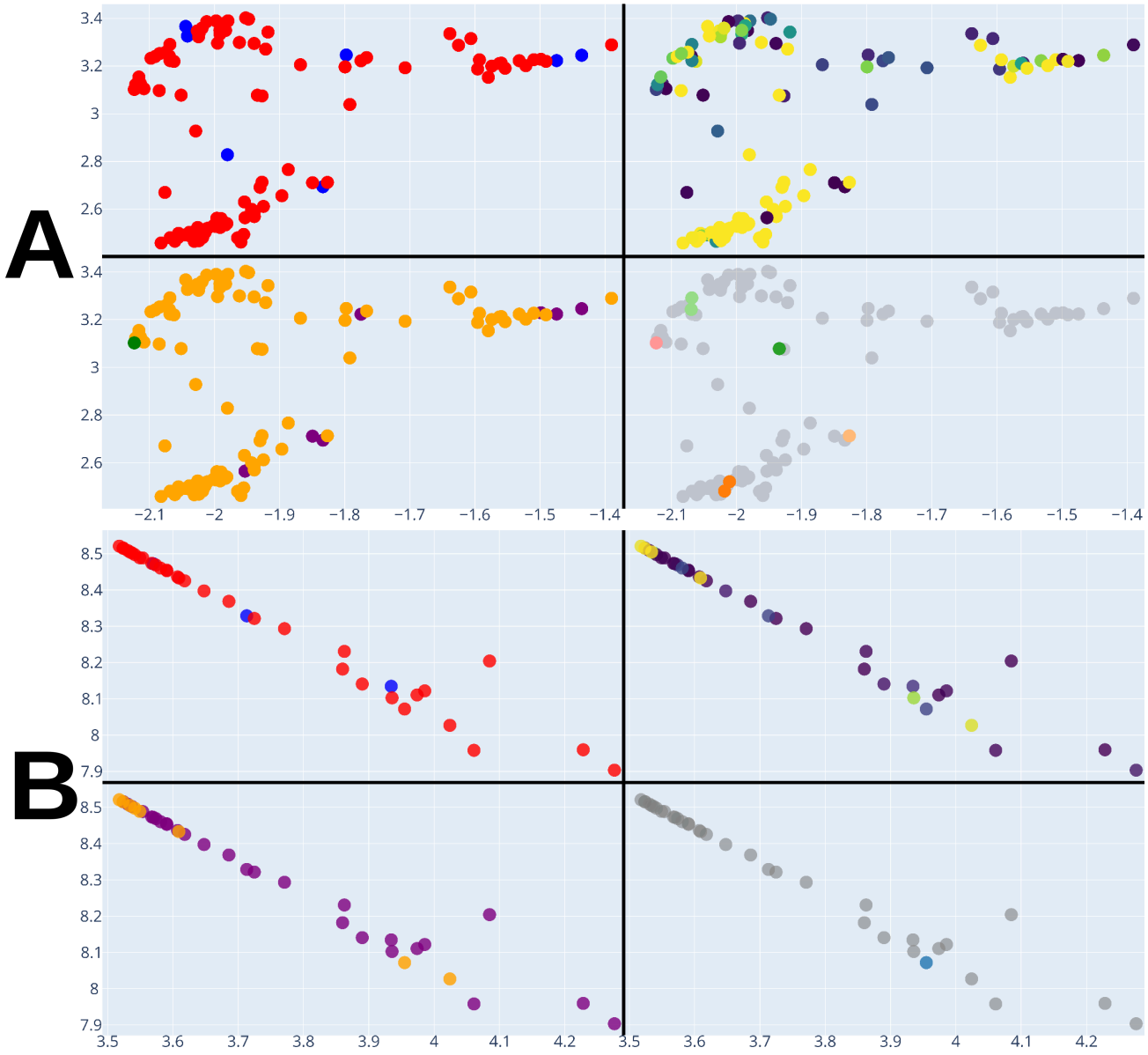


Figure 8: Analyse approfondie des groupes 1 et 2 issus de la deuxième projection UMAP (cf. Figure 7).

Chaque groupe est représenté sous forme de quatre sous-figures correspondant aux paramètres suivants : orientation du brin (en haut à gauche : rouge positif, bleu négatif), degré de conservation (en haut à droite : gradient violet vers jaune, faible à forte conservation), outil de prédiction utilisé (en bas à gauche : orange pour Prodigal, violet pour Getorf, vert pour protéine validée expérimentalement) et annotation fonctionnelle MOG (en bas à droite), les axes représentent les deux dimensions issues de la projection, leurs valeurs sont sans unité et servent à visualiser la proximité relative des points dans l'espace d'origine. **(A)** Groupe 1 (109 séquences) : majoritairement des séquences du brin positif (102), avec sept séquences négatives dont quatre prédites par Prodigal. Présence d'une protéine surimprimée validée expérimentalement (protéine K). Ce groupe est fonctionnellement diversifié, avec des protéines annotées « RHH », « Spike », « External Scaffolding » ainsi que des séquences d'une catégorie non annotée (turquoise). **(B)** Groupe 2 (37 séquences) : 35 séquences du brin

CONCLUSIONS ET PERSPECTIVES

Cette étude exploratoire visait à déterminer si la combinaison d'embeddings issus de modèles de langage protéiques suivi d'une projection par UMAP, approche jusqu'ici principalement utilisée pour la classification fonctionnelle, pouvait améliorer la prédiction de gènes chez les microvirus en complément des approches existantes. Plusieurs résultats intéressants ont été obtenus.

Premièrement, les protéines prédites ayant une annotation connue, et donc réellement codantes, ont des embeddings distincts de ceux des ORFs qui représentent pour la très grande majorité des séquences non codantes. Ce résultat valide la pertinence de l'approche, même si cela mérite d'être affiné. Comme attendu, les protéines ayant une même annotation (capside, réplication, etc.) sont regroupées par la projection, et ce même pour les protéines n'ayant pas de similarités de séquences détectées. C'est notamment le cas pour les protéines Pilot, pour lesquelles celles encodées par les Bullavirus et les Gokushovirus, sont ici regroupées alors qu'aucune trace de similarités n'est trouvée pour ces protéines, soit parce qu'elles ont beaucoup accumulé de mutations ou parce qu'elles ont eu une évolution convergente. La proximité de leurs embeddings est peut-être dûe au fait qu'elles sont constituées de motifs répétés aboutissant à une structure hélicoidale intervenant dans le mécanisme du "penton blooming" **[Bardy *et al*]**[10], qui fait que leurs embeddings sont similaires entre eux et différents de celui des autres protéines. De plus, plusieurs îlots (groupes 2 et 5) **(figure 7, Figure 8-B et Figure 9-B)** contiennent des ORFs et des protéines prédites rapprochées sur la projection et faisant partie du brin positif, ce qui suggère que toutes ces séquences soient codantes et que ces ORFs correspondent à des gènes non détectés. Un résultat similaire, même si moins clair, a été observé pour les groupes 1 et 4, contenant plus de 100 séquences, chacune avec une minorité sur le brin négatif et la présence d'une protéine surimprimée validée dans chacune.

Des analyses avec d'autres méthodes, comme t-SNE (T-distributed Stochastic Neighbor Embedding) ou l'étude des distances cosine des vecteurs de ces séquences, pourraient sans doute permettre d'affiner ces résultats et de vérifier s'il est possible de discriminer de manière plus claire les séquences codantes des non codantes. Une autre piste pour améliorer cette discrimination serait de filtrer les ORFs non codants afin qu'ils soient moins nombreux et

positif et deux du brin négatif. Il est constitué principalement de protéines prédites par Prodigal, regroupées en majorité à une extrémité du groupe. Présence d'une protéine annotée « ssDNA binding » issue du phage phiX174.

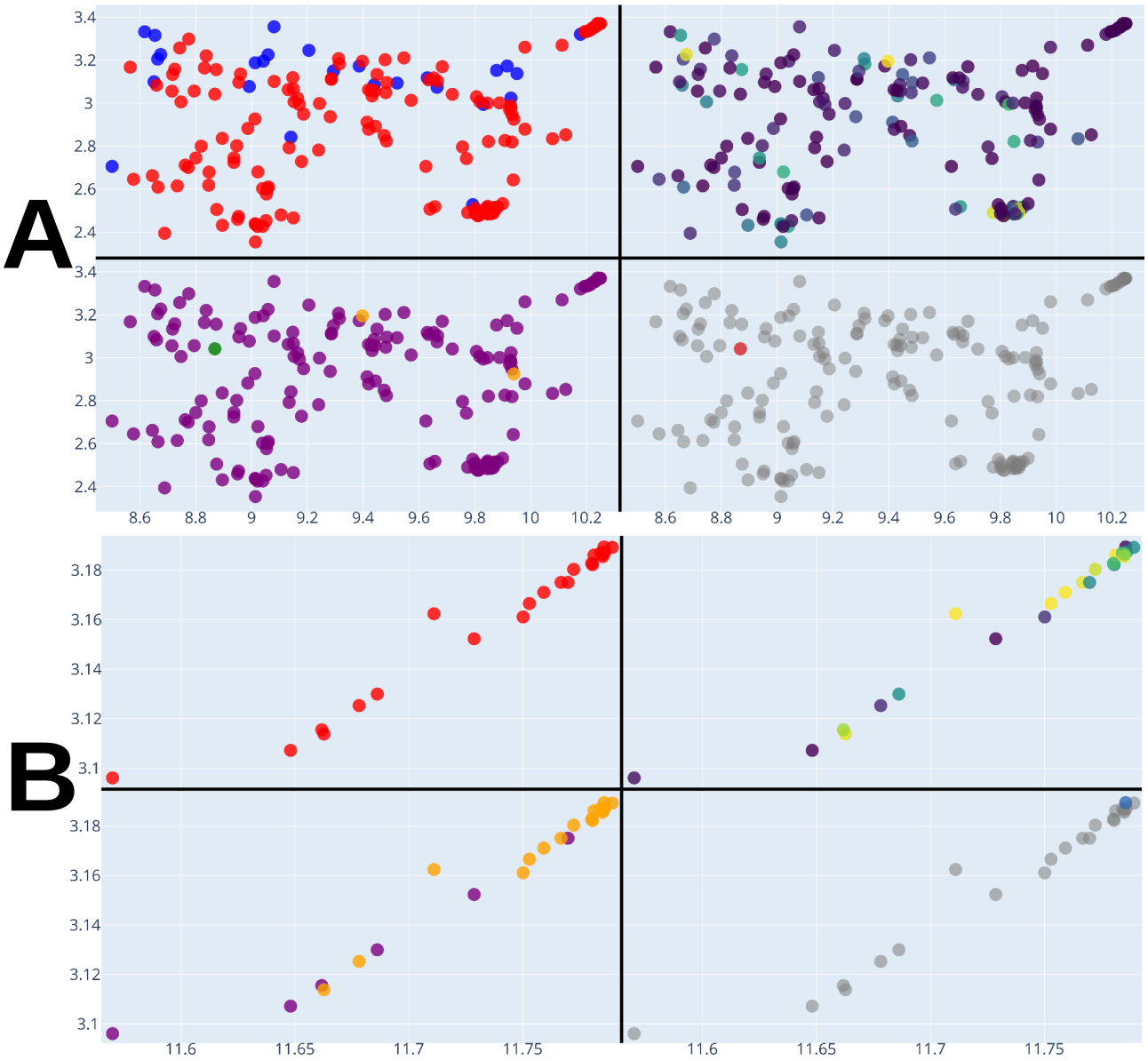


Figure 9: Analyse approfondie des groupes 4 et 5 issus de la deuxième projection UMAP (cf. Figure 7). Chaque groupe est représenté sous forme de quatre sous-figures correspondant aux paramètres suivants : orientation du brin (en haut à gauche : rouge positif, bleu négatif), degré de conservation (en haut à droite : gradient violet vers jaune, faible à forte conservation), outil de prédiction utilisé (en bas à gauche : orange pour Prodigal, violet pour Getorf, vert pour protéine validée expérimentalement) et annotation fonctionnelle MOG (en bas à droite). Les axes représentent les deux dimensions issues de la projection, leurs valeurs sont sans unité et servent à visualiser la proximité relative des points dans l'espace d'origine. **(A)** Groupe 4 (171 séquences) : en grande partie composé d'ORFs du brin positif (145), avec deux protéines prédites par Prodigal ainsi que une protéine surimprimée vérifiée expérimentalement. **(B)** Groupe 5 (23 séquences) : exclusivement constitué de

qu’ils aient un poids moindre. Nous pourrions par exemple les clusteriser par similarité de séquences et prendre uniquement une séquence représentative par cluster. Il serait également intéressant d’accroître la diversité des génomes considérés afin d’aider à ce que les projections soient mieux réparties, en évitant par exemple que les séquences codantes et non codantes comme celles de phiX174 ne soient regroupées par affiliation taxonomique.

séquences du brin positif, dont 17 protéines prédites par Prodigal et six ORFs identifiés par Getorf. Ce groupe compact à l'une des extrémités est fortement conservé. On y trouve une protéine annotée "Holin".

BIBLIOGRAPHIE

https://doi.org/10.1128/jb.188.3.1134-1142.2006

1. Chevallereau, A., Pons, B.J., Van Houte, S., and Westra, E.R. (2022). Interactions between bacterial and phage communities in natural environments. Nat. Rev. Microbiol. *20*, 49–62. https://doi.org/10.1038/s41579-021-00602-y.

2. Dion, M.B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. *18*, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

3. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

4. Anders, J., and Stadler, P.F. (2024). RNAcode_Web – Convenient identification of evolutionary conserved protein coding regions. J. Integr. Bioinforma. *20*, 20220046. https://doi.org/10.1515/jib-2022-0046.

5. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

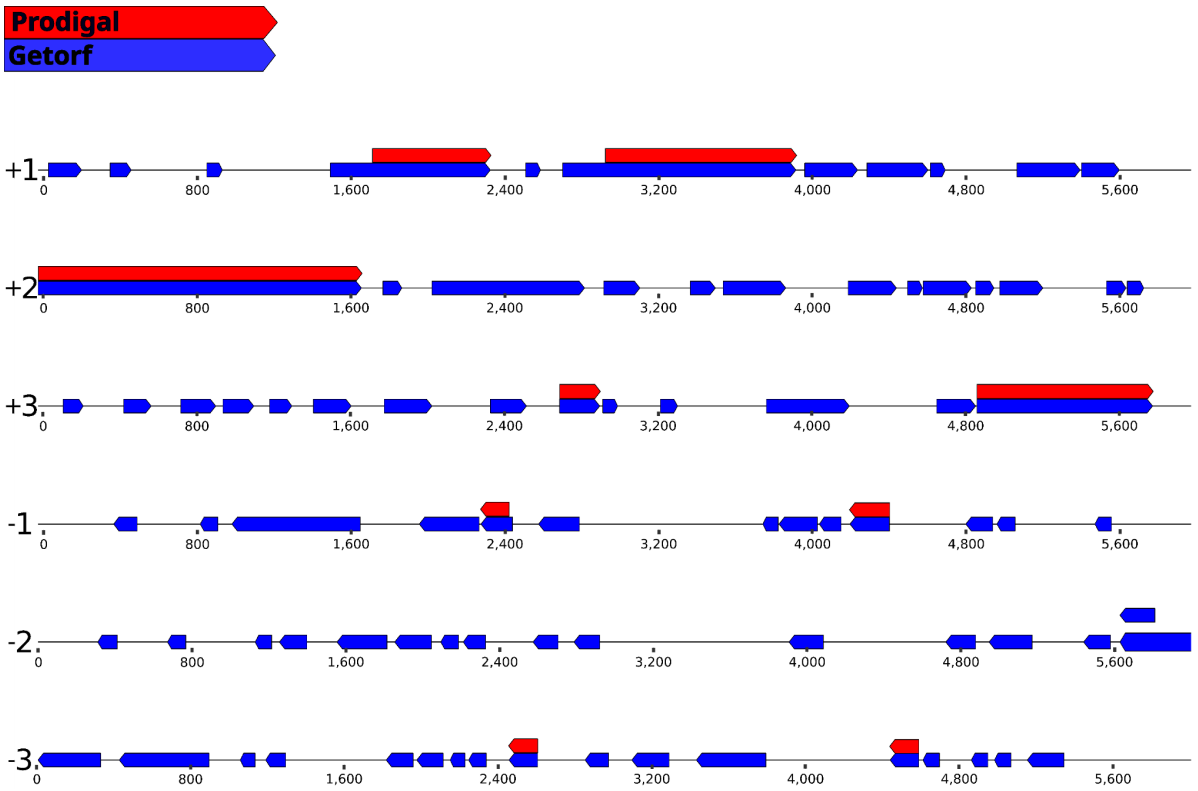
6. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

7. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. *44*, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

8. Nolet, C.J., Lafargue, V., Raff, E., Nanditale, T., Oates, T., Zedlewski, J., and Patterson, J. (2021). Bringing UMAP Closer to the Speed of Light with GPU Acceleration. Preprint at arXiv, https://doi.org/10.48550/arXiv.2008.00325 https://doi.org/10.48550/arXiv.2008.00325.

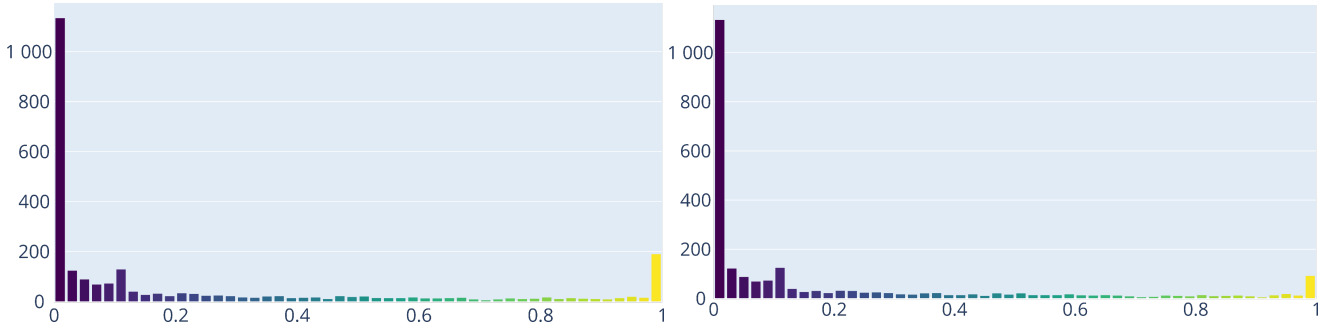
9. Rokyta, D.R., Burch, C.L., Caudle, S.B., and Wichman, H.A. (2006). Horizontal Gene Transfer and the Evolution of Microvirid Coliphage Genomes. J. Bacteriol. *188*, 1134–1142. https://doi.org/10.1128/jb.188.3.1134-1142.2006.

10. Bardy, P., MacDonald, C.I.W., Kirchberger, P.C., Jenkins, H.T., Botka, T., Byrom, L., Alim, N.T.B., Traore, D.A.K., Koenig, H.C., Nicholas, T.R., et al. (2025). Penton blooming, a conserved mechanism of genome delivery used by disparate microviruses. mBio *16*, e03713-24. https://doi.org/10.1128/mbio.03713-24.



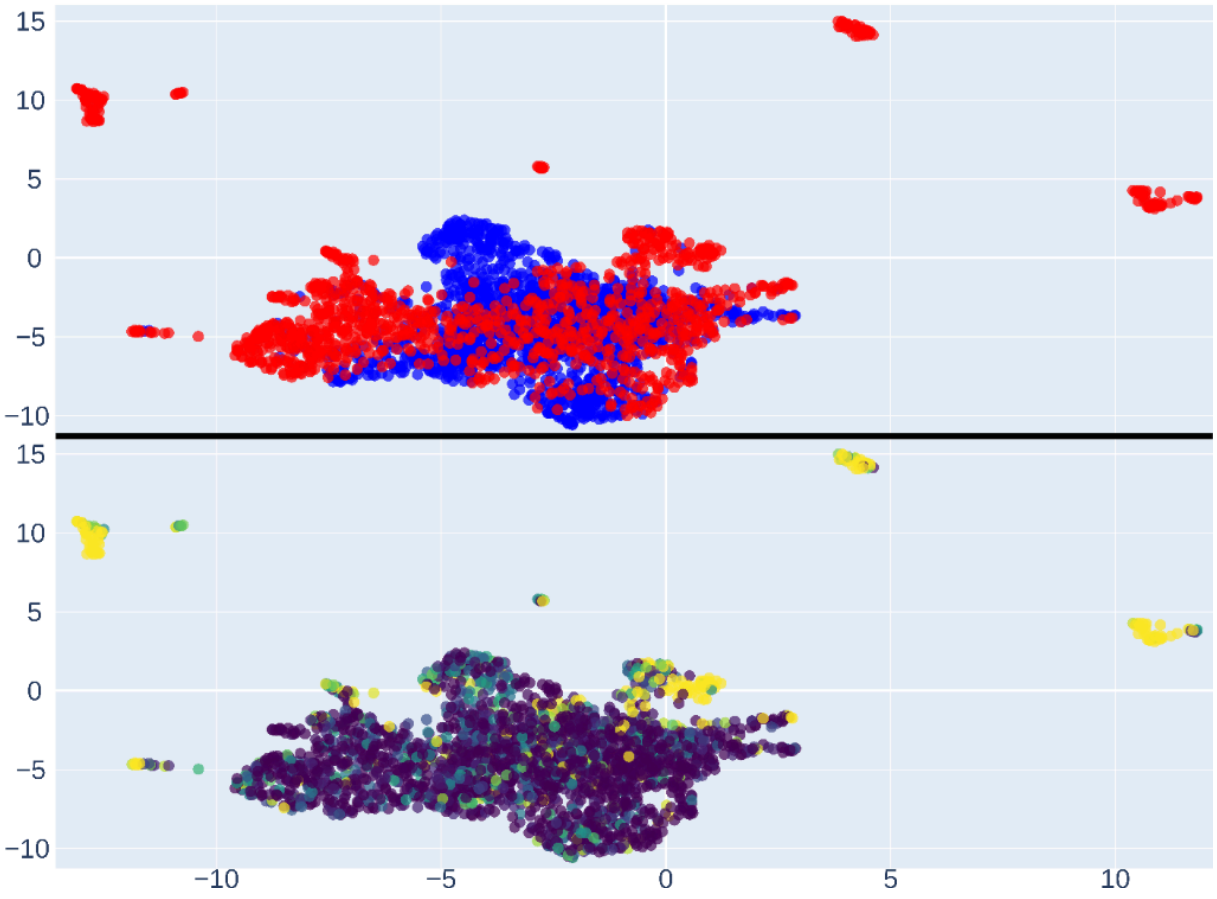
Annexe 1 : Carte génomique du phage Occultatum (6 200 nt).

Affichage des six cadres de lecture via DNAViewer. Les segments rouges (Protéines Prodigal) et bleus (ORFs Getorf). Est entouré un ORF non retenu, probablement codant car long et entre deux protéines codantes, sans doute écarté par Prodigal pour limiter les chevauchements. La prédiction de deux des quatre petites protéines sur le brin négatif et donc probablement pas codantes sont sûrement dû au fait que Prodigal cherche à combler les espaces vides générés par la non-prédiction de protéines chevauchantes.



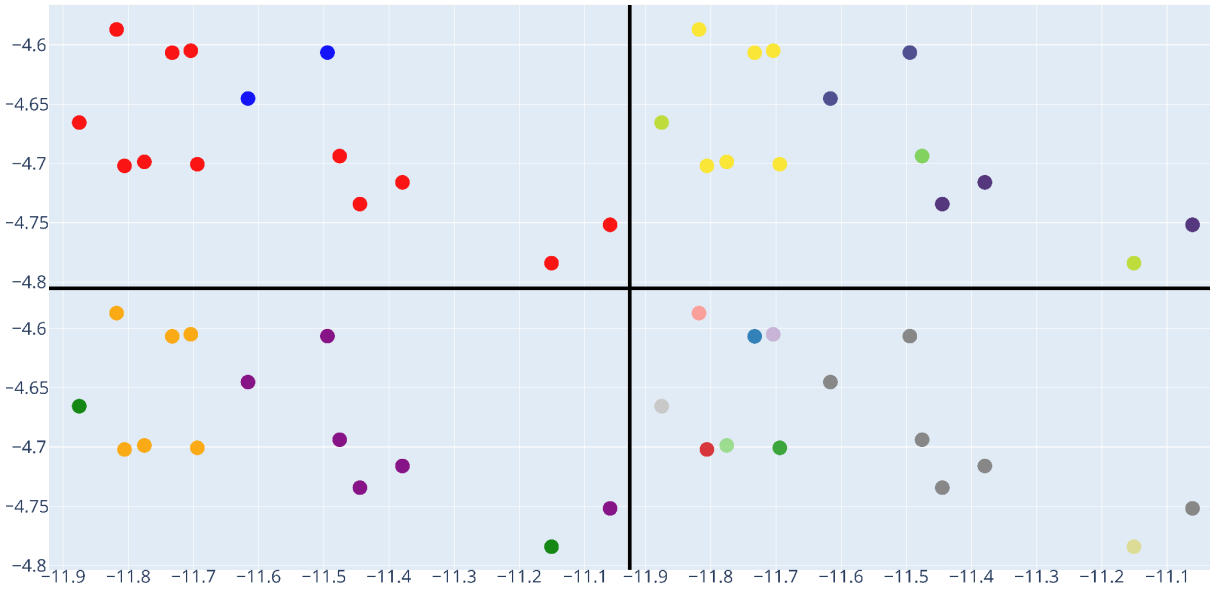
Annexe 2 : Histogrammes des valeurs de conservation des séquences.

Les deux histogrammes comparent les niveaux de conservation des séquences utilisées dans les projections UMAP. Celui de gauche correspond à l'ensemble des séquences de l'UMAP général, celui de droite à celles conservées dans l'UMAP centré sur le groupe principal. Dans les deux cas, la majorité des séquences présentent une faible conservation, avec une répartition marquée entre séquences peu et très conservées. La principale différence réside dans la diminution du nombre de séquences fortement conservées dans le second histogramme (d'environ 200 à une centaine), ce qui s'explique par le retrait des séquences fortement conservées présentes dans les îlots périphériques avant la réalisation du deuxième UMAP.



Annexe 3 : Complément à la figure 6.

Représentation du premier UMAP, avec une coloration alternative des points. En haut, les points sont colorés selon l'orientation du brin (rouge = brin +, bleu = brin –) ; en bas, selon le degré de conservation au sein de chaque cluster capsidique (indice de conservation, de mauve à jaune: 0 – 100 %).



Annexe 4 : Analyse approfondie du groupe 7 de la première projection UMAP (cf. Figure 6).

Les quatre sous-figures correspondant aux paramètres suivants : orientation du brin (en haut à gauche : rouge positif, bleu négatif), degré de conservation (en haut à droite : gradient violet vers jaune, faible à forte conservation), outil de prédiction utilisé (en bas à gauche : orange pour Prodigal, violet pour Getorf, vert pour protéine validée expérimentalement) et annotation fonctionnelle MOG (en bas à droite), les axes représentent les deux dimensions issues de la projection, leurs valeurs sont sans unité et servent à visualiser la proximité relative des points dans l'espace d'origine. Ce groupe ne contient que des protéines et ORFs issus du génome de PHIX174, on observe majoritairement des séquences du brin positif (12), avec deux séquences négatives. Présence de deux protéines surimprimées validées expérimentalement; protéine B "internal scaffolding" et protéine E "endolysin", on constate aussi la présence de 6 autres protéines associées à des MOGs représentées par des couleurs différentes dans la partie annotation fonctionnelle.