HAL open science

Détection de recombinaisons génomiques et protéomiques homologues par alignement multiple local et partiel

Benjamin Blanc

To cite this version:

Benjamin Blanc. Détection de recombinaisons génomiques et protéomiques homologues par alignement multiple local et partiel. Bio-informatique [q-bio.QM]. 2021. hal-03524403

HAL Id: hal-03524403

https://inria.hal.science/hal-03524403v1

Submitted on 13 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Rennes 1

Mémoire de 2ème année de master bio-informatique

Détection de recombinaisons génomiques et protéomiques homologues par alignement multiple local et partiel

Etudiant : Benjamin BLANC

Structures d'accueil : INRAE / INRIA

Tuteurs professionnels : Francois Coste

Marie-Agnès Petit

Tuteur académique : Olivier Dameron

18 Janvier 2021 — 17 Juillet 2021

Rapport du 8 Juin 2021

Master bioinformatique

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) Benjamin Blanc,
étudiant(e) en deuxième année de master bio-informatique
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 07/06/2021

Signature : Benjamin Blanc

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Table des matières

1 Introduction **3**

1.1 Les bactériophages . . . 3

1.2 Présentation du sujet . . . 4

1.3 Postulat de base . . . 6

2 Matériel et méthodes **7**

2.1 Paloma . . . 7

2.1.1 Alignement multiple partiel et local (PLMA) . . . 7

2.1.2 Approche de l'outil . . . 8

2.2 PHROGs . . . 11

2.3 Constitution des jeux de données . . . 11

2.3.1 Jeu de contrôle . . . 13

2.3.2 Jeu avec concaténation verticale . . . 14

2.3.3 Jeu avec concaténation horizontale . . . 15

3 Analyses et résultats **17**

3.1 Paramétrage sur jeu de contrôle . . . 17

3.2 Présence de régions répétées . . . 18

3.3 Présence de recombinaisons chez 3 phages . . . 18

3.4 Étude à l'échelle protéique des 33 paires de protéines de fibres . . . 21

3.5 Étude à l'échelle nucléotidique des répétitions . . . 22

4 Conclusion **25**

5 Bibliographie **26**

6 Résumé / Abstract **28**

1 Introduction

1.1 Les bactériophages

Les bactériophages (également appelés phages) sont des virus qui infectent spécifiquement les bactéries. Ils contiennent de l'acide nucléique (sous forme d'ADN ou d'ARN, selon le type du phage) et des protéines. Ils se distinguent des autres groupes viraux par leur vaste diversité morphologique et génomique. Cependant, on retrouve une structure de base chez la grande majorité des phages observés au microscope à ce jour. Comme illustré avec le phage T4 en figure 1, les phages les plus répandus possèdent une structure appelée tête ou capside, composée de nombreuses copies d'une ou plusieurs protéines. Son rôle primaire est de contenir et protéger le génome viral. Sous la capside se trouve une queue rigide ou contractile, en forme de tube. Elle est terminée par un plateau auquel sont ancrées des fibres de queue, qui permettront le contact avec l'hôte. La plupart des phages ont une structure (tête + queue) ayant une taille se situant entre 24 et 200 nm de diamètre.

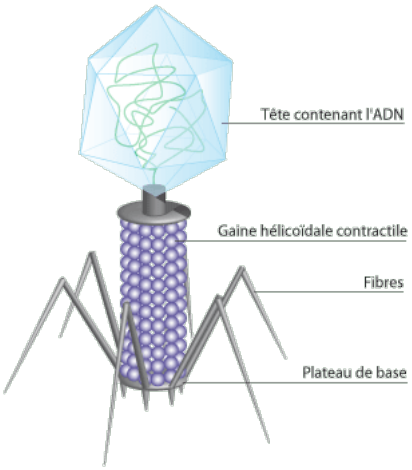


FIGURE 1 – Représentation du phage T4

Un phage va d'abord entrer en contact avec son hôte. Le phage adhère puis interagit alors avec des récepteurs spécifiques, exprimés à la surface cellulaire. Les fibres de queue sont en charge de cette reconnaissance. Si le phage reconnaît les récepteurs, il s'attache ensuite irréversiblement. Une fois le lien établi, le phage utilise sa queue, pour traverser la paroi bactérienne. Certains phages ont développé des enzymes de digestion pour affaiblir localement la paroi cellulaire, afin de mieux la transpercer sans endommager la cellule. Enfin, le phage injecte son matériel génétique dans la cellule, ne laissant qu'une capside vide à l'extérieur de la cellule.

On peut diviser les phages en deux grands groupes :

— Les phages virulents ne s'intègrent pas au génome bactérien et se contentent d'entrer dans la bactérie pour se multiplier en utilisant les ressources de son hôte en sa faveur. Après prolifération, ils ressortent de la bactérie en la tuant, et peuvent entamer un nouveau cycle d'infection. On parle de cycle lytique.

— Les phages tempérés ont un comportement plus évolué que les virulents. Lors de l'infection de leur hôte, si les conditions physiologiques pour réaliser un cycle lytique ne sont pas réunies, ils peuvent entrer en état d'hibernation au sein de leur hôte en intégrant le génome de la cellule infectée (état prophage) en attendant un moment plus propice pour se répliquer. La cellule, lors de sa division, dupliquera également le génome du phage. Cette cohabitation peut perdurer pendant plusieurs générations. On parle de cycle lysogénique. Le cycle lysogénique peut apporter des avantages à la bactérie et contribue ainsi à l'évolution bactérienne.

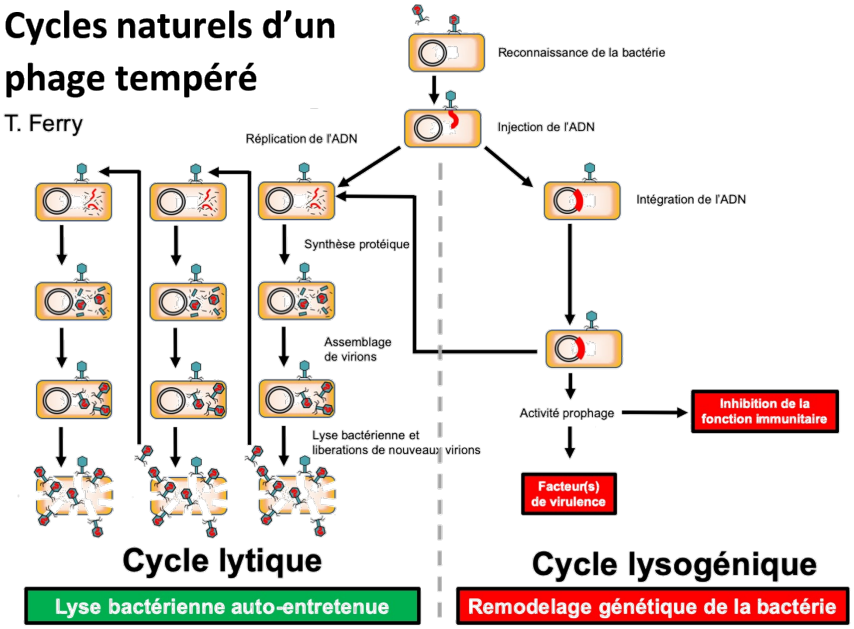


FIGURE 2 – Schéma des cycles naturels d'un phage tempéré, d'après la rubrique Phagotérapie du site de CRIOAc Lyon

Du point de vue évolutif, l'intégration de prophages et leur préservation dans le génome bactérien ont plusieurs conséquences. D'une part, les phages constituent un danger permanent pour la bactérie et menacent de réaliser un cycle lytique (lyser) dans l'hôte. D'autre part, la bactérie peut tirer avantage du prophage qui l'infecte. En effet, le génome des phages peut apporter à l'hôte des gènes non-essentiels qui augmentent sa capacité à s'adapter et à survivre dans son environnement. Ces fonctions confèrent un avantage sélectif favorisant le maintien du prophage. Il existe donc un lien chez les bactéries entre l'acquisition de prophages et l'émergence de nouveaux pathogènes.

1.2 Présentation du sujet

Les phages ont une évolution rapide de par la structure unique de leur génome. Une théorie modulaire de l'évolution des phages a été proposée pour la première fois il y a de nombreuses années [1]. Elle stipule que des ensembles de gènes proches spatialement peuvent être considérés comme des modules

fonctionnels qui sont mélangés par recombinaison, un échange de matériel génétique entre deux brins d'ADN. Ces recombinaisons donnent lieu à de nouvelles combinaisons de modules et donc à des phages potentiellement nouveaux et viables. C'est pour cette raison qu'on qualifie de mosaïques la grande majorité des génomes de phages séquencés à ce jour. Bien que ce phénomène ait été largement étudié, il reste difficile de prédire une recombinaison et d'en comprendre les causes à l'échelle nucléotidique. En 2014, De Paepe et al. démontrent que les mosaïques sont principalement générées par des mécanismes basés sur l'homologie [2]. En d'autres termes, les régions génomiques avec un fort taux de similarité pourraient être plus fréquemment sujettes aux recombinaisons.

La structure en mosaïque du génome des phages rend également difficile l'annotation de leurs protéines. Le nombre en constante augmentation de séquences de protéines disponibles grâce aux technologies de séquençage, a favorisé l'apparition de méthodes in-silico pour annoter fonctionnellement des protéines. L'approche la plus largement utilisée consiste à transférer les annotations des homologues identifiés, susceptibles de partager des structures similaires. La détection de l'homologie est rendue possible par le fait que le maintien de la fonction et de la structure contraint l'évolution des séquences protéiques, ce qui implique que les séquences homologues partagent des caractéristiques communes. Pour détecter ces caractéristiques communes, l'alignement multiple est une des méthodes les plus largement utilisées. Elle consiste à aligner collectivement un ensemble de séquences homologues suivant un algorithme. Cependant, il reste aujourd'hui compliqué d'aligner des séquences nucléotidiques ou protéiques de phages à cause du fort taux de mutations et de recombinaisons dont elles font l'objet. Les algorithmes d'alignement multiple classiques comme ClustalW [3] ou MUSCLE [4] ne parviennent pas à gérer ces recombinaisons, ce qui rend leur utilisation limitée dans le cadre de l'annotation de protéines de phages. Il convient donc de concevoir de nouveaux algorithmes ou d'en modifier d'actuels pour améliorer la détection de recombinaisons homologues. Dans le cadre de ce stage, nous testerons un algorithme, Paloma, qui repose sur le principe d'alignement multiple partiel et local [5][6]. Ce principe permet d'aligner une portion des séquences (alignement local), pour des sous-ensembles de séquences (alignement partiel). De plus, un mode expérimental de Paloma lui permet de représenter l'alignement des régions répétées. La combinaison de ces trois points font de Paloma un outil potentiellement adéquat pour représenter à travers un alignement la structure en mosaïque des génomes des phages. Le but de ce stage est de déterminer d'éventuelles recombinaisons entre les protéines de fibres de queue de plusieurs phages, et le cas échéant, d'essayer d'expliquer les occurrences de ce phénomène chez ces phages

Ce mémoire présentera l'utilisation des alignements multiples partiels et locaux pour comprendre le mécanisme de recombinaison chez les phages.. Dans un premier temps, il conviendra de familiariser le lecteur avec les travaux qui ont mené à l'analyse de fibres de queue chez certains phages. Nous introduirons ensuite le fonctionnement et les paramètres de l'algorithme Paloma, la base de données utilisée et la manière dont ont été conçus les différents jeux de données . Enfin, les alignements réalisés seront analysés.

1.3 Postulat de base

Mon travail prend racine dans les recherches menées par de Sordi et al. [7]. Cette étude porte sur un phage virulent, P10, apparenté au Myovirus FelixO1, et sa capacité à infecter plusieurs bactéries, dont son hôte premier, la souche pathogène Escherichia coli LF82. Il a été observé que P10 était capable d'évoluer pour contaminer de nouvelles bactéries, auparavant insensibles à ce phage. Cet avantage a été conféré par le biais d'un phénomène appelé "host jump". Il s'agit d'un gain de nouvelle fonction suite à une recombinaison ou à une mutation.

Pour comprendre le mécanisme de saut d'hôte, les génomes de deux phages ad_P10 (la version "évoluée" du phage P10) ont été isolés puis séquencés. Dans les deux génomes, un segment de 216 ou 280 nucléotides d'une région du gène de la fibre de queue gp91 a subi une duplication (Figure 3). Le segment s'est recombiné à l'intérieur du gène de fibre de queue long adjacent gp90, provoquant la délétion de 237 ou 273 nucléotides. De manière intéressante, de petites régions flanquantes homologues (environ 95% d'identité) de 17 à 22 paires de bases (figure 3) ont été trouvées, aux bornes de chaque duplication. La fréquence de cet évènement a ensuite été étudiée en séquençant la région correspondante dans 20 isolats ad_P10 choisis au hasard : elle était présente dans 10 des 20 isolats testés.

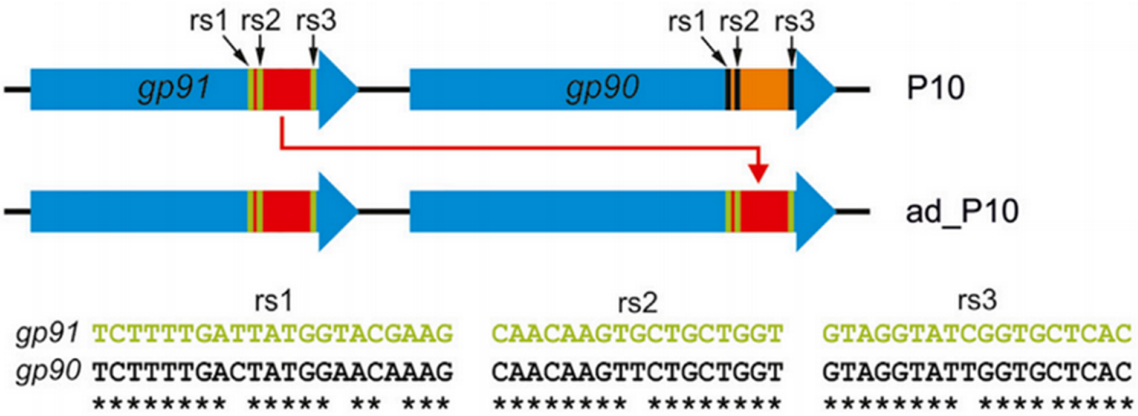


FIGURE 3 – Tiré de l'article *The Gut Microbiota Facilitates Drifts in the Genetic Diversity and Infectivity* of Bacterial Viruses [7]. Représentation schématique de la duplication (flèche rouge) de la région du gène gp91 (boîte rouge) dans gp90. Cette duplication est couplée à la délétion de la région correspondante dans gp90 (boîte orange) chez ad_P10. Les séquences des sites de recombinaison (rs) dans gp90 (en noir) et gp91 (vert) sont indiquées ainsi que leur emplacement (flèches noires).

La recombinaison étant un facteur majeur de l'évolution génomique, des événements similaires ont également été repérés dans d'autres génomes de phages (voir figure 4). Ainsi, trois autres phages de la famille FelixO1 présentaient la même organisation qu'ad_P10 : une duplication d'environ 200 nucléotides dans la séquence de deux gènes de fibre de queue adjacents. Cette découverte est le point de départ de mon travail.

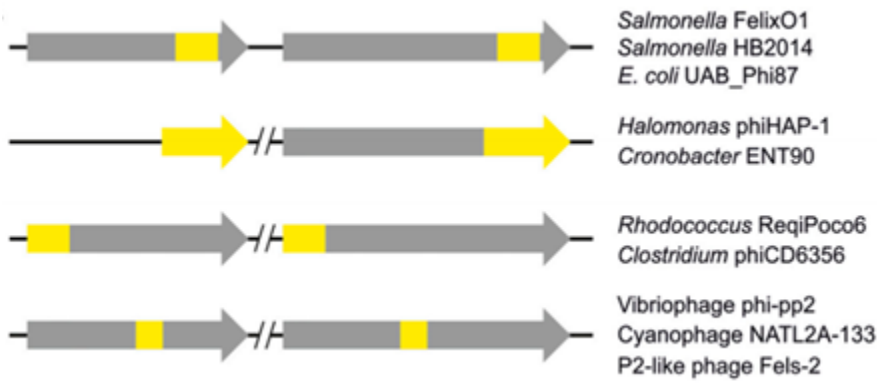


FIGURE 4 – Tiré de l'article *The Gut Microbiota Facilitates Drifts in the Genetic Diversity and Infectivity* of Bacterial Viruses [7]. Exemples de recombinaisons homologues intra-génomiques entre gènes de fibres de queue. Les régions dupliquées sont en jaune.

2 Matériel et méthodes

Une fois ce contexte posé, il convient maintenant d'utiliser les bons outils. Dans cette partie seront présentés le fonctionnement et les paramètres importants de l'algorithme de Paloma, ainsi que la base de donnée dont je me suis servi, Phrogs. La dernière partie expliquera comment on été constitués les différents jeux de données exploités.

2.1 Paloma

Paloma est un outil bio-informatique développé en 2008 dans le cadre d'une thèse [6]. Initialement créé pour l'apprentissage d'automate afin de modéliser des familles de séquences protéiques. Il a ensuite servi dans le cadre de plusieurs travaux scientifiques pour détecter des régions de conservation protéique [8] [9] [10]. Cette partie présente globalement l'approche des créateurs de Paloma [5][6]. L'outil est disponible sur la plateforme bio-informatique Genouest.

2.1.1 Alignement multiple partiel et local (PLMA)

L'un des objectifs principaux de l'alignement multiple est de mettre en valeur les caractéristiques communes d'un ensemble de séquences (protéiques dans le cadre de ce stage). Ce sont ces spécificités partagées qui font de ces séquences une famille. Si certaines familles embrassent une forte similarité, il est possible de les caractériser à l'aide d'un alignement global des séquences. Cependant, les phages échangent sans cesse du matériel génétique, si bien que ce type d'alignement n'est pas approprié. Pour une structure en mosaïque comme chez les phages, les domaines sont plus espacés, et plus court, sans pour autant que l'ordre des domaines soit conservé.

Paloma permet d'obtenir un alignement multiple local et partiel d'un ensemble de séquences. Chaque alignement local partiel (ou PLA pour Partial Local Alignment) correspond à un bloc de conservation qui est :

— Partiel : relâchant la contrainte que toutes les séquences doivent être alignées

— Local : relâchant la contrainte que toutes les positions doivent être alignées pour permettre d'aligner seulement des fragments de séquences.

Un ensemble de PLA compatibles constituent un alignement multiple local et partiel (ou PLMA pour Partial Local Multiple Alignment). Un exemple de PLMA est donné en figure 5.

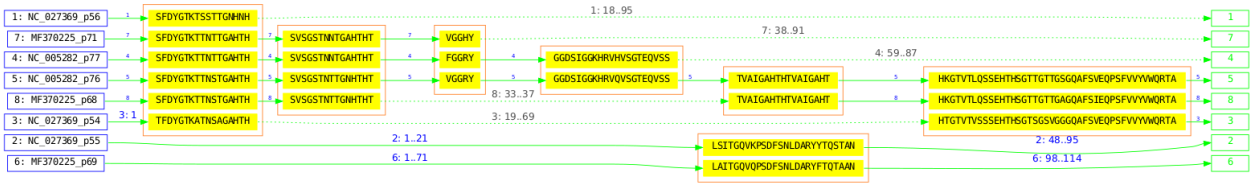


FIGURE 5 – Exemple d'un alignement multiple partiel local (PLMA) executé avec Paloma. Les fragments (définition 1) conservés sont affichés en jaune et les résidus impliqués sont indiqués. Les rectangles rouges désignent les PLAs. L'aspect local est visible horizontalement et l'aspect partiel verticalement.

2.1.2 Approche de l'outil

Nous allons maintenant expliquer comment Paloma aboutit à la production d'un PLMA. La première étape est une recherche de fragments (définition 1) similaires parmi toutes les séquences à disposition. Il est alors nécessaire d'introduire une mesure de similarité correspondant à des associations de fragments. Ces associations se font dans un premier temps par deux, puis par blocs de fragments dans un second temps. L'algorithme utilise le programme Dialign2 [11]. Ce programme, initialement employé pour obtenir des alignements multiples globaux, dispose d'une option permettant d'obtenir un ensemble de paires de fragments significativement similaires. Pour chaque paire, une similarité et un score de similarité sont ensuite calculés et seuls les SFPs (définition 3) sont conservés. Le calcul de la similarité d'une paire se base sur la somme des scores d'une matrice de substitution (BLOSUM par exemple). Le calcul de la significativité de cette similarité est présenté dans la définition 2

Définition 1. *Soit une séquence S = p1...pn. Un fragment F de S est un sous-ensemble de positions contiguës pi...pj de S.*

Définition 2. *Etant donné une paire de fragments (F1,F2) de même longueur l, la significativité de similarité w(F1,F2) est égale à -logP(s,l) tel que :*

— *P(s,l) : la probabilité pour une paire de fragments aléatoire de longueur l d'avoir une similarité supérieure à s*

— *s : la similarité de (F1,F2)*

Définition 3 (SFP (Significantly Similar Fragment Pair)**.** *Une paire de fragments (F1,F2) est significativement similaire pour un seuil t si w((F1,F2)) > t.*

Paloma intègre itérativement les SFP au PLMA suivant une heuristique favorisant la similarité des fragments et leur présence dans la famille. Cependant, des SFPs à intégrer peuvent venir perturber les alignements et diminuer la qualité des résultats en parasitant les SFP de forte similarité trouvés précédemment. Des contraintes ont alors été crées pour imposer une incompatibilité entre SFPs. La première contrainte est la préservation de fragment. C'est une contrainte qui permet, une fois un fragment identifié comme informatif, de préserver les positions de ce fragment en empêchant qu'elles puissent être alignées les unes aux autres. La seconde est la contrainte de consistance. Elle empêche deux positions d'une même séquence de se trouver associées à deux positions en ordre inverse sur une autre séquence. Dans le cadre de mon travail, cette contrainte de consistance pourra ne pas être imposée car l'observation d'une recombinaison entre deux séquences impliquent qu'il y ait des échanges croisés entre ces dernières, donc que leurs positions soient inversées.

Les fragments impliquées dans des SFPs permettent de créer un graphe. Chaque sommet est un fragment et les arêtes correspondent aux liens introduits par la mise en relation de deux fragments au sein d'une SFP. Chaque arête est ensuite pondérée par l'évaluation du score de la SFP correspondante. A partir du graphe produit, il est alors possible de construire des blocs de PLMA. Un bloc de PLMA est un ensemble de fragments et donc de positions alignables, contiguës et sur plusieurs séquences. Un bloc de PLMA peut donc attester d'une zone caractéristique dans un jeu de séquences biologiques. La construction d'un bloc de PLMA à partir du graphe des fragments revient à choisir une composante connexe (définition 4) ou une clique (définition 5) dans le graphe des fragments. Cela permet aux blocs de PLMA ainsi produits de respectivement représenter un consensus faible ou fort selon le contexte biologique. A chaque bloc de PLMA est également alloué un score de similarité. La mesure la plus classiquement utilisée est la somme des scores de similarités des paires de fragments participant à la construction du bloc.

Définition 4. *Une composante connexe CC dans un graphe G est un ensemble de sommets tel que pour tous sommets S1, S2 de CC, il existe un chemin C ayant comme extrémité S1 et S2 et ne passant que par des sommets de CC.*

Définition 5. *Une clique CL dans un graphe G est un ensemble de sommets tel que pour tous les sommets S1,S2 de CL, il existe une arête reliant S1 à S2.*

À la fin des itérations, on obtient un PLMA composé d'un ensemble de PLAs respectant toutes les contraintes imposées. La visualisation du PLMA produit est possible à l'aide d'un fichier produit au format ".dot" de Graphviz et du visualiseur de graphe interactif Xdot. Cette visualisation permet de choisir une coloration des séquences, qui nous sera utile pour les différencier et les annoter.

Paloma intègre un grand nombre de paramètres utilisés pour construire le PLMA à partir des séquences. Les principaux sont présentés dans ce rapport.

— **Seuil sur le score de similarité** : La significativité de la similarité des fragments est le principal paramètre sur lequel jouer et affecte principalement la hauteur des blocs. Pour rappel, la recherche des blocs de régions localement conservées est basée sur les paires de fragments significativement similaires calculées par Dialign [11]. Pour chaque paire de fragments, un score de poids lié au niveau de similarité de la paire de fragments peut être calculé. Seules les paires de fragments dont le poids est supérieur au seuil indiqué seront prises en compte pour construire les blocs PLMA. La valeur par défaut est fixée à 5.

— **Taille maximale des fragments** : En faisant varier la taille des fragments significatifs à considérer, on peut influencer la forme des PLMAs résultants : autoriser des fragments plus grands favorise des blocs plus longs et vice versa . Notons que limiter la taille maximale du fragment ne limite pas la taille maximale des blocs résultants qui peuvent être construits par chaînage des SFPs. Des valeurs plus grandes sont adéquates pour obtenir des caractérisations basées sur la similarité des domaines et des valeurs plus petites plus adaptées pour obtenir des caractérisations basées sur la similarité des acides aminés. La valeur par défaut est fixée à 15.

— **Poids des séquences** : Ce paramètre sert à pondérer chaque séquence. Par défaut, le poids est fixé à 1 pour chaque séquence. Il peut aussi être demandé de calculer le poids de chaque séquence en fonction de leur redondance dans le jeu de données. Des séquences redondantes auront alors un poids plus petit, car mieux représentées dans le jeu. Enfin, il est aussi possible d'indiquer manuellement les poids de chaque séquence.

— **Quorum** : Le quorum est le poids minimal de séquences dans un bloc nécessaire pour le maintenir dans le PLMA. Il peut être exprimé en pourcentage du poids total des séquences, ou directement comme le poids minimal des séquences. Si toutes les séquences ont un poids de 1, ce paramètre correspond alors au nombre minimal de séquences passant dans un bloc. On peut donc jouer sur ce paramètre pour simplifier le PLMA produit, en enlevant des blocs superflus.

— **Option avec répétition (expérimentale) / sans répetition** : L'outil dispose de deux alternatives entre lesquelles l'utilisateur doit choisir. L'option répétition autorise que deux fragments d'une même séquence soient alignés, et ainsi qu'il y ait des répétitions dans une séquence. Si on choisit cette option, la contrainte n'est alors pas imposée. Le mode no repeat (NR), activé par défaut, interdit une telle modélisation de ces répétitions.

— **Consensus** : La valeur sémantique biologique des blocs de PLMA va dépendre du consensus appliqué. S'il est défini comme fort, la construction des blocs se basera sur un principe de clique de fragments similaires tandis qu'un consensus faible d'appuiera lui sur un principe de composante connexe. Le consensus faible est la modalité par défaut.

Ainsi il est primordial de connaître le contexte biologique et l'objectif d'une utilisation de Paloma pour bien paramétrer l'outil. Des exemples de conseils de paramétrage sont par ailleurs données sur la plateforme Genouest, sur laquelle l'outil est disponible en ligne (`http://tools.genouest.org/tools/protomata/help`).

2.2 PHROGs

L'observation faite par Sordi et al. sur 4 phages proches de FelixO1 a été étendue lors de mon stage aux phages possédant des protéines homologues à gp90 et gp91. Pour les repérer, j'ai utilisé la base de données PHROGs(Prokaryotic virus Remote HOmologous Groups) [12]. Il s'agit d'une bibliothèque de familles de protéines virales qui ont été générées à l'aide d'une méthode de clustering en deux étapes, impliquant l'utilisation de la détection d'homologie à distance par des comparaisons profil-profil [13]. Cette base de données contient 38 880 groupes orthologues de protéines contenant plus de 860 000 protéines (similaires mais présentes chez des organismes différents) provenant de génomes complets de virus infectant des bactéries ou des archées. En sélectionnant plusieurs groupes orthologues de choix, il devient alors possible d'avoir des séquences de protéines homologues et comparables si annotés similairement.

2.3 Constitution des jeux de données

Le génome des phages n'est pas sujet aux recombinaisons de façon homogène. Les gènes codant pour la capside sont très conservés car leur fonction est vitale pour l'organisme, et ils ne sont pas soumis à évolution pour s'adapter à l'évolution de leur hôte. En revanche, les gènes de fibres de queue sont bien plus sensibles aux recombinaisons. Etant donné que les fibres de queue sont en charge de la reconnaissance des hôtes à infecter, des changements au sein de leurs gènes peuvent élargir l'éventail d'hôtes infectables. Il a été observé de nombreuses recombinaisons génomiques entre les fibres de queue [14] [15] [16]. Les jeux de données qui vont être construits se doivent d'être pertinents biologiquement parlant à aligner. Les séquences de fibres de queue de phages sont donc des données particulièrement adéquates pour étudier les recombinaisons. A l'inverse, les séquences issues de capsides sont adéquates pour observer une grande stabilité génomique (et créer un jeu de données contrôle par exemple). Une autre caractéristique d'un jeu de données adapté à ce cadre de travail est la proximité des gènes dont sont tirées les séquences. Comme présenté par Sordi et al. (figures 3 et 4), les recombinaisons observées se sont faites entre gènes adjacents. Cette proximité pourrait être une contexte favorable à un évènement de recombinaison.

En partant de ce principe, le jeu de données utilisé dans le cadre de ce stage se compose de deux ensembles de protéines de fibres de queue, respectivement homologues à gp90 et gp91. Chacun contient une protéine du phage FelixO1. Il s'agit des deux protéines sujettes à recombinaison, observées entre deux gènes voisins [7] chez ce phage (figure 4). Ainsi, ces deux protéines serviront de référence pour confirmer la présence d'une région conservée entre les deux protéines, tout en élargissant l'étude des recombinaisons chez les protéines de fibre de queue aux autres phages présents dans les deux groupes. Nous chercherons à voir si la duplication est souvent présente, et si ses bornes sont toujours les mêmes.

Le premier ensemble de protéines contient 72 protéines regroupées dans le phrog 2097. Le second ensemble en contient 32, réunies dans le phrog 4277. 32 phages ont au moins une protéine de fibre de queue présente dans chaque ensemble. Les gènes codant pour ces protéines sont adjacents dans le génome de ces phages,

comme chez FelixO1 [7]. Seules les 32 paires de protéines appartenant à ces phages seront conservées pour les analyses, pour observer les régions conservées entre protéines d'un même phage et comparer les divers évènements de recombinaisons entre phages. De plus, parmi les 32 phages communs aux deux groupes, j'ai observé que 3 codaient pour une deuxième protéine du phrog 2097. Il s'agit des phages Salmonella Phage Si3, Salmonella Phage ST11 et Escherichia Phage EC6 (dont les numéros d'accession sont respectivement KY626162, MF370225 et NC_027369 sur PHROGs). Les gènes codant pour ces protéines sont localisés à chaque fois de part et d'autres du gène lié au phrog 4277 (Cf. Figure 7) . Ainsi, 3 protéines sont disponibles pour ces 3 phages, contre 2 pour tous les autres phages.

En comparant ces trois paires de paralogues avec l'ensemble des protéines du phrog 2097, il a été constaté que chez ces 3 phages, une des de leurs deux protéines avait nettement divergé par rapport aux autres du phrog 2097. Cela pouvait conduire à de mauvaises interprétations des alignements. La protéine divergente chez ces 3 phages, est codé par le second gène appartenant au phrog 2097 par rapport au sens de transcription (figure 7). Ces trois protéines ont été retirées du jeu de données, constitué ainsi de 64 protéines.

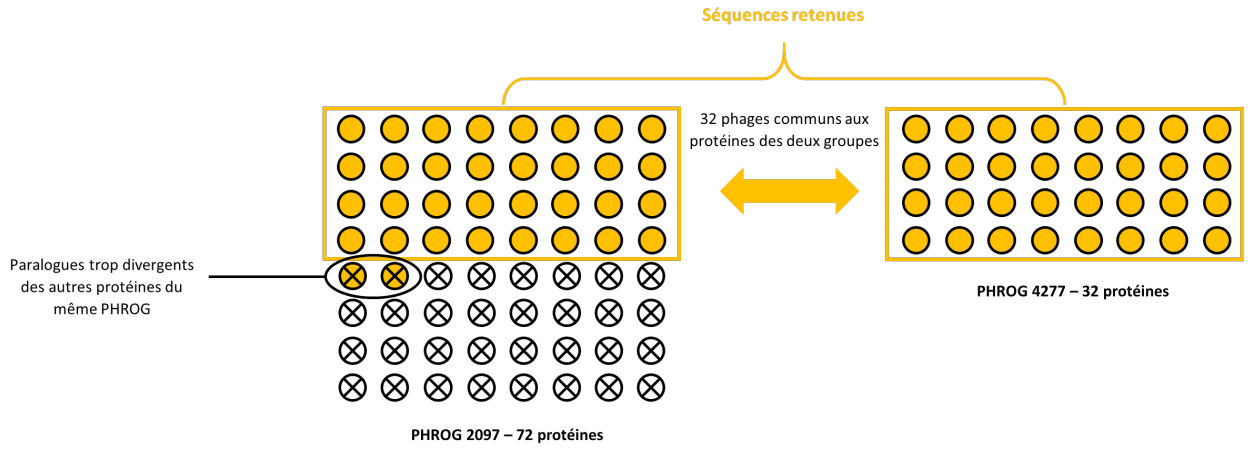
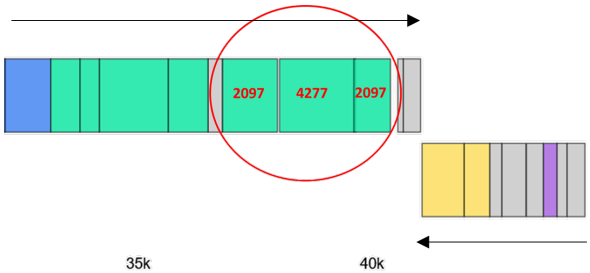


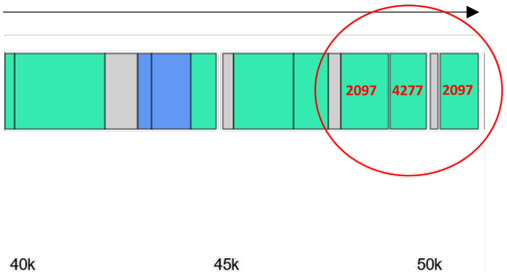
FIGURE 6 – Constitution du jeu de données utilisé pour réaliser les alignements multiples avec Paloma

Finalement, les 64 séquences protéiques restantes sont utilisées comme jeu de données initial. Les séquences ont été récupérées dans la base de données du NCBI, à partir des identifiants données sur la plateforme PHROGs. En plus de ce jeu de données, nous nous servirons également d'un jeu de contrôle.



(a) Escherichia phage EC6

(b) Salmonella phage Si3



(c) Salmonella phage ST11

FIGURE 7 – Extraits des génomes des phages Escherichia EC6, Salmonella Si3 et Salmonella ST11, respectivement de haut en bas. Un rectangle correspond à un gène. Les zones cerclées en rouge montrent le complexe de gènes d'intérêt. Le nombre correspondent au numéro du PHROG lié à ce gène. Les flèches en noir montrent le sens de transcription. Chez ces 3 phages, un gène lié à la famille 2097 est présent deux fois et celui présent en aval du gène lié à la famille 4277 est plus divergent.

2.3.1 Jeu de contrôle

Ce jeu permettra de vérifier que notre outil, Paloma, est fiable en observant un a priori biologique connu. Les paramètres à utiliser pour les analyses suivantes seront ajustés grâce à des alignements avec ce jeu, pour des configurations équivalentes.

La tête des phages est très conservée du point de vue génétique. Nous allons donc analyser des protéines de capside, pour s'assurer la non-présence des recombinaisons entre ces dernières. Deux nouveaux phrogs seront choisis comme contrôle négatif. Les deux familles de protéines retenues sont les phrogs 979, annotée "terminal large subunit" et 113, annotée "portal protein", contenant respectivement 151 et 789 protéines. On retrouve

chez ces deux groupes une protéine du phage d'intérêt FelixO1. Les protéines liées à ces familles sont très fréquemment situées côte à côte dans le génome des phages concernées (112 occurences de co-localisations).

Les visualisations de PLMAs contruites à partir de l'ensemble du jeu de données étant très chargées, il est beaucoup moins évident de repérer les changements résultant du paramétrage à l'oeil nu. On va alors chercher à réduire le jeu de données utilisé, de manière à ce que les sorties soit lisibles et faciles à interpréter.

Pour cela, on utilise MMSeqs2 (Many-against-Many sequence searching) [17] [18], une suite logicielle permettant de rechercher et de regrouper de gros ensembles de séquences de protéines et de nucléotides. En indiquant un seuil d'identité, MMSeqs2 est capable de regrouper un jeu de séquences en clusters, selon leur proximité séquentielle, ayant un pourcentage d'identité supérieur au seuil imposé. Chaque cluster contient une séquence représentant toutes les autres séquences du cluster et ce sont ces représentants que nous alignons. De cette manière, le nombre de séquences à aligner passent de 940 (151 + 789), à 140 (29 + 111). De plus, nous choisissons de ne conserver que les séquences de protéines qui ont phage commun aux deux phrogs). On passe ainsi à 30 séquences, issues de 15 phages. Afin d'avoir un phage référence, on rajoute les deux séquences issues du phage FelixO1 (une par phrog), pour un total de 32 séquences.

2.3.2 Jeu avec concaténation verticale

Nous avons vu précédemment qu'un des objectifs de ce travail était de détecter des recombinaisons homologues internes aux 32 phages étudiés, et d'observer la spécificité de ces recombinaisons, à savoir si une recombinaison s'effectue seulement chez un phage spécifique, ou si elles est commune à plusieurs ou à l'ensemble des phages d'études. Pour rappel, notre jeu de données est constitué de 64 séquences de protéines issues de deux familles de phrogs . Ces séquences représentent les protéines issues de 2 gènes codant pour les fibres de queue de 32 phages. Chaque phage a donc 2 séquences qui lui sont propres.

Deux types d'alignements ont été pensés pour obtenir des PLMAs biologiquement informatifs , mettre en avant de potentielles recombinaisons entre les deux familles d'intérêt et confirmer certaines recombinaisons homologues (phages FelixO1, UAB_Phi87 et vB_SPuM_SP116 constatées par Sordi et al.).

Le premier est un alignement avec concaténation verticale des séquences. Les séquences des deux familles sont regroupées dans un fichier fasta, puis Paloma est executé avec ce fichier fasta en entrée, en mode No Repeat et avec un paramétrage de couleur selon le phrog d'appartenance des séquences, soit 2 couleurs différentes. Le PLMA obtenu contiendra donc 64 séquences. L'objectif est de s'intéresser aux blocs par lesquels passent des séquences de couleurs différentes. Un exemple d'alignement est donné en figure 8. Comme au sein d'un phage, les deux gènes adjacents d'intérêt appartiennent aux deux familles étudiées, cela peut signifier qu'il y a eu échange de matériel génétique entre protéines. Il faut donc s'assurer que parmi les séquences qui passent par ce bloc, on en retrouve deux appartenant au même phage. Cette configuration permet de détecter facilement des recombinaisons propres à un ensemble de phages (1) ou à un phage unique (2). Enfin, même si cela ne concerne pas notre objectif, elle met aussi en évidence des régions conservées au sein d'une famille de

protéines (3). Il est donc plus facile d'identifier la "portée" d'une recombinaison au sein d'un ensemble de phages avec ce type de visualisation.

2.3.3 Jeu avec concaténation horizontale

Le second type d'alignement est réalisé avec une concaténation horizontale des séquences : pour chaque phage, on concatène ses deux séquences en une seule, toujours dans le même ordre (séquence du phrog 2097 puis séquence du phrog 4277 [1]), en les séparant par un cryptophone. Ce cryptophone est un motif facilement alignable par Paloma, qui le représentera par un bloc de conservation dans le PLMA. Ce bloc symbolisera alors la séparation entre les deux séquences de protéine d'un même phage. Pour notre étude, nous avons choisi d'utiliser le cryptophone "SSSSSSSSSS". Dans une matrice BLOSUM62, celle utilisée par défaut par Paloma lors du calcul de similarité, le S est la lettre avec la norme de match/mismatch la plus faible [2]. Un exemple d'alignement est donné en figure 8. On obtient donc 32 séquences, chacune correspondant à un phage. Paloma est lancé avec l'option repeat, et l'on pourra définir une couleur si nécessaire pour faire ressortir la séquence d'un phage en particulier lors de la visualisation du PLMA. L'avantage de cet alignement est de pouvoir suivre plus facilement les récombinaisons homologues intra-génomiques d'un phage. L'attention sera ici porté sur les transitions entre deux blocs enjambant le bloc du cryptophone dans l'alignement d'une séquence . Ces transitions peuvent symboliser des recombinaisons dans un génome, d'où l'éxécution de Paloma avec répétitions autorisées.

Après plusieurs essais avec ce jeu, il s'est avéré que la visualisation est très compliquée à interpréter. L'alignement avec concaténation horizontale, adéquat pour suivre l'alignement d'une séquence d'un phage particulier, sera alors utilisé sur un jeu plus restreint, contenant les 3 séquences de phages chez lesquels De Sordi et al. ont repéré une recombinaison (figure 4). Cette analyse permettra de confirmer la présence de cette recombinaison chez ces phages, et la visualisation sera bien plus aisée, n'ayant que 3 séquences à observer. Il apparaît également avec ce type d'alignement que certains arcs entre blocs manquent, ce qui ne facilite pas la lecture. Cela est dû au fait que l'option repeat reste encore expérimentale.

Ces différents jeux vont donc permettre de :

- Paramétrer l'outil à l'aide du jeu contrôle,
- Vérifier la présence de recombinaisons chez FelixO1 et deux autres phages grâce aux séquences concaténées horizontalement,
- Étudier de manière plus générale des recombinaisons à l'échelle d'un phage ou d'un ensemble de phages.

1. Cet ordre est défini pour faciliter la lecture du PLMA. Un autre fichier fasta, respectant l'ordre biologique d'apparition des gènes dans le génome des 32 phages, a également été contruit mais pas utilisé pour l'instant.

2. A l'inverse, un jeu de séquence avec un cryptophone "WWWWWWWWWW" a aussi été généré, le W étant la lettre avec la norme de match/mismatch la plus forte. Ce jeu n'a cependant pas été utilisé.

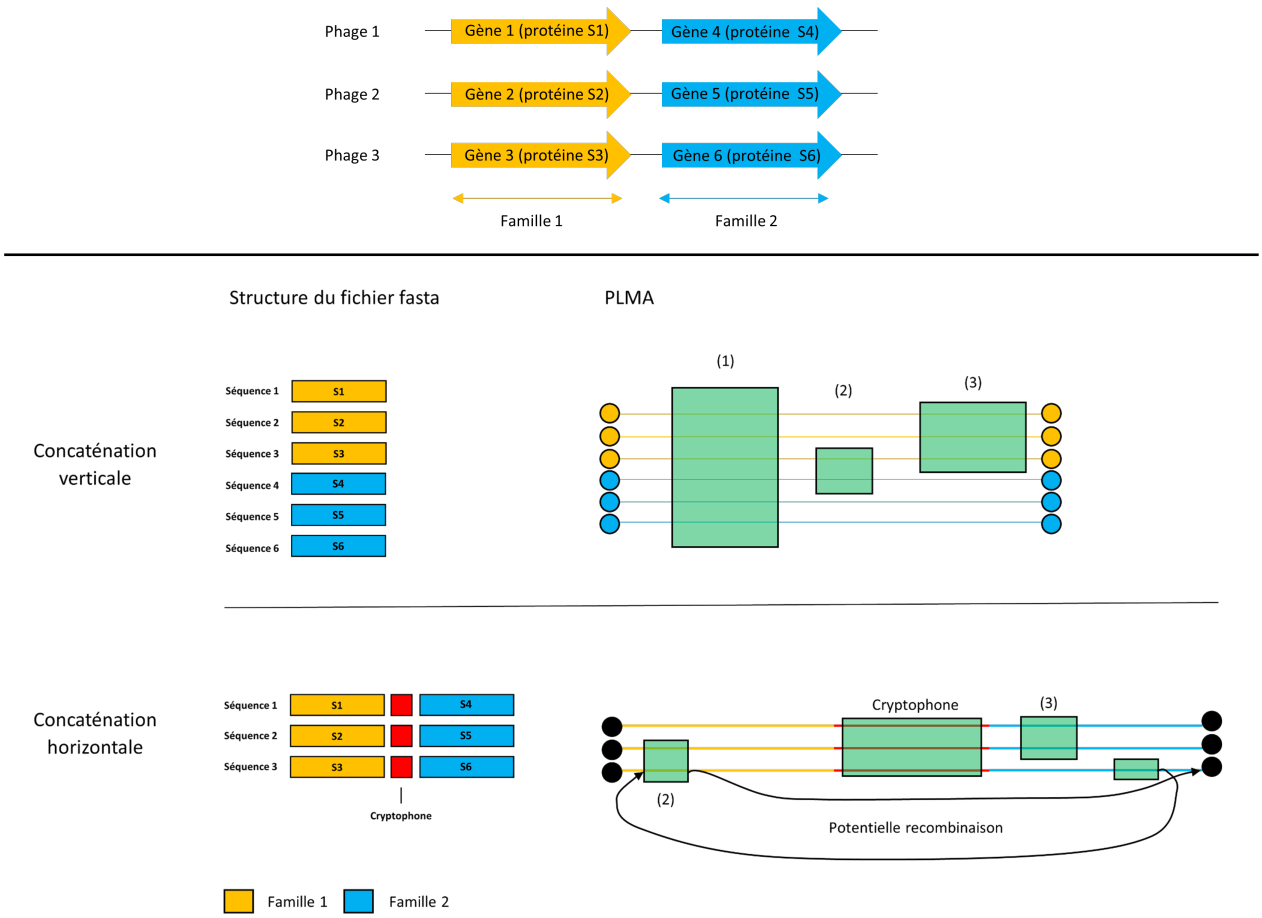


FIGURE 8 – **En haut :** définissons la famille 1 (en jaune sur la figure) composée des gènes homologues 1,2 et 3, codant pour les protéines S1, S2 et S3, et la famille 2 (en bleu sur la figure), composé des gènes homologues 4,5 et 6 codant pour les protéines S4, S5 et S6.

En bas : Dans le cas d'un alignement avec concaténation verticale, en haut, le fichier fasta aura 6 séquences. Si à la visualisation du PLMA généré, on observe un bloc laissant passer toutes les séquences (1), les séquences dans le bloc symbolisent une région conservée entre toutes ces séquences. Si un bloc laisse passer uniquement deux séquences d'un même phage (2), cela signifie que cette région conservée est uniquement présent chez ce phage. Si un bloc ne laisse passer que des séquences d'une même couleur (même famille) (3), on peut l'interpréter comme une conservation propre à cette famille de protéine. Dans le cas d'une concaténation horizontale, en bas, il n'y a que 3 séquences car elles ont été concaténées et séparées par un cryptophone (en rouge). Les répétitions étant autorisées avec cette stratégie là, des retours en arrière sont possible et si recombinaison il y a, cela se traduira sur la visualisation du PLMA par un retour en arrière vers la portion de protéine qui a été recombinée dans la première séquence, puis un grand bond en avant dans la visualisation du PLMA pour se situer juste après cette portion recombinée dans la seconde séquence. Dans cette configuration, on n'observe que des recombinaisons propres à un seul phage à la fois (2). On peut également toujours voir des régions conservées chez une famille de protéines (3).

3 Analyses et résultats

L'ensembles des analyses ont été réalisées avec la version initiale de Paloma. Quelques semaines avant le rendu de ce rapport, une nouvelle version Paloma-2, plus rapide à l'exécution mais avec moins d'options pour le moment, m'a été mise à disposition. Par manque de temps, je n'ai pas pu refaire ces analyses avec cette nouvelle version.

3.1 Paramétrage sur jeu de contrôle

Dans un premier temps, la fiabilité de l'outil sera vérifiée seulement sur un alignement avec concaténation verticale des séquences du jeu de contrôle, où la recombinaison n'est pas attendue. En effet, la visualisation de fausses recombinaisons sera plus facile avec cette configuration. Des alignements sont réalisés en faisant varier le seuil (1, 3 et 5) et la taille maximum des fragments (10 et 15), avec les autres paramètres par défaut. Pour une meilleure visualisation, des couleurs différentes sont attribués pour les séquences de chaque phrog dans le PLMA généré. Le nombre de blocs laissant passer des séquences de couleurs différentes, a alors été compté. Ces blocs symbolisent, dans un contexte biologique favorable, de potentielles recombinaisons d'un gène à l'autre. Cependant ici, ces blocs traduiraient plutôt des "faux-positifs" étant donné que la localisation des séquences (le collier des phages) ne se prêtent pas à l'apparition de recombinaisons. Les résultats de ces analyses sont présentées en table 1.

Seuil	Taille maximum	Nombre de blocs avec les deux phrogs
1	10	5
	15	3
3	10	1
	15	0
5	10	0
	15	0

Table 1 – Nombre de blocs incluant des séquences des deux familles de protéines différentes dans le PLMA obtenu avec Paloma en faisant varier la taille maximum des fragments et le seuil.

Un seuil insuffisamment stringent (= 1) peut entraîner quelques erreurs d'alignement, qui peuvent être problématique pour l'interprétation des PLMAs. En revanche, en augmentant le seuil et la taille, le nombre de mauvais alignement se réduit jusqu'à devenir nul.

Suite à ces analyses, il a été décidé, pour les analyses présentées sections 3.4 de prendre un seuil de 5 pour être suffisamment rigoureux dans nos alignements, et une taille maximum de fragments de 10 pour étudier la présence de régions flanquantes dans le génôme des protéines des 32 phages d'intérêt, car leur taille ne dépasse pas 10 acides aminés.

3.2 Présence de régions répétées

Pour rappel, des régions ont été repérées chez le phage LP10, se situant dans les gènes gp90 et gp91 de ce dernier 3. Ces régions, isolées, correspondent à une vingtaine de nucléotides. Le tableau 2 présente les séquence nucléotidiques et protéiques des répétitiions présentes dans gp90 et gp91. Pour la traduction, l'outil EMBOSS Transeq [3] (paramètres standards) a été utilisé. Par la suite, les séquences protéiques correspondant aux séquences nucléotidiques rs1, rs2 et rs3 seront respectivement appelées rp1, rp2 et rp3.

Région	Gène	Séquence nuclétidique	Séquence protéique
rs1	gp90	TCTTTTGACTATGGAACAAAG	SFDYGTK
	gp91	TCTTTTGATTATGGTACGAAG	SFDYGTK
rs2	gp90	CAACAAGTGCTGCTGGT	TSSAG
	gp91	CAACAAGTTCTGCTGGT	TSAAG
rs3	gp90	GTAGGTATCGGTGCTCAC	VGIGAH
	gp91	GTAGGTATTGGTGCTCAC	VGIGAH

Table 2 – Séquences nucléotidiques et protéiques correspondant à rs1, rs2 et rs3 chez le phage P10.

On remarque la présence d'un à trois mismatches dans les répétions nucléotidiques qui se traduisent également pour rp2 par un changement d'acide aminé.

A l'aide de l'outil EMBOSS Water [4], les séquences rp1, rp2 et rp3 sont bien retrouvées dans le même ordre dans les deux protéines des gènes gp90 et gp91 du phage P10 et ad_P10 (Figure 9). On retrouve également la duplication de la région entre rp1 et rp3, avec une mutation ponctuelle, chez ad_P10. Il y a bien eu recombinaison entre les deux gènes chez le phage évolué. Concernant FelixO1, rp1 et rp3 sont présents dans les deux séquences protéiques. Toutefois, il n'y a pas de traces de rp2. De plus, on observe une région bien conservée, d'une soixantaine d'acides aminés, commençant par rp1 mais s'arrêtant avant rp3. Il y a donc également eu un évènement de recombinaison et duplication d'une région, mais celle-ci n'est pas bornée à droite par la même répétition rp3 que dans ad_P10.

3.3 Présence de recombinaisons chez 3 phages

Sordi et al. avaient remarqué une la duplication d'une région entre deux gènes adjacents de fibres de queue chez 2 autres phages (UAB_Phi87 et vB_SPuM_SP116), en plus de FelixO1 (figure 4. Il serait possible d'utiliser à nouveau un alignement par paire, pour chacun de ces phages. Cependant, cela serait long et il serait épineux de comparer la conservation de ces régions entre phages à l'oeil. Plutôt que de réaliser

3. Cet outil traduit les séquences d'acides nucléiques en leurs séquences peptidiques correspondantes. Il peut traduire les trois trames avant et les trois trames arrière et a été développé par l'institut européen de bio-informatique (EMBL-EBI).

4. Cet outil d'alignement utilise l'algorithme de Smith-Waterman pour calculer l'alignement local de deux séquences.

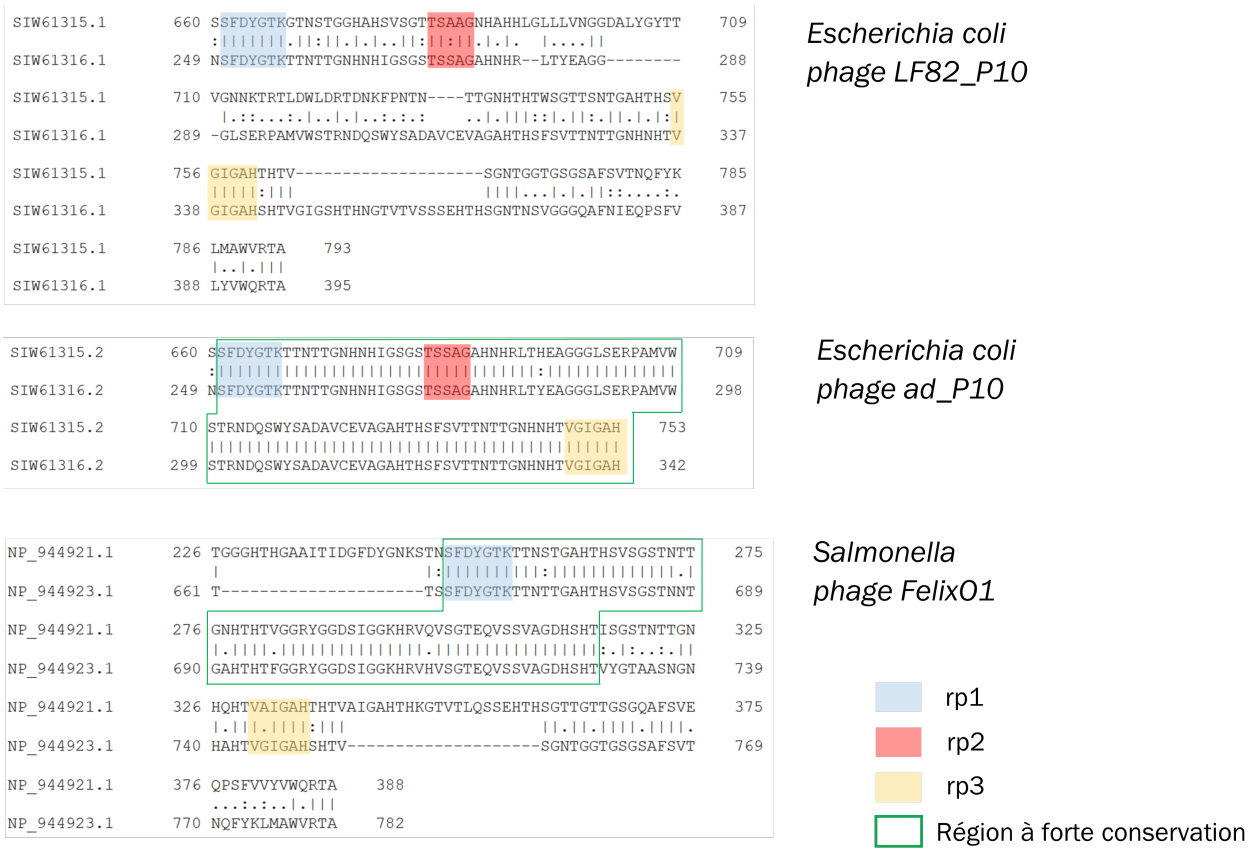


FIGURE 9 – Détection des séquences protéiques répétées rp1, rp2 et rp3 par alignement avec l'outil EMBOSS Water (paramètres standards) chez les phages LF82_P10, ad_P10 (LF82_P10 évolué) et FelixO1.

plusieurs alignements par paire, nous allons utiliser un alignement avec concaténation horizontale (présenté section 2.3.3). Les objectifs de cet alignement sont de montrer que Paloma est capable de détecter rp1 et rp3 chez FelixO1, comme observé en figure 9, de voir si rp1 et rp3 sont présents chez les deux autres phages alignés, et si l'on détecte une région bien conservée chez ces 3 phages, à partir de rp1 chez FelixO1. Cet alignemement est ici approprié car il permet sur un petit nombre de phages de visualiser des recombinaisons rapidement. Cet alignement portera sur les séquences des 3 phages FelixO1, UAB_Phi87 et vB_SPuM_SP116, avec un seuil de 5 et une taille de fragment de 10. Ces valeurs ont été choisies arbitrairement.

De plus, le quorum a été fixé à 4. Cela signifie qu'un bloc ne sera construit que s'il inclut au moins 4 séquences. Si une recombinaison a lieu entre deux génomes, la séquence recombinée sera dupliquée dans au moins un des deux génomes. Etant donné que les 3 séquences alignées ont des régions conservées homologues, on s'attend à voir des blocs de 3 séquences significativement similaires. En revanche, s'il s'agit de bloc de 4 séquences ou plus, cela indique la duplication d'une séquence conservée ailleurs dans le génome, qui est venu s'aligner avec les autres fragments de ce bloc. Les blocs de 4 séquences ou plus correspondent donc à de potentielles régions qui ont été recombinées. Il faut aussi noter qu'un quorum de 4 simplifie grandement la visualisation

du PLMA généré car il réduit drastiquement le nombre de blocs à afficher. Enfin, cette valeur de quorum ne fera pas apparaître le cryptophone dans le PLMA, car il s'agit d'une séquence présente une seule fois dans chaque séquence, donc elle apparaîtrait dans un bloc de 3 séquences au mieux.

Une partie de la visualisation du PLMA produit est présenté en figure 10. On retrouve les deux blocs incluant rp1 et rp3, contenant chacun 6 séquences. Cela signifie la présence de rp1 et rp3 dans les deux protéines de ces phages, donc d'une possibilité de recombinaison de ces séquences. On remarque entre rp1 et rp3 un ensemble de bloc, correspondant à des séquences successives dans les génomes protéiques. Mises bout à bout, ces séquences ont une taille qui varie entre 58 et 70 acides aminés, en commencant juste après rp1 et en terminant 12 acides aminés avant rp3. Il s'agit potentiellement de la longue région recombinée observée par Sordi et al. chez les 3 phages. Néanmoins, travaillant ici à l'échelle protéique et non nucléotidique, il est difficile de tirer des conclusions. Cela confirme également que notre outil est adapté à l'observation de potentielles recombinaisons.

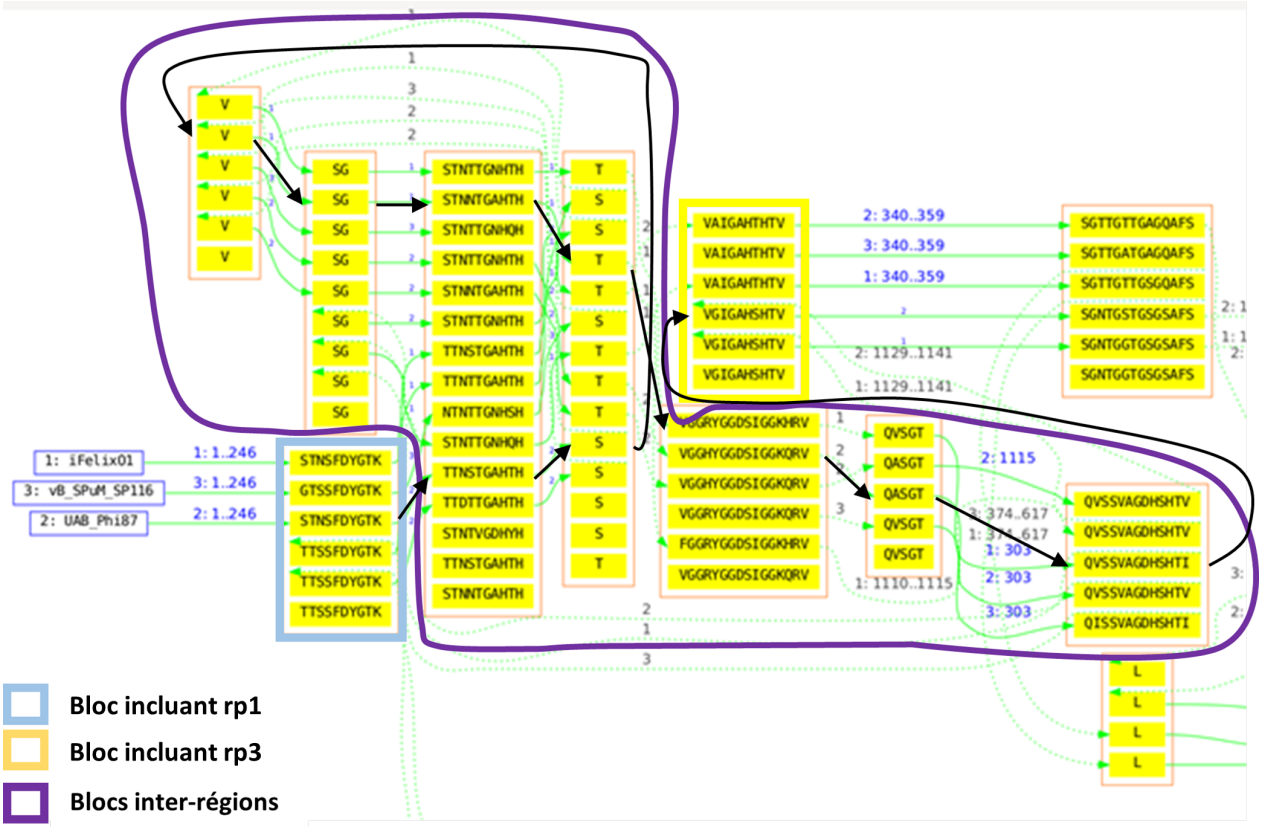


FIGURE 10 – Visualisation du PLMA généré par Paloma à partir d'un jeu de 3 séquences concaténées horizontalement. Les blocs entourées en bleu et en jaune correspondent respectivement aux séquences contenant rs1 et rs3. L'ensemble de blocs entourées en violet, une fois mis bout à bout dans l'ordre, donnent les séquences inter-régions, susceptibles d'avoir été recombinées. A titre indicatif, le chemin d'une séquence entre ces deux régions est donné.

3.4 Étude à l'échelle protéique des 33 paires de protéines de fibres

Après avoir paramétré correctement Paloma (section 3.1) et vérifié la présence d'une région très conservée entre les deux gènes des 3 phages (section 3.3) , nous avons montré que notre outil est en mesure, avec un bon paramétrage, de restituer des alignements interprétables biologiquement et pertinents pour détecter des recombinaisons homologues. Il est maintenant temps de l'utiliser sur un jeu de données plus large, contenant tous les phages communs aux phrogs 2097 et 4277 (soit 32). (cf. section 2.3.2). On cherchera ici à détecter rp1, rp2 et/ou rp3 chez les 32 phages et observer quelles sont les régions conservées entre les deux familles.

Un PLMA est construit en utilisant le jeu de données (présenté en section 2.3.2 avec les séquences protéique des deux gènes chez les 32 phages plus les deux séquences protéiques de P10 afin d'avoir une référence. Un seuil sur le score de 5 et un maximum pour les tailles de fragments de 10 ont été appliqués. Ces paramètres ont été choisis suite à l'analyse présentée section 3.1 .

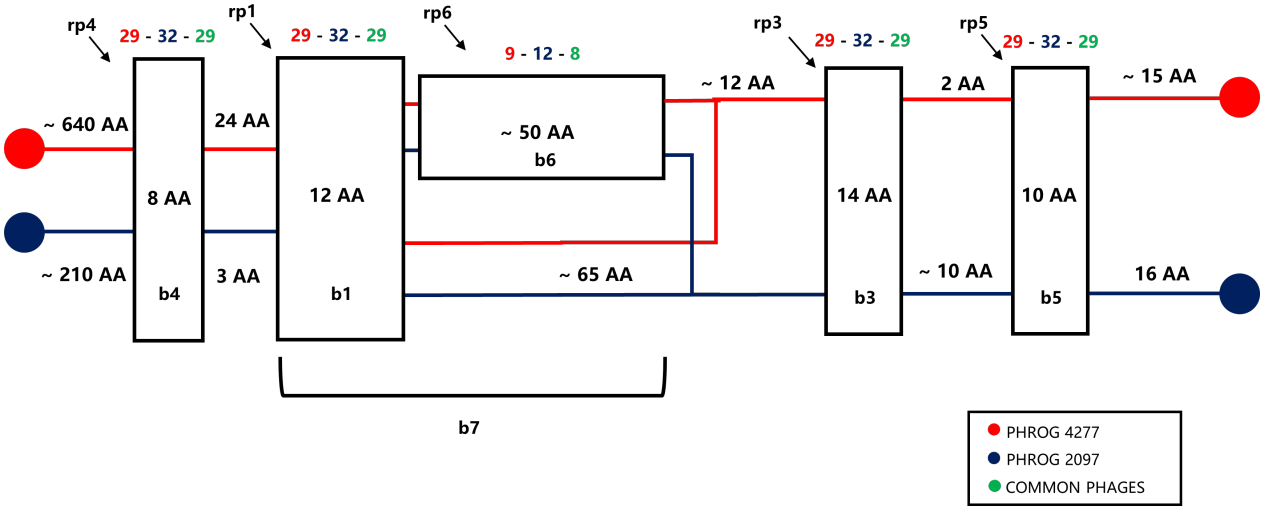


FIGURE 11 – Schéma simplifié de l'alignement de 66 séquences protéiques. Les traits en rouge symbolisent des séquences appartenant au phrog 4277, ceux en bleu des séquences appartenant au phrog 2097. Les blocs présentés peuvent être composé d'un seul ou d'un ensemble de blocs consécutifs dans le PLMA obtenu. Pour chaque bloc, les nombres en rouge, bleu et vert correspondent respectivement aux nombres de séquences du phrog 4277 passant par ce bloc, aux nombres de séquences du phrog 2097 passant par ce bloc et au nombre de phage dont les deux séquences protéiques passent par ce bloc. A titre informatif, les distances en acides aminés sont données entre les blocs.

Comme indiqué figure 11, les blocs qui englobent les séquences protéiques rp1, rp2, rp3, rp4, rp5 et rp6 seront dorénavant respectivement appelés b1, b2, b3, b4, b5 et b6. Le bloc b7 correspondra lui à la concaténation de b1 et b6.

Sur l'alignement produit, b1 et b3 englobent les séquences protéiques rp1 et rp3, avec quelques acides-aminés en plus. Ces deux blocs laissent passer les 32 séquences du phrog 2097 et 29 chez le phrog 4277. Les 3

séquences du phrog 4277 qui n'y figurent pas appartiennent aux 3 phages chez lesquels on observe deux gènes liés au phrog 2097 (Figure 7). Une longue région conservée d'une cinquantaine d'acides aminés est également détectée juste après b1. Elle correspond à la grande zone de conservation chez FelixO1 en figure 9. Cette grande région conservée entre les deux protéines de fibre de queue, appelée rp6, est présente chez 8 phages : Salmonella phage Felix 01, Enterobacteriaphage UAB_Phi87, Salmonella phage vB_SPuM_SP116, Salmonella phage BPS15Q2, Salmonella phage BPS17L1, Salmonella virus VSe11, Salmonella virus VSe102 et Staphylococcus phage SA1. On remarque toutefois que cette conservation, à partir de rp1 ne s'étend pas entièrement jusqu'à rp3, entre 10 et 15 acides aminés n'étant pas conservés juste avant rp3. rp2 n'est pas retrouvée dans un bloc. Cependant, l'apport d'une couleur particulière aux séquences du phage P10 dans l'alignement a permis de trouver un unique autre phage chez lequel on retrouve une conservation de rp2 : Escherichia phage vB_EcoM_Alf5.

Enfin, deux nouveaux blocs de conservations sont ressortis. Le premier, b4, se situe un peu avant b1 et fait 8 acides aminés. Le second, b5, d'une taille de 10 acides aminés, se situe juste après b3. La petite taille de ces blocs, leur conservation interne et leur position (un peu avant rp1 et un peu avant rp3) confortent l'idée qu'ils pourraient former une autre région répétée.

Par la suite, nous concentrerons notre analyse sur les 8 phages dont les deux séquences protéiques passent par b6. Le bloc b7 englobe probablement la recombinaison observée par De Sordi et al. (Figure 4) chez les 3 phages Salmonella FelixO1, Salmonella HB2014 et E.coli UAB_Phi87, ainsi que chez 5 nouveaux autres phages. Chez ces 8 phages, les deux séquences de protéines passent par les blocs b1 à b5, qui sont associés à des régions répétées. La conservation de b7 entre deux protéines du même phage oscille entre 76 et 94 % d'acides aminés identitiques pour ces 8 phages (table 3). Cette forte conservation conforte dans l'idée d'une recombinaison chez ces phages. Toutefois, il ne faut pas perdre de vue que ce sont les conservations à l'échelle nucléotidique qui déterminent si la recombinaison a lieu. Il est donc nécessaire de s'intéresser aux séquences codantes des protéines des 8 phages pour caractériser ces répétitions au niveau nucléotidique.

3.5 Étude à l'échelle nucléotidique des répétitions

Cette étude amène à proposer de nouvelles répétitions qui pourraient aussi servir de point de départ pour la recombinaison. Les séquences codantes sont récupérées sur GenBank à l'aide des identifiants des séquences protéiques sur PHROGs. Un alignement à concaténation verticale est effectué à partir des 32 séquences codantes avec un seuil de 5, une taille de fragments maximum de 20 et pas de répétitions. L'idée reste d'être stringent sur les alignements identifiés en conservant le seuil choisi avec le jeu de paramétrage, mais de favoriser des blocs de plus longues tailles pour se calquer sur les tailles de rs1, rs2 et rs3. Ces 3 régions observées par De Sordi et al. chez le phage P10 font une vingtaine de nucléotide tout au plus et leur alignement montre une conservation très forte (supérieure à 90%). Les séquences protéiques découvertes rp4 et rp5 correspondraient alors respectivement à 24 et 30 nucléotides. Les régions nucléotidiques flanquantes codant pour rp4 et rp5

<table>
<tr><th>Phage</th><th>b1</th><th>b2</th><th>b3</th><th>b4</th><th>b5</th><th>% d'identité dans b7</th></tr>
<tr><td>Salmonella phage Felix 01 *</td><td rowspan="9">Présente</td><td rowspan="8">Absente</td><td rowspan="9">Présente</td><td rowspan="9">Présente</td><td rowspan="9">Présente</td><td>92</td></tr>
<tr><td>Enterobacteriaphage UAB_Phi87 *</td><td>94</td></tr>
<tr><td>Salmonella phage vB_SPuM_SP116 *</td><td>82</td></tr>
<tr><td>Salmonella phage BPS15Q2</td><td>79</td></tr>
<tr><td>Salmonella phage BPS17L1</td><td>89</td></tr>
<tr><td>Salmonella virus VSe11</td><td>76</td></tr>
<tr><td>Salmonella virus VSe102</td><td>76</td></tr>
<tr><td>Staphylococcus phage SA1</td><td>80</td></tr>
<tr><td>Escherichia phage vB_EcoM_Alf5
Escherichia phage LF82_P10 *</td><td>Présent</td><td>\</td></tr>
</table>

Table 3 – Tableau récapitulatif de la présence de séquences protéiques (contenues dans des blocs) chez 10 phages. Les 8 phages en orange sont ceux dont les protéines partagent une grande région conservée de 66 AA à partir de rs1 d'après l'alignment obtenu avec Paloma. Les phage en bleu sont les seuls phages parmi les 33 alignés (32 + le phage P10) pour lesquels les deux protéines contiennent rp2. Enfin, les phages avec * sont les 4 phages présentés dans le papier Sordi et al. (Figures 3 et 4).

seront respectivement appelées rs4 et rs5. Pour essayer d'affiner les limites exactes de rs4 et rs5 , on part du postulat que ces séquences auront une forte similarité (supérieure à 90%, comme pour rs1, rs2 et rs3 chez le phage LP10).

Des alignements locaux ont été réalisés avec les séquences codantes de rp4 et rp5 chez Salmonella FelixO1, avec l'outil EMBOSS Water (paramètres standards). L'objectif est de définir les limites d'une région très conservée et de garder ses positions dans l'alignement pour pouvoir comparer les conservations des séquences codantes entre les 4 séquences protéiques (rp1,rp3, rp4 et rp5) chez les 8 phages. Ces positions sont déduites des alignements obtenus avec Paloma. Les régions nucléotidiques retenues pour définir rp4 et rp5 sont présentées en figure 12. Toutefois, rs4 semble petite et sa conservation pas suffisamment forte. Il est peu probable qu'elle puisse être caractérisée comme une région répétée.

Ensuite, pour chaque phage, les deux séquences de rs1, rs2, rs3 et rs4 sont alignées entre elles. La conservation de chaque région est donnée en table 4. Bien que les conservations ne soient pas aussi fortes que ce qui a été constaté chez le phage LP10 (plus de 90 % d'identité), elles restent élevées et au-dessus de la conservation moyenne du génome de ces phages. Nous concluons que les gènes étudiés disposent de 4 régions répétées qui pourraient servir de point d'ancrage pour la recombinaison homologue.

FIGURE 12 – Alignements des séquences codantes de rp4 et rp5 chez le phage Salmonella FelixO1. La région entourée en vert correspond à celle la mieux conservée d'après l'alignement.

Région nucélotidique	rs1	rs3	rs4	rs5
Taille (en nuclétoides)	21	18	14	24
Salmonella phage Felix O1	76 %	89 %	92 %	86 %
Enterobacteriaphage UAB_Phi87	76 %	67 %	79 %	79 %
Salmonella phage vB_SPuM_SP116	76 %	61 %	79 %	79 %
Salmonella phage BPS15Q2	90 %	61 %	83 %	79 %
Salmonella phage BPS17L1	86 %	67 %	88 %	86 %
Salmonella virus VSe11	62 %	83 %	88 %	71 %
Salmonella virus VSe102	62 %	83 %	88 %	79 %
Staphylococcus phage SA1	81 %	61 %	83 %	79 %

TABLE 4 – Tableau récapitulatif des alignements réalisés sur les 4 régions des séquences codantes de 8 phages. Pour chaque région est indiqué sa taille et son pourcentage de conservation dans les séquences codantes du phage en question.

4 Conclusion

En conclusion, nous avons pu voir que Paloma est un outil utilisable et utile à l'étude des recombinaisons à l'échelle protéique et nucléotidique. Nous avons retrouvé, à l'échelle protéique, la traduction de rs1 et rs3 chez les deux protéines homologues de 29 phages, contre 3 auparavant (cf. section 1.3). Sur ces 29 phages, 8 partagent une longue région conservée, d'une soixantaine d'acides aminés, commençant à partir de rp1 mais s'arrêtant avant rp3, contrairement à la région recombinée chez P10. De plus, nous avons détecté deux nouvelles petites régions partagées entre les 29 phages, rp4 et rp5. En étudiant les séquences codantes de ces protéines, nous nous sommes aperçus que rs1, rs3, rs4 et rs5 présentaient une forte similarité entre protéines homologues d'un même phage, comme rs1, rs2 et rs3 chez P10.

Sordi et al. avaient observé des recombinaisons intra génomiques chez d'autres phages. Etudier ces autres phages pourrait être source de nouvelles découvertes ou compléter les conclusions de mon travail. Ainsi durant mon dernier mois de stage, je m'appliquerai à étudier le phage Rhodococcus ReqiPoco6.

Ce travail pose les jalons d'un plus gros projet. Il est prévu d'améliorer la caractérisation et la recherche de protéines homologues chez les phages en construisant des protomates à partir des PLMAs obtenus par Paloma en utilisant les autres outils de la suite Protomata Learner pour tenter de prédire par homologie la fonction d'un ensemble de protéines phagiques. À terme, l'objectif est d'améliorer l'annotation de protéines phagiques, et d'aider à mieux comprendre le phénomène de recombinaison.

5 Bibliographie

Références

1. Botstein, D. A Theory of Modular Evolution for Bacteriophages*. *Annals of the New York Academy of Sciences* **354.** _eprint : https ://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1980.tb27987.x, 484-491. issn : 1749-6632 (1980).

2. De Paepe, M. *et al.* Temperate Phages Acquire DNA from Defective Prophages by Relaxed Homologous Recombination : The Role of Rad52-Like Recombinases. *PLoS Genetics* **10.** issn : 1553-7390 (6 mar. 2014).

3. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673-4680. issn : 0305-1048 (11 nov. 1994).

4. Edgar, R. C. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792-1797. issn : 1362-4962 (2004).

5. Coste, F. & Kerbellec, G. *Learning Automata on Protein Sequences* in *JOBIM* (éd. Denise, A. *et al.*) (Bordeaux, France, juil. 2006), 199-210.

6. Kerbellec, G. *Apprentissage d'automates modélisant des familles de séquences protéiques* thèse de doct. (Université Rennes 1, 19 juin 2008).

7. De Sordi, L., Khanna, V. & Debarbieux, L. The Gut Microbiota Facilitates Drifts in the Genetic Diversity and Infectivity of Bacterial Viruses. *Cell Host & Microbe* **22,** 801-808.e3. issn : 1934-6069 (13 déc. 2017).

8. Tartaglia, N. R. *et al.* Extracellular vesicles produced by human and animal Staphylococcus aureus strains share a highly conserved core proteome. *Scientific Reports* **10.** Number : 1 Publisher : Nature Publishing Group, 1-13. issn : 2045-2322 (21 mai 2020).

9. Coste, F., Garet, G., Groisillier, A., Nicolas, J. & Tonon, T. *Automated Enzyme classification by Formal Concept Analysis* in. ICFCA - 12th International Conference on Formal Concept Analysis (Springer, 10 juin 2014).

10. Bretaudeau, A. *et al.* CyanoLyase : a database of phycobilin lyase sequences, motifs and functions. *Nucleic Acids Research,* 6 (21 nov. 2012).

11. Morgenstern, B. DIALIGN 2 : improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics (Oxford, England)* **15,** 211-218. issn : 1367-4803 (mar. 1999).

12. Terzian, P. *et al.* Prokaryotic Virus Remote Homologous Groups (PHROGs) : clustering proteins from viruses of prokaryotes using remote homology detection.

13. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)* **21,** 951-960. issn : 1367-4803 (1er avr. 2005).

14. Sandmeier, H. Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. *Molecular Microbiology* **12,** 343-350. issn : 0950-382X (mai 1994).

15. Haggård-Ljungquist, E., Halling, C. & Calendar, R. DNA sequences of the tail fiber genes of bacteriophage P2 : evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. *Journal of Bacteriology* **174,** 1462-1477. issn : 0021-9193 (mar. 1992).

16. Sandmeier, H., Iida, S. & Arber, W. DNA inversion regions Min of plasmid p15B and Cin of bacteriophage P1 : evolution of bacteriophage tail fiber genes. *Journal of Bacteriology* **174,** 3936-3944. issn : 0021-9193 (juin 1992).

17. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9.** Number : 1 Publisher : Nature Publishing Group, 2542. issn : 2041-1723 (29 juin 2018).

18. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35.** Number : 11 Publisher : Nature Publishing Group, 1026-1028. issn : 1546-1696 (nov. 2017).

19. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve : multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14,** 1394-1403. issn : 1088-9051 (juil. 2004).

20. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve : multiple genome alignment with gene gain, loss and rearrangement. *PloS One* **5,** e11147. issn : 1932-6203 (25 juin 2010).

21. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1,** 31-45. issn : 2159-7073 (2011).

22. Nolan, J. M., Petrov, V., Bertrand, C., Krisch, H. M. & Karam, J. D. Genetic diversity among five T4-like bacteriophages. *Virology Journal* **3,** 30. issn : 1743-422X (23 mai 2006).

23. Nelson, D. Phage taxonomy : we agree to disagree. *Journal of Bacteriology* **186,** 7029-7031. issn : 0021-9193 (nov. 2004).

6 Résumé / Abstract

Benjamin BLANC - Master de bioinformatique de l'Université de Rennes 1

Les phages sont des micro-organismes viraux qui infectent les bactéries. Leur génome est structuré sous la forme de modules fonctionnels, fréquemment mélangés par recombinaison. Ces recombinaisons, favorisées par l'homologie entre séquences, donnent lieu à de nouvelles combinaisons de modules et donc à des phages potentiellement nouveaux et viables, mais les rend difficile à annoter avec des méthodes in-silico. Plus particulièrement, les méthodes classiques d'alignement multiple n'arrive pas à prendre en charge les recombinaisons. Le but de mon stage est d'utiliser un nouvel outil d'alignement multiple, Paloma, pour étudier les recombinaisons entre protéines homologues de 32 phages, dont certaines ont déjà fait l'objet de recombinaison d'après la littérature. La visualisation des alignements généré a permis de retrouver des régions recombinées chez 8 phages décrites par le passé chez 3 phages, et la présence de 4 séquences conservées entre ces 8 phages autour de la région recombinée. Il pourrait s'agir d'empreintes de recombinaison. Ces résultats sont encourageants et démontre la capacité de Paloma à tenir compte de recombinaisons lors d'un alignement multiple. À terme, il est prévu de coupler Paloma avec un outil d'apprentissage d'automates pour prédire la fonction inconnue de protéines de phage.

Mots clés : protéines, homologie, recombinaisons, séquences, Paloma

Phages are viral microorganisms that infect bacteria. Their genome is structured in many functional modules, frequently mixed by recombination. These recombinations, encouraged by the homology between sequences, result in new combinations of modules and thus in potentially new and viable phages, but make them difficult to annotate with in-silico methods. In particular, classical multiple alignment methods are unable to handle recombinations. The aim of my internship is to use a new multiple alignment tool, Paloma, to study recombinations between homologous proteins of 32 phages, whose some have already been recombined according to the literature. The visualization of the generated alignments allowed to find recombined regions in 8 phages described by the past in 3 phages, and the presence of 4 conserved sequences between these 8 phages around the recombined region. These could be recombination fingerprints. These results are encouraging and demonstrate the ability of Paloma to take into account recombinations in a multiple alignment. In the future, it is planned to couple Paloma with an automata learning tool to predict the unknown function of phage proteins.

Keywords : proteins, homology, rearrangement, sequences, Paloma