Développement de modèles spécifiques aux séquences génomique virales

Louise-Amelie Schmitt

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

SPÉCIALITÉ INFORMATIQUE

Par Louise-Amélie SCHMITT

DÉVELOPPEMENT DE MODÈLES SPÉCIFIQUES AUX SÉQUENCES MÉTAGÉNOMIQUES VIRALES

Sous la direction de : Guillaume BLIN

Soutenue le 19 juillet 2017

Membres du jury :

M. BLIN, Guillaume	Professeur des Universités, Université de Bordeaux	Président
Mme. GASPIN, Christine	Directeur de Recherche, INRA Toulouse	Rapporteur
M. PETERLONGO, Pierre	Chargé de Recherche, Inria Rennes	Rapporteur
Mme BEURTON-AIMAR, Marie	Maître de Conférences, Université de Bordeaux	Examinateur

Titre : Développement de modèles spécifiques aux séquences génomique virales

Résumé :

Le séquençage ADN d'échantillons complexes contenant plusieurs espèces est une technique de choix pour étudier le paysage viral d'un milieu donné. Or les génomes viraux sont difficiles à identifier, de par leur extrême variabilité et la relation étroite qu'ils entretiennent avec leurs hôtes. Nous proposons de nouvelles pistes de recherche pour apporter une solution spécifique aux séquences virales afin de répondre au besoin d'identification pour lequel les solutions génériques existantes n'apportent pas de réponse satisfaisante.

Mots clés :

métagénomique, virologie, signature, k-mers, classification supervisée, assignation taxonomique, phylogénie, environnement, apprentissage machine

---

Title : Developing viral genomic data-specific classification models

Abstract :

DNA sequencing of complex samples containing various living species is a choice approach to study the viral landscape of a given environment. Viral genomes are hard to identify due to their extreme variability and the tight relationship they have with their hosts. We hereby provide new leads for the development of a viruses-specific solution to the need for accurate identification that hasn't found a satisfactory solution in the existing universal software so far.

Keywords : [3 minimum]

Metagenomics, virology, signature, k-mers, supervised classification, taxonomic assignment, phylogeny, environment, machine learning

---

Unité de recherche

LaBRI, UMR CNRS 5800
351 cours de la Libération
330405 Talence

« *Nous finissons toujours par avoir le visage de nos vérités.* »

Albert Camus, *Le mythe de Sisyphe*

Remerciements

Cette thèse n'aurait jamais vu le jour sans les nombreuses personnes suivantes, qui m'ont offert un soutien sans faille, même lors des moments les plus difficiles. Je tiens à remercier particulièrement Guillaume Blin, qui m'a apporté une aide inestimable pour l'organisation et la rédaction de mon manuscrit et qui a eu la générosité d'adopter l'orpheline sans la moindre hésitation ; Marie Beurton-Aimar, qui a su me pousser à sortir des sables mouvants et dont l'aide précieuse et désintéressée ne sera jamais oubliée ; Patricia Thébault, qui a dépassé ses limites et m'a offert un soutien d'une compréhension et d'une humanité remarquables. Ma gratitude envers eux peine à trouver ses mots.

Je remercie très vivement Christine Gaspin et Pierre Peterlongo pour avoir rapporté ce manuscrit, mais également pour leur bienveillance et leur patience, qui m'ont permis de soutenir sereinement malgré les circonstances.

Je remercie également Thierry Candresse et Sébastien Theil, pour leurs données, mais également pour leurs conseils et leur expertise en virologie, ainsi que Thomas Hume, Marie Gasparoux et Émeric Sevin qui ont généreusement participé au développement du pipeline et ont offert un regard neuf sur le projet.

Je remercie l'École Doctorale Mathématique et Informatique pour avoir fait preuve de compréhension et de patience, m'ayant ainsi permis de finir et soutenir cette thèse, ainsi que pour leur implication dans l'avenir des futurs doctorants de l'École, afin qu'ils puissent construire leur avenir et celui de la science en sécurité.

Je remercie bien entendu mes courageux relecteurs, qui ont eu la patience de corriger mes bourdes dans un manuscrit pas toujours évident à lire lorsqu'on ne travaille pas dans l'informatique ; notamment mes parents, qui ont toujours cru en moi et m'ont aidée à rationaliser l'irrationel, ainsi que mon conjoint, qui m'a supportée et soutenue toutes ces années, qui a su être mes jambes lorsque les miennes cédaient. Ma meilleure amie également, qui, de surcroît, m'a ouvert des opportunités nouvelles lorsque la précarité de notre situation le demandait.

Je terminerai par un petit clin d'œil aux étudiants du département informatique de l'IUT de Bordeaux à qui j'ai eu le plaisir d'enseigner, ainsi qu'à l'équipe pédagogique grâce à qui j'ai eu une expérience mémorable.

Merci infiniment à tous.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Le monde viral est ubiquitaire et son évolution est indissociable de l'évolution du vivant. Les virus interagissent en permanence avec l'ensemble des espèces de manière bénéfique, nocive ou toutes les combinaisons intermédiaires. Ils font, défont, régulent et protègent les populations vivantes et, pourtant, restent encore très mal connus à l'heure actuelle.

La nature des virus est parfois difficile à cerner pour le grand public et l'immense majorité d'entre eux reste encore à découvrir. Nous connaissons déjà bien l'importance de certaines espèces virales pour la santé et l'agriculture mais, à l'ère de la globalisation des échanges, leur existence soulève de nouvelles problématiques qu'il est essentiel de comprendre et d'explorer. Cela passe par l'étoffement des connaissances sur cette immense biomasse que constituent les espèces virales dont nous ignorons encore tout. Cette thèse s'inscrit dans ce besoin d'identification et de classification. Nous y explorons une approche d'apprentissage automatique pour y répondre.

Nous voulons construire cette approche de manière spécifique aux séquences virales afin de prendre en compte leurs grandes différences avec le vivant. En effet, les virus possèdent des caractéristiques bien particulières qui les rendent difficiles à étudier. Ce sont des parasites obligatoires non vivants qui ne possèdent pas de gène commun à toutes les espèces qui puisse être utilisé pour les comparer. Certains s'insèrent dans le génome de leur hôte et participent ainsi à une pollution des données de référence, ce qui brouille les pistes pour leur identification. Ils sont difficiles à isoler expérimentalement et leurs taux de mutation très élevés rend leur étude difficile avec des outils généralistes construits autour des caractéristiques plus stables du vivant.

Nous espérons que cette exploration puisse apporter de nouvelles informations utiles à l'avancement de l'effort de compréhension du monde viral et de ses multiples intéractions avec l'environnement. Les enjeux économiques et sanitaires de cet effort sont multiples : prévention des épidémies, amélioration des stratégies agricoles, étude et sauvegarde de la biodiversité, promotion et mise en place de démarches écologiques et durables.

Heureusement, les nouvelles technologies de séquençage nous offrent à présent de nouvelles possibilités d'identification des espèces présentes dans des échantillons

complexes, offrant un regard sans précédent sur la biodiversité. Il est à présent possible d'obtenir des données sur l'ensemble des espèces présentes dans un milieu donné à un instant précis. Ces espèces incluent les espèces virales et offrent de nouvelles perspectives pour étoffer l'ensemble des connaissances sur les virus et leurs intéractions avec leur milieu. Cela passe par une identification en deux étapes : savoir reconnaître si une séquence appartient où non à un virus et, le cas échant, évaluer de quel type de virus il s'agit.

Cette thèse a pour objectif d'aborder ces deux étapes en utilisant des techniques d'apprentissage automatique. Nous avons choisi d'adopter une approche globale basée sur la composition nucléotidique des séquences sans utiliser d'annotation fonctionnelle, afin d'explorer la validité d'une approche "à l'aveugle" qui ne nécessite pas de se référer à la réalité biologique que ces séquences représentent.

Nous commencerons par détailler le contexte scientifique et les enjeux de cette thèse, en incluant l'ensemble des notions de base en biologie nécessaires à la compréhension du sujet, étant donné que cette thèse s'adresse à un public issu de la recherche en informatique. Ensuite, nous ferons un tour d'horizon des algorithmes de classification de séquences aujourd'hui disponibles, puis une évaluation de la difficulté de la tâche lorsque les données analysées sont d'origine virale. Enfin, nous présenterons les travaux effectués au cours de l'exploration des deux étapes d'identification et l'analyse de leurs résultats.

Chapitre 1

Contexte scientifique

Sommaire

Ce premier chapitre a pour but de présenter les notions de base sur lesquelles repose ce projet de thèse. Y sont détaillés le contexte biologique, comprenant les connaissances nécessaires à la compréhension des données manipulées et les problématiques scientifiques et environnementales ayant motivé la construction du projet, ainsi que le contexte bioinformatique, décrivant l'environnement technique dans lequel s'inscrivent ces problématiques.
Cette thèse étant une thèse d'Informatique, ce chapitre contient notamment l'ensemble des connaissances de bases en biologie moléculaire nécessaires à la compréhension du sujet. Le lecteur les ayant déjà acquises par ailleurs pourra commencer sans problème la lecture de ce chapitre à la section 1.1.2: [Virologie végétale et environnement].

1.1 Biologie et écologie du monde viral

1.1.1 Acides nucléiques et information génétique

1.1.1.1 Structure de la cellule vivante

La cellule est l'unité de base de tout être vivant. Il s'agit d'un système autonome qui contient tous les éléments nécessaires à sa propre fabrication à partir d'éléments présents dans le milieu dans lequel elle évolue. D'un point de vue thermodynamique, il s'agit d'un système ouvert qui se maintient à un état stationnaire de non-équilibre, grâce à des échanges de matière avec l'extérieur et des transformations irréversibles de cette matière à l'intérieur.

Cet état de déséquilibre permanent garantit les échanges énergétiques nécessaires à la perpétuelle évolution de son contenu. Il est à opposer à l'état d'équilibre chimique dans lequel la vitesse des réactions chimiques est nulle, et qui est associé au concept de mort.

La structure permettant le maintien de ce déséquilibre est la membrane plasmique qui délimite le milieu intérieur et le milieu extérieur de la cellule (cf. Fig. 1.1: [Structure d'une cellule végétale]). Le milieu intérieur d'une cellule est le protoplasme. Son contenu varie selon l'origine de la cellule. Chez de nombreuses espèces, appelées *prokaryotes*, il est uniquement composé du cytoplasme, une émulsion colloïdale dans laquelle se trouvent tous les composants de la cellule. Certaines espèces, en revanche, comme les animaux ou les plantes, possèdent un noyau dans lequel se trouvent enfermés les chromosomes. Ces dernières sont appelées *eukaryotes*. Dans ce cas, le protoplasme est composé de cytoplasme et d'un ou plusieurs noyaux.

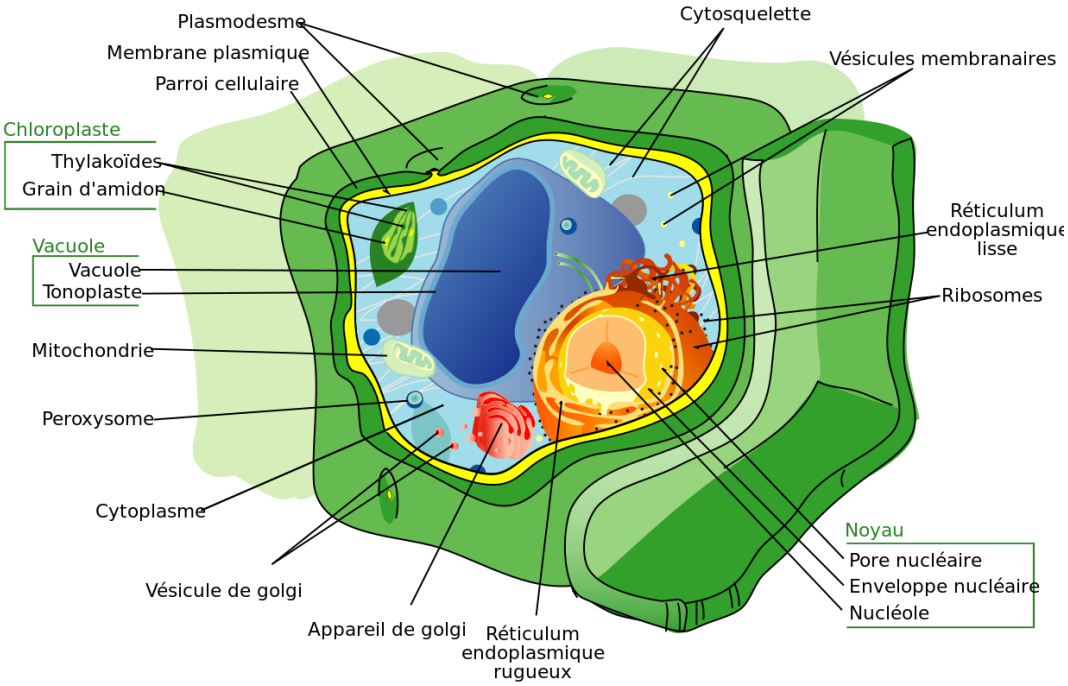


FIGURE 1.1 – Structure d'une cellule végétale. *Source : Wikimedia Commons*

Ce sont les chromosomes qui contiennent l'information génétique. Il s'agit de l'information nécessaire à la cellule pour la construction de ses propres composants. Ils sont composés d'*acide désoxyribonucléique* (ADN) dont la structure est un biopolymère, une chaîne linéaire d'éléments répétés qui forment une molécule ininterrompue qui peut atteindre plusieurs centimètres une fois dépliée (cf. KORNBERG et BAKER [2005]).

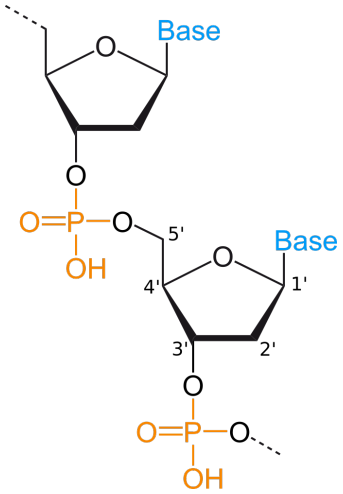


FIGURE 1.2 – Structure de base de l'ADN. En noir : les riboses ; en bleu : les bases azotées ; en orange : les phosphates liant les nucléosides.

Ces éléments répétés sont des nucléosides, liés entre eux par un phosphate afin de former une chaîne (cf. Fig. 1.2: [Structure de l'ADN]). Chaque nucléoside est composé d'un pentose, un sucre à 5 carbones : le ribose, ainsi que d'une base azotée dont la nature varie d'un nucléoside à l'autre. Chaque atome de carbone des riboses est numéroté de 1' à 5'. Les bases sont liées aux riboses par le carbone 1', et les phosphates lient les nucléosides du carbone 5' du ribose d'un nucléoside au carbone 3' de celui du nucléoside suivant.

Les nucléosides (cf. Fig. 1.3: [Nucléotides]) se trouvent également sous une forme libre phosphatée en 5' : le nucléotide. L'ADN étant polymérisé à partir de nucléotides triphosphates libres sur l'extrémité 3' d'une molécule, la longueur d'un brin d'ADN est généralement mesuré en nombre de nucléotides.

L'ADN est synthétisé à partir de quatre types de nucléotides de base, différenciés par la base azotée qui leur est liée (cf. Fig. 1.3: [Nucléotides]) et dont les nucléosides sont les suivantes :

— La désoxyadénosine, dont la base azotée est l'adénine, notée A

— La désoxycytidine, dont la base azotée est la cytosine, notée C

— La désoxyguanosine, dont la base azotée est la guanine, notée G

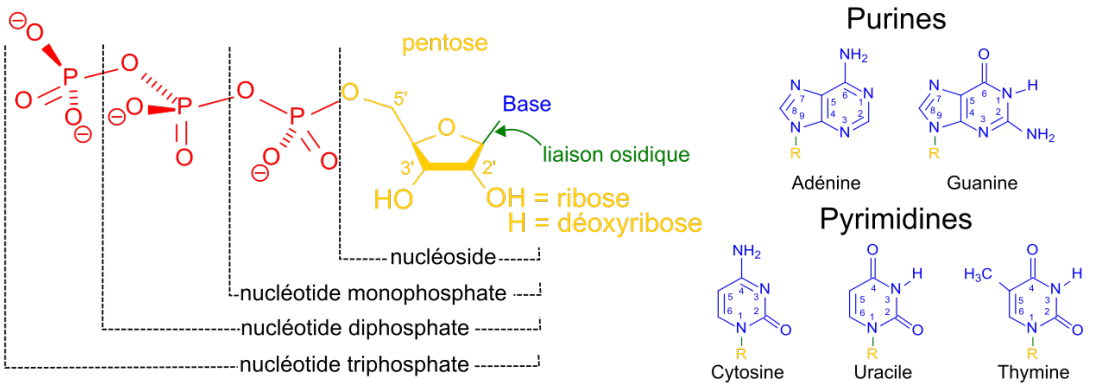


FIGURE 1.3 – Nucléotides et bases azotées. À gauche : les différentes variantes des nucléotides présentes dans le vivant ; à droite : les différentes bases azotées utilisées lors de la synthèse des acides nucléiques. *Source : Wikimedia Commons*

— La désoxythymidine, dont la base azotée est la thymine, notée T

L'uracile (cf. Fig. 1.3: [Nucléotides]) n'est pas présent dans les nucléotides utilisés lors de la synthèse de l'ADN. Il est en revanche présent dans un autre type d'acide nucléique que nous verrons plus tard (cf. 1.1.1.2: [Information génétique et expression génique]).

Chaque chromosome est constitué de deux molécules ininterrompues d'ADN se faisant face, attachées l'une à l'autre par des liaisons de faible énergie entre les bases azotées de chaque brin (cf. Fig. 1.4: [Paires de bases]). Les bases forment naturellement des liens entre elles selon leur morphologie, à la façon des pièces de puzzle dont la complémentarité géométrique permet une liaison spécifique. De cette manière, la structure de l'adénine est complémentaire à celle de la thymine, et la structure de la guanine est complémentaire à celle de la cytosine. La séquence des bases d'un des brins d'un chromosome est ainsi complémentaire à celle de l'autre brin. Le corollaire de cette propriété est qu'il suffit de connaître la séquence des nucléotides d'un des brins pour en déduire celle de l'autre.

Les deux brins d'ADN du chromosome s'enroulent l'un sur l'autre autour des bases liées entre elles sous la forme d'une double hélice dont la structure et le sens de rotation varient selon les organismes et les phases de la vie cellulaire. Ils sont également en contact avec des sels et des protéines assurant leur structure et leur maintenance (cf. Fig. 1.5: [Chromosome]). Les prokaryotes, qui ne possèdent pas de noyau, présentent en général un chromosome unique, la plupart du temps circulaire. En revanche, les eukaryotes en possèdent plusieurs, linéaires et regroupés dans le ou les noyaux de la cellule.

Lors de la division cellulaire, les chromosomes sont répliqués. Les deux brins sont écartés l'un de l'autre, et un brin complémentaire à chacun d'entre eux est

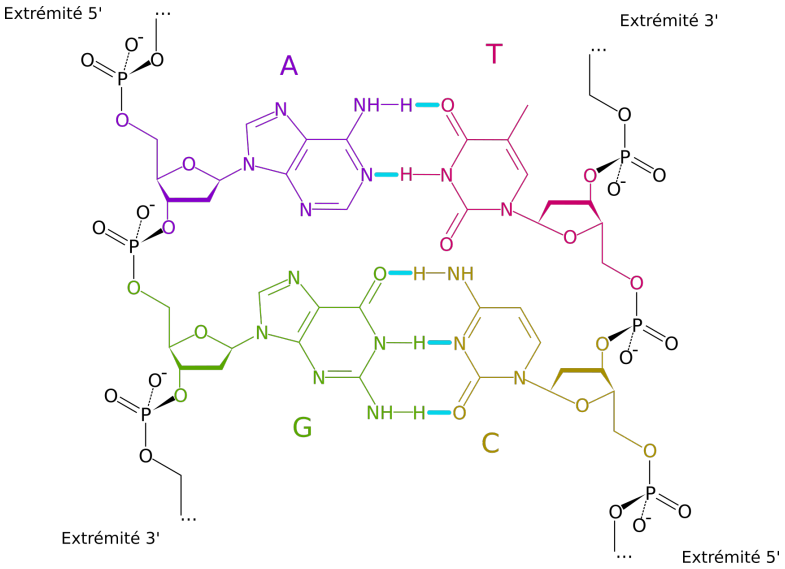


FIGURE 1.4 – Liaisons de faible énergie entre paires de bases azotées. En violet : la désoxyadénosine ; en rouge : la désoxythymidine ; en vert : la désoxyguanosine ; en jaune : la désoxycytidine ; en bleu : les liaisons hydrogène de faible énergie.

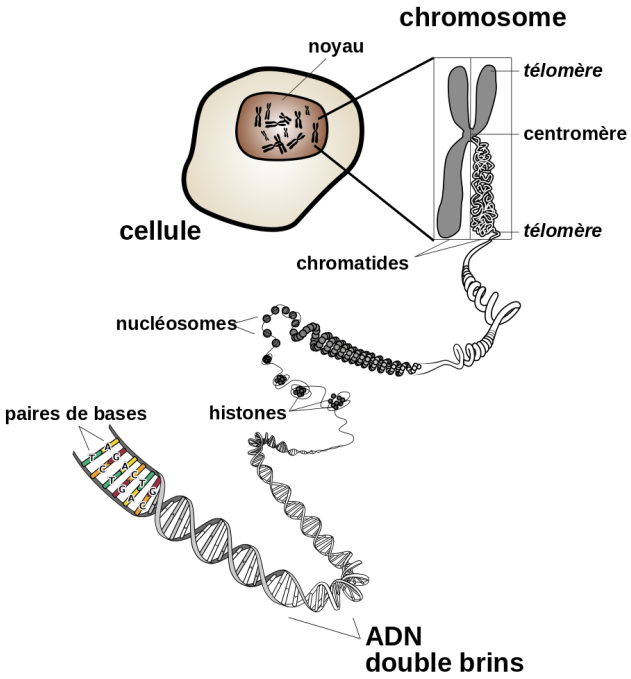


FIGURE 1.5 – Représentation artistique de la structure d'un chromosome chez les eukaryotes. *Source : Wikimedia Commons*

synthétisé au fur et à mesure de leur écartement (cf. Fig. 1.6: [Réplication]). Une fois la réplication effectuée, deux chromosomes en principe identiques sont obtenus, chacun possédant un brin appartenant au chromosome d'origine. Ils seront ensuite distribués dans les deux cellules filles issues de la division cellulaire. C'est ainsi que l'information génétique est transmise d'une génération à l'autre dans les cellules vivantes.

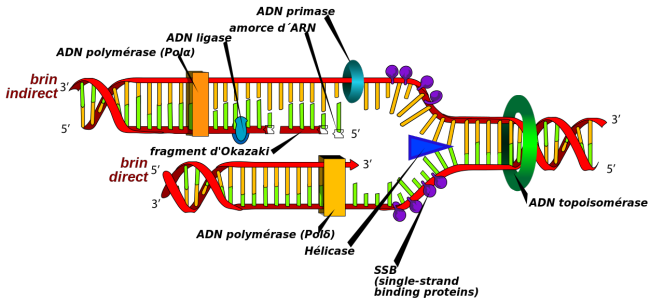


FIGURE 1.6 – Représentation schématique des mécanismes de réplication de l'ADN. Un ensemble de protéines est chargé du dépliement, de l'écartement, de la synthèse de l'ADN, ainsi que l'assurance de la continuité de chaque brin. *Source : Wikimedia Commons*

1.1.1.2 Information génétique et expression génique

La séquence des nucléotides sur les brins des chromosomes est porteuse d'information. Cette information est organisée sous forme de régions spécifiques qui codent chacun pour un élément du fonctionnement cellulaire. Ces régions spécifiques sont les *gènes* et sont répartis sur les deux brins de l'ADN des chromosomes. L'ensemble de l'information portée par l'ensemble des chromosomes est le *génome* de l'individu.

Afin d'extraire et exploiter l'information d'un gène, des structures spécifiques se fixent sur les motifs particuliers signalant son point de départ. L'ADN est ensuite déplié, les deux brins écartés et un nouveau brin est synthétisé sur toute la longueur du gène : c'est l'étape de transcription (cf. Fig. 1.7: [Transcription]). Ainsi, de nombreuses copies du gène sont faites les unes à la suite des autres pour une utilisation ultérieure.

Le brin d'acide nucléique synthétisé au cours de cette étape n'est pas un brin d'ADN, il s'agit ici d'*acide ribonucléique* (ARN). Les nucléosides de l'ARN sont proches des nucléosides de l'ADN dans leur composition, mais le pentose qui font leur base n'est ici pas du désoxyribose : il s'agit de ribose, qui porte un atome d'oxygène sur son carbone 2' (cf. Fig. 1.3: [Nucléotides]). L'autre différence majeure de l'ARN est que la thymidine en est absente : le nucléoside complémentaire à l'adénosine est l'uridine, dont la base azotée est l'uracile, notée U.

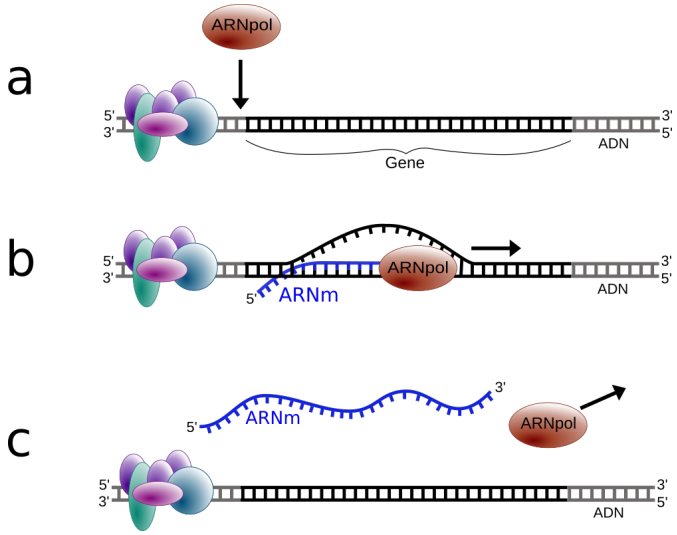


FIGURE 1.7 – Représentation schématique des mécanismes de transcription. (a) Phase d'initiation : L'ARN polymérase se fixe sur le début du gène. (b) Phase d'élongation : Le brin d'ARN complémentaire au brin lu est synthétisé. (c) Phase de terminaison : le motif de fin de gène est atteint, la polymérase et le brin d'ADN sont libérés. *Source : Wikimedia Commons*

Suivant le gène transcrit, les brins d'ARN ainsi produits peuvent avoir plusieurs fonctions. Certains portent une séquence de nucléosides qui entraîne un repliement de la molécule sur elle-même, formant des tiges et des boucles lorsque ses propres bases s'apparient les unes avec les autres. Ce repliement leur confère des propriétés chimiques particulières qui leur permet de jouer un rôle actif dans la vie cellulaire. Par exemple, les ribozymes (cf. Fig. 1.8: [Ribozyme]) sont des molécules d'ARN qui possèdes des propriétés catalytiques leur permettant de favoriser des réactions chimiques et de transformer le matériel cellulaire.

D'autres servent à transporter l'information qu'ils contiennent vers d'autres structures qui vont à leur tour lire leur séquence pour synthétiser un autre type de biopolymère : les protéines. Ce sont les ARN messagers (ARNm), la première catégorie d'ARN à avoir été décrite (BRENNER et collab. [1961]).

Les protéines sont des molécules complexes de haute spécificité. Ce sont les composants cellulaires les plus abondants et les plus divers, tant en terme de structure que de fonction. Elles jouent un rôle crucial dans presque tous les processus biologiques, que ce soit pour la catalyse des réactions chimiques (on les appelle dans ce cas des enzymes), le stockage et le transport de matières biologiques, la structure et la motilité cellulaires, ou encore les mécanismes de défense de l'organisme.

Il s'agit d'un autre type de biopolymère, à l'instar des acides nucléiques. Ce sont de longues chaînes polymérisées à partir d'acides aminés liés bout à bout par une

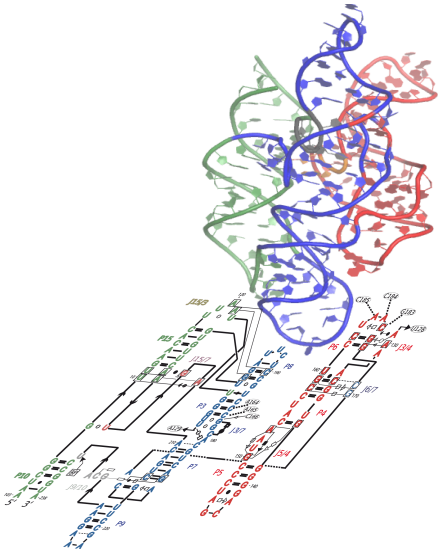


FIGURE 1.8 – Représentation 2D et 3D du repliement du ribozyme DiGIR1, un ARN impliqué dans la régulation de l'expression de vertains gènes. *Source : Wikimedia Commons*

liaison amide entre le groupe amide d'un acide aminé et le groupe carboxyle du suivant (cf. Fig. 1.9: [Protéine]).

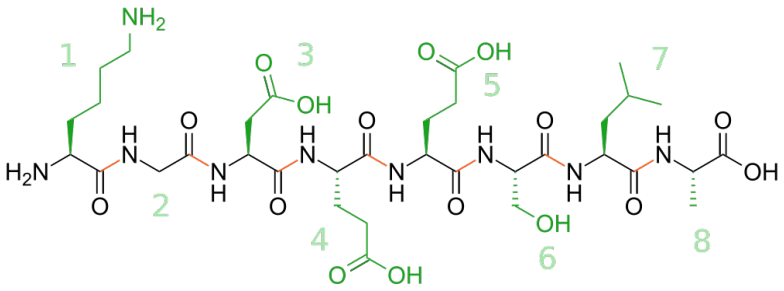


FIGURE 1.9 – Exemple de chaîne protéique : le peptide BMP, un composé aromatique spécifique à la viande de bœuf. En rouge : les liaisons amide entre les différents acides aminés. En vert : les chaînes latérales des acides aminés, dans l'ordre (1) Lysine, (2) Glycine, (3) Aspartate, (4,5) Glutamate, (6) Sérine, (7) Leucine (8) Alanine. *Source : Wikimedia Commons*

Les acides aminés utilisés lors de la synthèse des protéines sont au nombre de 22, dont 20 sont universellement présents (cf. Fig. 1.10: [Acides aminés]). Ils sont différenciés par une chaîne latérale qui varie d'un acide aminé à l'autre, en longueur, en composition et en propriétés chimiques, et qui n'est pas impliquée dans la liaison amide. La glycine est une exception à cette règle car elle ne possède pas de chaîne latérale, ce qui a la particularité de lui conférer une symétrie selon le plan formé par les deux carbones et l'azote.

La cellule passe de la séquence des nucléosides d'un ARNm à la protéine corres-

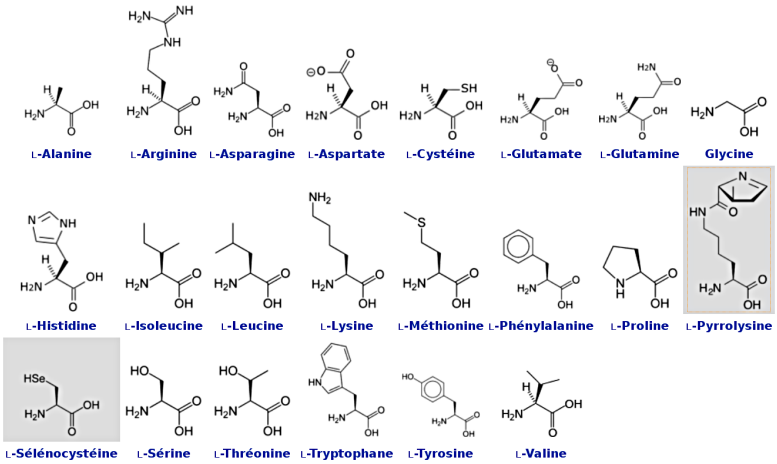


FIGURE 1.10 – Formule chimique des 22 acides aminés utilisés lors de la synthèse des protéines. Les acides aminés en gris ne se trouvent que chez certains organismes. *Source : Wikimedia Commons*

pondante en associant un acide aminé à chaque triplet de nucléoside "lu" par des complexes formés de molécules d'ARN et de protéines appelé *ribosomes*. Ces derniers se fixent sur l'ARNm au début du gène et construisent la protéine correspondante au fur et à mesure qu'ils se déplacent sur la molécule : c'est l'étape de *traduction* (cf. Fig. 1.11: [Traduction]).

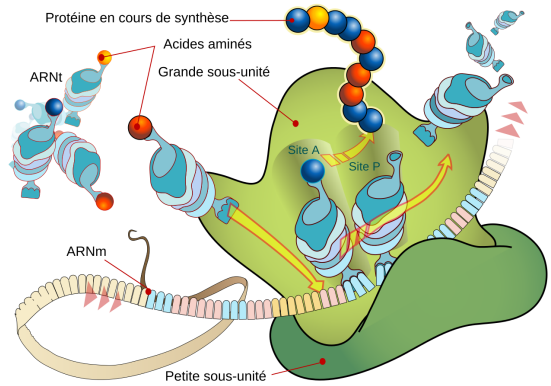


FIGURE 1.11 – Ribosome en cours de traduction d'un ARNm. *Source : Wikimedia Commons*

Chaque ribosome est composé de deux sous-unités qui s'assemblent autour de l'ARNm en début de lecture. La grande sous-unité est formée par plusieurs molécules d'ARN et la petite par une seule, ce sont les ARN ribosomiques (ARNr). Ces structures sont cruciales pour la survie de la cellule et varient très peu d'une espèce à l'autre.

L'élongation de la protéine en construction se fait au cœur de cette structure, grâce à l'intéraction entre le ribosome, l'ARNm et un autre type d'ARN : les ARN de transfert (ARNt), qui ont pour particularité de transporter un acide aminé et de

présenter à leur surface trois bases azotées consécutives. Au fur et à mesure que le ribosome avance sur de nouvelles bases de l'ARNm, l'ensemble des trois bases en cours de lecture s'apparie avec l'ARNt qui lui est complémentaire (cf. Fig. 1.5: [Chromosome]). L'acide aminé transporté par cet ARNt est ensuite lié à la protéine naissante par une liaison amide, et le ribosome peut passer aux trois bases suivantes de l'ARNm.

Ces ensembles de trois bases sont appelés *codons* et sont au nombre de $4^3 = 64$. Les ARNt définissent une correspondance précise entre chacun de ces codons et un acide aminé : c'est le *code génétique* (cf. Fig. 1.12: [Code génétique]).

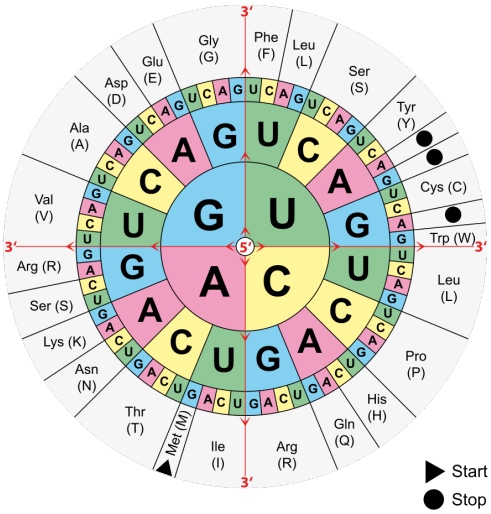


FIGURE 1.12 – Code génétique universel. *Source : Wikimedia Commons*

Trois codons signalent la fin de la lecture et correspondant à des ARNt particuliers permettant la terminaison de la protéine, sauf cas particuliers. Mais dans le cas général, comme il n'y a que 20 acides aminés utilisés lors de la traduction, il y a une redondance dans la correspondance entre chacun des 61 codons restants et les acides aminés qui leur sont associés. On peut par exemple observer que tous les codons commençant par "CU" correspondent à des ARNt transportant tous de la leucine (cf. Fig. 1.12: [Code génétique]). Certains acides aminés, en revanche, n'ont qu'un codon correspondant, comme la methionine ("AUG"), qui a en outre la particularité de signaler l'endroit où doit débuter la lecture de l'ARNm.

Le code génétique est généralement très conservé d'une espèce à l'autre, même si on peut observer des variations mineures. Par exemple, deux des codons de terminaison peuvent exceptionnellement correspondre aux ARNt transportant la pyrrolysine et la sélénocystéine vues précédemment (cf. Fig. 1.10: [Acides aminés]).

L'ensemble de ces mécanismes constitue le cœur du vivant. Ce sont eux qui garantissent la production des différents composants cellulaires, leur entretien et leur adaptation au milieu. Ils permettent en outre d'assurer la transmission de l'information d'une génération à l'autre ainsi que son exploitation.

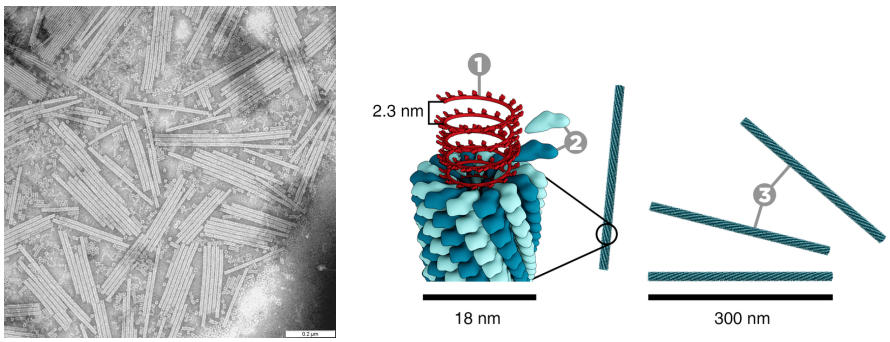
1.1.2 Virologie végétale et environnement

1.1.2.1 Les virus

Les virus sont des parasites obligatoires. Ce sont des molécules d'acides nucléiques (ARN ou ADN, cf. 1.1.1.1: [Structure de la cellule vivante], 1.1.1.2: [Information génétique et expression génique]) ne faisant pas partie des chromosomes d'un organisme, mais capables d'utiliser la machinerie cellulaire de leur hôte afin d'exprimer et répliquer leur contenu.

Il existe d'autres types d'acides nucléiques non chromosomiques qui peuvent être présents à l'intérieur d'une cellule et qui contiennent des gènes, mais la particularité des virus est qu'ils possèdent une forme extracellulaire leur permettant d'être transmis de manière horizontale (hors division cellulaire) entre des cellules vivantes sans avoir recours à des mécanismes cellulaires et en passant par le milieu extérieur. Cette forme est appelée *virion*. Elle est constituée du virus entouré d'une structure proteique protectrice appelée *capside* (cf. Fig. 1.13: [Mosaïque du tabac]).

Certains possèdent d'autres structures additionnelles, mais la forme la plus élémentaire d'un virion est un acide nucléique encapsidé. Cette forme est métaboliquement inerte et permet la conservation durable des virus dans des milieux divers.



(a) Virions vus par microscopie électronique

(b) Structure du virion : (1) Virus, ici une molécule d'ARN (2) Protéines de la capside (3) Virions

FIGURE 1.13 – Exemple de virus : le virus de la mosaïque du tabac (TMV), un virus pathogène des solanacées (pétunia, datura, belladone, mandragore, tabac, etc.). *Source : Wikimedia Commons*

Les génomes viraux sont très courts et contiennent peu d'information. Le plus petit virus connu est un virus satellite de l'hépatite B : le virus Delta. Il ne mesure

que $1,68.10^3$ nucléotides, ne contient qu'un seul gène et nécessite une infection préalable par l'hépatite B pour pouvoir se répliquer dans l'hôte (cf. http://viralzone.expasy.org/viralzone/all_by_species/175.html).

À l'autre extrémité de l'échelle on trouve le Megavirus, qui mesure $1,26.10^6$ nucléotides, contient 1120 gènes et dont la capside mesure 440nm de diamètre (ARSLAN et collab. [2011]). À titre de comparaison, la plus petite bactérie connue présente un diamètre de 250nm (MOROWITZ et collab. [1962]).

Les génomes viraux codent principalement pour des fonctions qui ne peuvent pas être assurées par l'hôte (MADIGAN [2007]), leur petite taille étant généralement incompatible avec toute redondance. Ainsi, toutes les étapes d'expression et de réplication sont assurées par les structures issues du génome de l'hôte : lors de leur phase intracellulaire, les virus se comportent comme des acides nucléiques cellulaires, ce qui force l'hôte à répliquer leur génome et à synthétiser les composants viraux à partir desquels de nouveaux virions sont assemblés (PRESCOTT et collab. [2003]).

Les virus possèdent des structures très variées. Il ne s'agit pas uniquement de leur taille, leur génome ou la manière dont les protéines de la capside s'assemblent. La nature même de leur matériel génétique présente une variabilité très importante. Certains se présentent directement sous forme d'ARNm, prêts à être traduits, mais d'autres types existent, comme des espèces à ARN moins directes ou des espèces à ADN.

Une classification des virus sur la base de leur structure moléculaire et les étapes nécessaires à la production de l'ARNm nécessaire à l'expression des gènes viraux a été mise en place dans les années 70 : la classification de Baltimore (BALTIMORE [1971], cf. Fig. 1.14: [Classification de Baltimore]). Cette classification rend compte de la diversité des mécanismes mis en œuvre lors de l'infection de l'hôte, et permet d'apercevoir la grande variété du monde viral.

Les différentes classes sont définies de la manière suivante :

— **I - Génomes à ADN double brin :** Ce sont les virus dont la structure et le comportement sont les plus similaires aux chromosomes de l'hôte. Leur expression se déroule de la même manière (cf. 1.1.1.2: [Information génétique et expression génique]).

— **II - Génomes à ADN simple brin :** Contrairement aux chromosomes de l'hôte qui se présente sous forme d'une double hélice formée par deux brins d'ADN enroulés l'un sur l'autre (cf. 1.1.1.1: [Structure de la cellule vivante]), ces virus ne possèdent qu'un brin et doivent subir une étape intermédiaire lors de laquelle un brin d'ADN complémentaire est synthétisé sur le virus lui-même pour retrouver une structure similaire exploitable par la cellule.

— **III - Génomes à ARN double brin :** L'un des deux brins d'ARN du virus doit être transcrit pour produire l'ARNm viral, mais la cellule ne possède aucune struc-

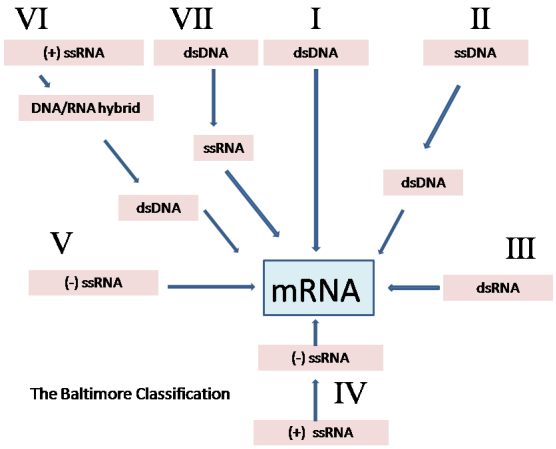


FIGURE 1.14 – Schéma représentant les classes définies dans la classification de Baltimore, en relation avec les étapes nécessaires à la production de l'ARNm viral. *Source : Wikimedia Commons*

ture permettant de synthétiser un ARNm à partir d'une double hélice d'ARN : elle ne sait "lire" que l'ADN. Lors de l'infection, le problème est contourné par l'injection simultanée du virus et de l'enzyme capable d'accomplir cette tâche, tous deux contenus dans le virion.

— **IV - Génomes à ARN simple brin de polarité positive :** La polarité d'un brin désigne l'orientation relative d'un brin par rapport au sens de lecture de l'ARNm dont il porte la séquence. Un ARN de polarité positive possède la même orientation qu'un ARNm et est en configuration directe pour l'étape de traduction. La cellule le reconnaît directement comme un ARNm et entame sa traduction dès le début de l'infection.

— **V - Génomes à ARN simple brin de polarité négative :** À l'inverse, les virus à ARN de polarité négatives possèdent la séquence complémentaire à l'ARNm qui doit être traduit. De la même manière que pour la classe III, ils nécessitent donc une étape de transcription afin d'obtenir des ARNm disponibles à la traduction.

— **VI - Génomes à ARN simple brin se répliquant avec un intermédiaire à ADN (rétrovirus) :** Ce sont des virus à ARN de polarité positive, comme la classe IV. La différence avec ces derniers est que le virus n'est pas reconnaissable comme un ARNm par la cellule et ne peut donc pas être traduit directement. Ils nécessitent une étape de *transcription inverse* : à l'inverse de la transcription lors de laquelle un brin d'ARN est synthétisé par complémentarité avec un brin d'ADN, ici un brin d'ADN est synthétisé sur le brin d'ARN viral. Cette étape supplémentaire nécessite une enzyme virale, *la transcriptase inverse*, également injectée dans la cellule avec le virus. Le brin d'ADN ainsi synthétisé subira ensuite une

transcription classique afin de produire des ARNm.

— **VII - Génomes à ADN double brin se répliquant avec un intermédiaire à ARN :** Il s'agit de la seule classe qui n'est pas définie par la stratégie d'expression de ses membres, mais par le mécanisme avec lequel ils se répliquent. Leur mode d'expression est identique à la classe I, très proche de l'expression naturelle du génome de l'hôte. En revanche, là où les virus de classe I sont répliqués comme les chromosomes de l'hôte (cf. 1.1.1.1: [Structure de la cellule vivante]), les virus de classe VII ont une stratégie de réplication surprenante qui inclut une étape de transcription inverse similaire à la classe VI. Suite à l'entrée du virus dans la cellule, ce dernier se circularise et est transcrit par les enzymes de l'hôte. Certains des ARN produits subissent ensuite une transcription inverse afin de produire en ADN de nouvelles copies du virus.

Les virus ne sont pas tous pathogènes. C'est encore par la diversité de leurs intéractions avec l'hôte dont ils dépendent que s'illustre leur originalité. Il est important de comprendre le rôle essentiel qu'ils jouent en faveur des espèces qu'ils infectent afin d'expliquer leur présence et leur nombre.

Il existe un biais d'étude en faveur des virus pathogènes, d'une part parce qu'ils sont facilement identifiables, et d'autre part parce qu'ils constituent une menace économique et sanitaire (cf. 1.1.2.2: [Importance de la virologie]). Il n'est donc pas surprenant qu'ils ne soient massivement connus que pour le seul danger que certains d'entre eux représentent (ROOSSINCK [2010]).

Ces relations hôte-parasite sont pourtant extrêmement variées et dépendent fortement du type d'organisme infecté. Les virus affectant les plantes illustrent particulièrement bien cette grande diversité. On distingue parmi eux quatre grandes catégories de modes de vie (ROOSSINCK [2010]) :

— **Mode de vie persistant :** Ce sont des virus généralement non pathogènes, qui présentent un très faible titrage (nombre de copies) cellulaire et qui appartiennent tous à la classe III. Ils ne possèdent pas de phase extracellulaire leur permettant de se transmettre de manière horizontale (entre deux cellules distinctes) et sont uniquement transmis de manière verticale : ils sont présents dans toutes les cellules et sont transmis d'une génération de plantes à l'autre par reproduction sexuée. Il est difficile de comprendre complètement toutes les conséquences de leur présence, car il est presque impossible de trouver une plante non infectée. Ils constituent les virus les plus courants, qui peuvent atteindre 70% des espèces virales concernées par certaines familles de végétaux.

— **Mode de vie aigu :** Contrairement aux virus persistants, ceux-ci sont presque exclusivement transmis de manière horizontale. Ils peuvent atteindre des titrages très élevés, qui évoluent généralement par cycles. L'infection aiguë peut se terminer de trois façons différentes. Dans le premier cas, elle est fatale pour

l'hôte. Dans le deuxième, les défenses immunitaires de l'hôte permettent un rétablissement complet de l'organisme. Enfin, dans le troisième, l'infection subit une conversion vers un mode de vie chronique.

— **Mode de vie chronique :** Ces virus restent présents dans l'hôte durant de longues périodes, contrairement aux infections aiguës, et peuvent parfois causer des maladies. Ils conservent la capacité à se transmettre horizontalement et certains peuvent retrouver un mode de vie aigu sous certaines conditions.

— **Mode de vie endogène :** Il s'agit ici d'une catégorie qui se trouve à la frontière du monde viral. Ces virus se sont insérés dans le génome de l'hôte et sont répliqués avec le reste des chromosomes. Une grande majorité d'entre eux sont inactifs et ont subi des transformations qui rendent leurs propres gènes défectueux. Certains de ces virus latents peuvent retrouver une activité sous certaines conditions, mais nombreux sont ceux qui n'en possèdent plus l'information nécessaire. Les dégâts subis rendent parfois difficile de poser la limite entre génome viral et génome de l'hôte.

Tous les virus ne représentent pas nécessairement un danger pour leur hôte, mais même les virus pathogènes jouent un rôle important dans l'évolution des organismes qu'ils infectent, ou avec lesquels leurs hôtes cohabitent.

Leur nature pathogène joue un rôle dans la régulation des populations et influence grandement l'équilibre des écosystèmes dans lesquels ils évoluent. Les virus ayant un mode de vie aigu, par exemple, sont particulièrement nombreux et actifs dans les zones où les hôtes sont concentrés. On peut en effet observer des épidémies particulièrement sévères dans les monocultures, les zones d'élevage intensif et les grands centres urbains, dont l'origine persistante a été démontrée dans certains cas (cf. 1.1.2.2: [Importance de la virologie], VILLARREAL et collab. [2000]).

Leur influence se joue également au niveau moléculaire. Les acides nucléiques ne sont pas des molécules parfaitement stables et les protéines intéragissant avec eux ne se comportent pas toujours de manière attendue. Par conséquent, il se produit parfois des événements qui sortent de ce qui est strictement nécessaire lors d'une infection virale.

La *transduction* est un processus courant, lors duquel des fragments du génome de l'hôte sont encapsidés à la place du virus, libérés dans le milieu, puis injectés dans un nouvel hôte. La capside étant responsable de l'injection du matériel génétique, l'origine des acides nucléiques qu'elle contient importe peu. Les fragments ainsi transmis d'une cellule peuvent être inexploitables, mais dans certains cas des gènes entiers peuvent être transférés. Il s'agit d'un processus bien connu de transfert horizontal d'information génétique, en particulier chez les prokaryotes (GRIFFITHS et collab. [2000], cf. Fig. 1.15: [Transduction]). Mais plusieurs observations de processus similaires chez les eukaryotes suggèrent que ce mécanisme est présent chez

l'ensemble du vivant (SAVORY et collab. [2015], JAMAIN et collab. [2001], SOANES et RICHARDS [2014]) et ont directement inspiré de nombreux projets de thérapie génique.

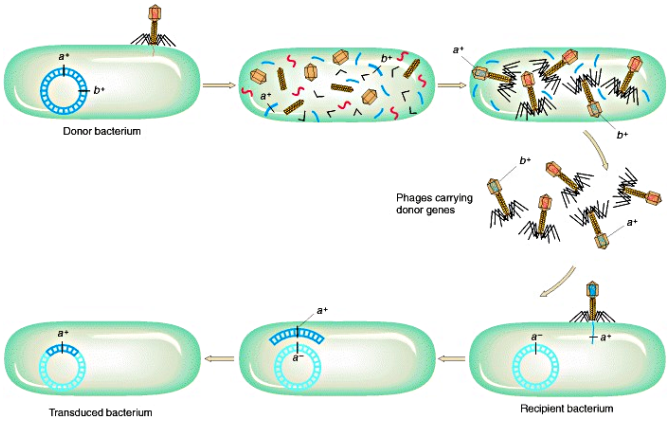


FIGURE 1.15 – Exemple de transduction : transduction généralisée chez les bactéries. Dans ce cas, n'importe quel fragment du génome du donneur peut être transmis au receveur, contrairement à la transduction spécialisée au cours de laquelle seules certaines parties du génome peuvent être transduites. *Source :* GRIFFITHS *et collab. [2000]*

La présence d'un virus peut parfois être bénéfique pour l'hôte. C'est très souvent le cas pour les virus endogènes, qui ne peuvent assurer leur survie sur le long terme si leur présence se fait au détriment de l'hôte. Certains offrent ainsi à l'hôte une résistance à d'autres infections virales, souvent causées par des virus qui leur sont proches.

Mais parfois les intéractions sont moins directes et dépendent de l'endroit où le virus s'est inséré dans le chromosome. Selon sa position par rapport aux autres gènes, il peut en modifier le niveau d'expression. Un des exemples les mieux étudiés est le cas d'un rétrovirus endogène humain dont l'insertion aurait entraîné l'expression de l'amylase dans les glandes salivaires en plus du pancréas, améliorant ainsi notre capacité à utiliser les glucides à longues chaînes (COFFIN et collab. [1997]).

Virus et hôtes sont intimement liés. La complexité et la variété de leurs intéractions témoignent d'une longue histoire d'influences mutuelles et du rôle qu'elles ont joué au cours de l'évolution. Leur présence dans l'ensemble du vivant ainsi que des traces très anciennes au niveau moléculaire suggèrent qu'ils sont apparus très tôt dans l'histoire de la vie, et ont participé de manière active à la construction du monde actuel. Bien que considérés comme des parasites obligatoires, leur complexité bouscule sans cesse les limites théoriques entre parasitisme, commensalisme et symbiose.

1.1.2.2 Importance de la virologie

La virologie est la discipline scientifique consacrée à l'étude des virus. Elle inclut un effort de description de leurs structures et de leurs mécanismes de reproduction et d'intéraction avec leurs hôtes, ainsi que l'étude des maladies qui leurs sont associées, mais également une approche plus globale de leur impact sur l'environnement. Elle possède également une forte dimension appliquée, notamment à travers la modification de souches virales à des fins thérapeutiques (e.g. vaccins, thérapie génique).

Cette thèse est fortement inspirée des travaux de l'équipe de virologie végétale de l'UMR1332 à l'Institut National de Recherche Agronomique (INRA). Cette équipe s'emploie notamment à développer de nouvelles approches pour la mise en évidence, la caractérisation, la détection et l'étude de la diversité d'agents viraux. Leurs travaux s'intéressent autant au milieu agricole qu'à l'environnement non anthropisé des îles Kerguelen.

Les pathologies virales ont un impact économique et environnemental considérable. Leur étude est essentielle afin de comprendre et de modéliser leurs mécanismes et ainsi pouvoir anticiper leur apparition et en maîtriser les conséquences. Chez les plantes, de nombreuses pathologies émergentes ont été récemment identifiées et près de la moitié d'entre elles sont causées par des agents viraux (cf. Fig. 1.16: [Pathologies végétales émergentes]). L'ampleur de ce phénomène est préoccupante et souligne l'importance des efforts académiques et industriels dans l'étude et la maîtrise des paramètres de ces épidémies d'un point de vue global.

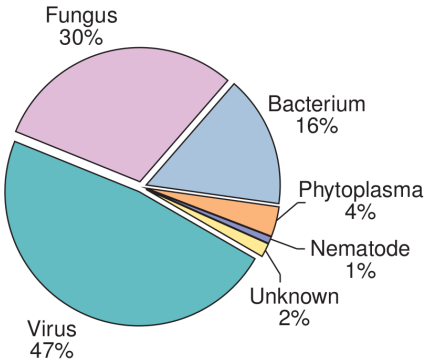


FIGURE 1.16 – Les virus causent près de la moitié (47%) des pathologies émergentes chez les plantes. Des proportions similaires sont retrouvées chez l'humain (44%) et chez les animaux sauvages (43%). *Source :* ANDERSON *et collab. [2004]*

D'un point de vue économique, ces épidémies représentent des pertes annuelles pouvant atteindre plusieurs milliards de dollars. Les monocultures extensives favo-

risent en particulier leur propagation, car la faible diversité des espèces utilisées ne favorisent pas l'apparition de résistances. Plus de 40% de la production mondiale est constituée de quatre aliments de base : le blé, le riz, le maïs et la pomme de terre. La diversité génétique des souches utilisées est progressivement diminuée par l'industrie de la biotechnologie agricole et amplifie davantage ce phénomène (ANDERSON et collab. [2004]).

Mais le problème ne se limite pas à ces quatre produits de base. De nombreuses productions secondaires, comme les agrumes, la banane, le café, le cacao, le tabac et la sylviculture, jouent un rôle important dans les pays en voie de développement en générant des revenus, de l'emploi et des échanges commerciaux. Ces productions sont également parfois durement touchées par des épidémies virales. Certaines d'entre elles jouent un rôle essentiel dans l'alimentation locale et sont d'autant plus critiques. C'est le cas du manioc, par exemple, qui nourrit une grande partie de l'Afrique sub-saharienne, soit plus de 200 millions de personnes. La production de manioc souffre grandement de l'émergence de la mosaïque du manioc (cf. Fig. 1.17: [Mosaïque du manioc]), une épidémie virale qui a fait chuter la production de 35.7% en Uganda entre 1989 et 1996, avec des pertes estimées à 60 millions de dollars par an entre 1992 et 1996. L'épidémie s'est rapidement répandue dans le continent et a nécessité une intervention internationale afin d'empêcher une famine généralisée (ANDERSON et collab. [2004]).

FIGURE 1.17 – Symptômes de la mosaïque du manioc africaine : (A) modérés (B) sévères (C) non infecté. *Source :* SCHOLTHOF *et collab. [2011]*

Il est nécessaire d'avoir une vision d'ensemble pour lutter contre ces épidémies. Il existe naturellement un fort biais d'étude en faveur des agents infectieux qui touchent l'économie et la santé de manière directe. En conséquence, seules 3186 espèces virales sont actuellement reconnues (ICTV 2014 Master Species List v3, 18 mars 2015), bien que de récents travaux de séquençage (cf. 1.1.3.1: [Techniques de séquençage]) suggèrent qu'il s'agit d'une sous-estimation dramatique du nombre total des espèces virales sur Terre (WREN et collab. [2006]).

Pourtant, la compréhension de l'émergence de ces épidémies nécessite un large champ de connaissances qui dépasse la simple description des agents pathogènes. Il est au contraire essentiel d'élargir les connaissances à de nombreuses espèces virales

et non virales afin de construire des modèles paramétrés pour un grand nombre de facteurs biologiques, écologiques et environnementaux.

En effet, si l'appauvrissement progressif de la diversité génétique des cultures aggrave l'impact des épidémies virales, il ne suffit pas à en expliquer l'ampleur ou l'émergence. La microbiologie s'est très fortement basée sur les postulats de Koch et les critères de Bradford-Hill depuis leur formulation aux XIXe et XXe siècles, respectivement. Ils définissent une relation de causalité entre pathogène et pathologie et ont permis de grandes avancées dans la compréhension des mécanismes de nombreuses maladies, ainsi que leur traitement. Mais la découverte de la complexité des intéractions entre microbes, comme ces virus dont la simple présence offre à leur hôte une résistance contre certaines infections ou peut, au contraire, le fragiliser selon l'espèce et le contexte, remet constamment en cause ces règles d'or (VAYSSIER-TAUSSAT et collab. [2014]).

Chez les plantes, nombre de ces pathologies émergentes trouvent leur origine dans la perturbation de l'équilibre entre les différentes espèces virales suite à des échanges commerciaux de plants et semences destinés à l'agriculture. Ces derniers contiennent naturellement une population virale spécifique à leur milieu d'origine, dont la co-évolution avec l'hôte et l'environnement assurent un équilibre contextuel. Leur introduction dans un nouveau milieu implique à la fois l'appauvrissement de leur propre microbiome, mais également la rencontre avec celui des autres espèces environnantes. Ainsi, ils sont plus susceptibles de succomber à des pathologies endémiques contre lesquelles ils ne possèdent pas de protection, et les espèces virales qu'ils véhiculent représentent également un danger pour la flore locale (ANDERSON et collab. [2004]).

Les virus font partie intégrante de la vie de l'hôte et son environnement. Au-delà du pouvoir pathogène de certains virus, nous avons montré la profondeur des relations qu'ils entretiennent avec le vivant. Dans ce contexte, il est difficile d'envisager de comprendre le fonctionnement d'un individu sans avoir accès aux virus qu'il véhicule et qui influencent son devenir, tant au niveau de son organisme que dans les échanges avec l'ensemble de son écosystème.

Cette thèse s'inscrit dans cet effort de recherche d'une vision globale, d'un inventaire des populations virales propres à chaque individu, au-delà de la maladie, afin de combler progressivement les lacunes dont souffrent les modèles hôte-pathogène classiques.

1.1.3 Métagénomique et nouvelles perspectives

1.1.3.1 Techniques de séquençage

Le séquençage d'une molécule d'ADN est le procédé visant à déterminer l'ordre des nucléosides qui la composent. Ce procédé permet de passer de la molécule à une représentation littérale de sa séquence sous forme de chaîne de caractères utilisant l'alphabet {A,C,G,T} représentant les quatre nucléosides de l'ADN (cf. 1.1.1.1: [Structure de la cellule vivante]). Les premières méthodes de séquençage ont vu le jour dans les années 1970. L'une d'entre elles, développée par l'équipe de Frederick Sanger au Royaume-Uni, est encore utilisée aujourd'hui.

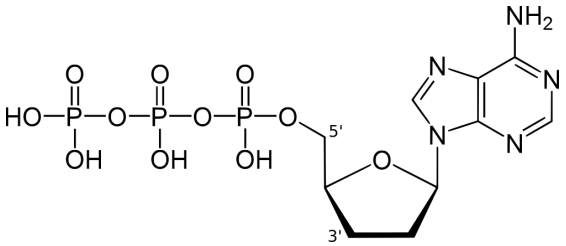


FIGURE 1.18 – Didéoxyadénosine triphosphate. L'absence d'oxygène sur le carbone 3' ne permet pas l'élongation du brin d'ADN sur lequel le didéoxynucléotide est fixé. *Source : Wikimedia Commons*

La méthode Sanger repose sur l'utilisation de nucléotides "bloquantes" au cours de la synthèse d'un brin complémentaire au brin à séquencer (cf. Fig. 1.6: [Réplication]). Ces nucléotides particuliers sont des *didéoxynucléotides*, qui ont pour particularité de ne pas posséder d'oxygène sur leur carbone 3' (cf. Fig. 1.18: [Didéoxyadénosine triphosphate]). Lors de la synthèse de l'ADN, c'est sur cet oxygène que se fait l'élongation du brin (cf. Fig. 1.2: [Structure de l'ADN]) et son absence implique l'arrêt complet de la synthèse.

Le principe est le suivant (cf. Fig. 1.19: [Méthode Sanger]) : Dans quatre tubes séparés sont mises de très nombreuses copies du brin à séquencer. On y ajoute également une amorce : un petit bout d'ADN complémentaire à l'extrémité du brin qui se fixe naturellement dessus et fournit un support pour le début de la synthèse. Chacun des tubes contient également un mélange des quatres déoxynucléotides nécessaires à la synthèse de l'ADN. Mais en plus de ces dernières, sont ajoutées un type de de didéoxynucléotide bloquant différent dans chaque tube.

La synthèse se déroulera normalement, mais de temps en temps, un nucléotide bloquant sera utilisé à la place d'un nucléotide normal, et une partie des brins s'arrêtera de grandir. Sachant que chaque tube bloquera à un type de nucléotide bien précis, il suffit ensuite de comparer la taille des brins afin d'en déduire la séquence. Ils sont séparés par taille en migrant à travers un gel, qui ralentira leur progression proportionnellement à leur taille. La séquence du brin synthétisé correspond au tube d'origine des brins lus sur le gel, du plus petit au plus grand.

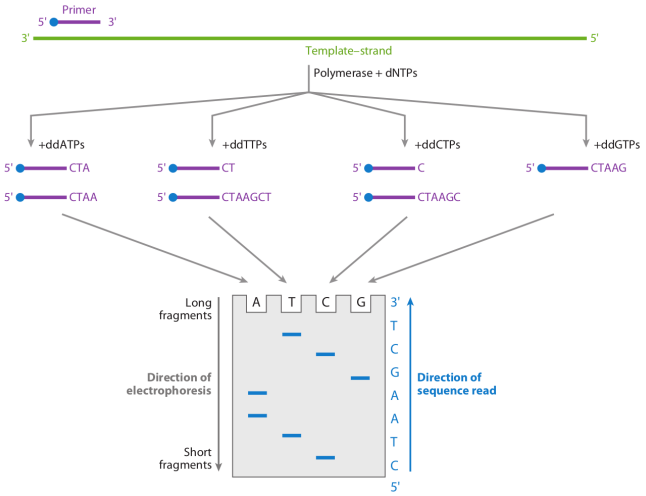


FIGURE 1.19 – Méthode Sanger : La position des différents brins synthétisés sur le gel d'électrophorèse indiquent la séquence "CTAAGCT". *Source :* MARDIS *[2013]*

Cette méthode est longue et coûteuse, mais permet la lecture de longues séquences avec une excellente précision. Elle a depuis vu de nombreuses améliorations, et sa fiabilité en fait encore aujourd'hui une méthode de choix lors de projets minutieux.

De nombreuses nouvelles technologies ont depuis vu le jour. La première amélioration majeure fut apportée par le laboratoire de Leroy Hood à l'Institut Technologique de Californie. Ils ont mis au point un instrument permettant l'automatisation de la lecture des gels d'électrophorèse, grâce à l'utilisation d'amorces fluorescentes et d'un laser pour la révélation. Cet instrument, commercialisé en 1986 par Applied Biosystems Inc., a permis la suppression de nombreuses étapes manuelles, réduisant ainsi le temps de manipulation et supprimant les erreurs humaines lors de la lecture des résultats.

Une décennie plus tard, en 1999, l'apparition de deux nouveaux instruments sur le marché a permis la disparition de l'étape de migration sur gel, coûteuse en temps et en travail manuel, grâce à l'introduction d'un système de capillaires permettant une séparation rapide des brins synthétisés.

C'est en 2005 que de nouvelles approches révolutionnaires ont bouleversé le rendement du séquençage. Ces améliorations marquent le début de ce qui est considéré comme le séquençage de nouvelle génération (*Next-Generation Sequencing*, ou *NGS*). Il n'est à présent plus nécessaire de faire de nombreuses copies des brins d'ADN à séquencer en amont du protocole : les machines automatisent à présent cette étape. L'élongation des brins synthétisés ne requiert également plus d'intervention manuelle et est entièrement gérée par la machine, qui apporte elle-même

les réactifs sur des supports fixes sur lesquels l'ADN est attaché. Les nucléotides ne bloquent plus la synthèse, mais sont modifiés par l'ajout de marqueurs permettant leur identification au fur et à mesure qu'ils sont polymérisés. Ainsi, ces méthodes ont ouvert la voie à une paralellisation massive du séquençage, permettant la lecture simultanée de centaines de millions de séquences (MARDIS [2013]).

Ces récentes avancées technologiques s'accompagnent d'une baisse spectaculaire du prix de revient de chaque séquence produite. En contrepartie, les limitations techniques impliquent une faible taille des séquences produites et une qualité variable de la lecture. Neanmoins, elles ont permis une explosion du volume de données obtenues : il est à présent possible de séquencer des échantillons complexes contenant l'ADN de nombreuses espèces.

Traditionnellement, le rendement des méthodes de séquençage ne permettaient pas d'étudier plus d'un génome à la fois. Les chromosomes étaient isolés, puis fragmentés de manière aléatoire afin d'obtenir des molécules de longueur compatible avec les protocoles de séquençage. Il s'agissait du seul moyen d'obtenir les séquences de suffisamment de fragments pour couvrir la majeure partie d'un génome complet. Cela implique de mettre en culture les organismes à séquencer et de pouvoir en isoler un nombre suffisant.

Or, il a été montré dans les années 1980 que les organismes cultivables en laboratoires ne représentent qu'une faible proportion du monde vivant, en particulier chez les microorganismes. Il est estimé, par exemple, que seules 0,1 à 1% des bactéries vivant dans la terre sont cultivables en milieu standard. Cette découverte n'est pas arrivée seule : simultanément, il a été mis en évidence que de nombreux phénomènes, dont certaines conditions pathologiques chez les organismes supérieurs traditionnellement imputés au stress, étaient très fortement liés aux microorganismes avec lesquels ils sont en contact (HANDELSMAN [2004]).

C'est dans ce contexte que la communauté scientifique a commencé à s'intéresser au séquençage d'échantillons complexes composés d'espèces multiples. Les nouvelles technologies de séquençages, bien plus rapides et abordables que les premières méthodes, ont permis de franchir ce pas dans les années 2000.

1.1.3.2 Métagénomique

La *métagénomique* désigne le procédé visant à étudier le contenu génétique de ces échantillons complexes. On ne s'intéresse pas ici à un organisme unique, mais à tous les organismes présents dans un échantillon donné prélevé dans la nature. Il peut s'agir d'eau de mer, de terre, de glace, d'un échantillon de flore intestinale, un morceau de fruit, n'importe quel sous-ensemble d'un milieu donné. Le contenu de l'échantillon est ensuite filtré, afin d'éliminer les particules indésirables et sélectionner les éléments d'intérêt. Leur ADN est extrait et préparé pour le séquençage (cf. Fig.

1.20: [Métagénomique]). L'ensemble des séquences ainsi produites constitue un jeu de données métagénomique.

Ces données représentent une fenêtre à travers laquelle il est possible d'observer des fragments de l'ensemble des génomes constituant l'échantillon, y compris les organismes non cultivables. Il s'agit d'une opportunité sans précédent pour l'étude des intéractions inter-espèces au niveau moléculaire, y compris dans le domaine de la virologie. En effet, il est possible de récupérer les particules virales lors de la filtration afin d'en obtenir le contnu génétique.

À la sortie du séquenceur, ces fragments sont mélangés et ne portent aucune indication des espèces desquelles ils proviennent. Il n'est néanmoins pas toujours nécessaire d'avoir accès à cette information, selon l'approche choisie pour analyser les données.

Certains projets utilisant ce procédé ne s'intéressent pas aux espèces en présence à proprement parler, mais à l'activité du milieu dans son ensemble en vue de la découverte de nouvelles biomolécules, ou de nouvelles variantes de gènes connus. L'échantillon est considéré comme un ensemble cohérent capable d'opérer des réactions chimiques qui lui sont propres. L'étude du rôle des génomes en présence dans l'ensemble de ces réactions ne nécessite pas l'inventaire des espèces qui le composent (SIMON et DANIEL [2011]).

D'autres approches, en revanche, s'intéressent à la biodiversité du milieu dont l'échantillon est issu. Cela implique l'identification de l'origine des différents fragments séquencés, à partir de l'ordre des caractères qui composent leur séquence. Cette thèse s'intéresse à ce type d'analyse pour l'identification des espèces virales (cf. 1.1.2.2: [Importance de la virologie]), en explorant de nouvelles approches pour effectuer cette étape de recensement.

1.1.4 Classification des espèces et biodiversité

1.1.4.1 Distance évolutive et homologie

L'évolution est la transformation des espèces vivantes au cours des générations, dont le moteur principal est un ensemble de processus qui modifient et réorganisent l'information génétique. Ces changements permettent l'apparition de variants au sein des populations, ce qui augmente la probabilité que des individus survivent à des changements et contraintes environnementaux, appelés pression de sélection. Ces modifications, ou mutations, s'opèrent au niveau moléculaire et peuvent être causées par des facteurs extérieurs (e.g. rayonnements, radiations, molécules exogènes, radicaux), mais peuvent également être endogènes (e.g. crossing-overs, mutations spontanées, éléments mobiles).

Ces changements se produisent de manière régulière et ont des conséquences variables en fonction de leur localisation dans le génome. Ils sont conservés au sein

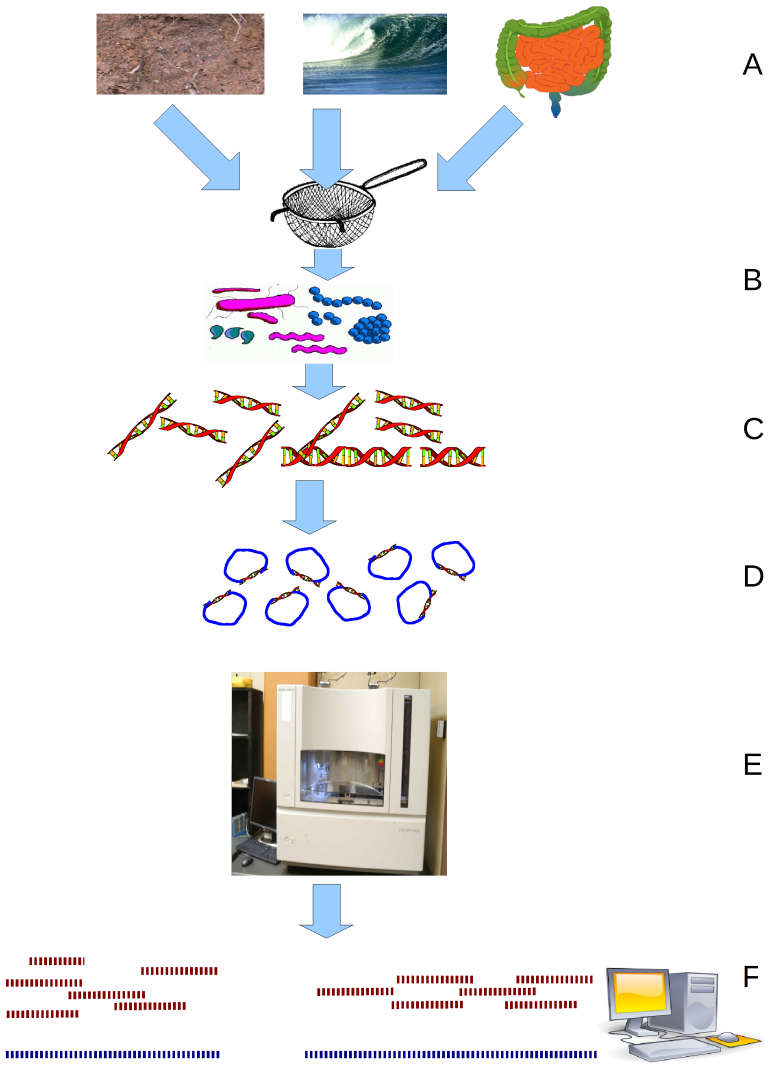


FIGURE 1.20 – Principe de la métagénomique : (A) Prélèvement d'un échantillon. (B) Filtration de l'échantillon. (C) Extraction et fragmentation de l'ADN toutes espèces confondues. (D) Préparation des fragments pour le séquençage. (E) Séquençage. (F) Assemblage et analyse *in silico* des séquences. *Source :* WOOLEY *et collab. [2010]*

des populations si les individus variants parviennent à se reproduire au cours de leur vie. Certaines parties du génome sont peu sensibles aux mutations car elles ne portent pas d'informations essentielles pour la survie cellulaire. En revanche, certaines zones sont critiques et ne tolèrent que très rarement des changements. C'est le cas, par exemple, des gènes codant pour les composantes de la machinerie de base de la cellule, comme les ARNr des ribosomes (cf. 1.1.1.2: [Information génétique et expression génique]).

Il est ainsi possible de trouver des indices sur l'histoire évolutive des espèces en comparant des séquences appartenant à plusieurs individus vivants. Cela est rendu possible grâce au séquençage (cf. 1.1.3.1: [Techniques de séquençage]) et a permis de faire la lumière sur de nombreux éléments essentiels de la *systématique* d'aujourd'hui.

La systématique est la science de l'inventaire du vivant. Elle se divise en deux grandes tâches : la description des êtres vivants et leur classification. Depuis la découverte des mécanismes de l'évolution, la classification suit un schéma respectant ses théories fondamentales : les espèces sont classifiées de manière arborescente, permettant de remonter les liens de parenté qui les lient jusqu'à *LUCA* (Last Universal Common Ancestor), l'organisme le plus récent dont est issue la totalité des espèces actuellement observables. L'ensemble de ces liens de parenté est la *phylogénie*.

Historiquement, la systématique se basait sur des critères observables, tels que les propriétés morphologiques (e.g. squelette, composants cellulaires) des différents organismes afin de les décrire et les comparer entre eux. L'apparition du séquençage a permis d'augmenter considérablement la quantité d'informations utilisables dans ce domaine. Les arbres ainsi construits avec les informations contenues dans des données de séquençage sont des *arbres phylogénétiques* (cf. Fig. 1.21: [Classification phylogénétique du vivant]).

Ces arbres représentent les relations de parenté inter-espèces. Chaque nœud interne correspond à l'ancêtre commun partagé par toutes les feuilles du sous-arbre dont il est la racine. Les feuilles peuvent être des espèces, des sous espèces ou des souches, selon les arbres, et les sous-arbres sont appelés *taxons*. Chaque taxon représente l'ensemble des espèces issues de l'ancêtre qui l'enracine.

Afin de construire un arbre phylogénétique, il faut comparer des séquences. De la même manière que lorsqu'on compare la morphologie de plusieurs organismes pour en repérer les similarités et différences, et en déduire les liens de parenté, il est possible de faire l'inventaire des événements mutationnels entre plusieurs séquences afin d'en déduire l'histoire évolutive.

Cependant, de la même manière qu'il n'est pas possible de comparer une nageoire de dauphin à une feuille de poirier, il est nécessaire que les séquences compa-

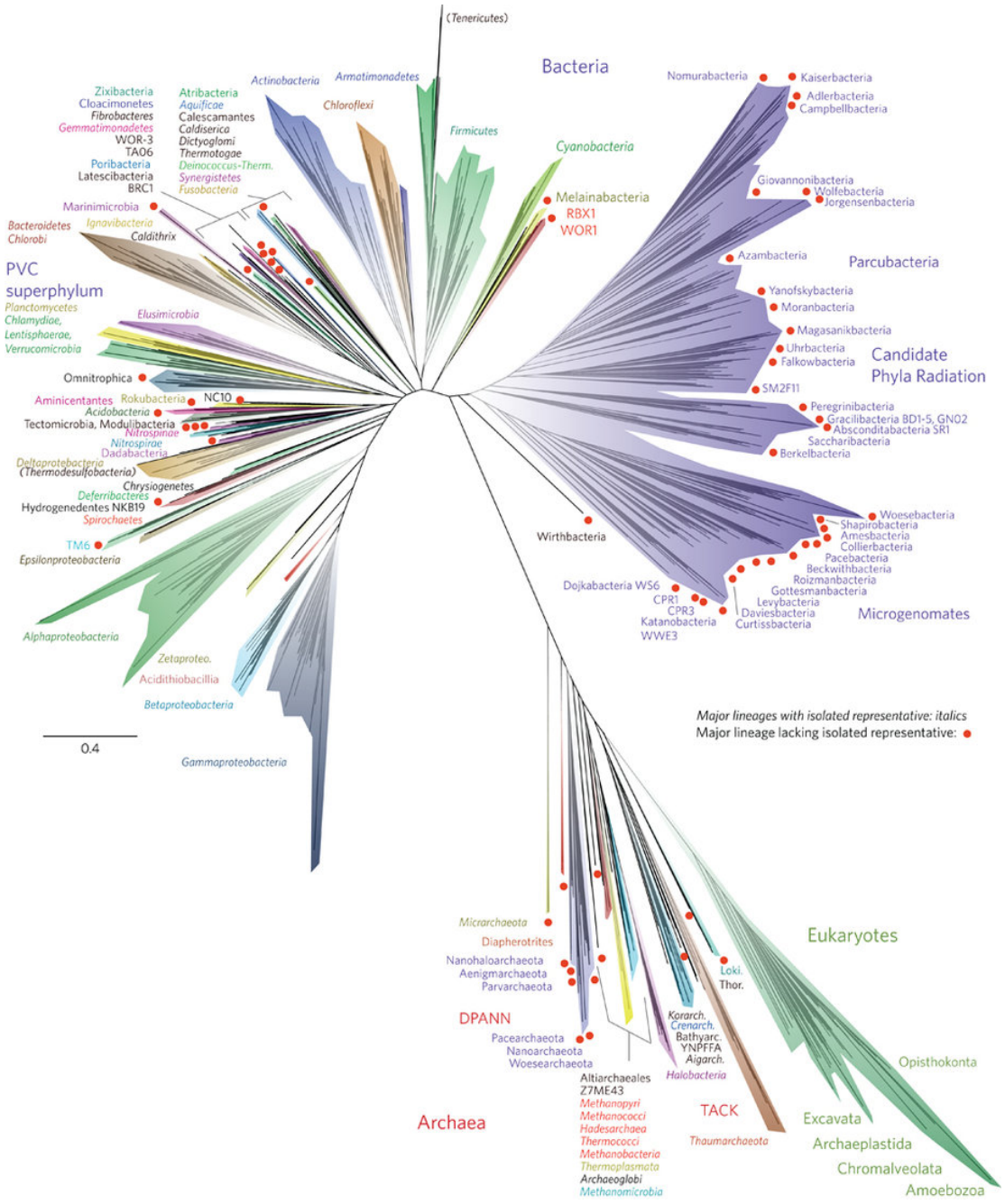


FIGURE 1.21 – Classification phylogénétique de l'ensemble du vivant connu. On peut distinguer quatre grands règnes : les eukaryotes, les archées, les bactéries, ainsi qu'un nouveau groupe prokaryote assimilé aux bactéries (en mauve à droite). *Source :* HUG *et collab. [2016].*

rées soient issues d'une origine commune identifiable. On parle alors de séquences *homologues* : deux séquences sont homologues si et seulement s'il est possible d'affirmer qu'elles sont issues d'une série d' événements de réplication (cf. Fig. 1.6: [Réplication]) dont le premier s'est opéré sur la même molécule d'ADN (ou d'ARN chez certains virus).

Il est possible de repérer une homologie entre plusieurs séquences si l'on observe une conservation des nucléotides en les alignant les uns par rapport aux autres (cf. Fig. 1.22: [Alignement de séquences]). Il est ainsi possible de faire l'inventaire des événements de *substitution* (changements de nucléotides), d'*insertion* (apparition de nucléotides supplémentaires) et de *délétion* (disparition de nucléotides) qui se sont produits depuis leur plus proche ancêtre commun.

```
AATATAGAACCAGGGA--TGAGGACTAGGTTAGTAG
||||||| ||| ||||  ||||||| ||| ||| ||
AATATAGGACC-GGGATATGAGGACCAGGATAGGAG
```

FIGURE 1.22 – Alignement de deux séquences homologues. En rouge : les nucléotides ayant subi des mutations.

Les séquences à comparer sont à choisir en fonction de l'âge de la divergence. Si l'on compare des espèces proches (e.g. pêcher et prunier), il est possible d'utiliser de nombreux gènes, car leurs génomes respectifs partagent de nombreuses caractéristiques génétiques et n'ont divergé que très récemment. En revanche, si l'on veut comparer des espèces très éloignées (e.g. une plante et un champignon), dont le plus proche ancêtre commun est très ancien, et dont les gènes les moins sensibles aux mutations ne sont plus comparables, les homologies sont rares et le choix est beaucoup plus restreint. Il se limite alors aux gènes dits *ubiquitaires*, les gènes présents dans toutes les espèces vivantes et qui ne varient que très peu au cours du temps car les mutations qui ne mettent pas la survie cellulaire en péril sont très rares.

Parmi les gènes ubiquitaires, on peut retrouver des gènes codant pour la structure cellulaire (e.g. l'actine, impliquée dans la structure et la mobilité cellulaires), la gestion et l'expression du matériel génétique (e.g. certaines polymérases chargées de l'élongation des acides nucléiques), la structure des chromosomes (e.g. les histones, acteurs majeurs du repliement de l'ADN), ainsi que toutes les autres fonctions de base de la machinerie cellulaire la plus élémentaire. Mais le gène ubiquitaire les plus couramment utilisé en phylogénétique en tant que gène "marqueur" est le gène codant pour la petite sous-unité des ribosomes que nous avons vus précédemment : l'ARNr 16S chez les prokaryotes et 18S chez les eukaryotes (cf. 1.1.1.2: [Information génétique et expression génique]).

Cela n'exclut pas l'utilisation de gènes plus spécifiques lors de l'utilisation d'un domaine phylogénique particulier, ce qui permet dans ce cas d'obtenir une résolu-

tion plus adaptée. Un pipeline d'analyse phylogénique (AMPHORA, WU et EISEN [2008]) propose ainsi l'utilisation de 31 gènes marqueurs spécialement sélectionnés pour le règne des bactéries, dans lequel ils sont présents dans tous les génomes, en général en une seule copie, et rarement sujets aux transferts horizontaux (i.e. échanges entre cellules n'impliquant pas la division cellulaire).

1.1.4.2 Notion d'espèce

La notion d'espèce est une des notions les plus controversées. Il s'agit du taxon de base de la systématique, le rang le plus bas qui représente les feuille de l'arbre phylogénétique, bien que dans certains cas des taxons supplémentaires peuvent être rajoutés selon les besoins (e.g. souche). Pourtant, définir et séparer les espèces n'est pas toujours une tâche aisée, car il est pratiquement impossible de le faire selon les même critères d'un groupe phylétique à l'autre.

La définition la plus répandue a été énoncée par Ernst Mayr en 1942 de la manière suivante : *"Les espèces sont des groupes de populations naturelles, effectivement ou potentiellement interfécondes, qui sont génétiquement isolées d'autres groupes similaires"* (MAYR [1942]). Le critère utilisé ici est l'interfécondité des individus dans des conditions naturelles. Cela implique la capacité de reproduction sexuée ainsi que l'existence de barrières naturelles interdisant le contact entre des populations potentiellement interfécondes. Cette définition est en effet efficace dans le règne animal, mais devient difficile à généraliser chez les végétaux par exemple, chez qui la reproduction sexuée n'est pas systématique, voire complètement inapplicable chez de nombreux groupes phylétiques, comme les bactéries, ou encore invérifiable chez les espèces fossiles.

De nombreuses définitions ont vu le jour au cours de l'histoire (cf. Annexe A.3: [Summary of 26 species concepts]) afin de répondre à diverses contraintes théoriques, sans pour autant offrir de consensus satisfaisant l'ensemble du vivant. Leur grand nombre témoigne de la difficulté de la tâche et de l'ampleur de la controverse qui entoure la question, mais également de la grande variété des espèces vivantes. Leur application conserve néanmoins une efficacité locale et permet malgré tout un découpage utile de la biodiversité.

Les virus, en revanche, ne font pas partie du vivant. [1]Leur grande variabilité rend l'étude de leur origine très difficile, mais de récentes études sur la conservation de certains domaines protéiques suggèrerait qu'il serait issus de multiples événements anciens qui se seraient produits au sein de cellules ancestrales possédant des génomes à ARN segmentés, qui auraient coexisté avec les ancêtres des cellules modernes (NASIR et CAETANO-ANOLLÉS). Ils ne peuvent donc pas trouver leur place parmi la phylogénie sus-citée (cf. Fig. 1.21: [Classification phylogénétique du vivant]), car ils ne partagent pas de lien de parenté avec LUCA, sa racine actuelle. Une extension

de cet arbre est envisagée afin de les inclure mais n'est pas encore d'actualité. Il est néanmoins possible de définir des espèces virales et d'identifier, dans de nombreux cas, une histoire évolutive commune jusqu'au niveau de leur famille (SIMMONDS [2015]). En revanche, ils ne possèdent pas de gène ubiquitaire et sont par conséquent impossibles à classifier selon des critères phylogénétiques globaux.

Ils font donc l'objet de classifications spécialisées, dont les niveaux hiérarchiques n'adhèrent à des propriétés évolutionnistes que parmi les plus bas d'entre eux. Les niveaux hiérarchiques les plus élevés sont souvent construits à partir de données structurales globales (nature des acides nucléiques, structure des virions, composition et géométrie des capsides), mais aussi d'habitudes et de traditions propres au domaine. Néanmoins, des efforts de modernisation sont en cours afin de synchroniser la taxonomie virale et les avancées de la recherche (GIBBS [2013], PETERSON [2014]).

La classification virale considérée comme la classification officielle est éditée par le *Comité International de Taxonomie des Virus* (ICTV), et fait autorité au sein de la communauté des virologues. L'ICTV édite un rapport chaque année depuis 1971 et met à disposition leur phylogénie virale en ligne (`http://ictvonline.org/virusTaxonomy.asp`). Le niveau hiérarchique le plus élevé proposé est l'Ordre, mais de nombreuses bases de données publiques (e.g. le *Centre Américain pour les Informations Biotechnologiques* ou NCBI : `http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi`) ajoutent un niveau hiérarchique supplémentaire en utilisant les grandes classes virales définies par Baltimore (BALTIMORE [1971], cf. Fig. 1.14: [Classification de Baltimore]).

Ces différentes classifications sont essentielles pour l'identification des espèces. Elles constituent un outil essentiel en métagénomique, non seulement pour la reconnaissance des espèces qui ont déjà été décrites et classifiées, mais elle fournissent également une base utile pour la découverte de nouvelles espèces, car elles offrent un outil de comparaison permettant de situer un élément inconnu par rapport à une organisation globale du vivant.

1. Ce sujet fait également l'objet d'une controverse et dépend fortement de la définition que l'on donne au *vivant*, qui varie également selon les sources. De même que la notion d'espèce, il s'agit de poser des limites théoriques permettant de segmenter et décrire le monde, afin de le rendre intelligible pour l'esprit humain, mais la nature ne suit pas de règles si strictes que de telles limites puissent être à toute épreuve.

1.2 L'outil informatique au service de la classification de séquences nucléiques

1.2.1 Apprentissage automatique

1.2.1.1 Principe et applications générales

Nous vivons dans un monde de données numériques. La production de nouvelles données est aujourd'hui sans précédent, en biologie (cf. Fig. 1.23: [Croissance de GenBank]) comme dans tous les domaines de la recherche et de l'industrie. Cette explosion crée inévitablement de nouveaux défis en terme d'utilisation et d'analyse des donnée produites, mais a également permis l'essor des méthodes d'*apprentissage automatique*, ou *machine learning*.

Il s'agit d'un champ d'étude de l'intelligence artificielle dans lequel de grandes quantités d'informations sont utilisées afin d'entraîner et faire évoluer un modèle par un processus systématique, afin de remplir une tâche. L'intérêt de ce genre de méthode est double : d'une part, les quantités de données à traiter rendent certaines tâches difficiles, voir impossibles à effectuer par des moyens algorithmiques classiques ; d'autre part, ces méthodes fonctionnent d'autant mieux que la quantité d'informations disponible est importante. Par conséquent, les méthodes d'apprentissage automatique sont un moyen idéal d'absorber et tirer profit de l'afflux grandissant de données numériques.

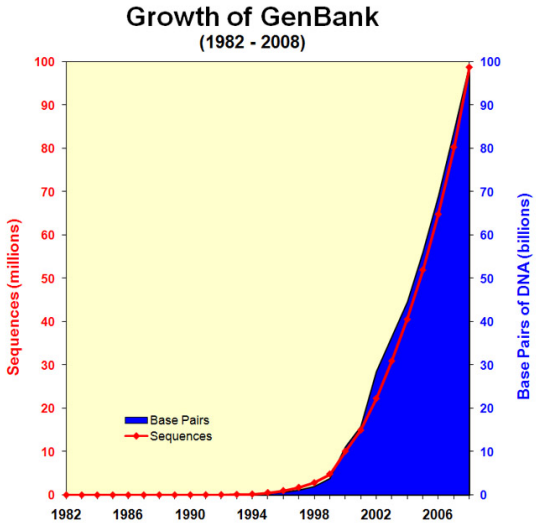


FIGURE 1.23 – Croissance de la banque de données génomiques GenBank jusqu'en 2008. Des statistiques fréquemment mises à jour sont disponible via `http://www.ncbi.nlm.nih.gov/genbank/statistics` *Source : http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008*.

Les tâches que sont capables de remplir ces méthodes sont simples, mais permettent de répondre à des questions complexes. Il s'agit généralement de classification, d'estimation de valeurs, de reconnaissance de formes, etc. Cette simplicité leur confère une grande versatilité, dont est témoin leur grand nombre de domaines d'application (moteurs de recherche, reconnaissance vocale et faciale, aide au diagnostic, détection des fraudes, etc.). Il n'est donc pas surprenant, compte tenu notamment de l'augmentation exponentielle des données de séquençage, et de la complexité des informations qu'elles contiennent, que l'apprentissage automatique ait trouvé de nombreuses applications dans le domaine de la biologie.

1.2.1.2 Pertinence en bioinformatique

Les données biologiques actuelles sont multiples. Les séquences d'acides nucléiques ne s'ont qu'un exemple parmi tant d'autres. Les protéines (cf. Fig. 1.9: [Protéine]) font également l'objet de techniques de séquençage particulières afin de déterminer leur séquence en acides aminés, et fournissent des données complémentaires essentielles en biologie moléculaire. Les données d'imagerie sont également nombreuses et constituent des sources d'informations complexes et nombreuses autant au niveau moléculaire (microarrays) que cellulaire (microscopie). L'évolution des technologies dans ces domaines, à l'instar du séquençage d'acides nucléiques, favorise grandement le rendement de production de données et offre de nombreuses applications possibles à l'apprentissage automatique. On le trouve ainsi utilisé pour répondre à des questions très variées, allant de la prédiction de structures tridimentionnelles des protéines à la modélisation de réseaux biologiques, en passant par la reconstruction d'arbres phylogénétiques (LARRAÑAGA et collab. [2006]).

Les influences interdisciplinaires ne sont pas unidirectionnelles. L'apprentissage automatique, comme toutes les autres disciplines informatiques dont tire partie la biologie, a pavé de nombreuses voies dans l'analyse des données biologiques. Elle offre de nouvelles possibilités et influence la manière de traiter les données. À l'inverse, les systèmes biologiques servent d'inspiration constante dans le domaine de l'informatique et ont permis la naissance de méthodes d'apprentissages automatique directement inspirées du vivant (e.g. réseaux de neurones, algorithmes génétiques). Il existe donc un lien très fort entre les deux disciplines, et il n'est donc pas surprenant de voir l'importance qu'a acquis l'apprentissage automatique dans la sphère biomédicale aujourd'hui (CHEN [2005], JENSEN et BATEMAN [2011]).

Les séquences de biopolymères sont particulièrement simples à traiter *in silico*. Qu'il s'agisse de protéines ou d'acides nucléiques, ils sont représentés par de simples chaînes de caractères utilisant un alphabet restreint. Les méthodes d'apprentissage automatiques sont par conséquent utilisables de manière assez directe, tout en tirant

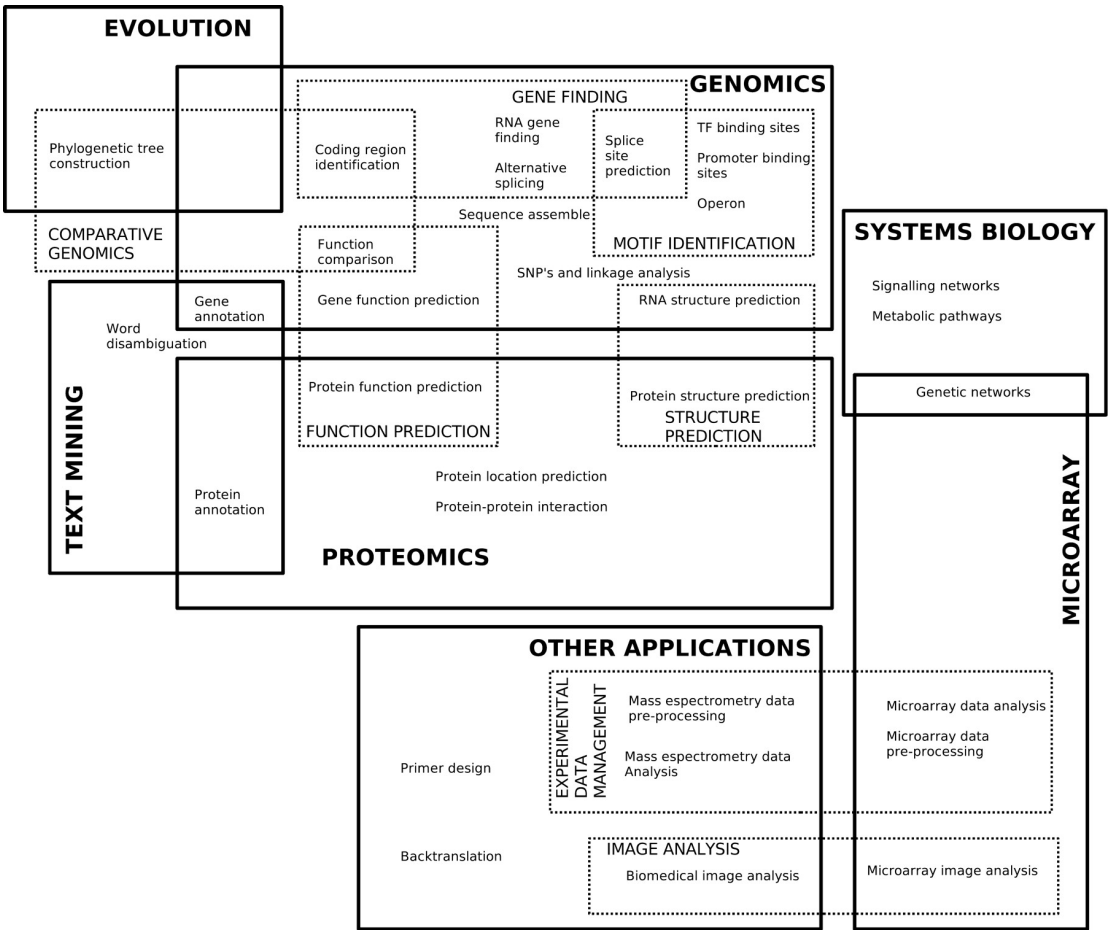


FIGURE 1.24 – Applications biologiques de l'apprentissage automatique *Source :* LARRAÑAGA *et collab. [2006]*.

parti des travaux effectués dans d'autres domaines utilisant des chaînes de caractères comme données d'entrée. Ces avantages leur confèrent une popularité indéniable dans ce domaine, en particulier en métagénomique où de grandes quantités de texte à traiter sont impliquées.

1.2.2 Classification supervisée de chaînes de caractères

1.2.2.1 Principe et spécificités

l'*apprentissage supervisé* est une classe d'algorithmes d'apprentissage automatique. Ces algorithmes répondent à un besoin particulier : identifier les classes auxquelles appartiennent des objets à partir d'attributs descriptifs, étant donnés un ensemble prédéfini de classes à utiliser, ainsi qu'un ensemble d'objets auxquels sont déjà attribuées les classes sus-citées.

Il s'agit de la catégorie d'algorithmes d'apprentissage automatique qui correspond à la question posée dans cette thèse (cf. 1.1.3.2: [Métagénomique]) : l'*assignation taxonomique* de séquences métagénomiques. Il s'agit d'assigner un taxon à des séquences inconnues grâce à l'information contenue dans leur séquence étant donné un ensemble de séquences préalablement étudiées dont on connait l'origine. Nous avons ainsi :

— S un ensemble de séquences génomiques sur lequel va s'effectuer l'apprentissage et dont on connaît le taxon d'origine

— D un ensemble de signatures de séquences génomiques permettant leur identification, construites à partir de l'information contenue dans les séquences elles-mêmes

— $\{T_1, \ldots, T_n\}$ l'ensemble des taxons dans lesquels on souhaite classer les séquences

— $X : S \rightarrow D$ la fonction qui construit une signature à partir de l'information contenue dans chaque séquence de S

— $Y : S \rightarrow \{T_1, \ldots, T_n\}$ la fonction qui associe à chaque séquence de S son taxon d'origine

L'apprentissage consiste donc à déterminer la fonction de classification $C : D \rightarrow \{T_1, \ldots, T_n\}$ associant un taxon à chaque signature, de manière à ce que C(X) approche au mieux Y. Cette fonction C, une fois acquise, permet ainsi de prédire l'origine d'une séquence inconnue.

1.2.2.2 Évaluation des résultats

Lors de l'étude d'échantillons métagénomiques, l'origine des séquences est par nature inconnue. Il n'y a donc aucun moyen de vérifier l'exactitude des prédictions a posteriori. Il est donc nécessaire, lors de la construction du modèle, de vérifier son

efficacité sur des données d'origine connue. On utilise pour cela des techniques de validation croisée sur les données destinées à l'apprentissage. Le principe commun à ces techniques est d'isoler une partie de ces données, afin que le modèle n'en ait pas connaissance au moment de l'apprentissage, puis de les utiliser comme données de test. Il est ainsi possible de comparer la classe réelle d'origine et la classe prédite, et de quantifier les erreurs commises.

Les trois grandes méthodes de validation croisées sont les suivantes : La méthode "holdout", la plus simple, consiste à simplement séparer les données en deux sous ensembles comme expliqué précédemment. On conserve généralement plus de 60% des données pour l'apprentissage et on teste sur le reste. La seconde, la méthode "k-fold", consiste à séparer les données entre k échantillons de tailles égales, et d'effectuer l'apprentissage et la prédiction k fois en utilisant tour à tour chaque échantillon comme ensemble de test. La dernière, la méthode "leave-one-out", est un cas particulier de la précédente, où $k = card(\mathrm{S})$.

La méthode "holdout" possède un défaut majeur : si l'échantillon choisi pour l'apprentissage n'est pas représentatif, l'évaluation des résultats ne pourra pas rendre compte de l'efficacité réelle du modèle. Il est donc préférable, dans un souci d'efficacité, d'opter pour l'une des deux autres. En revanche, si les données sont nombreuses et complexes, et par conséquent l'apprentissage long à effectuer, la méthode "leave-one-out" peut être irréalisable dans des temps raisonnables. La méthode "k-fold" constitue donc un compromis entre temps de calcul et efficacité de l'évaluation.

Afin de quantifier l'efficacité d'un classifieur, plusieurs indicateurs peuvent être utilisés. Dans le cas d'un classifieur multi-classes, comme le cas présent avec les différents taxons utilisés, toutes ne sont pas applicables. Il est néanmoins possible de calculer simplement le taux d'erreur de la même manière que pour un classifieur binaire :

$$\mathrm{TE(C)} = \frac{\sum_{i=1}^{card(\mathrm{S})} (\mathrm{C(X}(s_i) \neq \mathrm{Y}(s_i))}{card(\mathrm{S})} \tag{1.1}$$

D'autres indicateurs sont disponibles. La première étape consiste à construire une matrice de confusion (cf. Tab. 1.1: [Matrice de confusion]) afin d'organiser les résultats de façon à comparer les classes d'origines (classes réelles) et les classes prédites pour chaque séquence de test.

L'exactitude des résultats représente la proportion de prédictions correctes et peut être calculée à partir d'une matrice de la manière suivante :

$$\mathrm{E}xactitude = \frac{\sum_{i=1}^{n} t_i^i}{\sum_{i,j=1}^{n} t_i^j} \tag{1.2}$$

TABLEAU 1.1 – Matrice de confusion. Les colonnes représentent les taxons d'origine $\{T_1, \ldots, T_n\}$. Les lignes sont les taxons prédits pour chaque séquence. Chaque valeur t_j^i de la matrice correspond au nombre de séquence d'origine T_i dont le taxon prédit est T_j. Les prédictions correctes sont cettes pour lesquelles $i = j$.

		Réel					
		T_1	T_2	…	T_i	…	T_n
Prédit	T_1	t_1^1	t_1^2	…	t_1^i	…	t_1^n
	T_2	t_2^1		…		…	
	⋮	⋮	⋮		⋮		⋮
	T_i	t_i^1		…	t_i^i	…	
	⋮	⋮	⋮		⋮		⋮
	T_n	t_n^1		…		…	

Il est également possible de calculer la sensibilité, ou *rappel* du modèle. Il permet d'avoir une mesure de l'efficacité du classifieur pour retrouver les séquences appartenant à un taxon donné :

$$\mathrm{R}appel(\mathrm{T}_i) = \frac{t_i^i}{\sum_{j=1}^{n} t_j^i} \tag{1.3}$$

Le rappel sur l'ensemble des taxons peut être évalué en effectuant la moyenne du rappel de tous les taxons :

$$\mathrm{R}appel = \frac{\sum_{i=1}^{n} \mathrm{R}appel(\mathrm{T}_i)}{n} \tag{1.4}$$

La *précision* est l'indicateur offrant une mesure de l'efficacité du classifieur pour ne pas attribuer un taxon donné à des séquences ne lui appartenant pas. Il permet d'avoir une mesure du bruit des résultats et n'est pas sans rappeler la spécificité des classifieurs binaires. Elle est calculée de la manière suivante :

$$\mathrm{P}récision(\mathrm{T}_i) = \frac{t_i^i}{\sum_{j=1}^{n} t_i^j} \tag{1.5}$$

La précision peut également être évaluée sur tous les taxons par une moyenne :

$$\mathrm{P}récision = \frac{\sum_{i=1}^{n} \mathrm{P}récision(\mathrm{T}_i)}{n} \tag{1.6}$$

Le Kappa (κ) de Cohen est une mesure statistique permettant de mesurer la concordance entre classes réelles et classes prédites par rapport à ce que pourraient

être les résultats s'ils étaient obtenus de manière alatoire. Il permet ainsi d'avoir une idée de la pertinence de la classification, et est calculé de la manière suivante :

$$\kappa = \frac{\mathrm{E}xactitude - \mathrm{P}_e}{1 - \mathrm{P}_e} \tag{1.7}$$

avec P_e l'estimation théorique de l'exactitude observée dans le cas où les résultats seraient complètement dus au hasard :

$$\mathrm{P}_e = \sum_{k=1}^{n} \left(\frac{\sum_{i=1}^{n} t_k^i}{\sum_{i,j=1}^{n} t_j^i} \times \frac{\sum_{i=1}^{n} t_i^k}{\sum_{i,j=1}^{n} t_j^i} \right) \tag{1.8}$$

1.2.2.3 Exemples d'algorithmes

Méthode des k plus proches voisins : Cette méthode, aussi appelée *k-nearest neighbors* ou *kNN* consiste à déterminer, selon une métrique à définir, quels sont les k éléments les plus proches d'un élément à classifier parmi les données d'apprentissage. La classe prédite est choisie en fonction des classes des k éléments voisins. L'influence de chaque voisin peut être équivalente, mais elle peut également être pondérée, par leur éloignement par exemple.

Classification naïve bayésienne : Il s'agit d'une méthode de classification probabiliste basée sur le théorème de Bayes. Il suppose une forte indépendance des attributs descriptifs des objets à classifier. Contrairement au kNN, il n'utilise pas directement les données d'apprentissage pour effectuer des prédictions, mais nécessite la construction d'un modèle intermédiaire qui représente la structure des données. En pratique, la construction des modèles bayésiens naïfs utilise plus souvent le maximum de vraisemblance que de réelles probabilités bayésiennes.

AdaBoost : L'*adaptative boosting* est une méthode qui repose sur la sélection de plusieurs classifieurs faibles, pondérés en fonction de leur efficacité sur les données d'apprentissage, afin d'obtenir en les combinant une approximation d'une fonction complexe.

Machines à vecteurs de support : Aussi appelées *séparateurs à vaste marge* ou *SVM*, ces méthodes séparent l'espace des données de manière à ce que la marge entre la séparation et les données les plus proches soit maximale. La classification se fait ensuite au regard de ces séparations, et la classe attribuée à un nouvel objet est celle qui a permis la séparation du sous-espace dans lequel il se trouve. Dans le cas où la séparation ne peut pas être efectuée directement de manière linéaire, l'espace

est préalablement transformé par une fonction "noyau" (*kernel*, cf. Fig. 1.25: [Transformation de l'espace]) qu'il est possible de choisir en fonction de la structure des données.

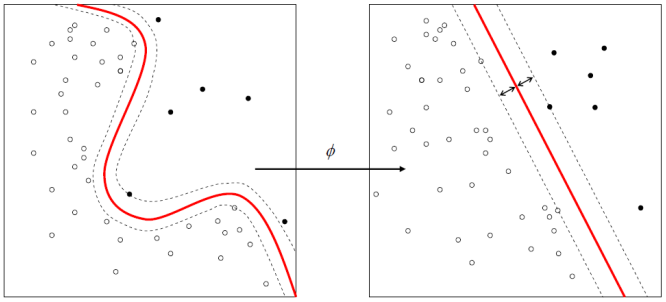


FIGURE 1.25 – Exemple de transformation de l'espace avec une fonction noyau (ϕ) dans le cas d'un classifieur binaire. La séparation des classes de manière linéaire n'est pas possible avant transformation (gauche) mais devient faisable une fois l'espace transformé (droite). *Source : Wikimedia Commons*

Forêt d'arbres décisionnels : Cette méthode repose sur l'utilisation d'arbres de décision. Les feuilles de ces arbres sont les différentes classes et chaque nœud décrit un test sur une variable d'apprentissage. Lors de l'apprentissage du modèle, une multitude d'arbres différents sont construits. Lors de l'étape de prédiction, la classe assignée à un nouvel objet est la classe la plus fréquemment proposée par l'ensemble des arbres. Les arbres individuels sont peu efficaces et souvent sujets au surapprentissage, mais la forêt tend à corriger ces défauts.

1.2.3 Application aux données biologiques

1.2.3.1 Données génomiques : nature et hétérogénéité

Les séquences sont représentées sous forme de chaînes de caractères. Chaque nucléotide est représenté par une lettre. Les nucléotides identifiés sans équivoque sont notés par les lettres classiques (cf. 1.1.1.1: [Structure de la cellule vivante]) :

— A pour l'adénosine
— C pour la cytidine
— G pour la guanosine
— T pour la thymidine (dans l'ADN uniquement)
— U pour l'uridine (dans l'ARN uniquement)

Cependant, les travaux de séquençage ne permettent pas toujours d'identifier la totalité des nucléotides composant une molécule d'acide nucléique. De plus, certaines séquences peuvent représenter un consensus parmi plusieurs exemplaires

possibles. Dans ce cas, les ambiguités et segments indéterminés sont notés par des caractères complémentaires (cf. Tab. A.5: [Nucléotides ambigus selon l'IUPAC]). Afin de simplifier les calculs et d'exclure toute source d'incertitude, nous ne travaillerons ici qu'avec des séquences ADN ne comportant que des caractères standards parmi l'alphabet {A, C, G, T}.

La représentation *in silico* des séquences génomiques nécessite un formalisme permettant de repérer le début et la fin des séquences, ainsi que de stocker des métadonnées, même si elles se limitent à un simple identifiant. Plusieurs formats de fichier ont vu le jour, chacun répondant à des besoins particuliers. Mais le format le plus couramment utilisé pour stocker des séquences biologiques complètes et non compressées est le format FASTA. Il s'agit d'un format qui a été proposé avec la parution de la suite d'outils FASTA dédiée à la comparaison de séquences biologiques (PEARSON et LIPMAN [1988]). Sa simplicité en fait un format flexible et facile à parser qui s'est rapidement imposé comme un standard en bioinformatique.

Un fichier au format FASTA est un fichier plat qui peut contenir une ou plusieurs séquences. Chaque séquence est représentée par au minimum deux lignes. La première ligne est une ligne descriptive et commence invariablement par le caractère ">". Ce dernier est immédiatement suivi de l'identifiant de la séquence, et parfois de commentaires séparés de l'identifiant par un espace. La ou les lignes suivantes ne sont composées que par les caractères représentant la séquence (cf. Fig. 1.26: [Format FASTA]). Il est possible de stocker la totalité de la séquence sur une seule ligne mais le programme FASTA ne supporte pas les lignes de plus de 120 caractères. Il fait donc partie des bonnes pratiques de ne pas dépasser cette limite. De même, afin de faciliter la lisibilité des fichiers, il est recommandé de laisser une ligne vide entre deux séquences.

```
>contig06156  length=430   numreads=13
ATACTCTTTGTGAGATATGGCAAAaTCaCCACTACTTTATCTCTgaTTATTTTtCTATTT
tCTGTATTTtGCCTAGCTTATTATAAATTGAATCTTCCGTTGAATTTACCGGGGGAACCT
TGATAGAAACTCAGTTTtCTCATCTAGTGGAaCTAAACGATGTGCGTCAAAAAaTTGATC
AATTAAagCTGGGATCTACACAAGTACAATCTCTTGGTAATTCTCGCAATATTTTGATTC
GCCTACAGAATTCTGCTAaCAAAAaGCCTGAGACTTTGGCAAATTTAGTGTCTAATCAGC
TAAGAACATTGGATCCTACTATGCAGGTCAAACAGACTGAATTTGTCGGCCCGCAGGTAG
GAAAAGAATTACTAAGCAATGGTTtATTAGCTCTTTTtATGGTATGTATTGGCATTATCA
TTTATTTGGC
```

FIGURE 1.26 – Exemple de format de fichier FASTA : contig issu d'un assemblage métagénomique (cf. 1.1.3.2: [Métagénomique]). Les caractères minuscules représentent des zones de faible couverture génomique.

La taille des séquences produites dépend de la technologie utilisée. Chaque séquence produite en sortie du processus de séquençage s'appelle *read* et correspond à un fragment d'ADN lu par la machine. Un read peut faire de plusieurs dizaines à

plusieurs centaines de nucléotides (METZKER [2010], LIU et collab. [2012], BARBA et collab. [2014]). La taille des reads est un des critères principaux dans le choix de la technologie de séquençage. En effet, les reads courts (≤ 400nt) sont généralement déconseillés pour des échantillons métagénomiques (WOMMACK et collab. [2008a]).

Il est néanmoins possible de travailler avec des séquences plus longues. Parmi les reads produits, certains sont "chevauchant". Il s'agit de reads correspondant à une même région génomique, qui vont fournir la séquence de fragments partageant une région d'origine commune. En repérant ces régions communes, il est possible de mutualiser l'information contenue dans les reads afin de révéler la séquence de segments plus longs d'un génome. Ce processus d'élongation s'appelle *assemblage* et permet notamment de retrouver progressivement les séquences de génomes complets dans le cas d'espèces isolées.

FIGURE 1.27 – Première étape du processus d'assemblage. (a) Reads alignés. (b) Séquence déduite de l'alignement des reads ou *contig*.

En métagénomique, il est possible d'effectuer une étape d'assemblage afin de regrouper des reads appartenant au même organisme et de pouvoir travailler en aval sur des séquences plus longues, donc contenant davantage d'informations (cf. Tab. 1.2: [Tailles d'éléments génomiques]). Il s'agit de la première étape du processus d'assemblage, qui regroupe les reads chevauchants en *contigs* .

Cette technique possède néanmoins plusieurs limites. L'assemblage de génomes complets à partir de données métagénomiques est impossible sauf dans certains cas où l'échantillon de départ est extrêmement pauvre (WOOLEY et collab. [2010]). Certains organismes sont trop mal représentés pour bénéficier d'un assemblage robuste : si la *couverture génomique* (nombre de reads partageant la même position dans le génome) est trop faible, il est difficile de dépister les éventuelles erreurs d'amplification et de séquençage. Dans certains cas, le chevauchement est même trop court pour affirmer une origine commune. De plus, certains organismes peuvent posséder la même séquence sur certaines régions de leurs génomes respectifs, ce qui entraîne des incertitudes lors de l'alignement. À cela s'ajoute la difficulté intrin-

TABLEAU 1.2 – Différentes tailles de séquences d'ADN et le type d'information qu'il est possible d'y trouver. *Souce :* WOOLEY *et collab. [2010]*

Longueur de séquence (nt)	Éléments génomiques
25 - 75	SNPs, petits décalages du cadre de lecture
100 - 400	Petites signatures fonctionnelles
500 - 1 000	Domaines protéiques entiers, gènes monodomaines
1 000 - 5 000	Petits opérons, gènes multidomaines
5 000 - 10 000	Grands opérons, éléments de contrôle en cis
> 100 000	Prophages, îlots de pathogénicité, éléments mobiles variés
> 1 000 000	Organisation complète de chromosomes prokaryotes

sèque de séquencer certaines régions génomiques qui possèdent des motifs répétés de nombreuses fois. Toutes ces difficultés limitent fortement la taille des contigs obtenus et laissent certains reads impossibles à assembler.

Malgré tout, et à plus forte raison si les reads sont courts (WOMMACK et collab. [2008b]), il est toujours intéressant d'augmenter la quantité d'information par séquence afin d'améliorer la précision des résultats. En effet, de récents résultats montrent que l'assemblage à l'état de contig améliore la robustesse du signal taxonomique contenu dans les reads individuels, même dans le cas d'une importante chiméricité (MENDE et collab. [2012], TEELING et GLÖCKNER [2012]). C'est pourquoi, dans notre évaluation expérimentale (cf. 3.3.1: [Pourquoi la classification par règne est-elle difficile ?], 4.1.1.2: [Choix du type de signatures]), nous travaillons exclusivement avec des longueurs comparables à celles de contigs qu'il est possible d'obtenir par assemblage standard lorsque les données proviennent de communautés biologiques complexes.

Dans tous les cas, il s'agit de données hautement hétérogènes. Contrairement à de nombreux domaines d'application de l'apprentissage automatique, où les données sont mesurées de manière contrôlée en nature et en quantité, les données génomiques proviennent de matériel vivant, intrinsèquement difficile à maîtriser de manière uniforme. Ces difficultés sont à prendre en compte lors de la prédiction de l'origine de séquences inconnues. Ces prédictions ne doivent pas être considérées au-delà de leur valeur estimative et ne peuvent s'affranchir d'une analyse plus ciblée si leurs résultats présentent un intérêt scientifique.

1.2.3.2 Génomes de référence et séquences d'apprentissage

Tout jeu de données d'apprentissage doit être constitué de données dont l'origine est connue. Afin d'avoir un jeu d'apprentissage qui représente l'ensemble du vivant, il faut avoir accès à de nombreux projets de séquençage dont les organismes d'origines sont multiples et couvrent tous les grands clades. Un tel travail ne peut être assumé par un laboratoire isolé, mais doit être le produit d'un effort collectif de mise en commun des travaux de séquençage et de systématique.

De nombreuses bases de données publiques spécialisées dans les séquences nucléiques existent, dans lesquelles les séquences sont accompagnées de nombreuses métadonnées, incluant leur espèce d'origine. Trois grands instituts concentrent les principales bases de dépôt primaires de séquences nucléiques :

— L'Institut de Bioinformatique Européen (*EMBL-EBI*, `http://www.ebi.ac.uk`), au Royaume-Uni, financé collaborativement par 21 états européens

— Le Centre National pour les Informations Biotechnologiques (*NCBI*, `http://www.ncbi.nlm.nih.gov`), aux États-Unis

— La Banque Japonaise de Données ADN (*DDBJ*, http://www.ddbj.nig.ac.jp), au Japon

Afin d'unifier ces sources de données, les trois instituts participent à un effort de mutualisation des données matérialisé par la Collaboration Internationale des bases de Données de Séquences Nucléotidiques (INSDC, http://insdc.org). Plusieurs de leurs bases sont synchronisées par type de données contenues (cf. Tab. 1.3: [INSDC - bases synchronisées]). Cet effort participe au succès de ces dépôts car il assure un accès commun à un travail communautaire dont la richesse est accessible quel que soit l'institut depuis lequel les bases sont consultées.

TABLEAU 1.3 – L'ensemble des bases synchronisées par l'INSDC. *Souce : http://insdc.org*

Type	**DDBJ**	**EMBL-EBI**	**NCBI**
Reads NGS	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Reads capillaires (Sanger)	Trace Archive		Trace Archive
Séquences annotées	DDBJ		GenBank
Échantillons	BioSample		BioSample
Projets d'étude	BioProject		Bioproject

Il existe en outre de nombreuses bases spécialisées (e.g. http://www.wormbase.org pour le ver qui prenait l'escargot comme taxi, http://www.antgenomes.org pour la fourmi, ou encore http://www.informatics.jax.org/ pour la souris) ou généralistes (e.g. http://genomesonline.org pour les projets de séquençage), mais elles hébergent peu souvent leurs propres séquences. Elles comprennent généralement un ensemble cohérent de métadonnées spécialisées et hébergent souvent tout ou partie des séquences associées par les bases de données de l'INSDC.

Ces bases de données sont alimentées par la communauté scientifique. Elles fournissent des règles de soumission permettant aux laboratoires d'enrichir les différentes bases avec les résultats de leurs travaux de séquençage, d'assemblage, d'annotation et d'analyse (e.g. description et règles de la base de données BioProject au NCBI : http://www.ncbi.nlm.nih.gov/books/NBK169438/).

1.2.3.3 Taxonomies de référence et classes

Les taxonomies de référence sont des arbres phylétiques. Leurs feuilles représentent généralement des espèces (cf. 1.1.4.1: [Distance évolutive et homologie]), mais il peut parfois s'agir de souches ou d'individus. Les nœuds de ces arbres sont appelés taxons. Chaque taxon correspond à un ancêtre partagé exclusivement et exhaustivement par tous les individus représentés par le sous-arbre dont il est la racine. Les taxons suivent des standards selon lesquels un certain nombre de niveaux hiérarchiques doivent être respectés (cf. Fig. 1.28: [Niveaux phylétiques]).

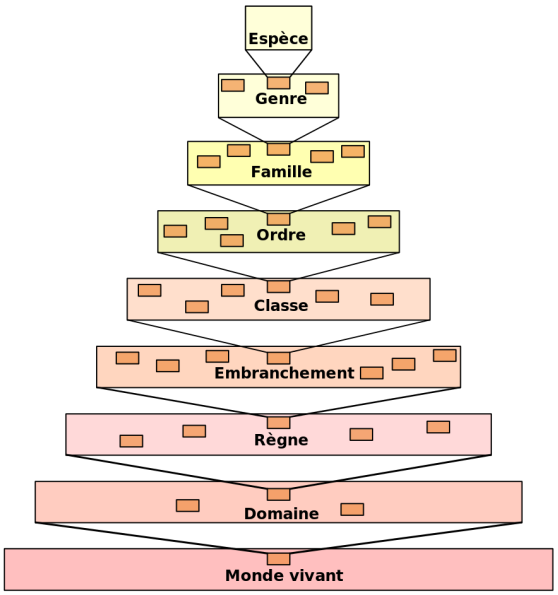


FIGURE 1.28 – Principaux niveaux phylétiques utilisés par les taxonomies de référence. Le niveau du monde vivant correspond à LUCA (cf. 1.1.4.1: [Distance évolutive et homologie]) et les domaines correspondent aux trois grands règnes du vivant : les eukaryotes, les archées et les bactéries (cf. Fig. 1.21: [Classification phylogénétique du vivant]) *Source : Wikimedia Commons*

Plusieurs bases de données taxonomiques publiques sont disponibles publiquement. On peut notamment citer le *Catalog of life* (`http://www.catalogueoflife.org`), né d'une collaboration entre l'*Integrated Taxonomic Information System* (`http://www.itis.gov/`) et *Species 2000* (`http://www.sp2000.org/`), mais également le *Tree of Life web project* (`http://tolweb.org`) ou encore *TreeBASE* (`http://www.treebase.org`). Mais la plus utilisée reste la base de données *taxonomy* du NCBI (`http://www.ncbi.nlm.nih.gov/taxonomy`) qui possède l'avantage d'être complètement intégrée dans l'environnement synchronisé de l'INSDC.

Ces bases de données proposent des taxonomies complètes ou partielles, uniques ou multiples, recoupant l'ensemble du vivant et du monde viral. En revanche, même si elles comprennent les virus, la base de donnée de l'ICTV (http://ictvonline.org/) reste la référence faisant autorité en virologie (cf. 1.1.4.2: [Notion d'espèce]).

Il est néanmoins possible de rencontrer des taxonomies simplifiées. Le cas le plus extrême est le nom latinisant par lequel on identifie les espèces. Il s'agit d'un système de nom binômial proposé par Carl von Linné au cours du XVIIIe siècle dans la dixième édition de son œuvre majeure, Systema Naturæ (LINNÉ, CARL VON et SALVIUS [1758]). Dans sa version actuelle, adaptée à la vision évolutive de la classification des espèces, sont utilisés le nom du genre et le nom de l'espèce. À titre d'exemple, le nom binômial de l'Arabette des dames, un organisme modèle utilisé en génétique depuis les années quarante, est *Arabidopsis thaliana*; *Arabidopsis* étant le nom du taxon générique, et *thaliana* étant le nom du taxon spécifique dans sa taxonomie.

Un autre exemple de taxonomie simplifiée se trouve dans la base de données *genome* du NCBI : Une section de cette base propose un résumé sur les projets de séquençage des génomes regroupés par grandes divisions taxonomiques (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/) contenant les liens vers les BioProjects correspondants (cf. Tab. 1.3: [INSDC - bases synchronisées]). Chaque entrée de ce résumé propose, parmi les métadonnées, un résumé de la taxonomie de l'organisme étudié constitué du nom du règne et de celui de l'embranchement. Pour les projets viraux, la taxonomie simplifiée est constituée de la classe de Baltimore (cf. Fig. 1.14: [Classification de Baltimore]) ainsi que de la famille.

Chapitre 2

État de l'art

Sommaire

Ce chapitre offre une définition détaillé de la problématique traitée et des différentes tâches de classification qui lui sont associées, ainsi qu'une présentation des méthodes disponibles à l'heure actuelle.

2.1 Cadre théorique

2.1.1 Définition du problème : Assignation taxonomique

Nous nous intéressons ici à l'application de techniques d'apprentissage automatique à l'identification de séquences génomiques inconnues. Il s'agit d'un problème de classification supervisée de chaînes de caractères (cf. 1.2.2.1: [Principe et spécificités]), dans lequel les objets manipulés sont la représentation de séquences génomiques sous forme de chaînes composées d'un alphabet de quatre lettres (cf. 1.2.3.1: [Données génomiques : nature et hétérogénéité]), et dont les classes utilisées sont des taxons (cf. 1.1.4.1: [Distance évolutive et homologie]).

Seuls seront considérés ici les classifieurs multiclasses. En effet, les classifieurs binaires (e.g. FACS, STRANNEHEIM et collab. [2010]) sont plus adaptés à la décontamination d'échantillons qu'à l'identification de séquences. Nous explorerons ici les différents types de méthodes utilisés par ces classifieurs.

Un certain nombre de méthodes bioinformatiques effectuent de manière efficace ce type de classification sur des séquences issues d'échantillons contenant principalement des espèces connues. Elles peuvent être organisées grossièrement en deux catégories majeures :

— Les méthodes par similarité (cf. 2.2: [Classification par similarité])
— Les méthodes par composition (cf. 2.3: [Classification par composition])

Il est possible de décomposer la question de l'assignation taxonomique en fonction des problématiques de départ. On définit ainsi deux grandes approches :

— **La classification par règne :** Il s'agit ici de déterminer à quel grand règne appartient une séquence. En plus des grands règnes du vivant (archées, bactéries, eukaryotes), nous ajoutons une classe supplémentaire pour les virus. Dans ce contexte, il s'agit d'arriver à isoler les virus du reste des séquences afin de procéder ensuite à l'approche suivante, tout en minimisant les erreurs. Cette approche s'apparente à la décontamination des échantillons, qui a pour but de supprimer les séquences n'appartenant pas à l'organisme d'intérêt. En revanche, il s'agit ici de conserver les séquences appartenant à tous les organismes d'un grand taxon.

— **La classification détaillée :** On s'intéresse ici à la taxonomie des organismes desquels proviennent les séquences. On détermine leur position dans un arbre donné, jusqu'à une profondeur déterminée soit à l'avance, soit en fonction d'un indicateur de fiabilité des résultats. Cette approche a pour but d'identifier l'origine des séquences en présence.

La plupart des outils évoqués dans ce chapitre s'intéressent en priorité à la seconde approche, car la tâche est facilitée par la filtration et la purification des échantillons en amont du séquençage (cf. 1.1.3.1: [Techniques de séquençage], HALL et collab. [2014]). Mais la première approche, qui s'intéresse à des données complexes dans lesquelles sont présentes des séquences d'origines multiples, est particulièrement difficile (cf. 3.3.1: [Pourquoi la classification par règne est-elle difficile ?]). Cette difficultée est d'autant plus importante lorsque des espèces virales inconnues y sont représentées, ce qui est souvent le cas (cf. 1.1.2.2: [Importance de la virologie]).

2.1.2 Classification par règne d'échantillons complexes

La séparation par règne des séquences composant un jeu de données métagénomique complexe représente un défi particulièrement difficile au niveau informatique. La difficulté majeure réside dans les séquences virales inconnues pour lesquelles aucun homologue n'a encore été identifié, caractérisé et répertorié dans les bases de données publiques.

Leur identification est pourtant un objectif majeur pour les biologistes. En effet, certaines espèces virales ne peuvent pas être isolées par filtrage, comme en témoigne la découverte du Mimivirus dont la taille ($\approx 400nm$, XIAO et collab. [2009]) dépasse largement les plus petites bactéries ($< 200nm$, LUEF et collab. [2015]). Néanmoins, leur découverte peut s'avérer essentielle pour l'identification de nouveaux pathogènes affectant l'homme, les plantes ou le bétail (ROOSSINCK [2012], LECUIT et ELOIT [2013], cf. 1.1.2.2: [Importance de la virologie]).

2.1.3 Classification détaillée de communautés bactériennes et virales

De l'autre côté du spectre, le problème de la caractérisation détaillée de jeux de données produits par séquençage ciblé a subi d'importants progrès au cours de ces dernières années. Contrairement à l'analyse d'échantillons métagénomiques non sélectifs, et donc plus complexes, des méthodes efficaces ont été développées dans le cas où certaines communautées (bactériennes ou virales) sont sélectionnées expérimentalement. Il s'agit d'un moyen de contourner la difficulté de la classification par règne, sans néanmoins y apporter de solution.

Pour les communautés microbiennes, la solution la plus efficace est l'utilisation de gènes marqueurs, tels que l'ARNr 16S pour les prokaryotes et l'ARNr 18S pour les eukaryotes (fungi). Dans ce cas, seuls ces gènes sont séquencés et l'information génomique utilisée est très partielle. Cela simplifie l'analyse pour deux raisons. D'une part, le volume de données reste raisonnable (pour une analyse à haut débit), et d'autre part, la classification taxonomique de ces gènes marqueurs est disponibles dans des bases de données de référence spécialisées telles que RDP (COLE et collab. [2009]) ou Greengenes (cf. DESANTIS et collab. [2006]). Les techniques par similarité de séquence combinées avec ces taxonomies de références sont extrêmement efficaces sur les distributions bactériennes connues (BAZINET et CUMMINGS [2012]). En revanche, ce type d'analyse souffre d'un défaut majeur : il ne fournit pas de moyen fiable de quantifier les espèces identifiées (ROUX et collab. [2011]).

Tandis que cette approche est faisable pour des populations bactériennes, elle n'est pas applicable lors de l'analyse de communautés virales car l'absence de gènes marqueurs semblables (EDWARDS et ROHWER [2005]) ne permet pas une comparaison directe de l'ensemble des espèces. L'étude de viromes se concentre sur la partie virale de l'échantillon et isole les particules virales encapsidées qui sont purifiées par filtration et (ultra)centrifugation. Cette approche à présent populaire réduit drastiquement la complexité des communautés, ce qui permet d'assembler systématiquement de plus grands contigs (10^3nt et plus), voire des génomes entiers dans le cas d'échantillons de faible complexité (COETZEE et collab. [2010], MINOT et collab. [2012]). En revanche, elle ne résoud pas vraiment le problème de la classification par règne, elle ne fait que l'éviter : suite à l'étape de purification, toutes les séquences sont généralement considérées "par définition" comme virales, jusqu'à preuve du contraire par des approches par homologie. De plus, cette stratégie n'est pas sans risque (pour plus de détails voir FANCELLO et collab. [2012]). Par exemple, les particules purifiées peuvent contenir des fragments de génomes cellulaires au lieu du génome viral (cf. 1.1.2.1: [Les virus]) à cause de la présence d'agents de transfert de gènes (LANG et BEATTY [2007]) ou en conséquence de transduction généralisée (pour une revue voir FROST et collab. [2005]). De plus, bien que filtrer les particules de plus de 220nm permet d'éviter la contamination par la plupart des cellules bactériennes, archéennes et eukaryotes, d'autres éléments contenant de l'ADN tels que des vésicules bactériennes (BILLER et collab. [2014]) peuvent être co-purifiées avec les virions. Une purification basée sur la filtration exclut également les plus gros virions, et ne permet pas d'obtenir une vision complète de la diversité virale. De plus, autant l'amplification LA (DUHAIME et collab. [2012]) que MDA ont leurs défauts (KIM et BAE [2011]). Dans le cas de la première, la ligation des adaptateurs n'est possible que pour les virus à ADN double brin, et par conséquent les génomes viraux à ADN simple brin sont massivement absents dans l'échantillon. Dans le cas de la seconde, l'amplification est préférentiellement effectuée sur des virus à ADN simple

brin circulaire au détriment des génomes à ADN double brin. Les conséquences de la présence de gènes cellulaires lors de l'analyse bioinformatique de données métagénomiques virales ont été décrites et des approches permettant de détecter leur présence ont été proposées (ROUX et collab. [2014]).

Malgré tout, l'étude des viromes a beneficié d'un large succès. Contrairement aux communautés bactériennes, les méthodes basées sur l'alignement de séquences ne semblent pas être les plus adaptées pour la classification virale. En effet, tel qu'il a été mentionné dans SUTTLE [2007], même dans le cas de reads viraux relativement longs, la fréquence d'homologie entre ces reads et les séquences protéiques au sein de la base de données Genbank n'est qu'autour de 30%. L'idée est d'éviter les fortes contraintes séquentielles imposées par les méthodes d'alignement sur la similarité nucléotidique, et de capturer un signal de similarité globale basé sur la composition des séquences (k-mers). Les techniques basées sur la composition semblent offrir des résultats satisfaisants pour la classification taxonomique détaillée d'échantillons viraux filtrés (YANG et collab. [2005], TRIFONOV et RABADAN [2010]).

2.2 Classification par similarité

2.2.1 Alignement de séquences

Cette première catégorie de classifieurs se base sur la comparaison de séquences par les techniques d'alignement (cf. Fig. 1.22: [Alignement de séquences]). Chaque séquence inconnue est confrontée aux séquences présentes dans une banque de données contenant des séquences connues, et à chaque alignement est attribué un score calculé en fonction des caractéristiques prises en compte. Parmi ces dernières, on peut trouver les *matches*, les *mismatches* et plus rarement les *gaps*, ces derniers pouvant être en outre pénalisés par leur longueur (cf. 1.1.4.1: [Distance évolutive et homologie]).

— *Match* : Deux nucléotides identiques face à face au sein de l'alignement. Partant du postulat que l'histoire évolutive des séquences correspond le plus souvent à la suite d' événements possibles la plus parcimonieuse, il en est déduit que la position observée n'a pas subi de mutation depuis la divergence des deux séquences (e.g. les 7 premières positions dans la figure).

— *Mismatch* : Deux nucléotides différents face à face. Partant du même principe, il est déduit que la position a subi un événement de substitution (e.g. A en face de G à la 8e position).

— *Gap* : Un ou plusieurs nucléotides ne trouvant pas de correspondance chez la séquence avec laquelle la séquence dont ils sont issus est alignée. Ces nucléotides supplémentaires peuvent être dus soit à un événement d'insertion dans

la séquence à laquelles ils appartiennent, soit à un événement de délétion dans l'autre séquence de l'alignement (e.g. A sans corresondance en position 12).

Contrairement aux algorithmes d'alignement globaux utilisés lors de la construction d'arbres phylogénétiques, au cours de laquelle des séquences sont comparées sur toute leur longueur, ces classifieurs utilisent des algorithmes d'alignement local, qui permettent de trouver la région de plus forte similarité entre deux séquences, sans forcer l'alignement des régions peu ou non similaires. Ce type d'algorithme est bien mieux adapté car il s'agit ici de comparer des fragments courts à des données de référence généralement plus importantes, comme des génomes complets. En effet, là où un alignement global sera généralement utilisé pour vérifier une homologie entre deux séquences de longueur comparable, par exemple pour déterminer si deux gènes ont un lien de parenté, l'alignement local sera plus pertinent lors de la recherche d'appartenance d'un fragment de séquence à une référence connue.

L'algorithme d'alignement local le plus célèbre est l'algorithme de Smith-Waterman (SMITH et WATERMAN [1981]). Il s'agit d'un algorithme optimal qui fournit l'alignement correspondant au meilleur score possible selon les valeurs de pondération choisies pour les matches, mismatches et gaps. Si son efficacité sur des petits jeux de données n'est plus à prouver, son exhaustivité implique néanmoins une quantité de calculs pouvant devenir problématique sur des données plus nombreuses (sa complexité étant de $O(n^2)$).

Comme il n'est pas raisonnable d'énumérer tous les alignements possibles pour les volumes de données actuellement produits, les logiciels d'alignement local font généralement usage de méthodes permettant de gagner du temps de calcul, en associant des heuristiques à la programmation dynamique (DURBIN et collab. [1998]).

C'est le cas de *Basic Local Alignment Search Tool* (BLAST, ALTSCHUL et collab. [1990], cf. Fig. 2.1: [BLAST]), une heuristique de recherche de similarité entre séquences biologiques, sur laquelle sont basées la plupart des méthodes de cette catégorie de classifieurs (BAZINET et CUMMINGS [2012]). Elle a l'avantage d'offrir, en plus d'un système d'attribution d'un score par alignement, le calcul de la probabilité et l'espérance mathématique d'obtenir ledit score en alignant la séquence au hasard dans la banque. Ainsi, elle donne une mesure de la robustesse des résultats par rapport aux données de référence.

La grande majorité des méthodes de cette catégorie utilise BLAST comme méthode de calcul de similarité. La plupart, comme MARTA (HORTON et collab. [2010]), MEGAN (HUSON et collab. [2007]), MTR (GORI et collab. [2011]), MG-RAST (MEYER et collab. [2008]) ou encore Sort-ITEMS (MONZOORUL HAQUE et collab. [2009]), s'en sert de manière très directe, et chacune propose une façon originale d'en exploiter les résultats (cf. 2.2.2: [L'algorithme LCA]). D'autres en proposent une utilisation

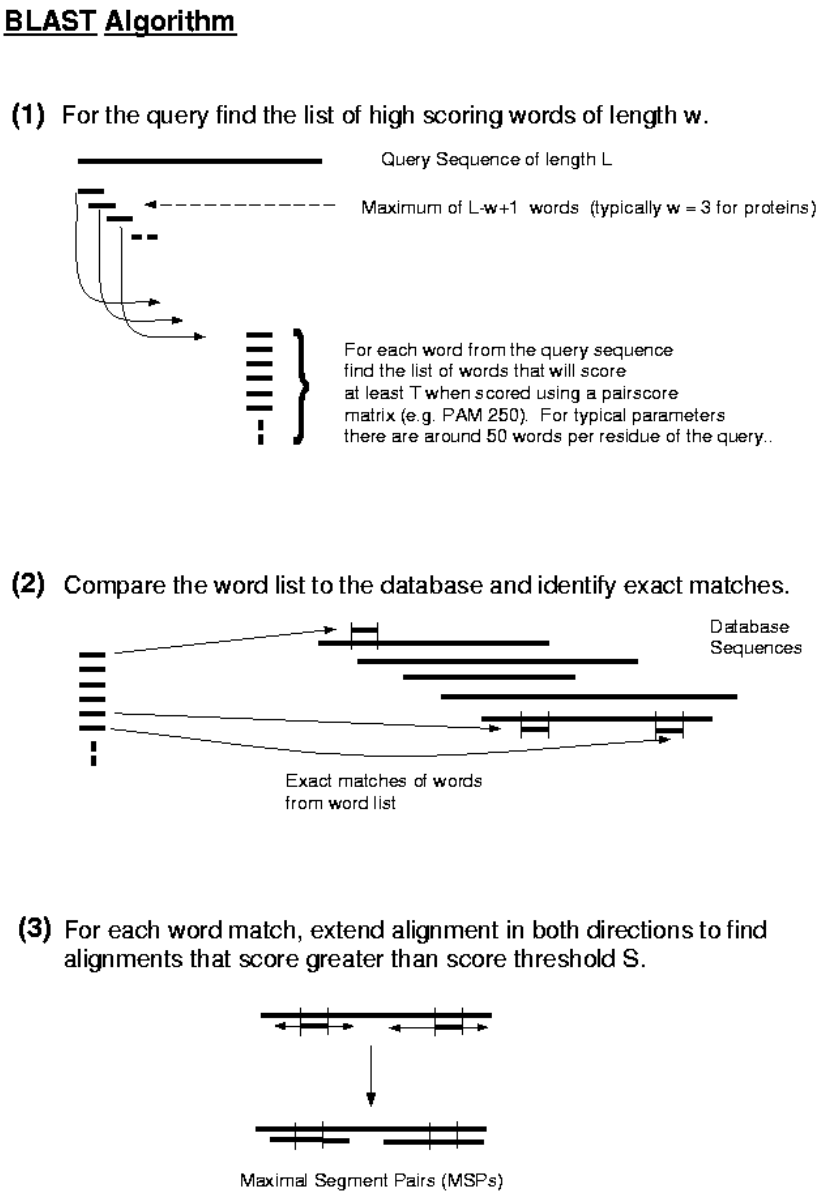


FIGURE 2.1 – Principe de fonctionnement de l'algorithme de BLAST. L'exemple illustré représente le cas d'alignement de séquences protéiques. *Source : Université de Can Tho*

.

moins directe. On peut ainsi citer MetaPhyler (LIU et collab. [2011]), qui inclut une étape d'apprentissage au cours de laquelle l'outil va déterminer les régles selon lesquelles un résultat BLAST pourra offrir une fiabilité raisonnable, sur un jeu de séquences de référence composé de 31 gènes ubiquitaires. Ensuite, au cours de l'étape de classification, le meilleur hit de chaque séquence est interprété selon ces règles.

Alternativement, d'autres méthodes d'alignement sont utilisées pour la classification, comme les automates de Markov à états cachés (*Hidden Markov Models*, HMM, EDDY [2004]). Ce sont des modèles statistiques surtout très utilisés depuis les années 70 pour la reconnaissance vocale. Leur utilisation la plus connue en bio-informatique est la prédiction de gènes, mais ils sont aussi utilisés pour faire de l'alignement de séquences. Il s'agit d'automates permettant de détecter si une séquence présente des enchaînements de lettres similaires à celles qui ont permis leur construction (DURBIN et collab. [1998]). L'implémentation la plus courante est HMMER3 (EDDY [2008], cf. Fig. 2.2: [HMMER]), notamment proposée comme méthode d'alignement alternative à BLAST dans CARMA3 (KRAUSE et collab. [2008]).

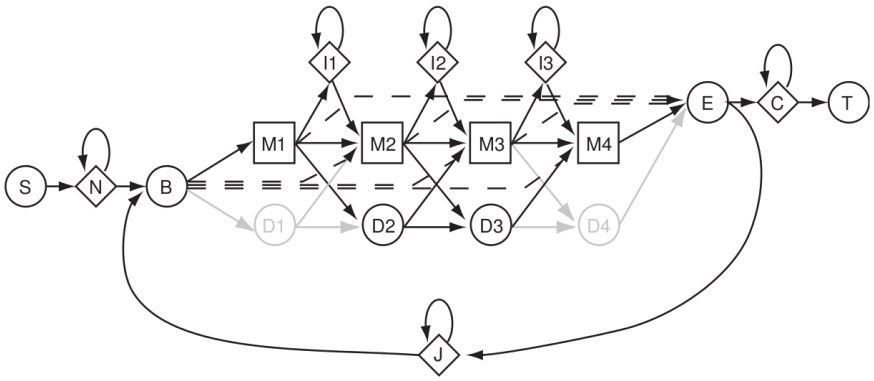


FIGURE 2.2 – Exemple de diagramme d'états d'un automate de HMMER. Les états nommés "M" représentent les positions conservées de l'alignement d'apprentissage. Les états d'insertion "I" permettent l'ajout de nouvelles positons dans le cas d'une insertion, et les états de délétion "D" sont silencieux, et permettent de sauter certaines positions si elles ne sont pas conservées par rapport au modèle. *Source :* JOHNSON *[2006]*

.

2.2.2 L'algorithme LCA

Afin d'exploiter les résultats de l'alignement d'une séquence aux séquences de référence d'une base de données, il faut déterminer une stratégie permettant de lui attribuer un taxon (cf. 1.1.4.1: [Distance évolutive et homologie]). Pour cette étape, de nombreux outils utilisent l'algorithme LCA (Lowest Common Ancestor), décrit pour la première fois par AHO et collab. [1973]. Le premier algorithme optimal fut décrit

par HAREL et TARJAN [1984], et il fut inclus dans un outil d'assignation taxonomique de séquences biologiques pour la première fois dans l'outil MEGAN (HUSON et collab. [2007]).

L'algorithme utilise la taxonomie du NCBI comme un arbre de classes (taxons). Il exploite les résultats d'alignement de la manière suivante :

— Il liste tous les hits pour chaque séquence

— Pour chaque séquence s, il calcule l'ensemble H de tous les taxons correspondant à la liste des hits de s

— Il cherche le nœud v de plus bas niveau dans l'arbre qui enracine un sous-arbre comprenant la totalité de H et assigne s au taxon représenté par v

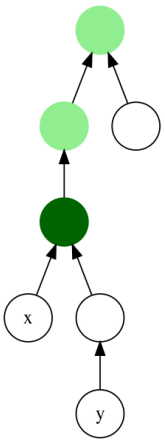


FIGURE 2.3 – Algorithme LCA : Exemple. Pour l'ensemble de séquences s tous les hits se produisent sur l'ensemble de taxons $H = \{x; y\}$. Parmi tous les ancêtres communs de l'ensemble H (en vert), celui de plus bas niveau est le LCA (en vert foncé). *Source : Wikimedia Commons*

.

Ainsi, le niveau du nœud correspondant à la classe attribuée à une séquence peut varier d'une séquence à l'autre. Plus une séquence est représentative de son taxon d'origine, plus le niveau du nœud sera bas, et plus le résultat sera précis. À l'inverse, certains gènes ubiquitaires, comme les gènes codant pour les ARNr, peuvent même se retrouver assignés à la racine de l'arbre.

Il existe des alternatives à l'algorithme LCA. On peut notamment citer la méthode proposée par MTR (*Multiple Taxonomic Ranks*, GORI et collab. [2011]) qui utilise ici une approche de clustering dans le but de traiter des reads courts (~100nt) sans avoir recours à une étape d'assemblage. Pour chaque niveau taxonomique, les reads sont regroupés en clusters selon les taxons représentés par les résultats de leur alignement sur une base de données protéiques par BLASTx. Chaque cluster est ensuite associé à

un taxon au niveau considéré, et les reads pour lesquels ce taxon n'est plus cohérent avec les taxons associés aux niveaux supérieurs sont assignés au taxon du niveau précédent et retirés du pool de reads pour les niveaux suivants.

2.2.3 Discussion

Les méthodes par alignement souffrent de deux limitations majeures : une faible vitesse de calcul et une faible sensibilité (BAZINET et CUMMINGS [2012], WOOD et SALZBERG [2014]). Récemment, de nouvelles solutions ont été apportées afin de dépasser ces limitations. Ces méthodes sont basées sur l'utilisation de longs k-mers et reposent sur le fait que si k est suffisamment grand, les k-mers deviennent très spécifiques. Par conséquent, le principe de ces méthodes est l'indexation des bases de données de référence par ces longs k-mers. Il s'agit du principe de base de Mega-Blast, un outil d'alignement généraliste de la suite d'outils BLAST, mais également de plusieurs méthodes spécifiques à l'assignation taxonomique telles que LMAT (AMES et collab. [2013]), Kraken (WOOD et SALZBERG [2014]) et CLARK (OUNIT et collab. [2015]). Le désavantage de ces approches est la sur-spécificité, qui rend problématique la classification de séquences provenant d'espèces inconnues.

Cette limitation peut être d'autant plus dramatique chez les virus, sachant leur très grande variabilité intraspécifique. Par exemple, les critères actuels de l'ICTV (cf. 1.1.4.2: [Notion d'espèce]) tolèrent jusqu'à 28% de divergence pour les gènes codant pour la polymérase ou les protéines de la capside pour des individus de la même espèce dans la famille des *Betaflexviridae* et un niveau de divergence similaire pour la totalité du génome chez les *Potyviridae* (KING et collab. [2011]). Par conséquent, la faible qualité des alignements entre séquences homologues appartenant à des membres d'une même espèce peut rendre la détection de cette appartenance particulièrement difficile.

Malgré tout, plueieurs outils spécialisés pour les séquences virales ont vu le jour comme PASC (BAO et collab. [2012, 2014]), qui utilise BLAST pour des alignements locaux et l'algorithme de Needleman-Wunsch pour des alignements globaux, ou plus récemment DEmARC (LAUBER et GORBALENYA [2012a,b,c]) qui utilise une mesure de divergence basée sur l'estimation du maximum de vraisemblance sur un alignement multiple de gènes marqueurs. Ces outils ne sont utilisable qu'au sein de familles virales particulières qui partagent des gènes communs et une histoire évolutive et, par conséquent, répondent à une problématique bien différente de la nôtre. Ils sont cependant très utilisés dans le cadre de l'effort d'amélioration de la taxonomie virale, et dans le cadre de l'étude de variants d'espèces d'intérêt.

Les méthodes d'alignement de séquences sont très dépendantes des relations d'homologie entre les séquences-requêtes et le contenu des bases de données auxquelles

elles sont comparées. Pour cette raison, plus les distances évolutives sont grandes entre les deux, plus il va être difficile d'établir une classification fiable. Il est possible d'être confronté à cette limitation si le contenu de la base de données ne contient pas de génomes proches de la requête, si elle appartient par exemple à un organisme inconnu pour lequel aucun proche parent n'a encore fait l'objet d'un effort de séquençage. Sachant la difficulté à isoler de nombreux génomes (cf. 1.1.3.2: [Métagénomique]), ainsi que la grande variabilité de certaines espèces, en particulier chez les virus, il n'est pas surprenant d'observer de mauvaises performances sur certains échantillons. En effet, une grande partie des reads produits par des projets de séquençage *de novo* d'espèces virales sont classifiés comme "inconnus" (BZHALAVA et collab. [2012], BZHALAVA et collab. [2013]).

L'algorithme LCA (cf. 2.2.2: [L'algorithme LCA]) utilise le parcours d'un arbre taxonomique pour trouver le plus petit ancêtre commun entre les taxons présentant un hit pour une séquence donnée. Cette méthode part du postulat que la taxonomie est construite selon les principes de la cladistique de sa racine à ses feuilles, c'est à dire qu'il y a des liens de parenté mesurables entre les différents nœuds de l'arbre. La taxonomie virale (cf. 1.1.4.2: [Notion d'espèce]) est un peu différente dans le fait qu'elle ne vérifie ces principes que de manière locale. Il y a donc une incompatibilité théorique fondamentale avec l'algorithme LCA dans le fait qu'il suppose l'existence d'un dernier ancêtre commun universel à la racine de l'arbre, que la taxonomie virale n'est à l'heure actuelle pas en mesure de fournir, ni même d'en affirmer l'existence.

Ce reproche est aussi applicable à MTR, mais s'y ajoutent également d'autres préoccupations. L'approche est intéressante si une étape d'assemblage n'est pas indispensable pour le reste de l'analyse des données, mais peut être handicapante si les reads ne sont plus accessibles. De plus, les auteurs comparent les deux méthodes sur des reads courts, mais ne précisent pas si l'avantage de MTR survit à une étape d'assemblage en amont de l'algorithme LCA.

2.3 Classification par composition

Une autre approche est basée sur l'analyse de la composition des séquences. La simplicité de l'alphabet qui compose les séquences d'ADN peut surprendre, lorsqu'on sait la complexité de l'information qu'elles contiennent. Observer et comparer les différentes régions de l'ADN permet de mettre en évidence une grande hétérogénéité de composition avec, par exemple, des motifs isolés très conservés spécifiques à certaines fonctions biologiques (e.g. séquences promotrices et régulatrices) et de longues zones contenant des motifs répétés très variables pouvant jouer un rôle structurel crucial (e.g. régions centromériques et télomériques).

Il s'agit ici de comparer les fréquences d'apparition de certains motifs entre les séquences, sans tenir compte de leur ordre d'apparition sur les brins. Contrairement à l'approche par similarité, il ne s'agit pas d'aligner les séquences mais d'extraire, à partir de leur composition, des signatures spécifiques et comparables entre lesquelles il est possible de calculer des distances.

Cette approche se base sur la notion de *biais compositionnel.* Il a été démontré que la composition des séquences varie non seulement d'une région génomique à l'autre, mais également d'un taxon à l'autre, et que ces biais sont porteurs d'information sur l'histoire évolutive des espèces (GAUTIER [2000]; MOOERS et HOLMES [2000]). Ainsi, à l'époque où le séquençage NGS n'existait pas encore (cf. 1.1.3.1: [Techniques de séquençage]), il était recommandé d'utiliser la température de dénaturation de l'ADN, à partir de laquelle il est possible de calculer la proportion de G-C dans les séquences, pour la définition d'unités taxonomiques dans le cadre de la systématique bactérienne (MOORE et collab. [1987]).

En revanche, même s'il s'agit d'un outil de mesure qui a fait ses preuves en phylogénétique bactérienne, la proportion de G-C se montre insuffisante à plus grande échelle et ne suffit plus à caractériser les taxons de manière suffisamment spécifique (FOERSTNER et collab. [2005]). Afin de remédier à ce problème, l'approche par composition étend le concept de biais compositionnel à des motifs plus spécifiques composés de plusieurs nucléotides.

De plus, contrairement aux méthodes d'alignement qui dépendent fortement de la conservation de la contiguité des segments homologues, les techniques utilisant la composition permettent de s'en affranchir et ainsi récupérer de la pertinence dans le cas de génomes très variables comme les génomes viraux (VINGA et ALMEIDA [2003]).

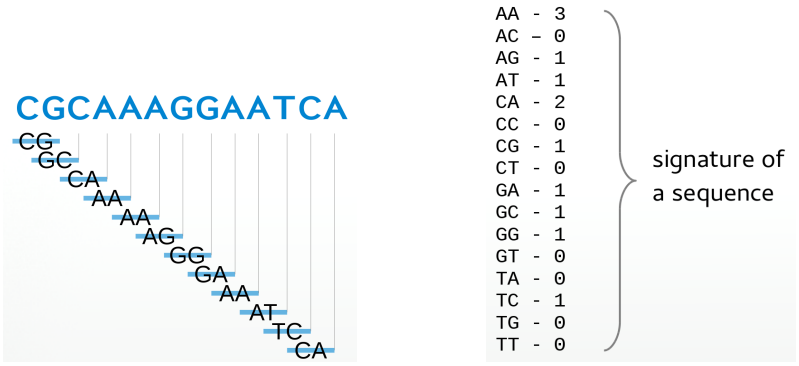
Ces méthodes reposent sur la décomposition des séquences en fréquences d'apparition de k-mers courts (non-uniques) et utilisent des techniques d'apprentissage automatique (e.g. SVM, kNN, Naïve Bayes, etc., cf. 1.2.1.1: [Principe et applications générales]) afin d'entraîner un classifieur sur une base de données de référence. L'assignation taxonomique de séquences inconnues est ensuite prédite par l'application de ce modèle pré-entraîné. Ces méthodes sont théoriquement plus adaptées à la classification d'espèces inconnues, car les distributions de k-mers courts sont moins sensibles au surapprentissage.

Il existe d'autres types d'approches sans alignement. Il est notamment important de citer NVR (YU et collab. [2013]) car il s'agit d'un outil spécialisé pour les séquences virales. Il repose sur un système de vecteur htérogène contenant 12 variables calculées à partir de la séquence. On y trouve le nombre d'occurrence de chaque lettre, leur

position moyenne au sein de la séquence, et un coefficient calculé en fonction des deux variables précédentes et de chaque position de la lettre courante. Il s'agit d'une approche très intéressante qui permet de raisonner très haut dans la taxonomie, mais qui ne fonctionne correctement que sur des génomes entiers. Elle ne répond donc pas à notre problématique mais est très utilisée pour tenter de compléter, prédire et corriger la taxonomie virale, et a déjà permis d'offrir une place à de nombreux génomes viraux dont la taxonomie était incomplète.

2.3.1 Distributions de k-mers et signatures

Les k-mers sont des mots de taille k lus sur une séquence à travers une fenêtre glissante que l'on déplace d'une base à chaque itération (cf. Fig. 2.4: [Signature d'une séquence]). L'alphabet A = {"A", "C", "G", "T"} utilisé dans les séquences génomiques étant de taille 4, le nombre total de mots possibles est 4^k.



(a) Décomposition d'une séquence en k-mers

(b) Nombre d'occurrences de chaque k-mer

FIGURE 2.4 – Extraction de la signature d'une séquence à partir de sa composition en k-mers. Nous prenons ici k=2 à titre d'exemple. (a) Tous les k-mers composant la séquence sont répertoriés avec une fenêtre glissante de taille k se déplaçant de 1 en 1. (b) Le nombre d'occurrences de chaque k-mer possible (4^k possibilités) est stocké dans un tableau. Ce tableau constitue la signature de la séquence, et le nombre d'occurrences peut être normalisé par la taille. On obtient alors un tableau de fréquences.

Des premiers travaux (KARLIN et BURGE [1995]) se sont rapidement intéressés aux fréquences d'apparition de dinucléotides ($k = 2$) afin de pallier au manque de spécificité des méthodes utilisant le contenu en G-C, et ont pu mettre en évidence une grande stabilité des fréquences des dinucléotides au sein d'un même génome, ainsi que d'importantes disparités entre différentes espèces (cf. Fig. 2.5: [Fréquences de dinucléotides chez *E. coli* et *C. elegans*]), confirmant ainsi l'intérêt d'utiliser des distributions de k-mers en tant que signatures spécifiques afin de différencier plusieurs espèces.

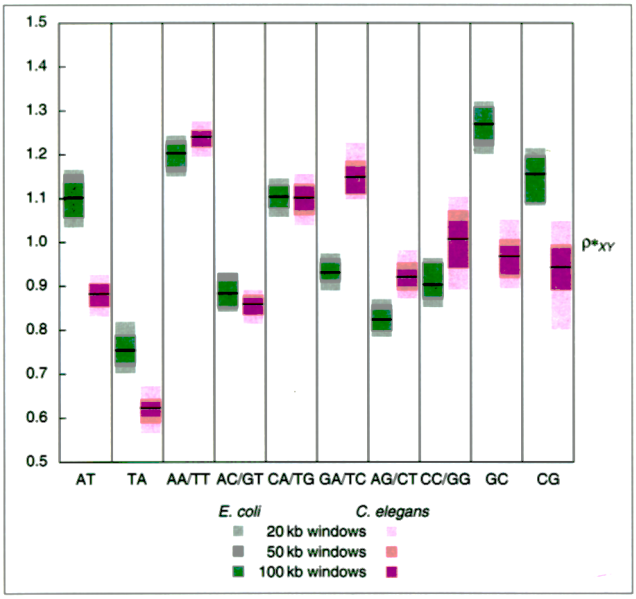


FIGURE 2.5 – Comparaison des fréquences relatives de certains dinucléotides entre *E. coli* et *C. elegans*. $\rho^*_{XY} = \frac{f_{XY}}{f_X f_Y}$ où f_X est la fréquence du nucléotide X et f_{XY} est la fréquence du dinucléotide XY dans la séquence et son complément. *Source :* KARLIN *et* BURGE *[1995]*

Un peu plus tard, grâce aux progrès de l'informatique offrant d'avantage de puissance de calcul et de stockage, plusieurs outils d'assignation taxonomique utilisant des k-mers courts (non uniques) ont vu le jour. Deux approches principales peuvent être distinguées :

— **L'approche par génome entier :** Cette approche consiste à comparer les fréquences de k-mers d'une séquence à classifier avec celles des génomes de référence dans leur intégralité. La plupart de ces méthodes utilisent des modèles de Markov, comme Swaap PH (PRIDE et collab. [2003]), ZOM/MCM (BOHLIN et collab. [2008]), Phymm/PhymmBL (BRADY et SALZBERG [2009]) ainsi que MGTAXA (`http://andreyto.github.io/mgtaxa`) qui est directement inspiré de la méthode de Phymm.

D'autres méthodes proposent différentes manières de calculer la distance entre la distribution de k-mers d'une séquence à classifier et celles des génomes de référence. La plus simple, proposée par BOHLIN et collab. [2008], consiste à calculer le coefficient de pearson entre chaque vecteur-requête et chaque vecteur de référence.

TACOA (DIAZ et collab. [2009]) propose des vecteurs de k-mers un peu plus complexes, contenant le ratio entre la fréquence d'apparition de chaque k-mer et sa fréquence attendue sachant le contenu en G-C de la séquence. Une

fonction discriminante est ensuite calculée taxon par taxon afin de calculer un score d'appartenance pour chaque séquence.

D'autres mesures de distance ont été explorées par TRIFONOV et RABADAN [2010], spécifiquement pour des données virales de classe V. Les auteurs ont notamment testé la distance de Manhattan, la distance euclidienne, le test du χ^2, ainsi qu'une version symétrisée de la divergence de Kullback-Leibler, une mesure de distances basée sur l'entropie de Shannon

— **L'approche par rééchantillonage :** Il s'agit ici de rééchantilloner les génomes de référence afin d'obtenir des fragments de ces génomes ayant des tailles comparables aux séquences à classifier. C'est à partir de ces fragments que les vecteurs de fréquences de k-mers sont calculés. Dans ce contexte, chaque génome de référence est représenté par un ensemble de vecteurs de fréquences différents les uns des autres. Ces vecteurs peuvent être considérés comme des coordonnées dans un espace à 4^k dimensions, et cette approche se base sur l'idée selon laquelle les vecteurs appartenant à un même taxon seront rapprochés dans l'espace.

PhyloPythia (MCHARDY et collab. [2007]) utilise cette représentation afin d'entraîner une machine à vecteur de support linéaire sur de long fragments (>1000nt). D'autres algorithmes d'apprentissage automatiques ont été utilisés dans la suite d'outils RDP (WANG et collab. [2007]) en utilisant des données d'ARNr comme gènes marqueurs : un classifieur naïf bayésien (RDP classifier) et l'algorithme des k plus proches voisins (RDP SeqMatch).

RAIphy (NALBANTOGLU et collab. [2011]) utilise une méthode différente. Tous les vecteurs de fréquences des fragemnts d'un même taxon sont utilisés afin de construire un unique vecteur, appelé Index d'Abondance Relative, qui contient, pour chaque k-mer, un score de représentation calculé avec une approche markovienne d'ordre 1. Ce score est positif si le k-mer est surreprésenté dans le taxon, et négatif s'il est sous-représenté. Pour chaque séquence à classifier, un score d'appartenance est calculé pour chaque taxon en faisant la somme de tous les scores de représentation de l'index du taxon, pondérée par les valeurs du vecteur de fréquences de k-mers de la séquence à classifier. Plus les fréquences d'apparition des k-mers de la séquence coïncideront avec l'abondance relative de ces mêmes k-mers au sein du taxon, plus le score d'appartenance sera élevé.

On peut également citer NBC (ROSEN et collab. [2011]), un outil à la limite entre alignement et composition utilisant un classifieur bayésien naïf, qui a la particularité de ne pas utiliser de génomes entiers ou de fragments rééchantillonnés, mais de reads issus de données réelles, d'une longueur moyenne de 230nt.

2.3.2 Discussion

Les méthodes par composition bénéficient depuis quelques années d'un certain succès. En effet, le nombre d'outils ayant vu le jour est témoin d'un intérêt indéniable pour ce type d'approche. S'affranchir de la séquentialité des nucléotides n'est pas chose aisée compte tenu de l'importance historique des méthodes par alignement dans l'étude des relations d'homologie, mais ouvre des portes en terme de gestion des réarrangements au niveau moléculaire, à travers une souplesse que ne permettent pas les outils d'alignement.

En revanche, même ces techniques ne parviennent pas à classifier autour de 50% des espèces qui ne sont pas représentées dans le jeu de données d'apprentissage (NALBANTOGLU et collab. [2011]). Cela est d'autant plus valable pour les séquences virales, dont la plupart ne parviennent pas à être assignées à un quelconque règne (ROSEN et collab. [2011]). En effet, pour les séquences virales (mais également eukaryotes), aucune des méthodes existantes ne produit une distribution taxonomique qui s'approche de la distribution attendue (BAZINET et CUMMINGS [2012]).

Ces méthodes sont très sensibles. De nombreux facteurs peuvent influencer leur efficacité de manière dramatique. Les paramètres utilisés par les algorithmes sont évidemment des acteurs importants, mais les données manipulées jouent également un rôle crucial. En effet, la longueur des k-mers, la taille des séquences, la présence ou non des organismes dans les données d'apprentissage ainsi que les espèces représentées peuvent faire grandement varier la qualité des résultats (TRIFONOV et RABADAN [2010]).

De plus, les génomes ne sont pas homogènes. Cela pose problème pour les deux approches. En effet, une signature unique par génome entier ne reflète pas les variations locales et souffre ainsi d'une perte d'information importante parmi les données d'apprentissage. De la même manière, si la couverture génomique des fragments rééchantillonés sur leurs génomes d'origine n'est pas suffisante, il est possible que certaines variations locales disparaissent des données d'apprentissage et constituent un angle mort pour le classifieur. Ce phénomène est bien illustré par BOHLIN et collab. [2008] qui montre de grandes variations d'autocorrelation locales sur une fenêtre glissante de 5000 nucléotides. Ces variations mettent également en évidence des régions de très faible autocorrélation correspondant aux ARNr mais également à de nombreux prophages, virus endogènes bactériens.

Il s'agit néanmoins d'approches prometteuses. Leur nature sensible ne font pas d'elles de bonnes candidates pour un outil universel mais montrent de bon résultats lorsqu'elles sont utilisées dans le cadre précis dans lequel elles ont été développées.

Chapitre 3

Identifier des séquences virales : une tâche difficile

Sommaire

Ce chapitre porte sur la difficulté liée à l'identification de séquences virales au sein de données métagénomiques. Sont d'abord détaillées les raisons théoriques de cette difficulté, du point de vue de la nature des données manipulées et des spécificités propres au monde viral, puis sont présentés les résultats de l'estimation *in silico* de cette difficulté, qui a fait l'objet d'une publication dans le journal Frontiers in Microbiology au cours de cette thèse (SOUEIDAN et collab. [2014]).

3.1 Données de référence

3.1.1 Propreté et pertinence des données

Les bases de données publiques sont alimentées en permanence par la communauté scientifique (cf. 1.2.3.2: [Génomes de référence et séquences d'apprentissage]). Cela permet l'assurance d'un accès aux données correspondant aux travaux de recherche les plus récents, ainsi que des bases de données qui reflètent l'évolution des connaissances dans leur ensemble. Néanmoins, l'afflux grandissant de données (cf. Fig. 1.23: [Croissance de GenBank]) pose de nombreux problèmes en terme de vérification de la validité des métadonnées et de la maintenance des différentes bases.

Par conséquent, les bases de données synchronisées par l'INSDC (cf. Tab. 1.3: [INSDC - bases synchronisées]) ne sont pas en mesure d'assurer une non-redondance des données, ni une vérification manuelle des métadonnées. Même RefSeq (`http://www.ncbi.nlm.nih.gov/refseq/about/`), une base de données locale vérifiée manuellement et réputée pour sa fiabilité, ne peut retenir la publication des données non vérifiées et signale le statut de chaque entrée dans les commentaires qui lui sont associés. Ces données publiques sont donc bruitées par nature (cf. 1.2.3.1: [Données génomiques : nature et hétérogénéité]), mais également du fait du système permettant leur mises en commun.

Les taxonomies de références ne sont pas des arbres complets ni des arbres parfaits. Un arbre complet est un arbre dont toutes les feuilles sont sur le même niveau. Or, même si la plupart des branches ont des espèces pour feuilles, il peut arriver qu'il y ait des niveaux supplémentaires comme des sous-espèces ou des souches.

Un arbre est parfait si tous ses niveaux sont remplis, mais ce n'est pas le cas ici. En effet, même si un ensemble de niveaux hiérarchiques standards est généralement respecté (cf. Fig. 1.28: [Niveaux phylétiques]), l'ensemble du vivant est bien trop hétérogène pour que de telles contraintes soient pertinentes dans tous les cas (cf. 1.1.4.2: [Notion d'espèce]). Ainsi, lorsque l'on recherche la phylogénie complète de

plusieurs espèces dans la phylogénie de référence du NCBI, il est possible d'obtenir un nombre très variable de niveaux phylétiques.

Par exemple, la bactérie modèle *Escherichia coli* (cf. Tab. 3.1: [Phylogénie : *E. coli*]) ne possède pas de règne, alors qu'à l'inverse, la plante modèle *Arabidopsis thaliana* (cf. Tab. 3.2: [Phylogénie : *A. thaliana*]) possède tous les niveaux standards auxquels ont été rajoutés de nombreux niveaux supplémentaires tels que la sous-classe et la tribu, ainsi que plusieurs niveaux ne possédant pas de dénomination.

Par conséquent, il est difficile de comparer l'ensemble du vivant sur un niveau phylétique donné, non seulement parce qu'ils ne sont pas toujours tous présent d'une espèce à l'autre, mais aussi parce qu'ils ne représentent pas systématiquement le même niveau de granularité, à plus forte raison lorsque les espèces sont éloignées les unes des autres. L'utilisation de ces bases de données est donc délicate et nécessite de prendre en compte l'ensemble des cas de figure et les incertitudes inhérentes à la systématique.

TABLEAU 3.1 – Phylogénie complète de la bacterie *Escherichia coli* selon la base de données *taxonomy* (`http://www.ncbi.nlm.nih.gov/taxonomy`) du NCBI. Les niveaux phylétiques notés *N/R* ne possèdent pas de nom (no rank).

Niveau phylétique	**Libellé**
N/R	Organismes cellulaires
Domaine	Bactéries
Embranchement	Protéobactéries
Classe	Gammaprotéobactéries
Ordre	Entérobacteriales
Famille	Entérobactériacées
Genre	Escherichia
Espèce	Coli

Il existe un très fort biais d'étude vers l'homme. En effet, les espèces représentant un intérêt industriel ou médical sont particulièrement bien représentées dans les bases de données de référence et déséquilibrent fortement les quantités de données entre les différents clades.

Ainsi, les eukaryotes sont particulièrement bien représentés car ils comprennent l'homme et la plupart des espèces utilisées dans l'industrie agroalimentaire (e.g. levures, bétail, poissons, cultures), de même que certaines espèces responsables d'épi-

TABLEAU 3.2 – Phylogénie complète de l'Arabette des Dames, ou *Arabidopsis thaliana* selon la base de données *taxonomy* (http://www.ncbi.nlm.nih.gov/taxonomy) du NCBI. Les niveaux phylétiques notés *N/R* ne possèdent pas de nom (no rank).

Niveau phylétique	**Libellé**
N/R	Organismes cellulaires
Domaine	Eukaryotes
Règne	Chlorobiontes
Embranchement	Streptophytes
N/R	Streptophytina
N/R	Embryophytes
N/R	Trachéophytes
N/R	Euphyllophytes
N/R	Spermatophytes
N/R	Magnoliophytes
N/R	Mésangiospermes
N/R	Eudicotylédones
N/R	Gunneridées
N/R	Pentapétalées
Sous-classe	Rosidées
N/R	Malvidées
Ordre	Brassicales
Famille	Brassicacées
Tribu	Camélinées
Genre	Arabidopsis
Espèce	Thaliana

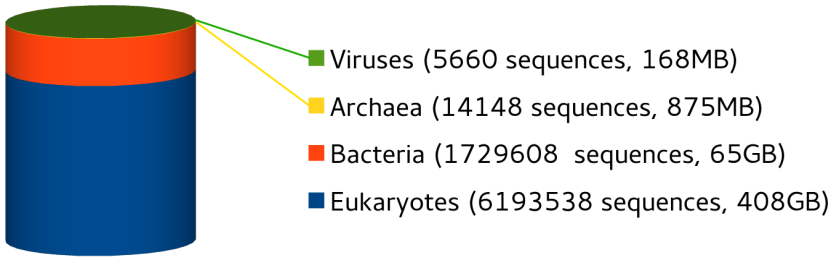


FIGURE 3.1 – Répartition par domaine des données de la base RefSeq (http://www.ncbi.nlm.nih.gov/refseq) du NCBI. L'ensemble de ces données représente 7 932 410 séquences comprenant 474.10^9 caractères au total. *Source : NCBI, RefSeq release 08/09/2014*

démies actuelles (e.g. moustiques vecteurs de la malaria) (cf. Fig. 3.1: [Données RefSeq par domaine]). Les bactéries également sont particulièrement bien étudiées, non seulement parce qu'elles comprennent de nombreux pathogènes, mais également pour l'intérêt industriel que représentent certaines espèces (e.g. traitement de la pierre, transformation de produits alimentaires, traitement des déchets, production de molécules dans l'industrie pharmaceutique).

À l'inverse, les archées sont très mal connues, si ce n'est pour la capacité de certaines espèces à vivre dans des milieux extrêmes (e.g. sources chaudes, lacs salés). Leur découverte est en effet très récente (WOESE et FOX [1977], BALCH et collab. [1977]). Elles sont néanmoins présentes dans de nombreux biotopes, mais ne représentent que peu d'intérêt hors du contexte académique, car il n'existe pas actuellement d'exemple d'archée pathogène connu. Elles sont pour la plupart difficiles à cultiver en laboratoire, à l'instar de la plupart des microorganismes, mais ne bénéficient pas encore du recul nécessaire au développement de méthodes de cultures aussi spécifiques que celles dont bénéficient actuellement l'étude des bactéries.

Les virus souffrent également d'un fort biais anthropocentrique. Nous avons vu que l'ensemble des virus connus sur lesquels nous disposons de données publiques est une sous-estimation dramatique de l'ensemble des espèces virales (cf. 1.1.2.2: [Importance de la virologie]), mais à cela s'ajoute un fort biais d'étude en faveur des pathogènes humains, ainsi que les agents responsables de pertes économiques. À titre d'exemple, une recherche sur le virus de la grippe dans la base de données GenBank (cf. 1.2.3.2: [Génomes de référence et séquences d'apprentissage]) offre 304 900 résultats tandis que le Clavavirus, un virus de classe I (cf. Fig. 1.14: [Classification de Baltimore]) affectant *Aeropyrum pernix*, une archée vivant obligatoirement en milieu salé à des températures proches de 100°C (SAKO et collab. [1996]), ne bénéficie que d'une unique entrée (oct. 2015).

Ces disparités sont d'autant plus difficiles à surmonter que les virus étant des parasites obligatoires, les contraintes de culture sont nécessairement nombreuses.

En effet, si l'hôte est lui-même difficile à isoler, ou si le titrage intracellulaire du virus est faible et qu'il ne possède pas de phase extracellulaire, comme c'est le cas des virus à mode de vie persistant (cf. 1.1.2.1: [Les virus]), leur étude représente un défi dont la difficulté impacte directement leur représentation dans les bases de données publiques.

3.1.2 Génomes viraux complets : une denrée rare

Les génomes viraux complets sont difficiles à obtenir. Nous avons déjà vu plusieurs raisons à cela : leur culture souffre de nombreuses contraintes, et leurs modes de vie ne permettent pas toujours leur visibilité avec les méthodes actuelles, ni de justifier l'étude approfondie d'espèces ne présentant pas d'intérêt médical ou industriel.

Fort heureusement, le développement de la métagénomique apporte un début de solution à la difficulté de détection et d'identification de nouveaux virus. En revanche, le manque de génomes viraux de référence représente un obstacle majeur au développement de solutions robustes, car la détection de génomes en métagénomique repose massivement sur la comparaison des séquences présentes dans un échantillon aux génomes déjà connus et identifiés. Ainsi, les génomes viraux inconnus d'espèces non cultivables qui ne peuvent être comparées aux espèces de référence (car elles en sont trop éloignées) restent un défi majeur pour l'effort systématique actuel, et créent des lacunes importantes dans l'ensemble des connaissances actuelles en virologie.

Il s'agit d'un problème qui s'auto-alimente en permanence et ne pourra être surmonté qu'au fur et à mesure des découvertes de nouvelles espèces virales, tant à travers l'amélioration des méthodes de détection en métagénomique que par l'effort continu de développement de nouvelles méthodes de culture et d'analyse.

Ainsi, la quantité de génomes viraux reste mineure par rapport au vivant. Il s'agit d'une contrainte qui, parce qu'elle ne bénéficie pas de solution immédiate, doit être soigneusement prise en considération tant lors du développement de nouvelles méthodes d'assignation taxonomique que lors de l'analyse des résultats obtenus par les méthodes existantes.

3.2 Difficultés structurelles propres aux virus

3.2.1 Absence de gène ubiquitaire

Chez le vivant, les gènes ubiquitaires offrent des points de repère fiables. Ils permettent de comparer entre elles même les espèces les plus éloignées (cf. 1.1.4.1:

[Distance évolutive et homologie]). Il n'est donc pas rare de découvrir de nouvelles souches ou espèces vivantes grâce à des outils d'alignement de séquences parmi des jeux de données métagénomiques, et de les replacer sur l'arbre phylogénétique du vivant en calculant leur distance évolutive par rapport aux espèces connues.

En revanche, les virus ne possèdent pas de gène ubiquitaire. Les outils de base permettant leur reproduction appartiennent à l'hôte qu'ils infectent (cf. 1.1.2.1: [Les virus]). Par conséquent, il n'est pas possible de faire une recherche exhaustive de toutes les espèces virales en présence dans un échantillon à partir d'un ensemble de gènes marqueurs. En effet, rien n'indique que toutes les espèces virales présentes partagent au moins un gène homologue aux gènes viraux connus.

Il est néanmoins possible d'effectuer ce type de recherche au sein de certaines familles spécifiques. Par exemple, les rétrovirus (classe VI, cf. Fig. 1.14: [Classification de Baltimore]) possèdent tous la particularité de se répliquer grâce à une étape de transcription inverse. Par conséquent, comme l'hôte ne possède pas de transcriptase inverse, il est possible de les identifier grâce à ce gène.

3.2.2 Ambiguïtés entre virus et hôte

Certains génomes viraux s'insèrent dans les chromosomes de l'hôte. En effet, nous avons vu que certains virus possèdent un mode de vie endogène (cf. 1.1.2.1: [Les virus]). Ils sont donc présent physiquement au sein des chromosomes lorsque le génome de leur hôte est séquencé et assemblé. Par conséquent, parmi les génomes de référence des espèces vivantes, de nombreuses séquences virales sont présentes et constituent une source de bruit non négligeable. Les rétrovirus ont longtemps été considérés comme les seuls capables d'acquérir un mode de vie endogène, mais il a été récemment montré que d'autres familles virales en avaient la capacité (HORIE et collab. [2010]), ce qui suppose que le bruit induit affecteraient plus d'espèces que ce qu'il était autrefois admis. En effet, une étude d'autocorrélation de certaines méthodes par composition montrent que les régions correspondant à des prophages, virus endogènes bactériens, présentent une composition très différente du génome qu'ils occupent (BOHLIN et collab. [2008]).

Les virus échangent parfois des fragments de séquences avec d'autres organismes. Il s'agit du phénomène de *recombinaison* (GRIFFITHS et collab. [1999]) qui se produit lorsque deux molécules d'acide nucléique se trouvant à proximité l'une de l'autre échangent une partie de leur matériel (cf. Fig. 3.2: [Recombinaison génétique]). Dans la plupart des cas, ces événements nécessitent une homologie de séquence entre les deux molécules, mais les recombinaisons dites illégitimes, ne nécessitant pas ou peu d'homologie, existent et sont sources de divers réarrangements (e.g. délétions,

insertions, translocations, multiplication du nombre de copies d'un gène, intégration virale).

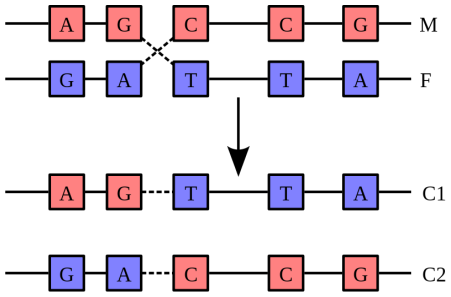


FIGURE 3.2 – Recombinaison génétique : réarrangement entre deux molécules d'acides nucléiques (ARN ou ADN) différentes (M, F), créant de nouvelles combinaisons génétiques (C1, C2). *Source : Wikimedia Commons*

Ce phénomène est un des éléments majeurs dans les mécanismes de l'évolution virale. Il est non seulement source de variabilité au sein d'une même espèce lorsque la recombinaison s'effectue entre deux particules virales, mais des échanges de matériel entre virus et hôte se produisent également (FILÉE et collab. [2008]). Il existe même des cas dans lesquels des virus infectant les eukaryotes ont acquis de nombreux gènes bactériens (FILÉE [2013]).

3.2.3 Hypervariabilité des génomes viraux

Les événements de recombinaison sont particulièrement courant chez les virus. Comme nous l'avons vu ci-dessus, les recombinaisons virales sont multiples et participent activement à l'évolution des espèces, mais elles sont surtout particulièrement fréquentes comparées à ce qu'il peut être observé chez le vivant (ONAFUWA-NUGA et TELESNITSKY [2009], SZTUBA-SOLIŃSKA et collab. [2011], LUKASHEV [2010], MARTIN et collab. [2011]).

Les taux de mutation sont également extrêmement élevés (DRAKE [1999], cf. Fig. 3.2: [Recombinaison génétique]). En effet, les polymérases virales sont beaucoup plus sujettes aux erreurs de réplication que les polymérases cellulaires, surtout chez les polymérases ARN (DUFFY et collab. [2008]). À cela s'ajoute le fait que les mécanismes de réparation cellulaires sont souvent incompatibles avec les structures virales. La plupart de ces erreurs restent donc non corrigées. Le plus souvent, les génomes produits sont déficients, mais entre le grand nombre de particules virales et le phénomène de sélection naturelle, l'apparition de nouveaux variants viables est extrêmement fréquente et rapide (LAURING et collab. [2013], PAFF et collab. [2014]).

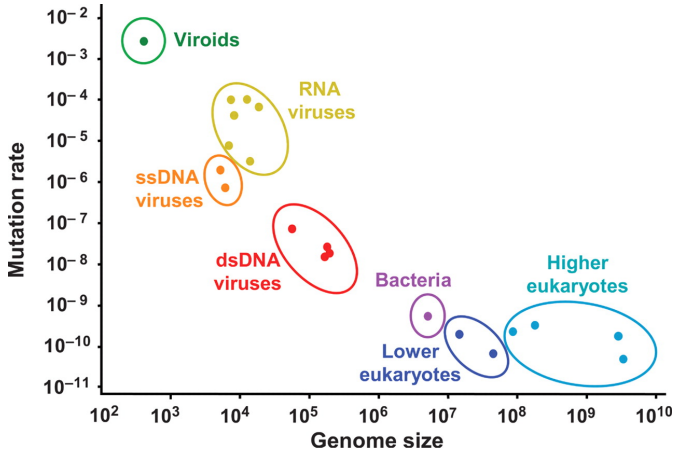


FIGURE 3.3 – Taux de mutation de différents types d'entités par rapport à la taille de leur génome. *Source : (*GAGO *et collab. [2009])*

3.3 Estimation in silico de la difficulté

L'assignation taxonomique est une tâche à multiples dimensions. Elle répond à un besoin simple : caractériser un échantillon en évaluant l'identité et la diversité des espèces en présence (cf. 1.1.3.2: [Métagénomique]). Mais la définition et l'application des différents protocoles permettant de l'effectuer peut varier grandement en fonction des questions de départ qui motivent sa mise en œuvre (cf. 2.1.1: [Définition du problème : Assignation taxonomique]), et les contraintes associées ne sont pas les mêmes.

Comme nous venons de le voir, différents problèmes peuvent entraver l'identification des séquences virales dans un jeu de données métagénomiques. L'absence de gène ubiquitaire, les ambiguïtés entre virus et hôte, ainsi que l'hypervariabilité des géomes viraux sont autant de sources d'erreur qui soulignent la nécessité d'apporter des solutions pour minimiser la fréquence et l'importance desdites erreurs.

Dans ce contexte, il est important de pouvoir quantifier la difficulté de la tâche. C'est dans cet objectif qu'ont été menés les travaux suivants, publiés dans le journal Frontiers in Microbiology (SOUEIDAN et collab. [2014])

3.3.1 Pourquoi la classification par règne est-elle difficile ?

Alors que les méthodes spécifiques à la classification détaillée d'échantillons métagénomiques bénéficient de nombreuses avancées offrant de bonnes performances (cf. 2.1.3: [Classification détaillée de communautés bactériennes et virales]), les méthodes utilisables pour la classification par règne peinent à offrir des résultats fiables sur l'ensemble des espèces (cf. 2.2.3: [Discussion], 2.3.2: [Discussion]). Cela soulève naturellement la question des raisons d'un tel fossé. Puisque la caractérisation

d'échantillons métagénomiques peut être formulée sous la forme d'une tâche d'apprentissage automatique supervisé, nous proposons ici l'utilisation de mesures de complexité des données afin de comparer la difficulté intrinsèque de chacune des approches dans le cadre de la classification d'échantillons métagénomiques.

Nous considérons ici trois tâches dont l'objectif est d'assigner une classe à chaque élément appartenant à un jeu de séquences. Les trois tâches décrites varient par la composition du jeu de séquences et par la portée des classes à assignées.

1. Pour un échantillon de séquences bactériennes donné, assigner chacune d'entre elles à un phylum (e.g. *Proteobacteria*) ou à une classe (e.g. *Gammaproteobacteria*)
2. Pour un échantillon de séquences virales donné, assigner chacune d'entre elles à une classe virale (e.g. dsDNA) ou à une famille (e.g. *Plasmaviridae*)
3. Pour un échantillon de séquences donné, assigner chacune d'entre elles à un règne (e.g. bactérie, archée, eukaryote ou virus)

Les tâches (1) et (2) sont des problèmes de classification détaillée et miment des études métagénomiques ciblées, tandis que la tâche (3) représente une classification par règne et mime l'analyse d'échantillons complexes non ciblés (cf. 2.1.1: [Définition du problème : Assignation taxonomique]). Comme nous nous intéressons à l'identification de nouvelles espèces dans de grands échantillons métagénomiques, nous avons adopté la représentation des séquences sous forme de vecteurs de fréquence de k-mers.

Nous avons analysé ces trois tâches de classification en utilisant une approche point par point de l'analyse de la complexité des données. Dans l'apprentissage automatique supervisé, la performance d'un classifieur dépend non seulement de l'algorithme d'apprentissage (e.g. SVM ou Naïve Bayes) mais également des données d'entraînement. Même si les métriques globales récapitulent les performances générales d'un classifieur, elles ne permettent pas d'indiquer si des performances modérées sont le résultat d'un mauvais ajustement de paramètres, d'un biais d'échantillonnage dans les données d'apprentissage ou de la difficulté intrinsèque de la tâche de classification. Neanmoins, de récentes publications sur la mauvaise classification d'instances de données montrent que pour une tâche de classification donnée, certaines instances de données sont intrinsèquement difficiles à classifier et que leur présence est révélatrice de la difficulté globale (SMITH et collab. [2013]). La plupart des études concordent sur la difficulté que représentent les données aberrantes ou les instances de données appartenant à une classe minoritaire, mais Smith a démontré que des métriques simples permettent vraiment de quantifier la difficulté intrinsèque d'une instance de données. L'une de ces métriques est le *k-Disagreeing Neighbors* (kDN), qui mesure, parmi les k plus proches voisins d'une instance donnée, le nombre d'instances qui ne partagent pas la même classe. Smith a démontré

que la mesure du kDN est très positivement corrélée avec les erreurs de classification d'une instance donnée sur un large panel d'algorithmes d'apprentissage et de rééchantillonnages de données d'entraînement.

Afin de comparer la difficulté de classification des trois tâches, nous avons généré, à partir d'un sous-ensemble représentatif d'organismes séquencés dans GenBank (téléchargement de septembre 2014, 25624 BioProjects, 100% des virus, archées et bactéries, 24 eukaryotes dont 18 plantes) 100 ensembles de 10000 fragments contigus choisis aléatoirement, d'une longueur moyenne de 500nt (correspondant à la taille moyenne de contigs métagénomiques afin de simuler une étape d'assemblage). Pour la tâche (1), seuls les génomes bactériens ont été considérés; pour la tâche (2), seuls les génomes viraux ont été considérés; pour la tâche (3), un mélange équilibré de virus, archées, bactéries et eukaryotes a été considéré. Chaque séquence a été représentée par un vecteur de fréquences d'apparition de 3-mers (i.e. le nombre de fois que chaque sous séquence de 3 nucléotides apparaît dans le contig) et nous avons défini la distance entre deux contigs comme étant la distance euclidienne entre leurs vecteurs à 64 (4^3) dimensions respectifs. Pour chaque contig, sa valeur de kDN est le nombre de contigs qui ne partagent pas la même classe que lui parmi ses 73 plus proches voisins. La difficulté de chaque classe correspondante est ensuite mesurée comme étant la valeur kDN médiane de tous les contigs de la classe donnée. Nous déterminons également si une valeur kDN médiane est significativement extrême (une valeur faible indique une classification facile, une valeur élevée correspond à une classe difficile), en estimant la distribution des valeurs kDN médianes sous l'hypothèse nulle de l'absence de relation entre les classes par des permutations aléatoires.

Dans la figure 3.4: [Distribution des kDN pour les trois tâches], nous résumons la distribution des kDN par classe pour chacune des trois tâches. Les voisins sont déterminés par rapport à la distance euclidienne dans l'espace des fréquences de 3-mers (cf. 3.3.1: [Pourquoi la classification par règne est-elle difficile?]). Par exemple, il y a plus de 6000 contigs archéens différents (barres rouges) qui ne possèdent pas un seul contig non-archéen parmi leurs 73 plus proches voisins (barre rouge correspondant à une valeur kDN de 0). La ligne pointillée représente la limite entre les contigs faciles à classifier correctement avec une majorité de voisins en accord avec eux (à gauche de la ligne) et ceux qui sont difficiles à classifier (à droite).

Le cadre du haut montre que pour la tâche de classification par règne, les contigs archéens et bactériens peuvent être facilement assignés à leur règnes respectifs, que cette classification est difficile pour les contigs eukaryotes, et encore plus difficile pour les virus. Lorsque la tâche est limitée aux bactéries seulement (cadres C1 et C2), la classification détaillée n'est pas difficile autant au niveau du phylum que de la classe. Pour les virus (cadres B1 et B2), la classification détaillée vers les classes virales (ssDNA, dsRNA, etc) est difficile, mais assigner une séquence virale au niveau de la famille est plus facile, même si ce n'est pas aussi aisé que pour les bactéries.

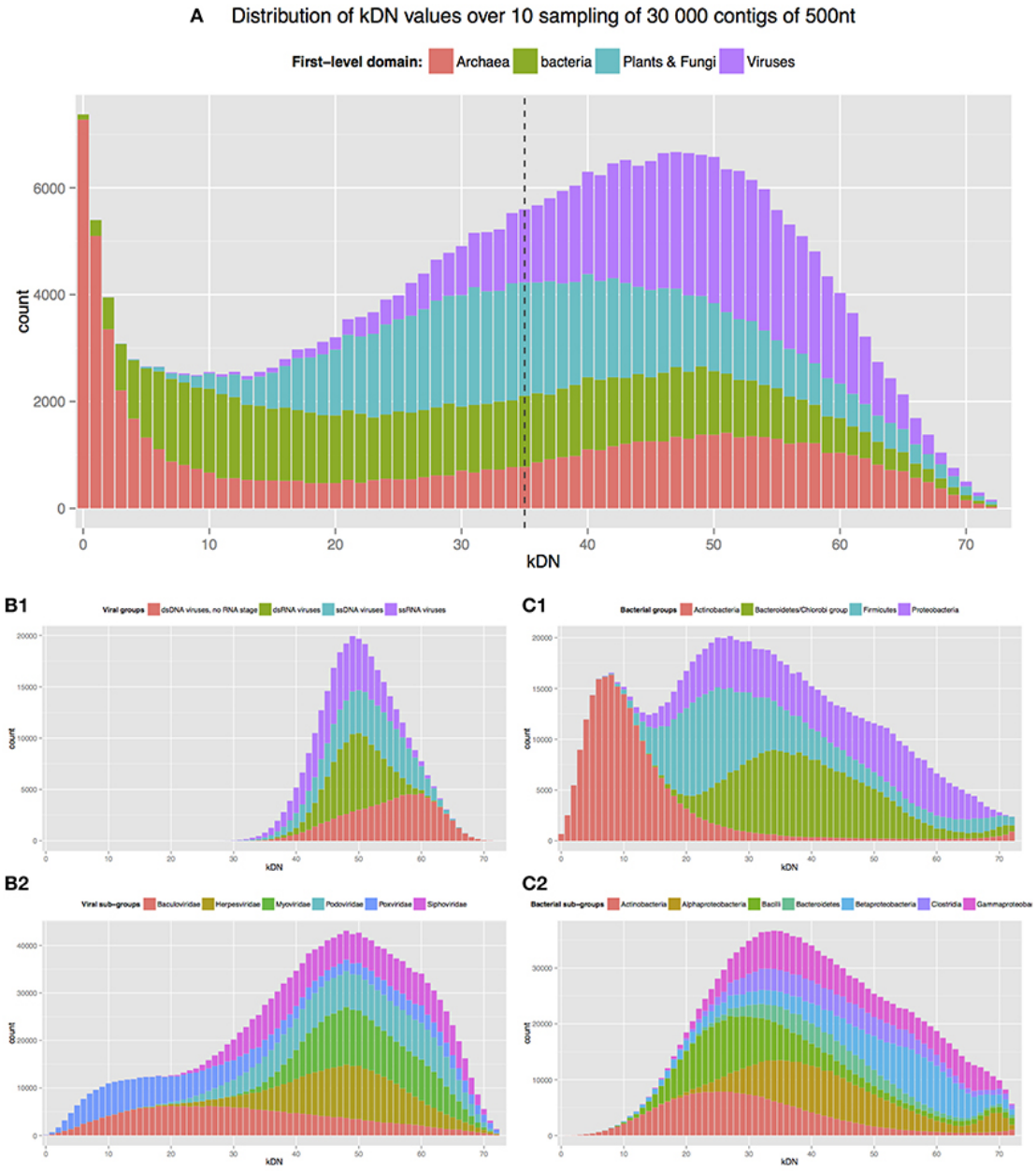


FIGURE 3.4 – Distribution des valeurs kDN par classe pour chacune des trois tâches. (A) Assignation de contigs de 500nt aux grands règnes (tâche 3) ; (B1,B2) Assignation de contigs de 500nt à une classe virale ou à une famille, respectivement (tâche 2) ; (C1,C2) Assignation de contigs bactériens de 500nt à un phylum ou à une classe, respectivelment (tâche 1). Chacun des 300000 contigs aléatoires est représenté par des vecteurs de fréquences de 3-mers. Les histogrammes indiquent combien de contigs (axe y) par classe (couleurs) possèdent un certain nombre de voisins (axe x) ne partageant pas leur propre classe, parmi les 73 plus proches. Seules les 4 classes les plus abondantes sont montrées pour (B1,C1) ; et 6 pour (B2,C2).

En accord avec des travaux antérieurs (MENDE et collab. [2012], TEELING et GLÖCKNER [2012]), nous avons vérifié que pour des contigs plus courts que 500nt, les distributions sont décalées vers la droite, ce qui correspond à une classification plus difficile. Inversement, pour des contigs plus longs que 500nt, les distributions sont décalées vers la gauche, ce qui correspond à une classification plus facile (cf. Fig. 3.5: [Distribution des kDN avec de plus grands contigs]).

Il a été précédemment observé que les signatures virales de 3-mers sont proches de celles de leurs hôtes (PRIDE et collab. [2006]). Néanmoins, des preuves contredisant cette observation ont été également proposées, par exemple pour de grands virus (MRÁZEK et KARLIN [2007]), ainsi que pour les virus des angiospermes (ADAMS et ANTONIW [2004]). Nous avons cherché à savoir si la difficulté de classification pouvait s'expliquer par la superposition des distributions de k-mers entre différents types d'hôtes et les virus qui les infectent. Dans ce but, nous avons échantillonné 4689 contigs depuis de grands groupes cellulaires (des génomes archéens, bactériens, végétaux et fongiques), ainsi que les virus connus comme les infectant. En utilisant une analyse en composante principale (PCA), nous avons projeté les vecteurs de fréquences de 3-mers de ces contigs sur deux dimensions. La Figure 3.6: [Projection 2D des fréquences de 3-mers] montre que les contigs viraux et cellulaires sont étalés uniformément sur ces deux dimensions, à l'exception des virus des plantes qui sont plus compacts. En utilisant une analyse de densité locale (cf. Fig. 3.7: [Projection 2D des fréquences de 3-mers, vue rapprochée]), nous avons observé (ellipses oranges) que les contigs des virus infectant les bactéries sont en effet proches de leurs hôtes (points 12 et 6, 13 et 7), mais qu'ils sont également proches des contigs archéens (points 13 et 3, 2 et 12). D'autre part, les virus des archées ne sont pas proches de leurs hôtes, tandis que les virus des plantes sont plus proches des bactéries et des archées que de leurs propres hôtes.

3.3.2 Discussion

Distinguer les séquences virales et cellulaires dans le cadre d'une étude environnementale non-ciblée reste un problème de classification sans solution à l'heure actuelle, particulièrement pour les espèces virales inconnues. Nous avons montré qu'une des raisons majeures pour lesquelles ce problème n'a toujours pas trouvé de solution satisfaisante est sa difficulté calculatoire intrinsèque. Elle réside dans le fait que les distributions de k-mers des séquences virales se superposent sans distinction avec celles des séquences cellulaires. Cela doit être mis en opposition avec la facilité relative avec laquelle cette tâche s'effectue chez les archées et les bactéries et qui explique certainement le succès des études portant sur l'assignation taxonomique de communautés bactériennes. La difficulté liée à la classification des séquences virales va s'alléger au fur et à mesure que les bases de données publiques de séquences génomiques s'enrichissent de données virales, mais cela ne résoudra pas suffisamment

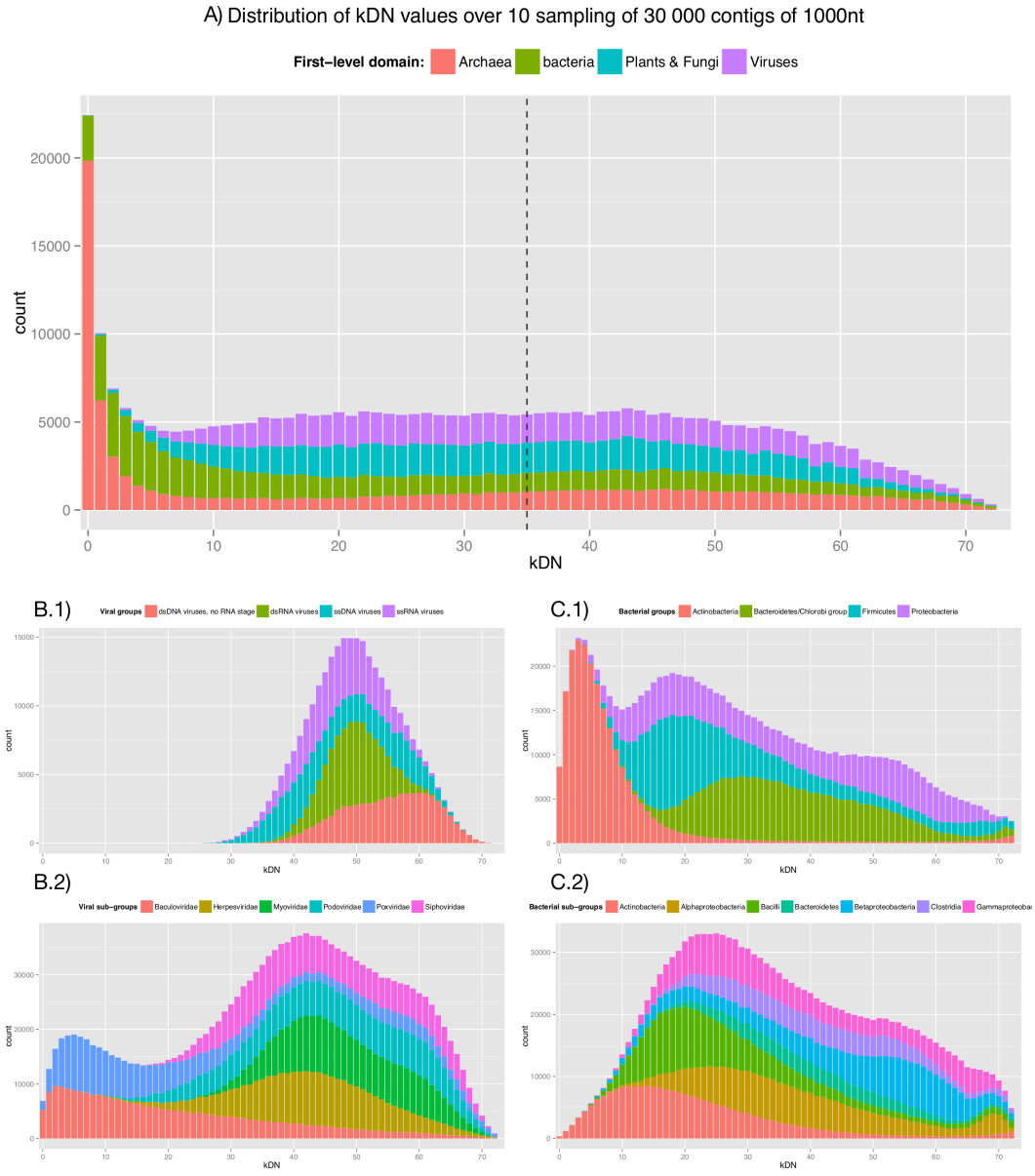


FIGURE 3.5 – Distribution des valeurs kDN par classe pour chacune des trois tâches, pour des contigs de 1000nt. Les cadres correspondent aux cadres de la Figure 3.4: [Distribution des kDN pour les trois tâches].

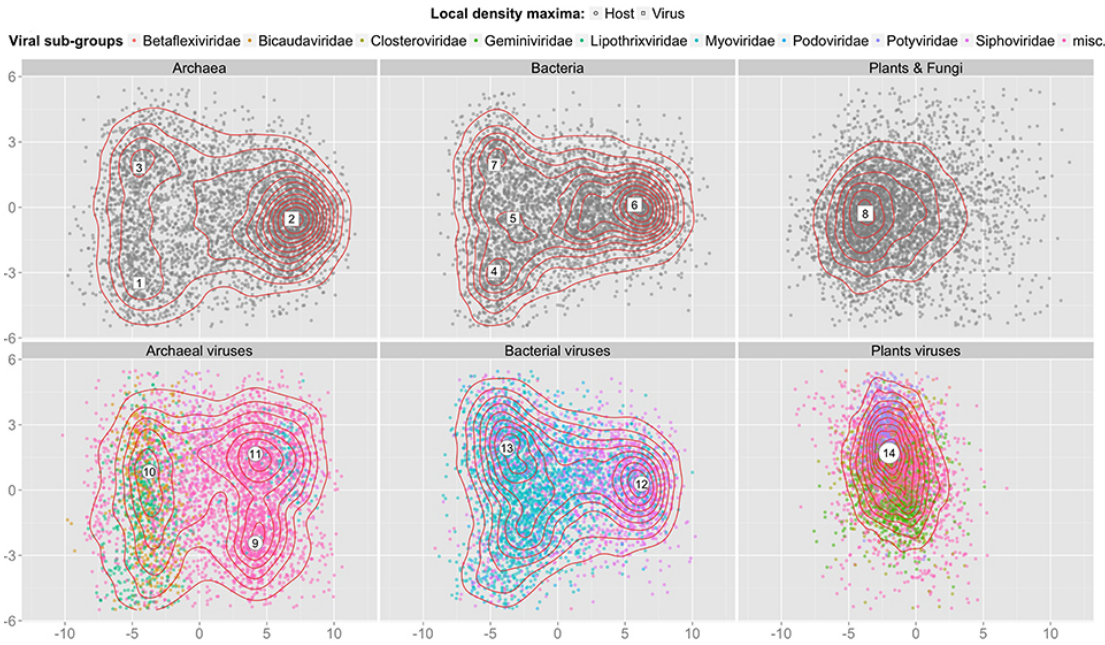


FIGURE 3.6 – Projection 2D des fréquences de 3-mers pour les contigs cellulaires et viraux. Les 2 premières dimensions de la réduction par PCA des 28134 contigs (points) d'une longueur moyenne de 500nt representées par des vecteurs de fréquences de 3-mers; échantillonnées en parts égales depuis des génomes venant des 3 grands règnes cellulaires (rangée du haut) et des grandes familles virales dont on sait qu'elles les infectent (rangée du bas). La dimension 1 (axe x) représente 30% de la variance, la dimension 2 (axe y) 8% de la variance. Pour chaque sous-cadre, l'estimation par noyau sur ces deux dimensions est représentée par des lignes de contour rouges et les maximums de densité locaux sont indiqués dans les pastilles blanches.

FIGURE 3.7 – Vue rapprochée de la Fig. 3.6: [Projection 2D des fréquences de 3-mers] avec tous les maximums de densité locaux. Les composantes principales ont été calculées une fois pour la totalité des contigs de tous les génomes à la fois. Les positions, coordonnées et axes de tous les cadres sont comparables.

le problème de la découverte de nouvelles espèces.

Nous pensons sincèrement qu'un choix minutieux des méthodes informatiques utilisées, ainsi que des efforts supplémentaires de recherche dans cette direction sont les clefs du progrès dans ce domaine. En l'état actuel des connaissances, nous recommandons l'adoption d'une stratégie d'assemblage à l'état de contig, combinée avec une analyse basée sur des fréquences de k-mers, pour l'identification des séquences virales dans des échantillons métagénomiques. En ce qui concerne le développement de nouvelles méthodes, l'allègement de la rigueur de l'indexation de longs k-mers semble être une piste prometteuse (cf. 2.2.3: [Discussion]).

Chapitre 4

Classification supervisée appliquée à la métagénomique virale

Sommaire

Ce chapitre présente les travaux effectués au cours de cette thèse, dans le cadre de la recherche de signatures efficaces afin de traiter des données métagénomiques en virologie. Il traite de la problématique de l'isolation des séquences virales ainsi que de leur identification.

4.1 Problématique

4.1.1 Cadre général

4.1.1.1 Rappel des objectifs

Comme nous l'avons vu précédemment (cf. 1.1.2.2: [Importance de la virologie], 2.1.1: [Définition du problème : Assignation taxonomique]), nous nous intéressons dans ces travaux à l'identification des séquences virales au sein d'échantillons métagénomiques. Cette identification a pour but de faire l'inventaire des espèces virales en présence dans un échantillon donné, mais également de découvrir d'éventuelles nouvelles espèces afin d'enrichir les connaissances actuelles dans le domaine. Nous avons également vu (cf. 1.2.2.1: [Principe et spécificités], 2.1.1: [Définition du problème : Assignation taxonomique]) qu'il s'agit d'un problème de classification multiclasse qui peut être traité par des algorithmes de classification supervisée de chaînes de caractères.

Il s'agit d'un problème de classification difficile de par la nature des données manipulées (cf. 3: [Identifier des séquences virales : une tâche difficile]) qui ne bénéficie actuellement pas de solution satisfaisante. De nombreux outils existent (cf. 2: [État de l'art]), mais n'ont généralement pas été développés pour traiter des séquences virales et ne prennent par conséquent pas en compte les contraintes de ce type de séquence.

Il existe donc un besoin réel pour un outil spécialisé. Cet outil doit répondre spécifiquement aux contraintes du monde viral et doit être basé sur des postulats qui lui sont spécifiques et/ou compatibles. Il est nécessaire que la méthode utilisée soit développée et testée sur des séquences virales en sachant que, si les méthodes adaptées au vivant ne fonctionnent pas correctement sur des données virales, il est également parfaitement possible que la réciproque soit vraie.

Parmi les contraintes à prendre en compte, il y a la séparation du problème en deux sous-problématiques distinctes : la classification par règne et la classification détaillée (cf. 2.1.1: [Définition du problème : Assignation taxonomique], 4.1.2: [Deux objectifs complémentaires]). Les deux étant complémentaires et répondant à des étapes distinctes du processus d'identification des séquences virales, il est important

de les traiter séparément afin de s'assurer que l'efficacité de la classification ne souffre pas du manque de spécificité d'une méthode générique.

Nous cherchons ici à développer un ensemble complémentaire de méthodes candidates, qui améliore les performances de l'assignation taxonomique de séquences métagénomiques virales.

4.1.1.2 Choix du type de signatures

Nous avons vu que les outils basés sur l'alignement de séquences (cf. 2.2: [Classification par similarité]) étaient fondamentalement inadaptés à la recherche de séquences virales car leur fonctionnement est basé sur un certain nombre de postulats (homologie forte, ordre des gènes très conservé) qui ne trouvent pas d'écho dans le monde viral. Pour cette raison, nous avons décidé d'explorer uniquement les méthodes de comparaison de séquences par composition (cf. 2.3: [Classification par composition]).

Il faut décomposer les séquences en sous-éléments constitutifs. La méthode la plus largement utilisée pour définir quels sont ces éléments de base est le découpage en k-mers (cf. 2.3.1: [Distributions de k-mers et signatures]). En revanche, il est parfaitement possible de choisir d'autres règles permettant de définir ces sous-éléments. En effet, les k-mers étant des sous-séquences chevauchantes et de taille constante, leur découpage ne répond à aucune réalité biologique et peut, par conséquent être la cause d'une perte d'information.

Pour cette raison, nous avons exploré une autre méthode de découpage des séquences : la segmentation des séquences selon leur composition (BRAUN et MÜLLER [1998]). Il s'agit d'un type d'approche permettant d'effectuer un découpage non chevauchant de chaînes de caractères de manière à ce que les segments obtenus conservent une certaine homogénéité de composition. L'objectif est de conserver un maximum de motifs informatifs sur l'origine et la fonction des séquences manipulées. En effet, de nombreux motifs spécifiques à une fonction présentent une redondance caractéristique de certains nucléotides (e.g. éléments cis-régulateurs tels que la boîte TATA ou la boîte de Pribnow, cf. Fig. 4.1: [Éléments cis-régulateurs]).

Nous avons donc étudié la composition des séquences selon une méthode de segmentation bayésienne (RAMENSKY et collab. [2000], MAKEEV et collab. [2001]). Malheureusement, nous nous sommes heurtés à deux limitations majeures de cette approche. La première est que seule une partie des petits segments obtenus sont utilisables. En effet, de nombreux segments sont trop longs, donc trop spécifiques et couvrent une grande partie des génomes utilisés, causant ainsi une perte conséquente d'information (cf. Fig. 4.2: [Comparaison des segments obtenus entre les génomes viraux de la famille des Bunyaviridae et des génomes aléatoires]). La seconde limitation concerne la stabilité du découpage. À partir de plusieurs séquences, nous

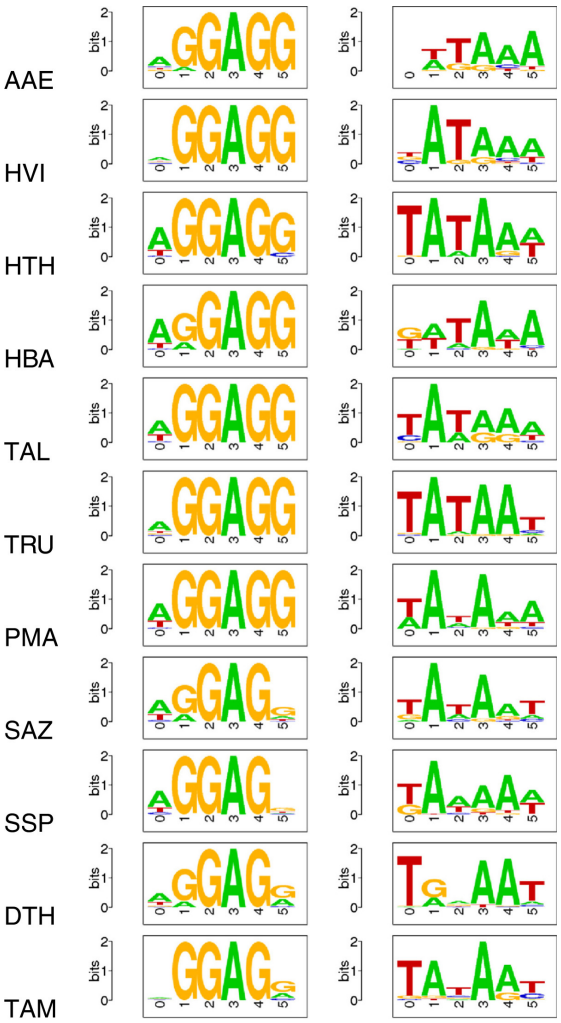


FIGURE 4.1 – Exemple d'éléments cis-régulateurs : composition de ces éléments à travers le génome de 11 espèces bactériennes. À gauche : la séquence Shine-Dalgarno, ou site de fixation du ribosome. À droite : la boîte de Pribnow, séquence promotrice de la transcription des gènes. *Source :* LECHNER *et collab.* [*2014*]

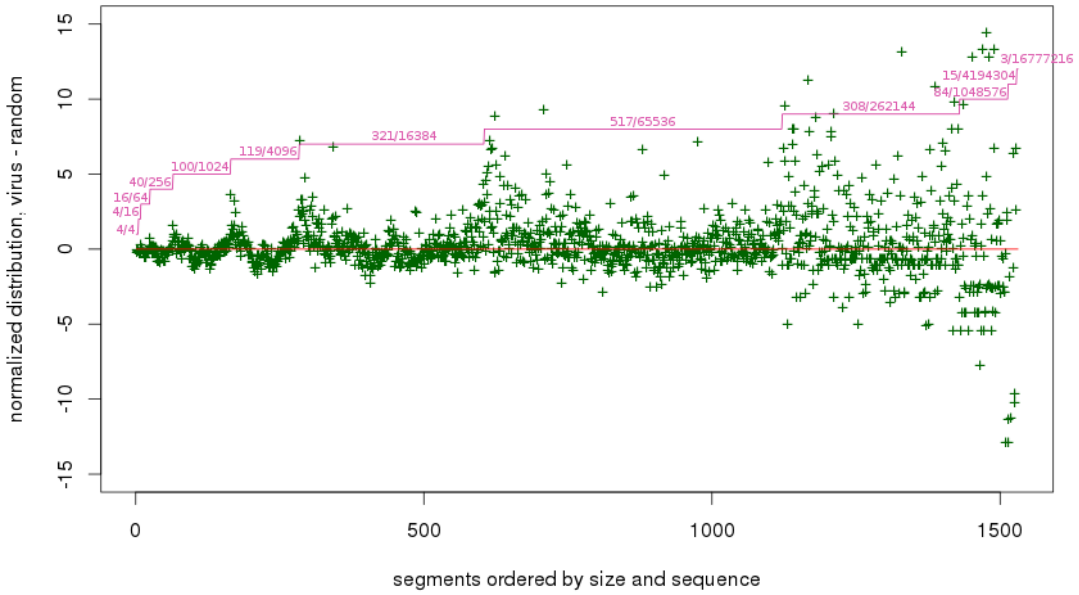


FIGURE 4.2 – Comparaison des segments communs obtenus entre les Bunyaviridae et des génomes aléatoires : X : segments ordonnées par taille puis par ordre alphabétique / Y vert : différences des distributions normalisées (par la taille des génomes et la taille des segments) dans l'ordre virus-aléatoire / Y violet : taille des segments avec, en annotation, le nombre de segments différents trouvés sur le nombre total de possibilités théoriques sur un alphabet de 4 lettres / Y rouge : y=0. Une partie des données dépasse le cadre du graphe.

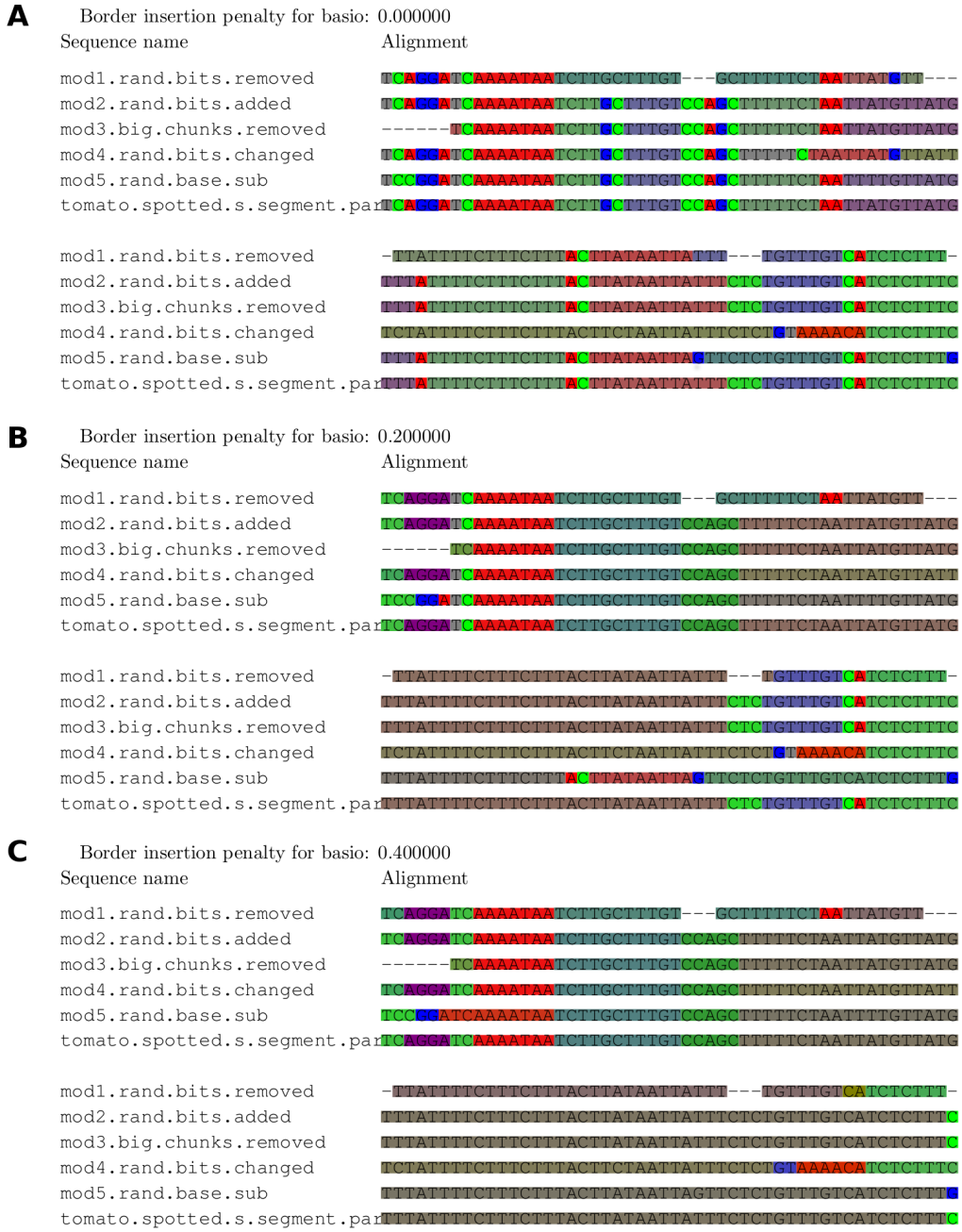


FIGURE 4.3 – Résultats de la segmentation bayésienne d'une partie du génome du virus de la maladie bronzée de la tomate. La séquence d'origine a été altérée de six manières différentes puis alignée (ligne du bas) avec les versions altérées avec ClustalW (lignes 1-5). Les différents segments obtenus sont colorés selon leur composition en nucléotides. La même séquence a été segmentée avec une pénalité d'insertion de bordure de 0.2 (B), 0.4 (C) ou sans pénalité (A).

avons artificiellement introduit des variations (insertions, délétions, substitutions) afin d'observer la robustesse des résultats (cf. Fig. 4.3: [Segmentation bayésienne]). Nous avons pu observer que même de petites variations pouvaient modifier la position des découpes sur plusieurs segments autour de la position modifiée. Nous avons contacté les auteurs à ce propos, qui ont confirmé ce comportement sur des segments de petite taille et conseillent une pénalité de découpage plus élevée pour obtenir des résultats robustes. Or, les segments obtenus avec une telle pénalité sont d'un ordre de taille comparable aux contigs obtenus par assemblage et ne permettent pas de découper ces derniers en sous-éléments constitutifs.

Nous avons donc choisi d'utiliser la composition des séquences en k-mers. Ainsi, chaque séquence est représentée par un tableau de valeurs de taille 4^k. En sachant que les données manipulées sont des contigs, dont la taille est variable, il est nécessaire de normaliser le nombre d'occurrences de chaque k-mer par la taille de la séquence. Nous obtenons ainsi pour chacune d'entre elles un tableau de fréquences (cf. Fig. 4.4: [Signatures génomiques par fréquences de k-mers : deux virus]).

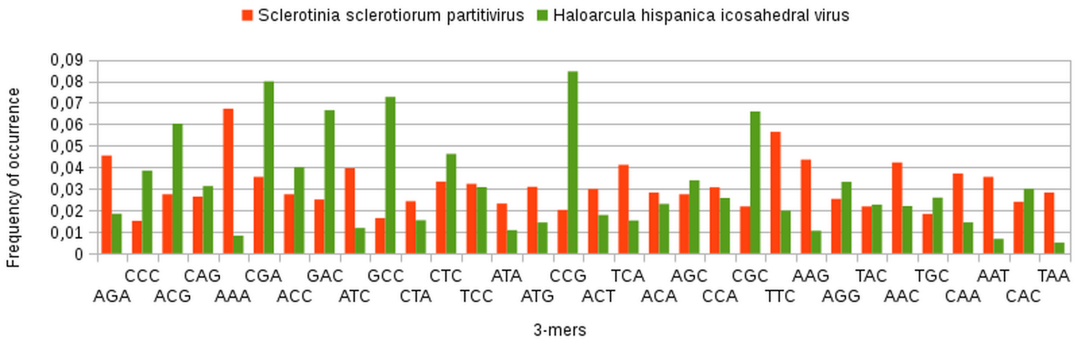


FIGURE 4.4 – Comparaison de deux signatures génomiques par fréquences de k-mers virales. L'ordre des k-mers représentés dans la figure est arbitraire. Tous les k-mers ne sont pas représentés : seuls sont présents ceux qui ont une fréquence non nulle chez les deux espèces.

4.1.1.3 Données d'entrée et rééchantillonnage

Afin d'effectuer l'apprentissage, il est nécessaire d'avoir des génomes dont l'origine est connue. Dans le but de minimiser les éventuelles incompatibilités et d'assurer la pérennité du système, nous avons choisi d'utiliser les bases de données du NCBI, qui bénéficient de l'unification permanente de l'INSDC (cf. 1.2.3.2: [Génomes de référence et séquences d'apprentissage]).

Les données à classifier sont des contigs. Il s'agit du produit de l'assemblage de reads chevauchants qui permettent d'obtenir des séquences plus longues (cf. 1.2.3.1:

[Données génomiques : nature et hétérogénéité]). Cependant, l'ensemble des biais structurels et expérimentaux est incompatible avec l'assurance d'un assemblage parfait. En effet, certaines ambiguïtés entre génomes ne permettent pas toujours de distinguer les chevauchements dûs à une origine commune de ceux qui sont issus d'une homologie fortuite, à plus forte raison sur des reads courts. De plus, la couverture génomique du séquençage n'est pas homogène et peut laisser des régions entières non séquencées, ou trop peu représentées pour effectuer un assemblage robuste. Par conséquent, les contigs sont de tailles très variables (cf. Fig. 4.5: [Distribution des tailles de contigs métagénomiques]) dont la plupart se trouvent autour de 500 nt, mais dont certains peuvent largement dépasser 10000 nt.

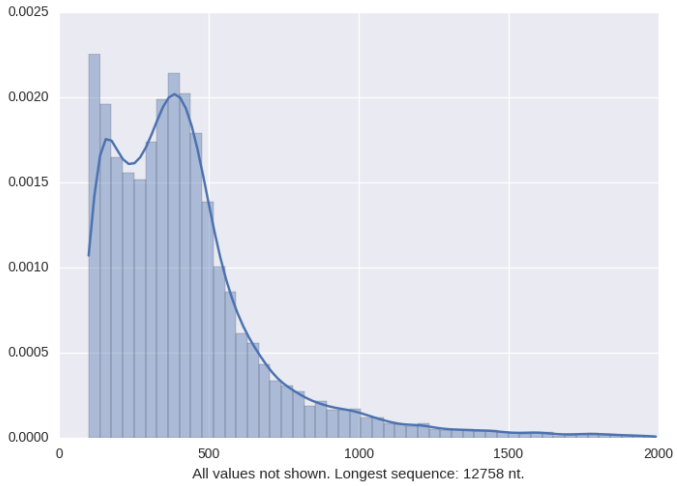


FIGURE 4.5 – Distribution relative des tailles de contigs obtenus après assemblage de séquences issues du séquençage d'un échantillon filtré expérimentalement pour ne conserver que les petits éléments d'un ordre de grandeur correspondant à la plupart des virions. La courbe correspond à l'estimation de la densité de probabilité par la méthode de Parzen-Rosenblatt. *Origine : INRA UMR1332, Îles Kerguelen*

Il s'agit d'une contrainte à prendre en compte lors de la construction des données d'apprentissage, car la composition des génomes n'étant pas homogène sur toute la longueur des chromosomes, il est impossible de comparer de manière fiable la composition d'un génome entier à celle d'une petite partie de ce dernier.

Nous avons donc choisi d'effectuer un rééchantillonage. Afin d'entraîner le classifieur sur des données comparables aux données à classifier, nous construisons les jeux de données d'apprentissage en extrayant des sous-séquences des génomes de référence, de longueurs comparables à ce qu'il est possible de trouver parmi des contigs

(cf. 4.3.2.1: [Rééchantillonage]). Nous obtenons ainsi un ensemble de "contigs artificiels".

4.1.2 Deux objectifs complémentaires

4.1.2.1 Classification par règne

La classification par règne a pour objectif de déterminer à quel grand groupe taxonomique appartient une séquence (cf. 2.1.1: [Définition du problème : Assignation taxonomique]). En effet, les échantillons métagénomiques contiennent du matériel génétique d'origines très variées, où tous les grands règnes sont souvent représentés, ce qui constitue un volume conséquent de données à traiter. Il est donc important, lorsqu'on s'intéresse exclusivement à un groupe particulier d'organismes, de filtrer le reste des populations présentes afin de les exclure des données à analyser.

En virologie, ce filtrage est souvent effectué en amont. Les échantillons biologiques sont filtrés expérimentalement afin de ne récupérer que les particules dont la taille est comparable à la plupart des virions. Mais cette technique possède ses limites. Certaines espèces bactériennes et virales sont de taille similaire, ce qui implique que le filtrage ne permettra pas la récupération de certains virus, et/ou ne pourra pas empêcher la pollution des données par des petits organismes unicellulaires (cf. 2.1.2: [Classification par règne d'échantillons complexes]). De plus, le filtrage ne permet pas d'avoir accès aux virus dont le mode de vie est persistant par exemple, car ces derniers ne possèdent pas de phase extracellulaire (cf. 1.1.2.1: [Les virus]).

Il est donc important de pouvoir trier les données par règne. Que ce soit parce que le filtrage n'est pas fiable à 100% ou parce qu'il n'est simplement pas souhaité, il existe un réel besoin d'isoler les séquences du grand groupe taxonomique d'intérêt, à plus forte raison en virologie où il existe une ambiguïté non négligeable entre virus et vivant qu'il est difficile de contourner (cf. 3.2: [Difficultés structurelles propres aux virus]).

4.1.2.2 Classification détaillée

L'objectif de la classification détaillée est de fournir des informations sur la taxonomie de l'organisme d'origine de chaque séquence (cf. 2.1.1: [Définition du problème : Assignation taxonomique]). Cependant, la taxonomie virale représente un défi à plusieurs niveaux. Entre les taux de mutations élevés, l'absence de gènes ubiquitaires (cf. 3.2: [Difficultés structurelles propres aux virus]), la pauvreté des génomes viraux connus, les biais de représentations de certains virus dans les bases de données publiques et le fait que les taxonomies virales ne sont pas entièrement

construites sur l'histoire évolutive des espèces (cf. 3.1: [Données de référence]), la classification détaillée de séquences virales pose encore de nombreux problèmes.

Dans ce contexte, il est par définition illusoire d'espérer classifier la plupart des séquences virales au niveau de l'espèce. En revanche, il est très informatif de voir s'il est possible d'attribuer une séquence virale à un niveau taxonomique un peu plus élevé et, ensuite, traiter les séquences au cas par cas, soit parce qu'elles ont été attribuées à une famille d'intérêt, soit parce qu'elles n'ont pas pu être attribuées à une famille et peuvent être issues d'une espèce virale inconnue. Ces étapes d'affinage des résultats requièrent une analyse des gènes qui nécessite l'exploitation d'annotations génomiques qu'il serait trop long à automatiser ici.

4.2 Workflow général

Ces deux objectifs, bien qu'abordés séparément, possèdent un mécanisme général commun. Il font tous deux l'objet d'un workflow similaire (cf. Fig. 4.6: [Workflow général]) qui a subi de nombreuses évolutions au cours de cette thèse.

4.2.1 Préparation des données de référence

La première étape du workflow (n°1 Fig. 4.6: [Workflow général]) consiste à télécharger les données de référence depuis la base de données publique utilisée. Il s'agit de récupérer les génomes à partir desquels le modèle est entraîné, mais également les informations relatives à la taxonomie de ces génomes. Le cas échéant, il peut être également nécessaire de nettoyer ces données, afin de supprimer les incohérences (e.g. les génomes dont l'origine ne correspond pas à l'organisme annoncé dans les métadonnées).

Ensuite, ces données sont organisées de manière logique dans un format exploitable par le reste du workflow, afin que les génomes soient accessibles à partir de leur origine taxonomique.

Enfin, nous rééchantillonnons les données taxon par taxon (n°2 Fig. 4.6: [Workflow général]) pour obtenir un jeu de séquences dont la longueur est de l'ordre du contig (500 nt en moyenne) de sorte que chaque taxon au niveau choisi pour constituer les classes soit représenté par le même nombre de séquences.

4.2.2 Apprentissage et classification

Nous représentons toutes nos séquences par des vecteurs de fréquences de k-mers (cf. 4.1.1.2: [Choix du type de signatures]). Une étape nécessaire consiste donc à extraire ces signatures pour chaque séquence utilisée, que ce soit les séquences destinées à l'apprentissage ou celles qui sont à classifier (n°3 Fig. 4.6: [Workflow général]).

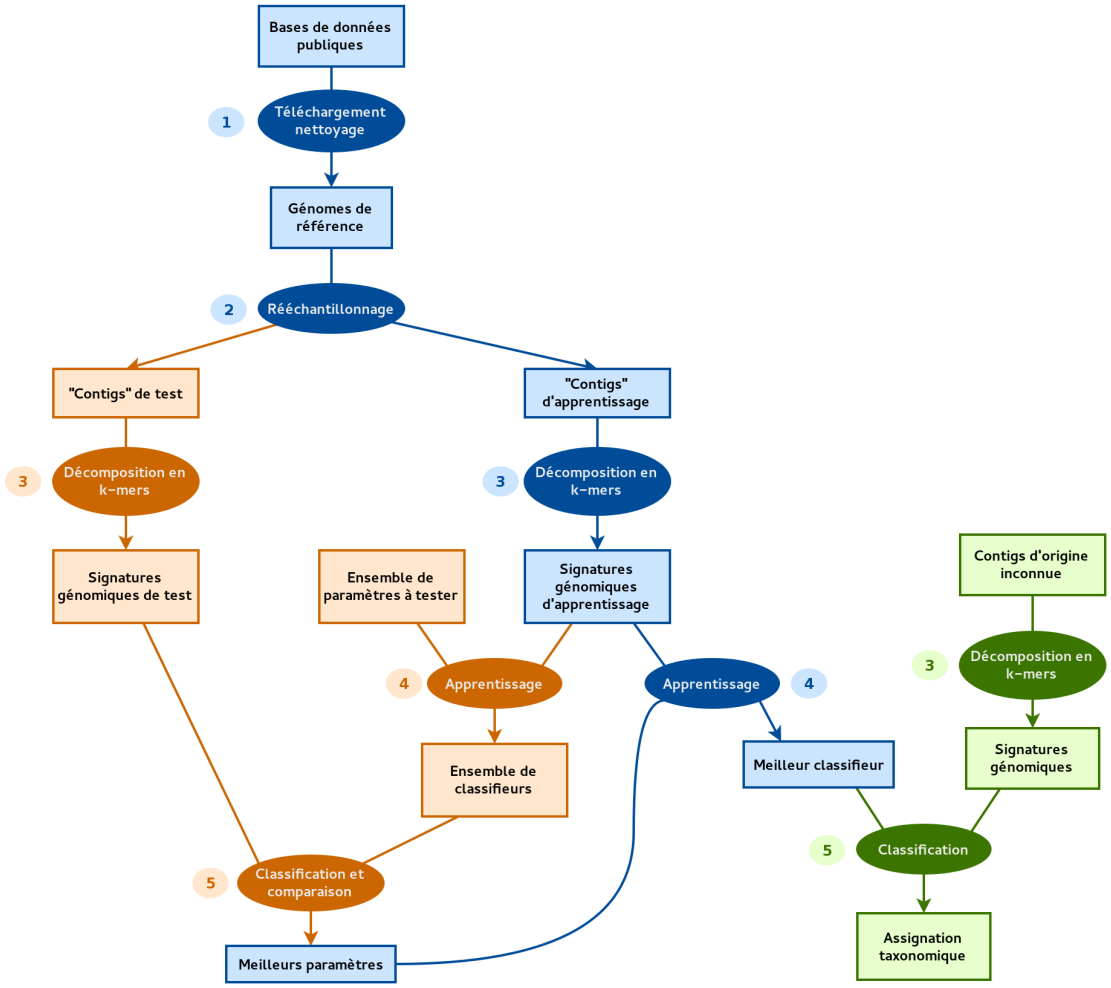


FIGURE 4.6 – Workflow général, applicable aux deux types de classification abordés. En bleu sont représentées les étapes qui sont exclusives à la création du classifieur pour une base de données spécifique. En orange sont les étapes qui sont exclusives à la recherche d'un ensemble de paramètres pour un nouvel ensemble de données publiques ou pour optimiser un ensemble de paramètres existants, en cas de mise à jour de la base de données utilisée par exemple. En vert sont les étapes relatives à l'utilisation finale du classifieur.

Ces signatures, qui sont des vecteurs de 4^k valeurs, peuvent être considérées comme des coordonnées dans un espace à 4^k dimensions.

C'est dans cet espace que nous effectuons notre apprentissage (n°4 Fig. 4.6: [Workflow général]), à partir des points obtenus depuis nos contigs artificiels. Chaque classifieur issu de cette étape d'apprentissage doit permettre ensuite d'associer un taxon à chaque séquence inconnue, à partir de la position dans l'espace de sa signature (n°5 Fig. 4.6: [Workflow général]).

4.2.3 Méta-apprentissage

Quel que soit l'algorithme utilisé, l'efficacité d'un classifieur dépend des données d'apprentissage et des paramètres utilisés lors de sa construction. En effet, un ensemble de paramètres peut être très adapté à un jeu de données mais être complètement inefficace pour un autre. C'est pourquoi il est nécessaire de prévoir la possibilité de trouver des paramètres efficaces pour un ensemble de données d'apprentissage. Il s'agit d'un ensemble d'étapes supplémentaires (étapes en orange Fig. 4.6: [Workflow général]) qui doivent être exécutées en cas de doute sur la validité des paramètres (en cas de mise à jour des données par exemple) ou dans le cas d'un changement de base de données.

Pour réaliser cette tâche, il faut déterminer un ensemble de valeurs à tester pour chaque paramètres et créer un classifieur pour chaque combinaison de valeurs. Un jeu de contigs artificiels supplémentaire, différent du jeu d'apprentissage, est créé afin de tester l'efficacité de chaque classifieur ainsi créé. La combinaison de valeurs utilisée par le classifieur présentant les meilleurs performances est retenue comme étant l'ensemble des meilleurs paramètres pour les données utilisées et sera choisie pour la construction du classifieur final.

4.3 Classification par règne

4.3.1 Données de référence

4.3.1.1 Bases de données

Les séquences de référence utilisées ici proviennent de plusieurs bases de données du NCBI :

— **Les données virales** sont un téléchargement complet du contenu de la partie virale de la base de données "genomes" du NCBI (`http://www.ncbi.nlm.nih.gov/genome/viruses`)

— **Les données non virales** sont un assortiment de génomes choisis de manière à avoir une quantité de données réduite afin de pouvoir effectuer des tests

rapides, tout en couvrant au maximum et de la manière la plus homogène possible tous les grands taxons du vivant. La liste complète est disponible dans le Tableau 4.1: [Séquences utilisées pour représenter les grands règnes du vivant.]. La base de données GOLD (`https://gold.jgi.doe.gov`) a été utilisée afin de faciliter ce choix.

La taxonomie n'a pas été détaillée pour les grands règnes du vivant. Les virus ont été séparés par catégorie de Baltimore puis par famille par des répertoires sur le système de fichiers sur lequel ils sont stockés. En effet, la taxonomie importe peu en deçà des grands règnes ici. Nous y attacherons en revanche un soin particulier dans la partie 4.4: [Classification détaillée].

4.3.1.2 Filtration des données virales

La classification par règne possède des contraintes spécifiques qui sont dues à la présence de séquences issues du vivant (cf. 3.2.2: [Ambiguïtés entre virus et hôte]). En effet, ces ambiguïtés sont source de bruit et ajoutent à la difficulté de la tâche.

Par conséquent, il y a naturellement de nombreuses séquences vivantes qui sont classifiées en tant que virus et vice-versa. Nous avons donc cherché un moyen de réduire ces erreurs.

Malheureusement, les ambiguïtés sont inévitables. Il est déraisonnable d'imaginer pouvoir améliorer les résultats pour les séquences virales sans impacter l'efficacité de la classification des séquences issues du vivant. Cela nous a amenés à réexaminer les objectifs que nous nous sommes fixés.

Ces travaux sont motivés par le besoin d'identifier les espèces virales en présence dans un échantillon (cf. 1.1.2.2: [Importance de la virologie]). L'objectif n'est donc pas d'obtenir une classification la plus juste possible, mais d'extraire les séquences virales à partir de données mixtes. Cela implique un certain degré de liberté concernant la qualité des résultats parmi les classes représentant l'ensemble du vivant. En revanche, la pollution des virus par des séquences appartenant au vivant est plus problématique.

En étant plus strict sur la classification virale et en favorisant les classes du vivant en cas d'ambiguïté, de nombreuses séquences virales seront nécessairement assignées à la mauvaise classe. En revanche, l'assignation d'une séquence à la classe des virus sera beaucoup plus fiable et reflétera davantage la réalité biologique du milieu d'origine.

De plus, il est tout à fait possible que, pour un virus donné, l'ambiguïté ne soit que partielle et qu'un fragment de son génome ne possède pas de similarité gênante pour sa classification. Dans ce cas, l'identification des fragments issus de cette espèce ne sera pas exhaustive, mais elle sera néanmoins représentée dans les résultats.

TABLEAU 4.1 – Séquences utilisées pour représenter les grands règnes du vivant.

Espèce	Sous-groupe	Numéros d'identification au NCBI
Eucaryotes :		
Plantes :		
Arabidopsis thaliana	dicotylédone	CP002684.1, CP002685.1, CP002686.1, CP002687.1, CP002688.1, Y08501.2, AP000423.1
Oryza sativa	monocotylédone	AP008207.2, AP008208.2, AP008209.2, AP008210.2, AP008211.2, AP008212.2, AP008213.2, AP008214.2, AP008215.2, AP008216.2, AP008217.2, AP008218.2, X15901.1, BA000029.3, D00293.1
Selaginella moellendorffii	lycophyte	NW_003314330.1, NW_003314286.1, NW_003315018.1, NW_003315017.1, NW_003315016.1, NW_003315015.1, NW_003315014.1, NW_003315013.1, NW_003315012.1, NW_003315011.1, NW_003315010.1, NW_003315009.1, NW_003315008.1, NW_003315007.1, NW_003315006.1, NW_003315005.1, NW_003315004.1, NW_003315003.1, NW_003315002.1, NW_003315001.1, NW_003315000.1, NW_003314999.1, NW_003314998.1, NW_003314997.1, NW_003314996.1, NW_003314995.1, NW_003314994.1, NW_003314993.1, NW_003314992.1, NW_003314991.1, NW_003314990.1, NW_003314989.1, NW_003314988.1, NW_003314987.1, NW_003314986.1, NW_003314985.1, NW_003314984.1, NW_003314983.1, NW_003314982.1, NW_003314981.1, NW_003314980.1, NW_003314979.1, NW_003314978.1, NW_003314977.1, NW_003314976.1, NW_003314975.1, NW_003314974.1, NW_003314973.1, NW_003314972.1, NW_003314971.1
Ostreococcus lucimarinus	algue verte	CP000581.1, CP000582.1, CP000583.1, CP000584.1, CP000585.1, CP000586.1, CP000587.1, CP000588.1, CP000589.1, CP000590.1, CP000591.1, CP000592.1, CP000593.1, CP000594.1, CP000595.1, CP000596.1, CP000597.1, CP000598.1, CP000599.1, CP000600.1, CP000601.1
Fungi (champignons) :		
Cryptococcus neoformans	basidiomycète	CP003820.1, CP003821.1, CP003822.1, CP003823.1, CP003824.1, CP003825.1, CP003826.1, CP003827.1, CP003828.1, CP003829.1, CP003830.1, CP003831.1, CP003832.1, CP003833.1, CP003834.1
Sporisorium reilianum	basidiomycète	FQ311430.1, FQ311441.1, FQ311452.1, FQ311463.1, FQ311470.1, FQ311471.1, FQ311472.1, FQ311473.1, FQ311474.1, FQ311431.1, FQ311432.1, FQ311433.1, FQ311434.1, FQ311435.1, FQ311436.1, FQ311437.1, FQ311438.1, FQ311439.1, FQ311440.1, FQ311442.1, FQ311443.1, FQ311444.1, FQ311445.1, FQ311469.1
Encephalitozoon cuniculi	microspore	AL391737.2, AL590442.1, AL590443.1, AL590444.1, AL590445.1, AL590446.1, AL590447.1, AL590448.1, AL590451.2, AL590449.1, AL590450.1
Fusarium oxysporum	ascomycète	CM000589.1, CM000590.1, CM000591.1, CM000592.1, CM000593.1, CM000594.1, CM000595.1, CM000596.1, CM000597.1, CM000598.1, CM000599.1, CM000600.1, CM000601.1, CM000602.1, CM000603.1
Magnaporthe oryzae	ascomycète	CM001231.1, CM001232.1, CM001233.1, CM001234.1, CM001235.1, CM001236.1, CM001237.1
Animaux :		
Apis mellifera	insecte	CM000054.5, CM000055.5, CM000059.5, CM000060.5, CM000058.5, CM000056.5, CM000063.5, CM000061.5, CM000062.5, CM000068.5, CM000057.5, CM000067.5, CM000064.5, CM000065.5, CM000066.5, CM000069.5, L06178.1
Bactéries :		
Escherichia coli	gram-	NC_002695.1
Agrobacterium radiobacter	gram-	CP000628.1, CP000629.1, CP000632.1
Rhizobium leguminosarum	gram-	AM236080.1
Bacillus thurigiensis	gram+	CP001903.1, CP001904.1
Streptococcus pneumoniae	gram+	AE007317.1
Staphylococcus aureus	gram+	HE579073.1, HE579074.1
Mycoplasma genitalium	mycoplasme	CP003773.1
Archées :		
Sulfolobus islandicus	crenarchaea	CP002425.1
Methanococcus mripaludis	euryarchaea	NC_009637.1
Nitrosopumilus maritimus	thaumarchaea	NC_010085.1
Korarchaeum cryptofilum	autre	NC_010482

Dans ce but, nous avons choisi de filtrer les séquences d'apprentissage. Nous utilisons MegaBlast (cf. ZHANG et collab. [2004], 2.2.1: [Alignement de séquences]) afin d'aligner chaque contig viral d'apprentissage sur l'ensemble des génomes entiers des classes du vivant. L'alignement est effectué avec une taille de mot réduite (16nt, minimum recommandé), afin d'améliorer la sensibilité inter-espèces et seuls les hits dont la e-value est inférieure à 0.0001 sont conservés. Ces valeurs permettent un bon compromis entre sensibilité (apparition de hits viraux dans des génomes vivants) et spécificité (hits correspondant à des ambiguïtés sources de bruit).

Les contigs viraux obtenant ainsi des hits sont retirés du jeu d'apprentissage. La construction du classifieur se fait ensuite sur ce jeu réduit.

4.3.2 Spécificités du workflow

4.3.2.1 Rééchantillonage

Le rééchantillonnage est ici effectué classe par classe de la manière suivante : Toutes les séquences de référence de la classe sont concaténées et n positions sont tirées aléatoirement sur toute l'étendue de la concaténation.

Ensuite, pour chaque position p de cet ensemble, une valeur de longueur l est tirée aléatoirement selon une distribution normale tronquée (moyenne : 500, écart-type : 200, min : 100, max : 4000), afin de rester proche d'une distribution expérimentale de tailles de contigs (cf. Fig. 4.5: [Distribution des tailles de contigs métagénomiques]). Si $p+l$ se trouve dans la même séquence d'origine que p, on conserve la sous-séquence entre p et $p+l$. Dans le cas contraire, on incrémente un compteur d'échec e.

Une fois que les n couples $\{p;l\}$, sont traités, il reste e coordonnées à extraire. Si certaines séquences ne sont pas encore représentées par les $n-e$ coordonnées retenues, on ne concatène que ces séquences afin d'y tirer e nouvelles positions et répéter les étapes précédentes. Si toutes les séquences de référence sont déjà représentées, les e nouvelles positions sont tirées sur la concaténation de l'ensemble des séquences de référence.

Ces étapes sont répétées jusqu'à l'obtention de n coordonnées. Ensuite, la séquence délimitée par chaque couple $\{p;l\}$ est extraite, avec une probabilité de 0.5 d'extraire la séquence complémentaire afin que le brin complémentaire soit également représenté (cf. 1.1.1.1: [Structure de la cellule vivante]). En effet, les génomes stockés dans les bases de données ne contiennent qu'un des deux brins de l'ADN, le brin complémentaire pouvant être simplement déduit de la séquence du premier.

Cette méthode ne garantit pas la représentation de tous les génomes, mais elle la favorise fortement. Lorsque n est suffisamment grand pour obtenir une couverture génomique proche de 1 en alignant tous les contigs sur leurs séquences d'origine (cf. 4.7: [Exemple de rééchantillonnage]), nous n'avons pas observé de séquence non

représentée.

FIGURE 4.7 – Exemple de rééchantillonage : alignement des contigs obtenus par le rééchantillonage du génome d'Arabidopsis thaliana (plante modèle). Il s'agit d'une vue partielle du chromosome 1, les contigs alignés sur la séquence de référence sont représentés par des flèches orientées à droite, et ceux qui sont alignés sur le brin complémentaire sont représentés par des flèches orientées à gauche. Visualisé avec le navigateur de génome *Integrative Genomics Viewer*.

4.3.2.2 Algorithme

Comme nous l'avons évoqué précédemment, chaque séquence est représentée sous forme d'un vecteur de fréquences qui est utilisable comme les coordonnées d'un point dans un espace à 4^k dimensions. Dans ce contexte, l'utilisation de nombreux algorithmes d'apprentissage automatique est très directe. Nous avons testé plusieurs algorithmes afin de déterminer le plus efficace sur les données manipulées.

Nous avons comparé cinq algorithmes : la méthode des k plus proches voisins (k-Nearest Neighbors ou kNN), la classification naïve Bayésienne (Naive Bayes), l'adaptative boosting (AdaBoost), une machine à vecteur de support (SVM), ainsi qu'une forêt d'arbres décisionnels (Random forest) (détails : cf. 1.2.2.3: [Exemples d'algorithmes]).

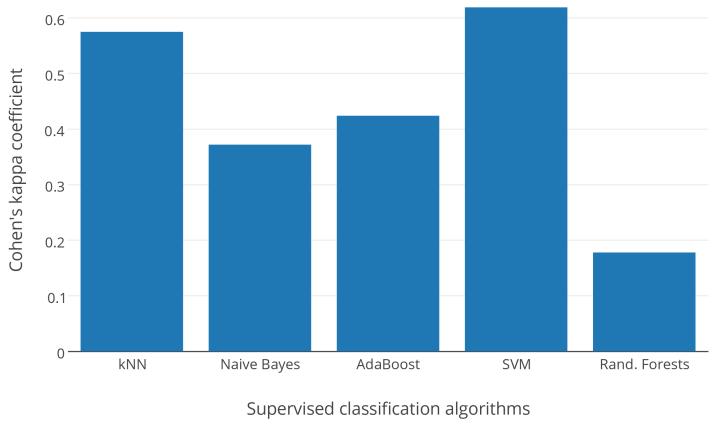


FIGURE 4.8 – Comparaison des performances des meilleurs classifieurs obtenus avec différents algorithmes d'apprentissage automatique. Les valeurs comparées sont obtenues en faisant la moyenne des valeurs de kappa sur une validation croisée de 50 échantillons.

Nous avons effectué l'apprentissage avec des tailles de k-mers variant de 3 à 6, et en testant de nombreuses combinaisons de paramètres pour chaque algorithme. Pour chacun, nous avons retenu le classifieur dont les résultats sont les meilleurs d'après le Kappa de Cohen (cf. 1.2.2.2: [Évaluation des résultats]). Les résultats sont présentés Fig. 4.8: [Comparaison de différents algorithmes].

Le classifieur offrant la meilleur performance est un SVM à noyau gaussien avec un gamma de 90 sur des 3-mers ($\kappa = 0.619$). En revanche, les pires résultats ont été également obtenus avec un classifieur SVM avec un gamma de 1 sur des 5-mers ($\kappa = 0.233$). Il s'agit donc d'un algorithme très sensible dont le choix des paramètres est crucial pour la qualité des résultats. Notre choix s'est néanmoins porté dessus et a permis de souligner l'importance du méta-apprentissage.

Nous avons également envisagé de transformer l'espace. Nous avons choisi de tester le *Large margin nearest neighbor* (LMNN), une pseudométrique destinée à "regrouper" les données de même classe pour améliorer les performances de la classification par kNN mais dont les bénéfices sont pertinents pour tout type de classification.

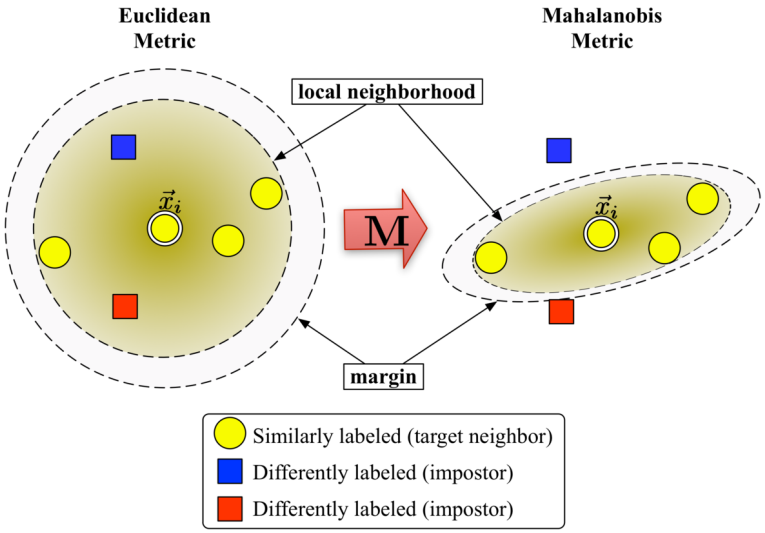


FIGURE 4.9 – Illustration de l'algorithme du *Large Margin Nearest Neighbor* (LMNN). Lorsque les données sont dispersées le long d'un axe, les plus proches voisins peuvent inclure de nombreux imposteurs. Le LMNN permet de transformer cet espace afin que les distances entre les données de même classe soit réduite par rapport aux données de classe différente.

Malheureusement, l'apprentissage du LMNN est extrêmement long et peut même prendre plusieurs semaines avec des k-mers de longueur $k > 3$, car le calcul est impossible à paralléliser. Nous avons donc été contraints d'abandonner l'idée.

4.3.2.3 Méta-apprentissage

Le méta-apprentissage n'a pas été automatisé. Nous nous sommes servis des résultats obtenus lors de la recherche de l'algorithme le plus efficace (cf. 4.3.2.2: [Algorithme]) et avons progressivement fait évoluer l'ensemble de paramètres ayant permis la construction du meilleur classifieur en fonction des résultats des différents tests effectués.

4.3.3 Résultats et discussion

Les résultats de la classification par règne sont présentés Figure 4.10: [Résultats de la classification par règne]. Il est possible d'effectuer plusieurs observations à partir de ces matrices :

— Augmenter la taille des k-mers améliore leur spécificité et donc, intuitivement, devrait améliorer les résultats. Or, l'inverse se produit : plus k est grand, plus la classification est mauvaise. Cet effet pourraît être dû au fait que plus k est grand, moins chaque k-mer est représenté dans les données et, par conséquent, plus les vecteurs de fréquences sont creux et donc difficilement comparables.

— Sans filtration des données d'apprentissage, on observe en effet de nombreux contigs viraux classifiés dans les classes du vivant et encore davantage de contigs vivants classifiés chez les virus. L'ambiguïté est donc vérifiée dans les données. Filtrer les données d'apprentissage permet en effet de décontaminer efficacement les résultats viraux, mais les appauvrit également. Cet appauvrissement augmente de manière drastique à mesure que k augmente.

Dans le contexte de cette thèse, les travaux présentés ici sont des travaux préliminaires. De nombreuses améliorations sont envisageables, mais ils ont permis de confirmer la possibilité de récupérer des informations taxonomiques à partir de la classification de vecteurs de fréquences de k-mers dans le cadre d'une classification par règne de données métagénomiques en virologie, domaine largement absent de la littérature.

Parmi les améliorations envisageables, nous pouvons citer :

— La méthode de rééchantillonnage utilisée ici extrait des fragments des génomes de référence avec une probabilité de 0.5 d'obtenir le complément inverse (correspondant au brin opposé dans le cas d'un chromosome double-brin). Or, certains virions sont encapsidés sous forme simple brin (cf. Fig. 1.14: [Classification de Baltimore]), et par conséquent, des contigs issus de techniques de séquençage sans amplification préalable (copie des fragments séquencés pour amplifier le signal) peuvent ne pas présenter de complément inverse. Il est tout à fait envisageable que conserver cette méthode de rééchantillonnage dans ce

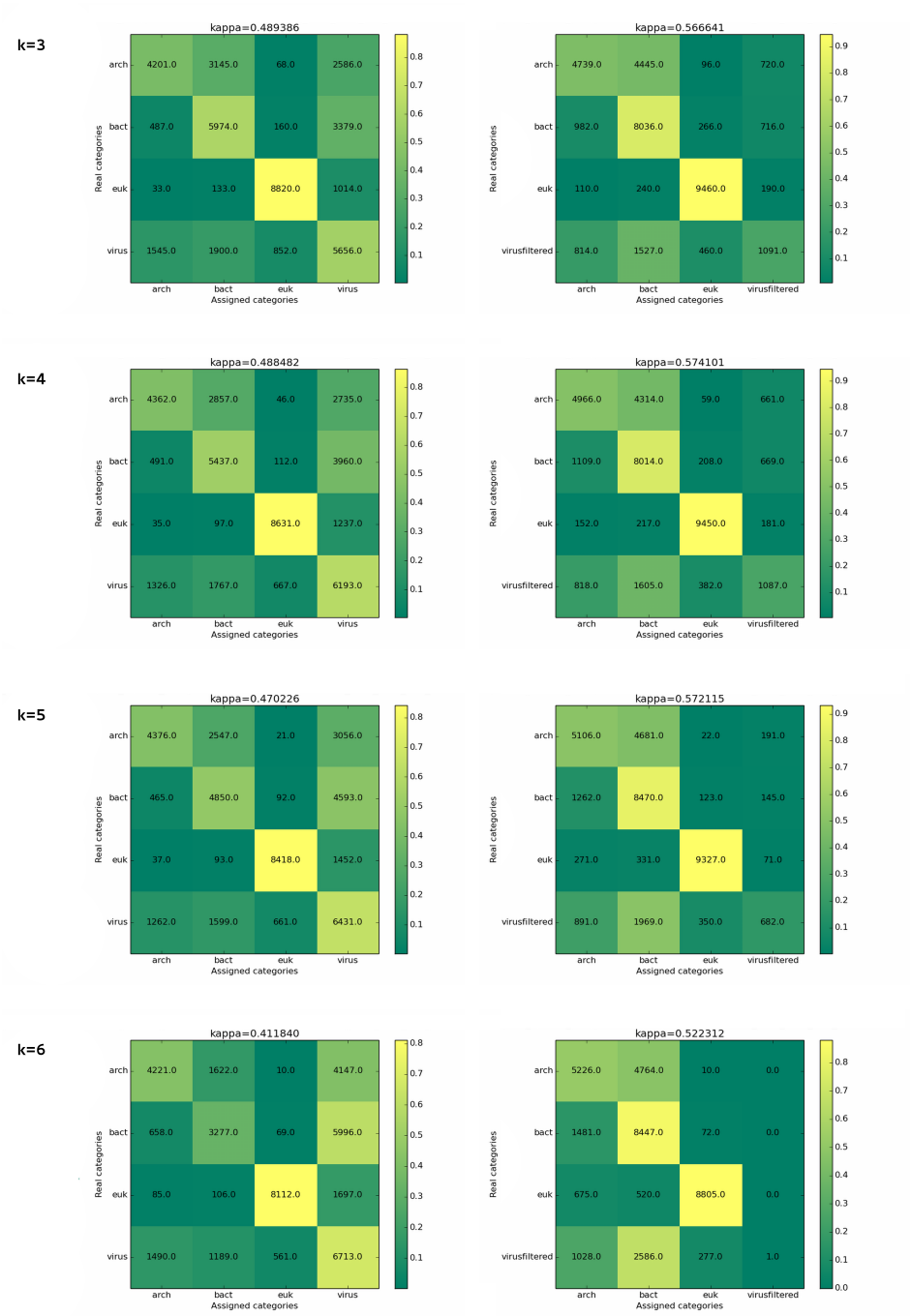


FIGURE 4.10 – Résultats de la classification par règne, pour des tailles de k-mers allant de 3 à 6 (une taille par ligne). Dans la colonne de gauche, nous avons les résultats d'un classifieur entraîné sur l'ensemble des contigs (10000 par classe). Dans la colonne de droite sont les résultats obtenus après filtrage des mêmes données d'apprentissage (cf. 4.3.1.2: [Filtration des données virales]). Les résultats sont présentés sous forme de matrices de confusion avec la valeur du kappa de Cohen indiquée au dessus.

cas puisse contribuer à une perte de signal. Corriger cela implique de conserver deux informations : le type de séquençage utilisé et la classe de Baltimore du virus à rééchantilloner, et désactiver l'extraction du complément inverse au besoin.

— Il s'agit ici de séparer les contigs viraux des contigs issus de matériel génétique non-viral ne représentant pas d'intérêt et pour lequel la filtration mécanique n'aurait pas été efficace. C'est une étape de décontamination qui peut être réduite à un problème de classification binaire, mais il n'est pas exclu que la classification multiclasse ne présente pas des avantages en termes de performances, surtout lorsqu'il s'agit de données aussi diverses. En revanche, il serait intéressant d'effectuer une comparaison avec les résultats obtenus avec une machine à vecteur de support équivalente, mais en regroupant l'ensemble des données non-virales en une seule et même classe.

— Les données d'apprentissage utilisées ici ne représentent pas l'ensemble des données publiques disponibles. Les utiliser dans leur totalité pose de nombreux problèmes. Parmi eux, on peut citer deux problèmes majeurs : le premier est lié à l'espace disque disponible lors de la création du classifieur. L'ensemble des génomes disponibles représente un volume de données considérable qui demande non seulement de mobiliser un espace de stockage important à chaque apprentissage (création du classifieur ou mise à jour afin de prendre en compte les avancées dans le domaine du séquençage), mais aussi d'importantes ressources de calcul en terme de temps (l'apprentissage n'est pas toujours parallélisable contrairement à la classification) et de mémoire disponible. Le second problème est lié aux biais d'études entraînant la sur-représentation de certains taxons, et même certains règnes (cf. Fig. 3.1: [Données RefSeq par domaine]), par rapport à d'autres. Utiliser l'ensemble des données disponibles implique d'évaluer au préalable l'impact de ces biais et de construire une méthode permettant d'en atténuer les effets. Sachant que les données sont extrêmement bruitées et contiennent de nombreuses erreurs humaines, il s'agit d'un problème complexe pouvant faire à lui seul l'objet d'un important projet de recherche.

Il s'agit néanmoins de résultats encourageants dans un contexte où les difficultés inhérentes à la problématique sont nombreuses et difficiles à contourner. Ils nous ont également permis d'acquérir une expérience utile pour aborder la problématique de la classification détaillée et ont permis de mettre au jour un certain nombre de pistes à explorer afin d'améliorer ce type de méthode.

L'étape offrant le plus de perspectives d'amélioration est le rééchantillonage. En effet, nous avons vu précédemment (cf. 2.3.2: [Discussion]) que l'efficacité d'un classi-

fieur par composition est très dépendante des données d'apprentissage. Deux pistes majeures se dessinent à ce sujet :

— L'information contenue par une séquence dépend directement de sa taille. Par conséquent, les profils de fréquences de k-mers vont être très différents selon la taille des fragments appris, ou des contigs manipulés. Sachant la grande variabilité de la taille des contigs expérimentaux, il serait intéressant de les séparer par taille et de les traiter différemment, en créant un classifieur spécifique par fourchette de longueur. Cela permettrait certainement d'améliorer la gestion des contigs très longs (4000nt) qui restent minoritaires et sont mal représentés dans les données d'apprentissage.

— Les différents taxons ne contiennent pas le même nombre de génomes séquencés et les génomes eux-mêmes présentent une grande variabilité en terme de longueur et de nombre de séquences. Cela implique un important biais de représentativité lorsque tous les taxons sont traités de la même manière à tous les niveaux, du rééchantillonage à l'apprentissage. Il serait intéressant d'explorer la possibilité de pondérer les données en fonction de leur nature et de leur nombre, afin de contourner ce biais.

L'étape de classification offre également des améliorations possibles. Les machines à vecteur de support restent des algorithmes très sensibles. Nous les avons choisis pour leur potentiel, mais il serait intéressant d'explorer des options plus stables, comme l'algorithme des k plus proches voisins. De plus, nous avons utilisé la distance euclidienne comme mesure de distance, mais d'autres mesures peuvent être explorées, comme le coefficient de Pearson ou des mesures basées sur l'entropie de Shannon.

4.3.4 Comparaison avec d'autres outils

Afin de comparer les performances de notre méthode à d'autres méthodes existantes, nous avons retenu un ensemble d'outils. Nous avons uniquement retenu des outils de classification multiclasses pour avoir des résultats comparables.

— Kraken : Un outil basé sur l'indexation de k-mers longs (WOOD et SALZBERG [2014])

— RAIPhy : Un outil par composition reposant sur des index d'abondance relative pour chaque k-mer (NALBANTOGLU et collab. [2011])

— NBC : Un outil à la frontière entre alignement et composition utilisant des données expérimentales pour l'apprentissage (ROSEN et collab. [2011])

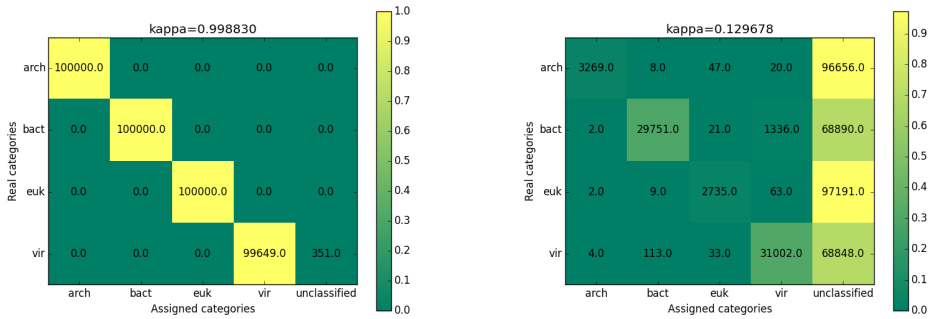
TaxyPro avait été envisagé mais il ne fournit que des statistiques sur les espèces représentées et ne permet pas de récupérer en sortie les résultats séquence par séquence. Il est donc impossible de calculer les différentes métriques permettant de comparer l'efficacité des résultats entre eux (cf. 1.2.2.2: [Évaluation des résultats]).

Nous avons utilisé les paramètres recommandés de ces outils car ils sont spécifiques aux méthodes utilisées. En effet, si nous cherchons à utiliser des paramètres proches des nôtres, les résultats sont systématiquement très mauvais. Il est donc juste, si nous utilisons les paramètres qui fonctionnent le mieux pour nous, d'en faire de même pour les outils comparés.

4.3.4.1 Kraken

Les résultats de Kraken sont présentés dans la Figure 4.11: [Résultats Kraken : classification par règne]. À gauche, nous pouvons voir qu'il classifie efficacement les données d'apprentissage, à l'exception de quelques séquences virales qu'il ne parvient pas à assigner à une classe. Cela montre encore une fois à quel point les virus sont problématiques lorsqu'il s'agit de les identifier de manière automatique. Même dans les meilleurs conditions possibles, il restent difficiles à traiter.

En revanche, la partie droite de la figure montre les résultats du classifieur sur un ensemble de données à classifier différent du jeu d'apprentissage, et la plupart des données se retrouvent non classifiées. Étonnamment, plus de virus sont correctement identifiés que toutes les autres classes mais, dans l'ensemble, le classifieur est victime d'un surapprentissage massif et est donc inutilisable dans ce contexte.



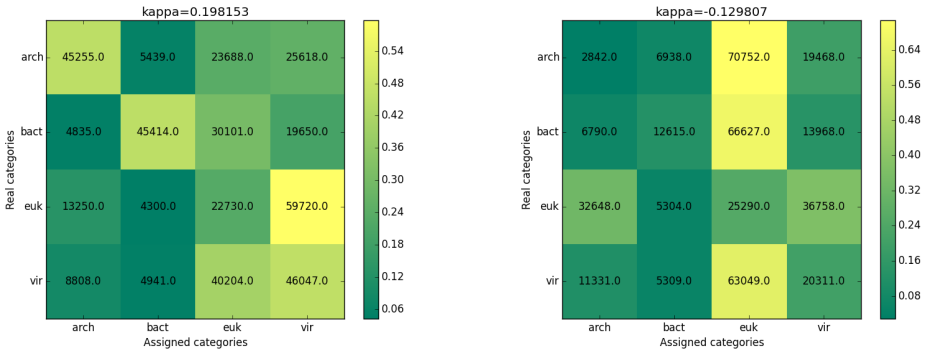
(a) Classification de l'échantillon d'apprentissage

(b) Classification d'un échantillon différent du jeu d'apprentissage.

FIGURE 4.11 – Résultats de la classification par règne avec Kraken. Taille de k-mers par défaut : $k = 31$

4.3.4.2 Naive Bayes Classifier

Les résultats du classifieur naïf Bayésien sont présentés dans la Figure 4.12: [Résultats NBC : classification par règne]. Malheureusement, il est difficile de les décrire. Même sur ses propres données d'apprentissage le classifieur donne des résultats presque aléatoires et, sur un jeu différent, il classe presque toutes les séquences dans les eukaryotes, sauf les eukaryotes eux-mêmes. Il est donc inutilisable également dans ce contexte.



(a) Classification de l'échantillon d'apprentissage

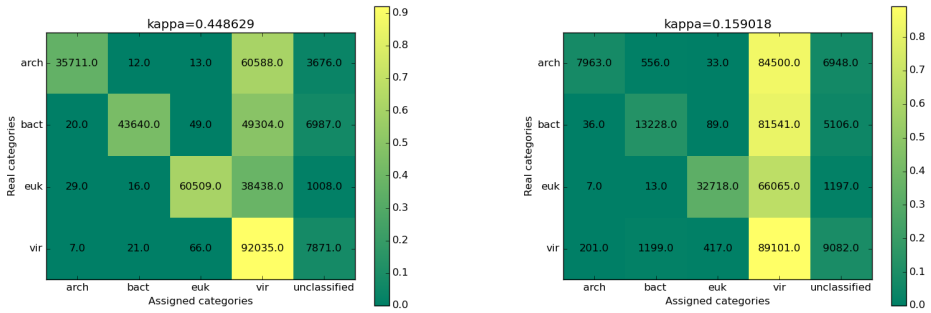
(b) Classification d'un échantillon différent du jeu d'apprentissage.

FIGURE 4.12 – Résultats de la classification par règne avec un classifieur naïf bayésien (NBC). Taille de k-mers par défaut : $k = 15$

4.3.4.3 RAIphy

Les résultats de RAIphy sont présentés dans la Figure 4.13: [Résultats RAIphy : classification par règne]. Il s'agit du classifieur offrant les meilleurs performances parmi les trois testés (kappa = 0.16 sur un jeu de données différent), mais souffre d'une classe virale très invasive. En effet, même sur les données d'apprentissage, elle phagocyte une bonne moitié des séquences.

Il est envisageable qu'en supprimant les virus des données utilisées, de bien meilleurs résultats seraient obtenus. Pourtant, les auteurs ont utilisé tous les génomes disponibles dans la base de données RefSeq pour leurs tests, mais leurs résultats suggèrent que les génomes viraux n'ont pas été utilisés (NALBANTOGLU et collab. [2011]), ce qui pourrait expliquer qu'ils posent tant problème ici. Ces résultats suggèrent donc qu'il s'agit d'un outil très efficace pour la classification par règne, tant que les virus sont écartés des données d'apprentissage et filtrés des échantillons à classifier, ce qui va à l'encontre de notre problématique.



(a) Classification de l'échantillon d'apprentissage

(b) Classification d'un échantillon différent du jeu d'apprentissage.

FIGURE 4.13 – Résultats de la classification par règne avec RAIphy. Taille de k-mers par défaut : $k = 7$

4.4 Classification détaillée

4.4.1 Données de référence

4.4.1.1 Bases de données

Nous nous intéressons ici à la classification détaillée de séquences virales. Cette partie repose donc sur le postulat selon lequel toutes les séquences manipulées ont été identifiées comme appartenant au monde viral. L'objectif étant d'obtenir des informations complémentaires sur la taxonomie des espèces virales dont les séquences sont issues, nous avons besoin ici de données sur la taxonomie détaillée des données d'apprentissage. Nous avons donc orienté notre choix de bases de données dans ce sens.

Données taxonomiques : Les taxonomies complètes n'étant pas des arbres complets et parfaits (cf. 3.1.1: [Propreté et pertinence des données]), nous avons choisi d'utiliser une taxonomie simplifiée afin de pouvoir séparer simplement les données en plusieurs classes, quel que soit le niveau hiérarchique considéré. En effet, le choix du niveau (cf. Fig. 1.28: [Niveaux phylétiques]) influe sur la résolution et sur l'efficacité du classifieur et doit pouvoir être varié en fonction de la problématique et des données. Par exemple, une résolution (i.e. niveau hiérarchique utilisé pour constituer les classes) trop importante peut être problématique lors de la classification d'espèces inconnues. Une taxonomie simplifiée limite les possibilités en terme de résolution de la classification mais permet l'automatisation du processus tout en fournissant des informations taxonomiques intermédiaires et en limitant les erreurs chez ces espèces non représentées dans les données d'apprentissage. En aval de cette analyse, le traitement des taxons d'intérêt doit être fait au cas par cas, afin d'affiner

l'identification et obtenir une précision que l'on ne peut pas encore atteindre de manière automatisée avec l'ensemble des données de référence actuellement disponibles.

Données génomiques : Nous avons décidé de continuer à utiliser les bases de données du NCBI afin d'assurer une pérennité et une bonne actualisation des sources de données.

Cette double contrainte nous a amenés à utiliser la taxonomie simplifiée inclue dans les Genomes Reports de la base de données *genome* du NCBI (cf. 1.2.3.3: [Taxonomies de référence et classes], `ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS`). Cette taxonomie comporte les niveaux phylétiques suivants :

— **Pour les virus :** classes de Baltimore (cf. Fig. 1.14: [Classification de Baltimore]), famille, espèce.

— **Pour les prokaryotes :** embranchement, classe, espèce.

— **Pour les eukaryotes :** groupes d'usage commun (animaux, plantes, mycètes, protistes, etc.), sous-groupes d'usage commun selon le NCBI (mammifères, oiseaux, ascomycètes, algues vertes, vers plats, sporozoaires, etc.), espèce.

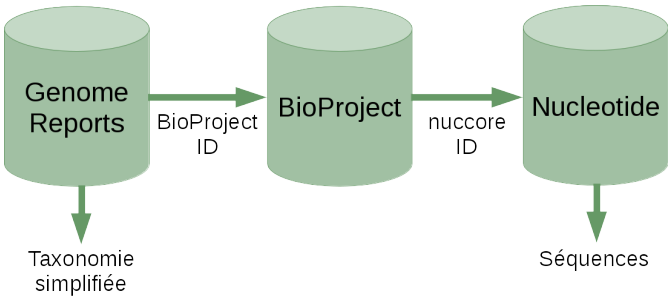


FIGURE 4.14 – Récupération des données d'apprentissage : Les Genome Reports fournissent la taxonomie simplifiée ainsi que les identifiants des projets correspondant à chaque organisme dans la base de données BioProject. Cette dernière contient elle-même les identifiants correspondant à toutes les séquences associées à chaque projet auxquelles il est possible d'accéder depuis la base *Nucleotide* (*nuccore*, `http://www.ncbi.nlm.nih.gov/nuccore`) du NCBI qui collecte les séquences depuis diverses sources, notamment RefSeq (`http://www.ncbi.nlm.nih.gov/refseq`) et GenBank (`http://www.ncbi.nlm.nih.gov/genbank`).

Nous récupérons l'ensemble des données disponibles et exploitables (e.g. pour lesquelles l'organisme d'origine des séquences récupérées correspond à l'organisme annoncé dans le Genome Report). Il est important de noter que la récupération de l'intégralité d'une base de données présente des avantages certains en terme de représentation des espèces, d'automatisation et de mise à jour, mais elle présente également quelques désavantages. Le désavantage principal est lié au biais d'étude des

TABLEAU 4.2 – Statistiques sur les Genome Reports viraux. Données du 02/06/2016.

	Nombre de segments	Nombre de séquences	Longueur en nuléotides
Minimum	1	1	220
Maximum	105	3125	247387
Moyenne	1.3	49.4	39401
Médiane	1	7	11131

virus. Par exemple, si on regarde les Genome Reports viraux (version du 02/06/2016), nous n'avons pas moins de 10 entrées pour la grippe, totalisant à elles seules 107 séquences et 133 411 nt, alors que la plus petite est une entrée unique ne contenant que 220 nt sur deux séquences au total. Il existe également des biais naturels liés à la structure des virus. La plupart des génomes viraux ne sont composés que d'un brin continu, mais certains génomes sont séparés en plusieurs segments. Pour reprendre l'exemple de la grippe, cette dernière en possède 8. Certaines espèces, comme celles appartenant à la famille des Polydnaviridae, poussent cette complexité à l'extrême. Les cinq espèces représentées sont respectivement constituées de 30, 15, 105, 56 et 23 segments et totalisent 477 séquences et 1 531 176 nt. Il suffit de regarder la différence entre les moyennes et médianes des différentes métadonnées pour se rendre compte de l'importance des conséquences de ces biais d'étude (cf. Tableau 4.2: [Statistiques sur les Genome Reports viraux. Données du 02/06/2016.]).

Il est néanmoins difficile d'envisager d'exclure des génomes pour réduire ces biais. Outre le fait que les génomes viraux sont une denrée rare, et que la difficulté de la tâche est en partie due à leur rareté, réduire le biais structurel serait contre-productif. Les virus ont une diversité de structure extraordinaire et diminuer la représentation des génomes les plus lourds et les plus complexes signifierait de réduire encore davantage nos chances d'identifier des séquences appartenant à des espèces partageant leur taxon. De la même manière, les virus les plus étudiés sont représentés par plusieurs souches, comme la grippe. Les taux de mutation extrêmes propres aux virus rendent les différentes souches de grippe difficilement comparables (cf. Fig. A.20: [Grippes A, B et C, segment 1]) et en favoriser une par rapport aux autres pourrait introduire un nouveau biais. De plus, les conséquences de ce biais sur l'efficacité de l'apprentissage restent encore à évaluer.

4.4.1.2 Téléchargement

Afin de faciliter la récupération et la maintenance des données, leur téléchargement est automatisé par un script qui, à partir des Genome Reports, consulte les bases de données publiques comme indiqué dans la Figure 4.14: [Récupération des données]. Il utilise le système global de recherches inter-bases de données du

NCBI Entrez (GIBNEY et BAXEVANIS [2011]). Ce système permet de récupérer les BioProjects puis, à partir des identifiants de séquences qu'ils contiennent, et de télécharger les séquences correspondantes au format FASTA. Il génère également la taxonomie simplifiée en croisant les données taxonomiques contenues dans les Genomes Reports et les identifiants contenus dans les BioProjects, afin de faire le lien direct entre chaque séquence récupérée et sa taxonomie d'origine. Les requêtes sont effectuées via l'API *Entrez* (https://www.ncbi.nlm.nih.gov/books/NBK25501/). Comme elles sont nombreuses et que les données récupérées sont volumineuses, la procédure de téléchargement souffre régulièrement d'erreurs de connexion. Afin de rendre cette tâche plus fluide et conviviale, une gestion automatique des erreurs HTTP a été mise en place.

4.4.2 Spécificités du workflow

4.4.2.1 Rééchantillonnage

La stratégie de rééchantillonnage ici est similaire à celle utilisée lors de la classification par règne, mais quelques modifications y ont été introduites afin de réduire certains biais et l'adapter aux nouvelles contraintes. L'équi-représentativité des classes est maintenue : chaque classe aura le même nombre de contigs dans le jeu de données de sortie. En revanche, les différences sont les suivantes :

— Précédemment, la taille des contigs était tirée au hasard après avoir fixé la position de départ de chaque contig. Dans ce contexte, et à plus forte raison avec des génomes courts tels que les génomes viraux, les longueurs les plus grandes sont pénalisées car elles ont une probabilité plus élevée de faire dépasser le contig du génome. Ainsi, la distribution des longueurs des contigs obtenus ne suit plus la loi normale selon laquelle les longueurs ont été tirées. Afin de pallier à ce biais, les longueurs sont ici tirées par avance et stockées avant de placer les positions de départ de chaque contig de manière aléatoire sur les génomes. De plus, les positions ne permettant pas d'utiliser la longueur à placer sont retirées des positions disponibles.

— La représentation de tous les génomes est ici forcée. En effet, même si la couverture génomique des contigs sur les génomes de référence est faible, chaque génome sera représenté par au moins un d'entre eux.

4.4.2.2 Algorithme

Les résultats que nous avons obtenus pour la classification par règne étant encourageants, nous avons décidé de conserver le même algorithme de classification (cf. 4.3.2.2: [Algorithme]). En revanche, nous avons ajouté une étape afin de répondre à un problème que nous n'avions pas considéré auparavant.

Le "fléau de la dimension" est un phénomène propre aux espaces de grande dimension. (BELLMAN [1961]) Il s'agit d'une conséquence de la croissance exponentielle du volume de l'espace à mesure que le nombre de dimensions augmente. Ainsi, les données deviennent éparses et il est de plus en plus difficile de les relier entre elles.

Or, toutes les dimensions ne sont pas aussi pertinentes les unes que les autres lorsqu'on souhaite identifier des données de même nature. Les points de même classe peuvent être rapprochés sur certaines dimensions et dispersés sur d'autres. Il est donc possible de réduire ce phénomène de dispersion en identifiant les dimensions les moins pertinentes et en les supprimant afin de réduire le volume de l'espace et de regrouper les données de même classe.

Nous utilisons une analyse en composantes principales afin d'identifier les n dimensions les plus pertinentes parmi l'ensemble des 4^k dimensions de l'espace, à savoir celles selon lesquelles les données de même classes ont une variance moindre. Ces n dimensions sont conservées pour l'apprentissage et la classification, les autres sont ignorées. Le nombre de dimensions à conserver n est ajustable en fonction des données d'apprentissage.

4.4.2.3 Méta-apprentissage

L'ensemble des paramètres à tester n'est pas très grand (le nombre de dimensions à conserver pour la PCA, ainsi que le paramètre de la fonction de coût de la marge souple de la SVM et le paramètre libre de la fonction à base radiale du noyau gaussien), mais choisir une grille de valeurs afin de tester exhaustivement toutes les combinaisons possibles peut s'avérer extrêmement long.

Afin de réduire le temps nécessaire à cette étape, nous avons choisi d'utiliser une recherche aléatoire. Une fonction aléatoire est fournie pour chaque paramètre et un nombre de tirages à tester est défini. Cette méthode, outre le gain de temps considérable qu'elle permet, est connue pour offrir des résultats souvent aussi bons que la recherche exhaustive par grille de valeurs, voire meilleurs dans certains cas (BERGSTRA et BENGIO [2012]).

Chaque combinaison obtenue après le tirage d'une valeur aléatoire pour chaque paramètre est ensuite testée plusieurs fois sur un jeu de données avec une validation croisée de type *k-fold*. La combinaison de valeurs obtenant les meilleurs résultats est enregistrée au format JSON, utilisable ensuite pour créer le classifieur final.

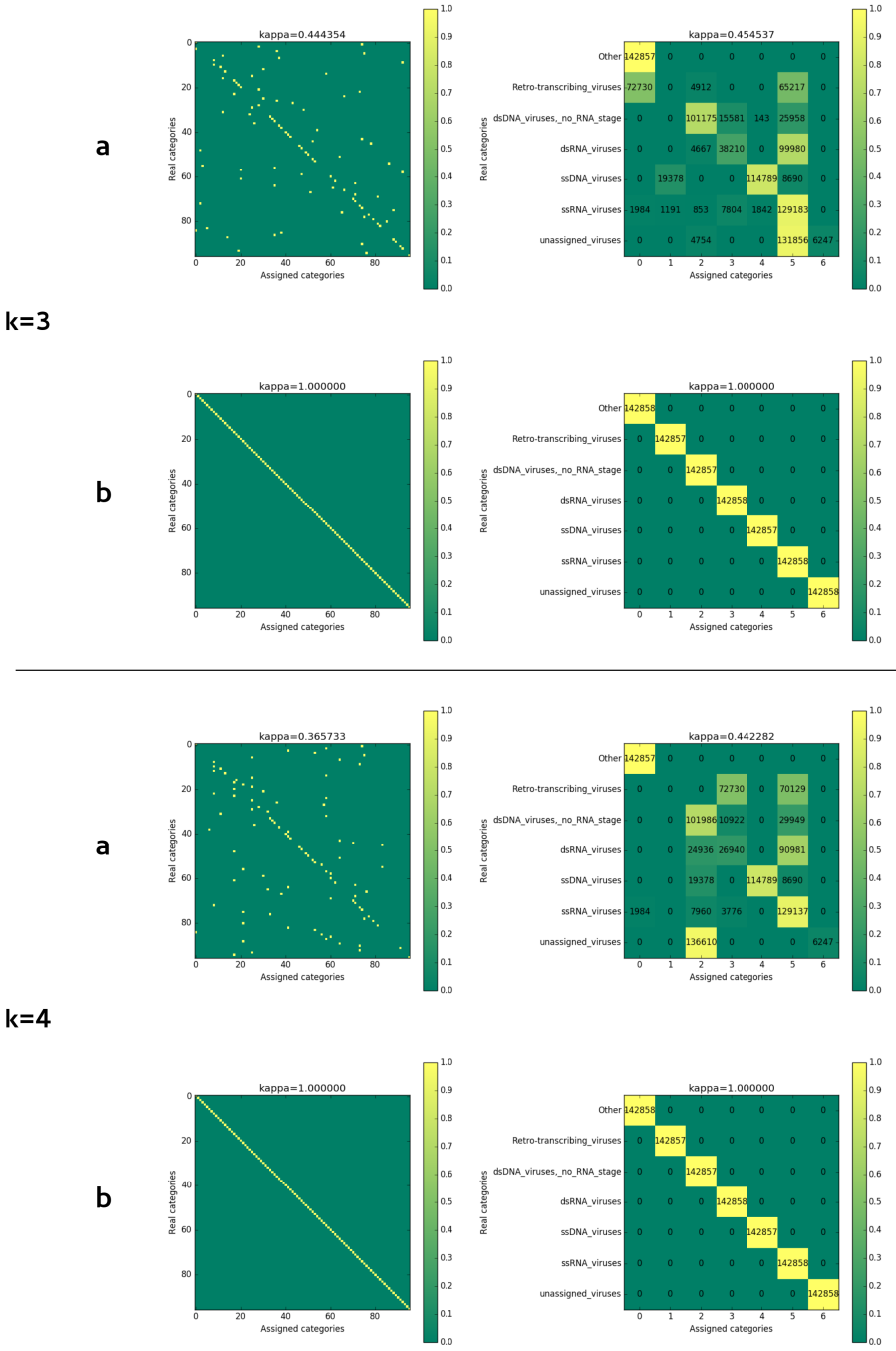


FIGURE 4.15 – Matrices de confusion des résultats de la classification détaillée, pour des k-mers de taille 3 (moitié haute de la figure) et 4 (moitié basse de la figure). Pour chaque taille de k-mer, les résultats en (a) sont obtenus sur un jeu de données à classifier différent du jeu de données d'apprentissage, et les résultats en (b) sont obtenus avec le même jeu de données pour l'apprentissage et la classification.

4.4.3 Résultats et discussion

4.4.3.1 Vue d'ensemble des résultats

Les résultats de la classification détaillée sont présentés dans la Figure 4.15: [Résultats de la classification détaillée]. Chaque sous-figure est une représentation de la matrice de confusion des résultats obtenus avec des 3-mers (haut) et des 4-mers (bas), de la classification des données ayant entraîné le modèle (b) ou d'un jeu de données issu d'un rééchantillonage différent (a). Les sous-figures de gauche sont une représentation directe des résultats avec, en abscisse, les familles virales auxquelles les contigs ont été attribués par le modèle et, en ordonnée, les familles virales d'origine de ces contigs. Les sous-figures de droite sont une représentation des mêmes matrices de confusion, obtenues en regroupant les contigs par classe de Baltimore. Les valeurs dans les cases ainsi que leur couleur correspondent au nombre de contigs. Nous pouvons faire plusieurs observations à partir de ces données :

— La classification des données d'apprentissage donnent une diagonale parfaite, toutes les séquenes sont attribuées à leur classe d'origine. En revanche, la classification de données nouvelles présente des caractristiques bien différentes. Pour une partie des familles, toutes les séquences sont correctement classifiées (47.28% pour k=3, 38.95% pour k=4). Les séquences appartenant aux autres familles sont attribuées à une classe différente de leur classe d'origine.

— Dans le cas des séquences mal classifiées, toutes les séquences appartenant à la même famille d'origine sont attribuées à la même classe. Ainsi, toutes les séquences appartenant à la même famille sont systématiquement regroupées ensemble, qu'elle corresponde à la famille d'origine ou non. On observe même des classes dans lesquelles sont classifiées des familles ne leur correspondant pas, alors que celle leur appartenant est classifiée ailleurs (cf. Fig. 4.16: [Destination des données de test]).

— Lorsqu'on regroupe les données par classe de Baltimore, on peut s'apercevoir que certains types de virus regroupent de nombreuses classes dans lesquelles les familles mal classifiées se retrouvent souvent. Il s'agit principalement des virus à ARN simple brin (ssRNA, classe IV), mais également, dans une moindre mesure, les virus à ADN double brin (dsDNA, classe I). Il est également intéressant de noter que les virus "autres", regroupant les espèces viroïdes ne correspondant à aucune classe de Baltimore, ne sont pas confondus avec les autres virus.

— De la même manière que pour la classification par règne (cf. Fig. 4.10: [Résultats de la classification par règne]), augmenter la valeur de k n'améliore pas la qualité des résultats.

Nous sommes confrontés ici à un problème de surapprentissage encore plus important que dans le cadre de la classification par règne, vu que le modèle construit

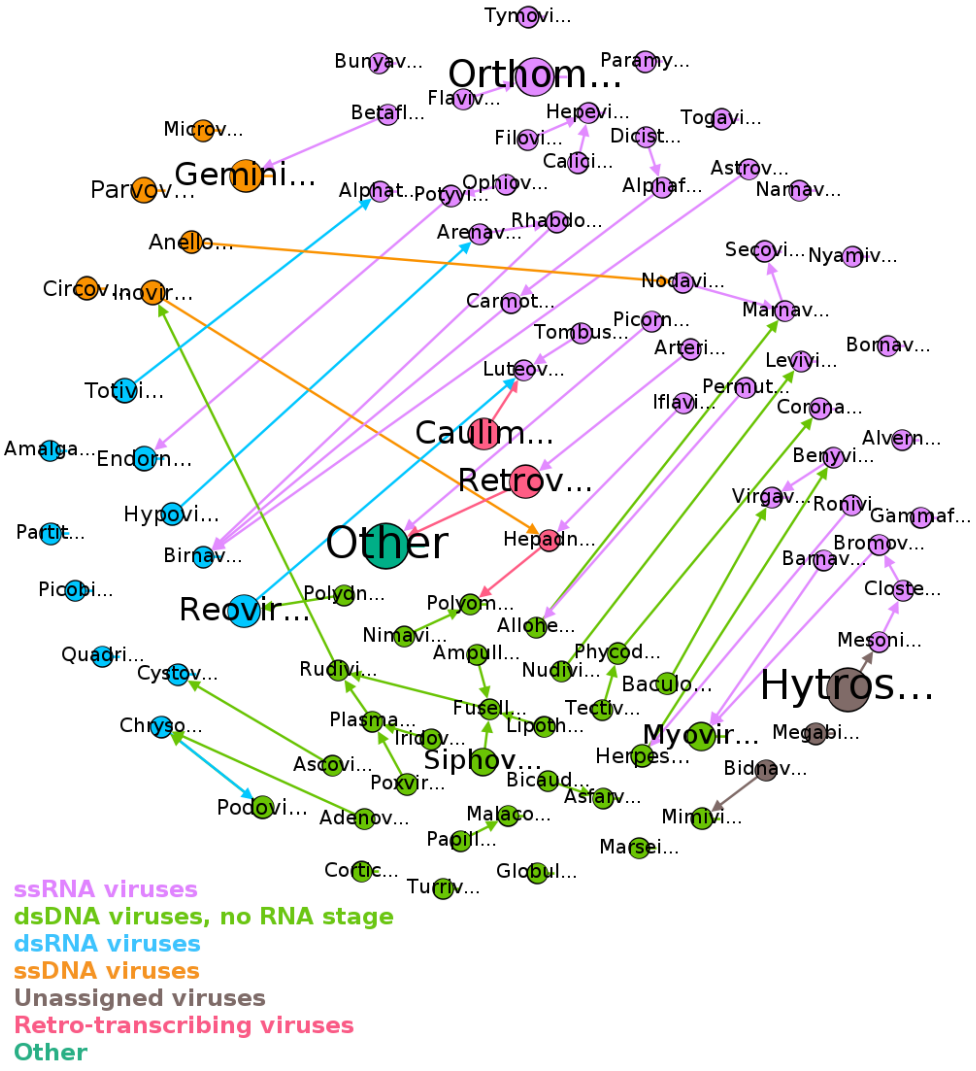
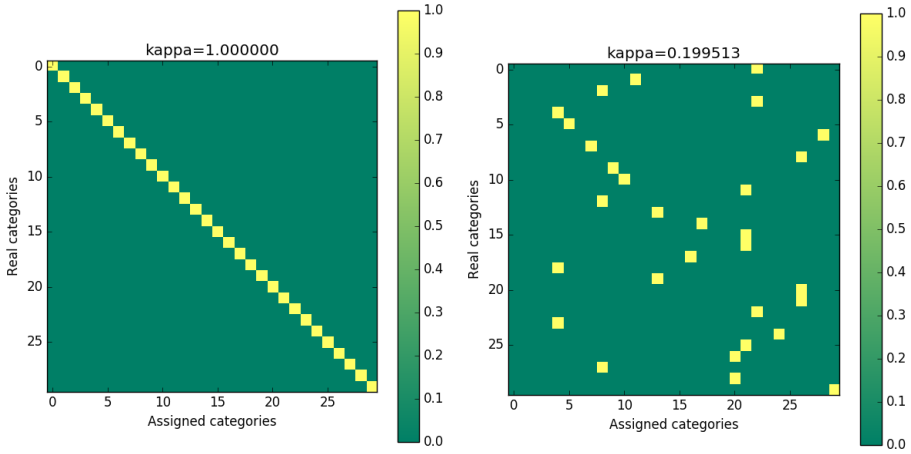


FIGURE 4.16 – Graphe représentant la destination des données de test pour $k = 3$

classe parfaitement le jeu d'apprentissage qui a permis sa construction, mais ne permet pas la classification correcte de plus de la moitié d'un jeu de données issu d'un autre rééchantillonage. Il est donc important d'explorer ces résultats plus en détail afin d'écarter d'éventuelles erreurs de méthode.

4.4.3.2 Cas des virus de classe I (dsDNA viruses, no RNA stage)

Nous avons pu observer de nombreux échanges entre classes de Baltimore dans la Figure 4.16: [Destination des données de test], ce qui peut soulever des questions relatives aux critères sur lesquels est construit le classifieur. Afin de vérifier si ces comportements se retrouvent dans des données plus comparables entre elles, nous avons isolé les virus de classe I (dsDNA viruses, no RNA stage) et avons appliqué la même méthode sur ce sous-ensemble de données. Les résultats obtenus sont présentés Figure 4.17: [Résultats pour les virus de classe I]. Nous n'obtenons ici que 21.36% de séquences et 30% de familles correctement classées, ce qui tend à montrer que limiter le nombre de classes et leur disparité ne permet pas d'améliorer les résultats. En effet, en entraînant le modèle avec toutes les classes de Baltimore, les familles de la classe I correctement classées sont également de 30%, représentant 36.15% des séquences de cette même classe.



(a) Classification des données d'apprentissage

(b) Classification de données de test

FIGURE 4.17 – Matrices de confusion des résultats de la classification des données issues de la classe I

Isoler une classe de Baltimore nous permet également d'écarter une possible influence de la couverture génomique (ici, la couverture des contigs sur les génomes de références utilisés pour leur création) sur la qualité des résultats. En effet, cette couverture varie grandement d'une classe de Baltimore à l'autre (cf. 4.3: [Couver-

TABLEAU 4.3 – Estimation de la couverture génomique des contigs d'apprentissage sur leurs génomes d'origine, par classe de Baltimore.

Classe de Baltimore	Couverture moyenne	Écart-type
dsDNA_viruses,_no_RNA_stage	0.74	0.07
ssRNA_viruses	2.26	0.35
ssDNA_viruses	42.59	5.40
dsRNA_viruses	51.99	10.13
unassigned_viruses	284.63	18.39
Retro-transcribing_viruses	98.16	9.81
Other	14793.27	N/A

ture génomique]) et une couverture élevée (e.g."unassigned viruses", dont seule une famille sur les trois est bien classifiée) ne garantit pas des résultats supérieurs. Une exception existe peut-être dans la classe "Other", qui ne possède pas de famille proprement dite et dont la couverture génomique est d'un ordre de grandeur bien supérieur aux autres.

Nous avons estimé la couverture génomique de chaque famille en faisant le rapport de la longueur cumulée des contigs d'apprentissage sur la longueur cumulée des génomes dont ils sont issus. Elle ne varie que très peu au sein de chaque classe (cf. 4.3: [Couverture génomique]) et le coefficient de Pearson entre cette couverture et l'exactitude des résultats n'est que de 0.07. Cela tend à confirmer qu'elle n'a que très peu d'importance dans l'efficacité du classifieur.

En ce qui concerne le devenir des contigs, nous pouvons observer dans la Figure 4.18: [Graphe des résultats des virus de classe I] que l'absence des autres classes de Baltimore modifie l'assignation pour certaines familles. Celles qui sont correctement classées (en vert) ne sont pas nécessairement les mêmes dans les deux cas. Pour certaines (e.g. *Baculoviridar, Herpesviridae, Podoviridae, Polydnaviridae*), l'absence des autres classes (en gris) permet de retrouver une classification correcte et d'améliorer les résultats. En revanche, quelques familles auparavant correctement classifiées (e.g. *Marseilleviridae, Mimiviridae, Myoviridae, Polyomaviridae*) se retrouvent attirées par d'autres familles. Néanmoins, certains motifs semblent rester stables (e.g les échanges entre *Siphovorodae, Ampullaviridae, Lipothrixviridae, Rudiviridae, Plasmaviridae, Iridoviridae* et *Poxviridae*) et plusieurs familles présentent de bon résultats dans les deux cas (e.g. *Turriviridae, Globuloviridae, Asfarviridae, Corticoviridae, Malacoherpesviridae*).

Ces résultats soulignent l'influence de la présence de certaines familles sur l'efficacité globale du classifieur, mais souligne également que la robustesse des résultats varie grandement d'une famille à l'autre. Les familles les plus représentées dans les

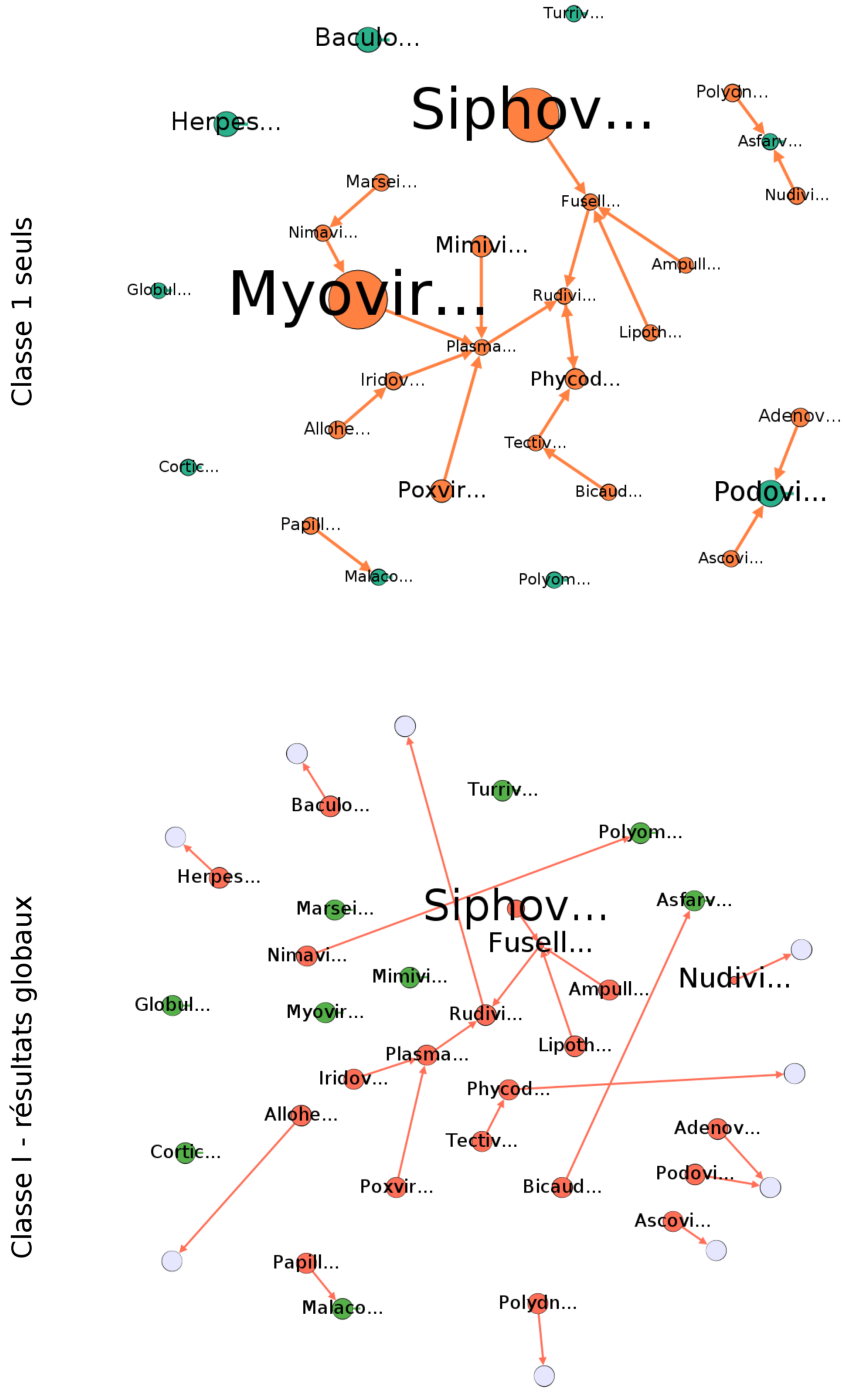


FIGURE 4.18 – Graphe représentant la destination des données de test pour $k = 3$ et pour les virus de classe I.

données d'apprentissage (i.e. *Myoviridae* et *Siphoviridae*) ne sont pas les plus stables, ce qui soulève la question de l'influence des effectifs des différentes familles dans les données d'apprentissage sur la capacité du classifieur à les reconnaître.

4.4.3.3 Rééchantillonage et effectifs

Le premier point à vérifier afin de répondre à la question sus-citée est la méthode de rééchantillonage, pour laquelle nous avons pris le parti d'utiliser des données dont les effectifs par famille dépendent de la quantité de données d'apprentissage dans ces mêmes familles. Par conséquent, les données d'apprentissage ne sont pas en quantité équivalente d'une famille à l'autre (cf. 4.4.2.1: [Rééchantillonnage]), un parti pris afin de limiter les disparités de densité entre les familles, mais dont les effets restent à évaluer.

Si l'on regarde les données d'apprentissage dans le cas des virus de classe I, les écarts de représentation sont très importants (cf. 4.4: [Évaluation statistique des effectifs des données de classe I]). Nous nous sommes d'abord intéressés aux familles les moins représentées dans les données d'apprentissage, afin d'évaluer l'influence de leur faible représentation :

— La plus petite famille est les *Corticoviridae* avec 11 contigs. Cette famille reste bien classifiée, que le modèle soit construit avec toutes les classes de Baltimore ou non, malgré sa très petite taille.

— Ensuite, nous avons les *Plasmaviridae*, avec 15 contigs. Cette famille est mal classifiée dans les deux cas, mais possède la particularité d'attirer un nombre important d'autres familles : les *Iridoviridae* et les *Poxviridae* de manière invariable, ainsi que les *Mimiviridae* et les *Myoviridae* lorsque l'apprentissage n'est effectué que sur les contigs de classe I. Ces résultats sont remarquables car ces familles attirées ont la particularité d'être parmi les mieux représentées (effectifs respectifs de 1673, 6544, 5383 et 44533 contigs).

— La famille suivante parmi les moins représentées est les *Ampullaviridae*, avec 20 contigs. Celle-ci se comporte de manière stable dans les deux cas et est systématiquement phagocytée par les *Fuselloviridae*. Elle n'attire aucune autre famille et fait partie d'un des motifs très bien conservés décrits dans la partie 4.4.3.2: [Cas des virus de classe I (dsDNA viruses, no RNA stage)].

Au regard de ces résultats, il est donc important de tester l'hypothèse intuitive selon laquelle les effectifs n'ont pas d'influence sur la qualité de la classification. En effet, nous pouvons observer dans la Figure 4.19: [Répartition des résultats par effectifs] que les familles les mieux classifiées ne sont pas nécessairement celles qui s'éloignent le plus de l'origine et que la qualité de la classification ne semble dépendre ni du nombre de contigs utilisés pour l'apprentissage, ni du nombre de séquences de référence. Le coefficient de Pearson entre la qualité des résultats (nombre de contigs

TABLEAU 4.4 – Évaluation statistique des effectifs des données de classe I

	Longueur cumul. génomes	Nombre contigs	longueur cumul. contigs
Min	10079	11	8877
Max	42952271	44533	32405511
Médiane	679054	721.5	521801.5
Moyenne	4635439.8	4761.9	3461344.6
Écart-type	10276768.04	10587.72	7700485.87

classifiés dans la bonne famille) et le nombre de contigs utilisés pour l'apprentissage n'est que de 0.05 (p-value = 0.59) et consolide fortement cette hypothèse.

Afin d'explorer cette dernière plus en détail, nous avons choisi, parmi nos trois plus petites familles, celles qui sont mal classifiées (i.e. *Plasmaviridae* et *Ampullaviridae*) et avons fortement enrichi les données d'apprentissage pour ces dernières. Tout en conservant les mêmes paramètres, nous avons construit deux classifieurs supplémentaires, chacun ayant reçu 10000 contigs supplémentaires d'une des deux familles. Ces contigs ont été obtenus via un rééchantillonage indépendant et ont été ajoutés aux données précédentes :

— Les *Ampullaviridae* (famille au comportement stable que les autres classes de Baltimore soient présentes ou non) restent mal classifiée malgré l'enrichissement. Tous les contigs issus de cette famille sont toujours attribués aux *Fuselloviridae* et elle n'attire toujours aucune autre famille.

— Les *Plasmaviridae*, en revanche, bénéficient de l'enrichissement : cette famille se retrouve correctement classifiée, sans attirer vers elles plus de familles qu'elle n'attirait déjà (i.e. *Iridoviridae*, *Poxviridae*, *Mimiviridae* et *Myoviridae*).

Il est par conséquent difficile de confirmer l'hypothèse selon laquelle l'effectif des familles, lors de l'entraînement du modèle, importe peu car, même si le coefficient de corrélation global est faible, nous pouvons voir que les effets de la variation de ces effectifs varient d'une famille à l'autre. Il S'agit donc certainement d'un phénomène plus complexe qui pourrait notamment inclure des effets combinés de représentation faible et de grande variabilité intragénomique. Dans ces conditions, il est difficile de raisonner de manière globale, même au sein d'une classe de Baltimore donnée.

4.4.3.4 Les limites de l'approche "non-biologique"

Les modèles produits par cette approche permettent indéniablement de détecter de l'information discriminante permettant de reconnaître près de la moitié des familles testées. De plus, le fait que les contigs appartenant à la même famille restent groupés met en évidence la reconnaissance d'un signal, même en cas d'échec

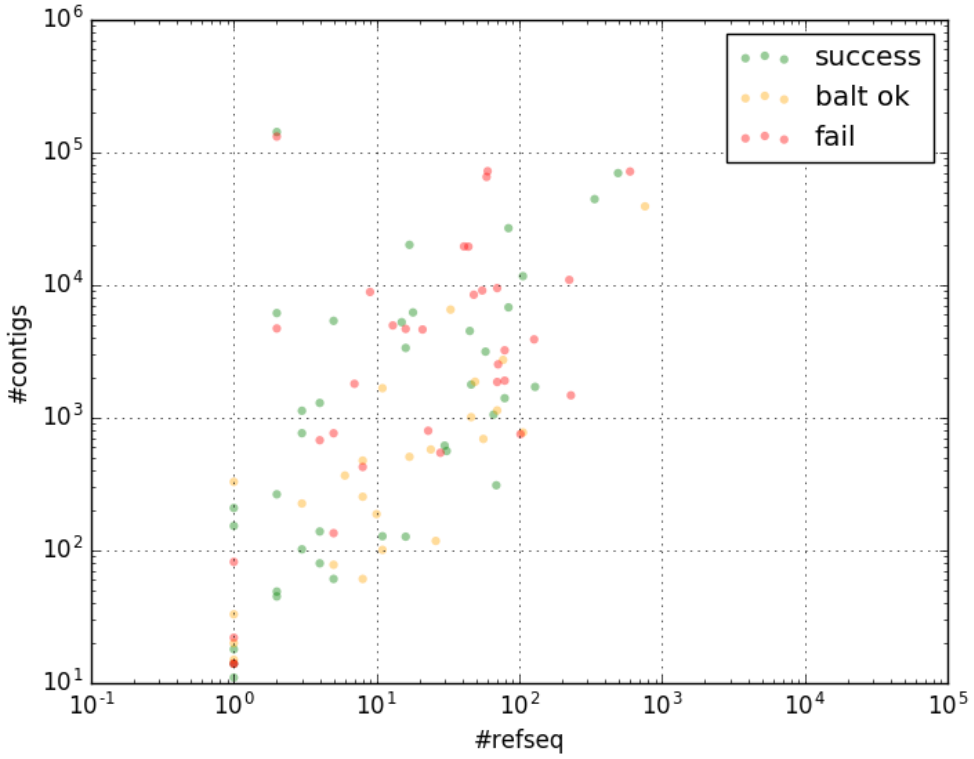


FIGURE 4.19 – Répartition des résultats pour chaque famille, par effectifs des données d'apprentissage (ordonnées, sur une échelle logarithmique) et par nombre de séquences de référence(abscisses, sur une échelle logarithmique). Les points en vert sont les familles correctement classifiées, les points en orange sont les familles mal classifiées à qui il est attribué une famille de la même classe de Baltimore et les points en rouge sont les familles mal classifiées à qui il est attribué une famille d'une classe de Baltimore différente.

de la classification, bien que le caractère systématique de ce comportement reste à expliquer.

De plus, il est tout à fait possible qu'il s'agisse d'un cas de figure dans lequel la recherche de paramètres par validation croisée ne permet pas d'éviter le surapprentissage d'une SVM. En effet, la théorie sur laquelle elle se base ne le permettent pas toujours (CAWLEY et TALBOT [2010]). Pourtant, une vérification rapide de la persistance de ce problème de surapprentissage a été effectuée sur Weka (`http://www.cs.waikato.ac.nz/~ml/weka/index.html`) avec une implémentation différente d'une SVM à noyau gaussien, ainsi que l'algorithme des k plus proches voisins suggère qu'il s'agirait d'un problème plus général, qui serait lié à la nature des données manipulées.

Cette observation, ainsi que la variabilité des comportements des différentes familles lors de la classification, met en évidence une limite majeure de ce type d'approche uniquement basée sur la composition des séquences. En traitant toutes les familles de la même manière, on met de côté plusieurs éléments cruciaux expliquant l'hétérogénéité des familles virales : il faut d'abord considérer l'explication évolutive de ces différences. Les familles virales ne sont pas issues d'une histoire évolutive unique, mais d' événements distincts au sein de plusieurs cellules ancestrales qui n'ont pas de descendants modernes (NASIR et CAETANO-ANOLLÉS). Il s'agit donc d'histoires évolutives multiples, parfois non linéaires, dont les distances n'évoluent pas de manière homogène au cours du temps. Les familles sous-représentées n'offrent que très peu d'information sur leur évolution ni sur les caractéristiques communes qui font leur spécificité. Afin d'accomoder ces différences et d'inclure des projets de séquençage récents de manière rétro-compatible, les taxons de même niveau dans la taxonomie de l'ICTV ne sont pas toujours construits sur les mêmes critères. Historiquement, la définition des familles reposait majoritairement sur l'annotation fonctionnelle de leurs génomes et la structure des capsides, sans nécessairement offrir d'alignement efficace, mais les espèces découvertes plus récemment tendent à être classées sur des critères fortement basés sur la similarité des séquences car leur afflux ne permet pas toujours une annotation complète (SIMMONDS [2015]; SIMMONDS et collab. [2017]).

Dans ce contexte, un modèle construit avec, d'une part, des données sans annotation et, d'autre part, des classes construites parfois sans annotation, parfois autour d'elles, est privé d'un ensemble d'informations importantes pour faire le lien entre ces données et les différentes classes auxquelles elles appartiennent.

Des efforts de recherche sont en cours afin de corriger cette taxonomie au fur et à mesure que de nouvelles espèces sont découvertes et améliorent progressivement son adaptabilité. Nous avons vu notamment l'outil NVR (YU et collab. [2013]) au cours de la section 2.3: [Classification par composition] qui montre une efficacité remarquable dans ce sens, à tous les niveaux taxonomiques. On peut donc espérer

avoir à l'avenir une taxonomie plus fidèle à la réalité biologique du monde viral, mais ces avancées n'enlèveront pas aux virus leur caractère hétérogène ni leur grande variabilité évolutive. Il est donc prudent d'envisager que l'assignation taxonomique virale ne peut que bénéficier d'un retour à l'information contenue par les génomes pour la découverte d'espèces proches des familles connues.

Conclusion

Dans un contexte économique et écologique où la biodiversité est une préoccupation centrale, les nouvelles technologies de séquençage permettent d'avoir un nouveau regard sur nos écosystèmes et les relations entre les différentes espèces qui les composent. Parmi ces espèces, les virus représentent encore aujourd'hui un défi scientifique majeur de par leur variabilité, la complexité de leurs intéractions avec leurs hôtes et leur environnement, ainsi que leur faible représentation parmi les projets de séquençage. Pourtant, la compréhension de ces intéractions est essentielle pour décrypter les mécanismes qui animent notre environnement et ainsi anticiper et corriger l'impact de l'homme sur les équilibres naturels dont il dépend pour sa survie.

Une étape majeure dans ce travail de décryptage est la découverte, l'identification et le recensement des espèces virales. Contrairement aux espèces vivantes, qui bénéficient déjà d'un certain recul, les virus restent encore très largement méconnus. La métagénomique, approche globale de séquençage de milieux complexes, permet depuis peu de combler petit à petit ces lacunes, mais souffre d'un manque d'outils d'analyse des données spécifiques au monde viral. Nous apportons ici des pistes de recherche dans le but de combler ce manque.

Il est possible de distinguer deux approches complémentaires pour l'identification des espèces virales. La première est la classification par règne, dont le but est d'identifier une séquence comme appartenant ou non à un génome viral. Elle permet de séparer les virus du vivant dans des échantillons complexes. La seconde est la classification détaillée, dont le but est d'identifier de manière fine l'identité de l'organisme auquel appartient une séquence. Dans le cas présent, il s'agit d'ifentifier la famille virale, en sachant que les séquences soumises ont déjà été traitées afin de ne conserver que les virus.

Nous avons choisi d'utiliser des machines à vecteur de support pour construire nos modèles et nos résultats montrent qu'un signal existe dans la composition des séquences, quelle que soit l'approche utilisée. De plus, en supprimant des données d'apprentissage les séquences présentant une ambiguïté forte avec le vivant, la pollution des résultats par les séquences non virales est notablement diminuée, au prix toutefois d'une perte de sensibilité. Pourtant, les résultats obtenus avec des outils gé-

néralistes très répandus (Kraken, RAIPhy et NBC) sont très inférieurs. En regardant de près le comportement de la classification détaillées, nous avons mis en évidence l'importance des spécificités de chaque famille virale. Ce résultat met en évidence qu'il est difficilement envisageable de suivre une approche de classification totalement déconnectée de la réalité biologique représentée par les séquences manipulées.

Nous avons montré qu'il est possible d'améliorer la détection et l'identification des virus, et l'échec des outils généralistes souligne la nécessité d'adapter l'outil au monde viral et à sa diversité. Les données étant intrinsèquement hétérogènes, il n'est pas surprenant de constater une différence d'efficacité des algorithmes en fonction des données d'intérêt. Les caractéristiques particulières des virus exacerbent cet effet. Il est donc nécessaire d'explorer des solutions spécifiques aux données virales, qui prennent en compte leurs particularités.

Ces travaux sont encourageants pour l'avenir de la métagénomique virale. Il reste de nombreuses possibilités techniques à explorer et de connaissances à enrichir afin de tirer le meilleur parti de la transversalité entre biologie et informatique de ce type d'étude.

Bibliographie

ADAMS, M. J. et J. F. ANTONIW. 2004, «Codon usage bias amongst plant viruses.», *Archives of virology*, vol. 149, nº 1, doi :10.1007/s00705-003-0186-6, p. 113–35, ISSN 0304-8608. URL http://www.ncbi.nlm.nih.gov/pubmed/14689279. 76

AHO, A. V., J. E. HOPCROFT et J. D. ULLMAN. 1973, «On finding lowest common ancestors in trees», dans *Proceedings of the fifth annual ACM symposium on Theory of computing - STOC '73*, ACM Press, New York, New York, USA, p. 253–265, doi :10.1145/800125.804056. URL http://dl.acm.org/citation.cfm?id=800125.804056. 55

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS et D. J. LIPMAN. 1990, «Basic local alignment search tool.», *Journal of molecular biology*, vol. 215, nº 3, doi :10.1016/S0022-2836(05)80360-2, p. 403–10, ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/2231712. 53

AMES, S., D. HYSOM et S. GARDNER. 2013, «Scalable metagenomic taxonomy classification using a reference genome database», ..., vol. 29, nº 18, doi :10.1093/bioinformatics/btt389, p. 2253–2260. URL http://bioinformatics.oxfordjournals.org/content/29/18/2253.short. 57

ANDERSON, P. K., A. A. CUNNINGHAM, N. G. PATEL, F. J. MORALES, P. R. EPSTEIN et P. DASZAK. 2004, «Emerging infectious diseases of plants : pathogen pollution, climate change and agrotechnology drivers.», *Trends in ecology & evolution*, vol. 19, nº 10, doi :10.1016/j.tree.2004.07.021, p. 535–44, ISSN 0169-5347. URL http://www.ncbi.nlm.nih.gov/pubmed/16701319. 20, 21, 22

ARSLAN, D., M. LEGENDRE, V. SELTZER, C. ABERGEL et J.-M. CLAVERIE. 2011, «Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, nº 42, doi :10.1073/pnas.1110889108, p. 17 486–91, ISSN 1091-6490. URL http://www.pnas.org/content/108/42/17486. 15

BALCH, W. E., L. J. MAGRUM, G. E. FOX, R. S. WOLFE et C. R. WOESE. 1977, «An ancient divergence among the bacteria», *Journal of Molecular Evolution*, vol. 9, nº 4, doi :10.1007/BF01796092, p. 305–311, ISSN 0022-2844. URL http://link.springer.com/10.1007/BF01796092. 68

BALTIMORE, D. 1971, «Expression of animal virus genomes.», *Bacteriological reviews*, vol. 35, nº 3, p. 235–41, ISSN 0005-3678. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=378387&tool=pmcentrez&rendertype=abstract. 15, 32

BAO, Y., V. CHETVERNIN et T. TATUSOVA. 2012, «PAirwise Sequence Comparison (PASC) and Its Application in the Classification of Filoviruses», *Viruses*, vol. 4, nº 12, doi :10.3390/v4081318, p. 1318–1327, ISSN 1999-4915. URL http://www.mdpi.com/1999-4915/4/8/1318/. 57

BAO, Y., V. CHETVERNIN et T. TATUSOVA. 2014, «Improvements to pairwise sequence comparison (PASC) : a genome-based web tool for virus classification.», *Archives of virology*, vol. 159, nº 12, doi :10.1007/s00705-014-2197-x, p. 3293–304, ISSN 1432-8798. URL http://www.ncbi.nlm.nih.gov/pubmed/25119676http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4221606. 57

BARBA, M., H. CZOSNEK et A. HADIDI. 2014, «Historical perspective, development and applications of next-generation sequencing in plant virology.», *Viruses*, vol. 6, nº 1, doi :10.3390/v6010106, p. 106–36, ISSN 1999-4915. URL http://www.mdpi.com/1999-4915/6/1/106/htm. 42

BAZINET, A. L. et M. P. CUMMINGS. 2012, «A comparative evaluation of sequence classification programs.», *BMC bioinformatics*, vol. 13, doi :10.1186/1471-2105-13-92, p. 92, ISSN 1471-2105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3428669&tool=pmcentrez&rendertype=abstract. 51, 53, 57, 63

BELLMAN, R. 1961, *Dynamic Programming*, Princeton University Press, ISBN 0486428095, 340 p., doi :862270. 108

BERGSTRA, J. et Y. BENGIO. 2012, «Random search for hyper-parameter optimization», *The Journal of Machine Learning Research*, vol. 13, nº 1, p. 281–305, ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2188385.2188395. 108

BILLER, S. J., F. SCHUBOTZ, S. E. ROGGENSACK, A. W. THOMPSON, R. E. SUMMONS et S. W. CHISHOLM. 2014, «Bacterial vesicles in marine ecosystems.»,

Science (New York, N.Y.), vol. 343, nº 6167, doi :10.1126/science.1243457, p. 183–6, ISSN 1095-9203. URL http://www.sciencemag.org/content/343/6167/183.abstract. 51

BOHLIN, J., E. SKJERVE et D. W. USSERY. 2008, «Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes.», *BMC genomics*, vol. 9, doi :10.1186/1471-2164-9-104, p. 104, ISSN 1471-2164. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2289816{&}tool=pmcentrez{&}rendertype=abstract. 61, 63, 70

BRADY, A. et S. L. SALZBERG. 2009, «Phymm and PhymmBL : metagenomic phylogenetic classification with interpolated Markov models.», *Nature methods*, vol. 6, nº 9, doi :10.1038/nmeth.1358, p. 673–6, ISSN 1548-7105. URL http://www.ncbi.nlm.nih.gov/pubmed/19648916http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2762791http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762791{&}tool=pmcentrez{&}rendertype=abstract. 61

BRAUN, J. V. et H.-G. MÜLLER. 1998, «Statistical methods for DNA sequence segmentation», *Statistical Science*, vol. 13, nº 2, doi :10.1214/ss/1028905933, p. 142–162. URL http://projecteuclid.org:80/Dienst/getRecord?id=euclid.ss/1028905933/. 83

BRENNER, S., F. JACOB et M. MESELSON. 1961, «An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis», *Nature*, vol. 190, nº 4776, doi :10.1038/190576a0, p. 576–581, ISSN 0028-0836. URL http://dx.doi.org/10.1038/190576a0. 10

BZHALAVA, D., J. EKSTRÖM, F. LYSHOLM, E. HULTIN, H. FAUST, B. PERSSON, M. LEHTINEN, E.-M. DE VILLIERS et J. DILLNER. 2012, «Phylogenetically diverse TT virus viremia among pregnant women.», *Virology*, vol. 432, nº 2, doi :10.1016/j.virol.2012.06.022, p. 427–34, ISSN 1096-0341. URL http://www.sciencedirect.com/science/article/pii/S0042682212003248. 58

BZHALAVA, D., H. JOHANSSON, J. EKSTRÖM, H. FAUST, B. MÖLLER, C. EKLUND, P. NORDIN, B. STENQUIST, J. PAOLI, B. PERSSON, O. FORSLUND et J. DILLNER. 2013, «Unbiased approach for virus detection in skin lesions.», *PloS one*, vol. 8, nº 6, doi :10.1371/journal.pone.0065953, p. e65 953, ISSN 1932-6203. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065953. 58

CAWLEY, G. C. et N. L. C. TALBOT. 2010, «On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation», *Journal of Machine Learning Research*, vol. 11, nº Jul, p. 2079–2107, ISSN ISSN 1533-7928. 118

CHEN, Y.-P. P., éd.. 2005, *Bioinformatics Technologies*, Springer-Verlag, Berlin/Heidelberg, ISBN 3-540-20873-9, doi :10.1007/b138246. URL http://www.springerlink.com/index/10.1007/b138246. 34

COETZEE, B., M.-J. FREEBOROUGH, H. J. MAREE, J.-M. CELTON, D. J. G. REES et J. T. BURGER. 2010, «Deep sequencing analysis of viruses infecting grapevines : Virome of a vineyard.», *Virology*, vol. 400, nº 2, doi :10.1016/j.virol.2010.01.023, p. 157–63, ISSN 1096-0341. URL http://www.sciencedirect.com/science/article/pii/S004268221000053X. 51

COFFIN, J. M., S. H. HUGHES et H. E. VARMUS. 1997, «Retroviruses», URL http://www.ncbi.nlm.nih.gov/books/NBK19376/. 19

COLE, J. R., Q. WANG, E. CARDENAS, J. FISH, B. CHAI, R. J. FARRIS, A. S. KULAM-SYED-MOHIDEEN, D. M. MCGARRELL, T. MARSH, G. M. GARRITY et J. M. TIEDJE. 2009, «The Ribosomal Database Project : improved alignments and new tools for rRNA analysis.», *Nucleic acids research*, vol. 37, nº Database issue, doi :10.1093/nar/gkn879, p. D141–5, ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686447&tool=pmcentrez&rendertype=abstract. 51

DESANTIS, T. Z., P. HUGENHOLTZ, N. LARSEN, M. ROJAS, E. L. BRODIE, K. KELLER, T. HUBER, D. DALEVI, P. HU et G. L. ANDERSEN. 2006, «Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.», *Applied and environmental microbiology*, vol. 72, nº 7, doi : 10.1128/AEM.03006-05, p. 5069–72, ISSN 0099-2240. URL http://aem.asm.org/content/72/7/5069.long. 51

DIAZ, N. N., L. KRAUSE, A. GOESMANN, K. NIEHAUS et T. W. NATTKEMPER. 2009, «TACOA : taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach.», *BMC bioinformatics*, vol. 10, nº 1, doi :10.1186/1471-2105-10-56, p. 56, ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/10/56. 61

DRAKE, J. W. 1999, «The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes.», *Annals of the New York Academy of Sciences*, vol. 870, p. 100–7, ISSN 0077-8923. URL http://www.ncbi.nlm.nih.gov/pubmed/10415476. 71

DUFFY, S., L. A. SHACKELTON et E. C. HOLMES. 2008, «Rates of evolutionary change in viruses : patterns and determinants.», *Nature reviews. Genetics*, vol. 9, nº 4, doi :10.1038/nrg2323, p. 267–76, ISSN 1471-0064. URL http://www.ncbi.nlm.nih.gov/pubmed/18319742. 71

DUHAIME, M. B., L. DENG, B. T. POULOS et M. B. SULLIVAN. 2012, «Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples : a rigorous assessment and optimization of the linker amplification method.», *Environmental microbiology*, vol. 14, nº 9, doi :10.1111/j.1462-2920.2012.02791.x, p. 2526–37, ISSN 1462-2920. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3466414&tool=pmcentrez&rendertype=abstract. 51

DURBIN, R., S. R. EDDY, A. KROGH et G. MITCHISON. 1998, *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, ISBN 113945739X. URL https://encrypted.google.com/books?id=HUUhAwAAQBAJ&pgis=1. 53, 55

EDDY, S. R. 2004, «What is a hidden Markov model?», *Nature biotechnology*, vol. 22, nº 10, doi :10.1038/nbt1004-1315, p. 1315–6, ISSN 1087-0156. URL http://www.nature.com/index.html?file=/nbt/journal/v22/n10/full/nbt1004-1315.html. 55

EDDY, S. R. 2008, «A probabilistic model of local sequence alignment that simplifies statistical significance estimation.», *PLoS computational biology*, vol. 4, nº 5, doi :10.1371/journal.pcbi.1000069, p. e1000069, ISSN 1553-7358. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000069. 55

EDWARDS, R. A. et F. ROHWER. 2005, «Viral metagenomics.», *Nature reviews. Microbiology*, vol. 3, nº 6, doi :10.1038/nrmicro1163, p. 504–10, ISSN 1740-1526. URL http://dx.doi.org/10.1038/nrmicro1163. 51

FANCELLO, L., D. RAOULT et C. DESNUES. 2012, «Computational tools for viral metagenomics and their application in clinical research.», *Virology*, vol. 434, nº 2, doi :10.1016/j.virol.2012.09.025, p. 162–74, ISSN 1096-0341. URL http://www.ncbi.nlm.nih.gov/pubmed/23062738. 51

FILÉE, J. 2013, «Route of NCLDV evolution : the genomic accordion.», *Current opinion in virology*, vol. 3, nº 5, doi :10.1016/j.coviro.2013.07.003, p. 595–9, ISSN 1879-6265. URL http://www.ncbi.nlm.nih.gov/pubmed/23896278. 71

FILÉE, J., N. POUGET et M. CHANDLER. 2008, «Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses.», *BMC evolutionary biology*, vol. 8, doi :10.1186/1471-2148-8-320, p. 320, ISSN 1471-2148. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2607284{&}tool=pmcentrez{&}rendertype=abstract. 71

FOERSTNER, K. U., C. VON MERING, S. D. HOOPER et P. BORK. 2005, «Environments shape the nucleotide composition of genomes.», *EMBO reports*, vol. 6, nº 12, doi :10.1038/sj.embor.7400538, p. 1208–13, ISSN 1469-221X. URL http://embor.embopress.org/content/6/12/1208.abstract. 59

FROST, L. S., R. LEPLAE, A. O. SUMMERS et A. TOUSSAINT. 2005, «Mobile genetic elements : the agents of open source evolution.», *Nature reviews. Microbiology*, vol. 3, nº 9, doi :10.1038/nrmicro1235, p. 722–32, ISSN 1740-1526. URL http://dx.doi.org/10.1038/nrmicro1235. 51

GAGO, S., S. F. ELENA, R. FLORES et R. SANJUÁN. 2009, «Extremely high mutation rate of a hammerhead viroid.», *Science (New York, N.Y.)*, vol. 323, nº 5919, doi : 10.1126/science.1169202, p. 1308, ISSN 1095-9203. URL http://www.sciencemag.org/content/323/5919/1308.abstract. 72

GAUTIER, C. 2000, «Compositional bias in DNA», *Current Opinion in Genetics & Development*, vol. 10, nº 6, doi :10.1016/S0959-437X(00)00144-1, p. 656–661, ISSN 0959437X. URL http://www.sciencedirect.com/science/article/pii/S0959437X00001441. 59

GIBBS, A. J. 2013, «Viral taxonomy needs a spring clean ; its exploration era is over.», *Virology journal*, vol. 10, nº 1, doi :10.1186/1743-422X-10-254, p. 254, ISSN 1743-422X. URL http://www.virologyj.com/content/10/1/254. 32

GIBNEY, G. et A. D. BAXEVANIS. 2011, «Searching NCBI Databases Using Entrez.», *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, vol. Chapter 6, doi :10.1002/0471142905.hg0610s71, p. Unit6.10, ISSN 1934-8258. URL http://www.ncbi.nlm.nih.gov/pubmed/21975942. 107

GORI, F., G. FOLINO, M. S. M. JETTEN et E. MARCHIORI. 2011, «MTR : taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks.», *Bioinformatics (Oxford, England)*, vol. 27, nº 2, doi : 10.1093/bioinformatics/btq649, p. 196–203, ISSN 1367-4811. URL http://bioinformatics.oxfordjournals.org/content/27/2/196.full. 53, 56

GRIFFITHS, A. J., W. M. GELBART, J. H. MILLER et R. C. LEWONTIN. 1999, *Modern Genetic Analysis*, W. H. Freeman. URL http://www.ncbi.nlm.nih.gov/books/NBK21248/. 70

GRIFFITHS, A. J., J. H. MILLER, D. T. SUZUKI, R. C. LEWONTIN et W. M. GELBART. 2000, *An Introduction to Genetic Analysis.*, W. H. Freeman. URL http://www.ncbi.nlm.nih.gov/books/NBK21760/. 18, 19

HALL, R. J., J. WANG, A. K. TODD, A. B. BISSIELO, S. YEN, H. STRYDOM, N. E. MOORE, X. REN, Q. S. HUANG, P. E. CARTER et M. PEACEY. 2014, «Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.», *Journal of virological methods*, vol. 195, doi :10.1016/j.jviromet.2013.08.035, p. 194–204, ISSN 1879-0984. URL http://www.ncbi.nlm.nih.gov/pubmed/24036074. 50

HANDELSMAN, J. 2004, «Metagenomics : application of genomics to uncultured microorganisms», *Microbiology and Molecular Biology Reviews*, vol. 68, nº 4, doi : 10.1128/MBR.68.4.669. URL http://mmbr.asm.org/content/68/4/669.short. 25

HAREL, D. et R. E. TARJAN. 1984, «Fast Algorithms for Finding Nearest Common Ancestors», *SIAM Journal on Computing*, vol. 13, nº 2, doi :10.1137/0213024, p. 338–355, ISSN 0097-5397. URL http://dl.acm.org/citation.cfm?id=373.411. 56

HORIE, M., T. HONDA, Y. SUZUKI, Y. KOBAYASHI, T. DAITO, T. OSHIDA, K. IKUTA, P. JERN, T. GOJOBORI, J. M. COFFIN et K. TOMONAGA. 2010, «Endogenous non-retroviral RNA virus elements in mammalian genomes.», *Nature*, vol. 463, nº 7277, doi :10.1038/nature08695, p. 84–7, ISSN 1476-4687. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2818285{&}tool=pmcentrez{&}rendertype=abstract. 70

HORTON, M., N. BODENHAUSEN et J. BERGELSON. 2010, «MARTA : a suite of Java-based tools for assigning taxonomic status to DNA sequences.», *Bioinformatics (Oxford, England)*, vol. 26, nº 4, doi :10.1093/bioinformatics/btp682, p. 568–9, ISSN 1367-4811. URL http://bioinformatics.oxfordjournals.org/content/26/4/568.long. 53

HUG, L. A., B. J. BAKER, K. ANANTHARAMAN, C. T. BROWN, A. J. PROBST, C. J. CASTELLE, C. N. BUTTERFIELD, A. W. HERNSDORF, Y. AMANO, K. ISE, Y. SUZUKI, N. DUDEK, D. A. RELMAN, K. M. FINSTAD, R. AMUNDSON, B. C. THOMAS et J. F. BANFIELD. 2016, «A new view of the tree of life», *Nature Microbiology*, vol. 1, doi :10.1038/nmicrobiol.2016.48, p. 16 048, ISSN 2058-5276. URL http://www.nature.com/articles/nmicrobiol201648. 29

HUSON, D. H., A. F. AUCH, J. QI et S. C. SCHUSTER. 2007, «MEGAN analysis of metagenomic data.», *Genome research*, vol. 17, nº 3, doi :10.1101/gr.5969107, p. 377–86, ISSN 1088-9051. URL http://genome.cshlp.org/content/17/3/377.full.html. 53, 56

JAMAIN, S., M. GIRONDOT, P. LEROY, M. CLERGUE, H. QUACH, M. FELLOUS et T. BOURGERON. 2001, «Transduction of the human gene FAM8A1 by

endogenous retrovirus during primate evolution.», *Genomics*, vol. 78, nº 1-2, doi : 10.1006/geno.2001.6642, p. 38–45, ISSN 0888-7543. URL http://www.ncbi.nlm.nih.gov/pubmed/11707071. 19

JENSEN, L. J. et A. BATEMAN. 2011, «The rise and fall of supervised machine learning techniques.», *Bioinformatics (Oxford, England)*, vol. 27, nº 24, doi :10.1093/bioinformatics/btr585, p. 3331–2, ISSN 1367-4811. URL http://bioinformatics.oxfordjournals.org/content/27/24/3331.full. 34

JOHNSON, S. 2006, *Remote Protein Homology Detection Using Hidden Markov Models*, thèse de doctorat. URL https://books.google.fr/books/about/Remote{_}Protein{_}Homology{_}Detection{_}Using.html?id=gVlCOAAACAAJ{&}pgis=1. 55

KARLIN, S. et C. BURGE. 1995, «Dinucleotide relative abundance extremes : a genomic signature.», *Trends in genetics : TIG*, vol. 11, nº 7, p. 283–90, ISSN 0168-9525. URL http://www.ncbi.nlm.nih.gov/pubmed/7482779. 60, 61

KIM, K.-H. et J.-W. BAE. 2011, «Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses.», *Applied and environmental microbiology*, vol. 77, nº 21, doi :10.1128/AEM.00289-11, p. 7663–8, ISSN 1098-5336. URL http://aem.asm.org/content/77/21/7663.full. 51

KING, A. M., M. J. ADAMS et E. J. LEFKOWITZ. 2011, *Virus Taxonomy : Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses*, ISBN 0123846846, 1327 p.. URL https://books.google.com/books?hl=fr&lr=&id=KXRCYay3pH4C&pgis=1. 57

KORNBERG, A. et T. A. BAKER. 2005, *DNA Replication*, ISBN 1891389440, 931 p.. URL https://books.google.fr/books/about/DNA_Replication.html?id=KDsubusF0YsC&pgis=1. 6

KRAUSE, L., N. N. DIAZ, A. GOESMANN, S. KELLEY, T. W. NATTKEMPER, F. ROHWER, R. A. EDWARDS et J. STOYE. 2008, «Phylogenetic classification of short environmental DNA fragments.», *Nucleic acids research*, vol. 36, nº 7, doi :10.1093/nar/gkn038, p. 2230–9, ISSN 1362-4962. URL http://nar.oxfordjournals.org/content/36/7/2230.full. 55

LANG, A. S. et J. T. BEATTY. 2007, «Importance of widespread gene transfer agent genes in alpha-proteobacteria.», *Trends in microbiology*, vol. 15, nº 2, doi : 10.1016/j.tim.2006.12.001, p. 54–62, ISSN 0966-842X. URL http://www.ncbi.nlm.nih.gov/pubmed/17184993. 51

LARRAÑAGA, P., B. CALVO, R. SANTANA, C. BIELZA, J. GALDIANO, I. N. INZA, J. A. LOZANO, R. ARMAÑANZAS, G. SANTAFÉ, A. PÉREZ et V. ROBLES. 2006, «Machine learning in bioinformatics.», *Briefings in bioinformatics*, vol. 7, nº 1, p. 86–112, ISSN 1467-5463. URL http://www.ncbi.nlm.nih.gov/pubmed/16761367. 34, 35

LAUBER, C. et A. E. GORBALENYA. 2012a, «Genetics-Based Classification of Filoviruses Calls for Expanded Sampling of Genomic Sequences», *Viruses*, vol. 4, nº 12, doi :10.3390/v4091425, p. 1425–1437, ISSN 1999-4915. URL http://www.mdpi.com/1999-4915/4/9/1425/. 57

LAUBER, C. et A. E. GORBALENYA. 2012b, «Partitioning the genetic diversity of a virus family : approach and evaluation through a case study of picornaviruses.», *Journal of virology*, vol. 86, nº 7, doi :10.1128/JVI.07173-11, p. 3890–904, ISSN 1098-5514. URL http://www.ncbi.nlm.nih.gov/pubmed/22278230http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3302503. 57

LAUBER, C. et A. E. GORBALENYA. 2012c, «Toward genetics-based virus taxonomy : comparative analysis of a genetics-based classification and the taxonomy of picornaviruses.», *Journal of virology*, vol. 86, nº 7, doi :10.1128/JVI.07174-11, p. 3905–15, ISSN 1098-5514. URL http://www.ncbi.nlm.nih.gov/pubmed/22278238http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3302533. 57

LAURING, A. S., J. FRYDMAN et R. ANDINO. 2013, «The role of mutational robustness in RNA virus evolution.», *Nature reviews. Microbiology*, vol. 11, nº 5, doi :10.1038/nrmicro3003, p. 327–36, ISSN 1740-1534. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3981611{&}tool=pmcentrez{&}rendertype=abstract. 71

LECHNER, M., A. I. NICKEL, S. WEHNER, K. RIEGE, N. WIESEKE, B. M. BECKMANN, R. K. HARTMANN et M. MARZ. 2014, «Genomewide comparison and novel ncRNAs of Aquificales.», *BMC genomics*, vol. 15, nº 1, doi : 10.1186/1471-2164-15-522, p. 522, ISSN 1471-2164. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-522. 84

LECUIT, M. et M. ELOIT. 2013, «The human virome : new tools and concepts.», *Trends in microbiology*, vol. 21, nº 10, doi :10.1016/j.tim.2013.07.001, p. 510–5, ISSN 1878-4380. URL http://www.ncbi.nlm.nih.gov/pubmed/23906500. 50

LINNÉ, CARL VON et L. SALVIUS. 1758, *Caroli Linnaei...Systema naturae per regna tria naturae*, vol. v.1, Impensis Direct. Laurentii Salvii,. URL http://www.biodiversitylibrary.org/item/10277#page/3/mode/1up. 47

LIU, B., T. GIBBONS, M. GHODSI, T. TREANGEN et M. POP. 2011, «Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences.», *BMC genomics*, vol. 12 Suppl 2, nº Suppl 2, doi :10.1186/1471-2164-12-S2-S4, p. S4, ISSN 1471-2164. URL http://www.biomedcentral.com/1471-2164/12/S2/S4. 55

LIU, L., Y. LI, S. LI, N. HU, Y. HE, R. PONG, D. LIN, L. LU et M. LAW. 2012, «Comparison of next-generation sequencing systems.», *Journal of biomedicine & biotechnology*, vol. 2012, doi :10.1155/2012/251364, p. 251 364, ISSN 1110-7251. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398667&tool=pmcentrez&rendertype=abstract. 42

LUEF, B., K. R. FRISCHKORN, K. C. WRIGHTON, H.-Y. N. HOLMAN, G. BIRARDA, B. C. THOMAS, A. SINGH, K. H. WILLIAMS, C. E. SIEGERIST, S. G. TRINGE, K. H. DOWNING, L. R. COMOLLI et J. F. BANFIELD. 2015, «Diverse uncultivated ultra-small bacterial cells in groundwater.», *Nature communications*, vol. 6, doi :10.1038/ncomms7372, p. 6372, ISSN 2041-1723. URL http://www.nature.com/ncomms/2015/150227/ncomms7372/full/ncomms7372.html. 50

LUKASHEV, A. N. 2010, «Recombination among picornaviruses.», *Reviews in medical virology*, vol. 20, nº 5, doi :10.1002/rmv.660, p. 327–37, ISSN 1099-1654. URL http://www.ncbi.nlm.nih.gov/pubmed/20632373. 71

MADIGAN, M. 2007, «Brock, biologie des micro-organismes», . 15

MAKEEV, V., V. RAMENSKY et M. GELFAND. 2001, «Bayesian approach to DNA segmentation into regions with different average nucleotide composition», *Computational . . .*, p. 57–73. URL http://www.springerlink.com/index/vtukvtgjw51kpmav.pdf. 83

MARDIS, E. R. 2013, «Next-generation sequencing platforms.», *Annual review of analytical chemistry (Palo Alto, Calif.)*, vol. 6, doi : 10.1146/annurev-anchem-062012-092628, p. 287–303, ISSN 1936-1335. URL http://www.ncbi.nlm.nih.gov/pubmed/23560931. 24, 25

MARTIN, D. P., P. BIAGINI, P. LEFEUVRE, M. GOLDEN, P. ROUMAGNAC et A. VARSANI. 2011, «Recombination in eukaryotic single stranded DNA viruses.», *Viruses*, vol. 3, nº 9, doi :10.3390/v3091699, p. 1699–738, ISSN 1999-4915. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3187698{&}tool=pmcentrez{&}rendertype=abstract. 71

MAYR, E. 1942, *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*, ISBN 0674862503, 334 p.. URL https://books.google.fr/books/

about/Systematics_and_the_Origin_of_Species_fr.html?id=mAIjnLp6r_MC&pgis=1. 31

MCHARDY, A. C., H. G. MARTÍN, A. TSIRIGOS, P. HUGENHOLTZ et I. RIGOUTSOS. 2007, «Accurate phylogenetic classification of variable-length DNA fragments.», *Nature methods*, vol. 4, nº 1, doi :10.1038/nmeth976, p. 63–72, ISSN 1548-7091. URL http://www.ncbi.nlm.nih.gov/pubmed/17179938. 62

MENDE, D. R., A. S. WALLER, S. SUNAGAWA, A. I. JÄRVELIN, M. M. CHAN, M. ARUMUGAM, J. RAES et P. BORK. 2012, «Assessment of metagenomic assembly using simulated next generation sequencing data.», *PloS one*, vol. 7, nº 2, doi :10.1371/journal.pone.0031386, p. e31 386, ISSN 1932-6203. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3285633&tool=pmcentrez&rendertype=abstract. 44, 76

METZKER, M. L. 2010, «Sequencing technologies - the next generation.», *Nature reviews. Genetics*, vol. 11, nº 1, doi :10.1038/nrg2626, p. 31–46, ISSN 1471-0064. URL http://www.ncbi.nlm.nih.gov/pubmed/19997069. 42

MEYER, F., D. PAARMANN, M. D'SOUZA, R. OLSON, E. M. GLASS, M. KUBAL, T. PACZIAN, A. RODRIGUEZ, R. STEVENS, A. WILKE, J. WILKENING et R. A. EDWARDS. 2008, «The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.», *BMC bioinformatics*, vol. 9, nº 1, doi :10.1186/1471-2105-9-386, p. 386, ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/9/386. 53

MINOT, S., S. GRUNBERG, G. D. WU, J. D. LEWIS et F. D. BUSHMAN. 2012, «Hypervariable loci in the human gut virome.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, nº 10, doi :10.1073/pnas.1119061109, p. 3962–6, ISSN 1091-6490. URL http://www.pnas.org/content/109/10/3962.full. 51

MONZOORUL HAQUE, M., T. S. GHOSH, D. KOMANDURI et S. S. MANDE. 2009, «SOrt-ITEMS : Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences.», *Bioinformatics (Oxford, England)*, vol. 25, nº 14, doi :10.1093/bioinformatics/btp317, p. 1722–30, ISSN 1367-4811. URL http://bioinformatics.oxfordjournals.org/content/25/14/1722.full. 53

MOOERS, A. Ø. et E. C. HOLMES. 2000, «The evolution of base composition and phylogenetic inference», *Trends in Ecology & Evolution*, vol. 15, nº 9, doi : 10.1016/S0169-5347(00)01934-0, p. 365–369, ISSN 01695347. URL http://www.sciencedirect.com/science/article/pii/S0169534700019340. 59

MOORE, W. E. C., E. STACKEBRANDT, O. KANDLER, R. R. COLWELL, M. I. KRICHEVSKY, H. G. TRUPER, R. G. E. MURRAY, L. G. WAYNE, P. A. D. GRIMONT, D. J. BRENNER, M. P. STARR et L. H. MOORE. 1987, «Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics», *International Journal of Systematic and Evolutionary Microbiology*, vol. 37, nº 4, doi :10.1099/00207713-37-4-463, p. 463–464, ISSN 1466-5026. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-37-4-463. 59

MOROWITZ, H. J., M. E. TOURTELLOTTE, W. R. GUILD, E. CASTRO, C. WOESE et R. C. CLEVERDON. 1962, «The chemical composition and submicroscopic morphology of Mycoplasma gallisepticum, Avian PPLO 5969», *Journal of Molecular Biology*, vol. 4, nº 2, doi :10.1016/S0022-2836(62)80041-2, p. 93–IN5, ISSN 00222836. URL http://www.sciencedirect.com/science/article/pii/S0022283662800412. 15

MRÁZEK, J. et S. KARLIN. 2007, «Distinctive features of large complex virus genomes and proteomes.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, nº 12, doi :10.1073/pnas.0700429104, p. 5127–32, ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1829274&tool=pmcentrez&rendertype=abstract. 76

NALBANTOGLU, O. U., S. F. WAY, S. H. HINRICHS et K. SAYOOD. 2011, «RAIphy : phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles.», *BMC bioinformatics*, vol. 12, doi :10.1186/1471-2105-12-41, p. 41, ISSN 1471-2105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3038895&tool=pmcentrez&rendertype=abstract. 62, 63, 101, 103

NASIR, A. et G. CAETANO-ANOLLÉS. «A phylogenomic data-driven exploration of viral origins and evolution», doi :10.1126/sciadv.1500527. URL http://advances.sciencemag.org/content/advances/1/8/e1500527.full.pdf. 31, 118

ONAFUWA-NUGA, A. et A. TELESNITSKY. 2009, «The remarkable frequency of human immunodeficiency virus type 1 genetic recombination.», *Microbiology and molecular biology reviews : MMBR*, vol. 73, nº 3, doi :10.1128/MMBR.00012-09, p. 451–80, Table of Contents, ISSN 1098-5557. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2738136{&}tool=pmcentrez{&}rendertype=abstract. 71

OUNIT, R., S. WANAMAKER, T. J. CLOSE, S. LONARDI, J. VENTER, K. REMINGTON, J. HEIDELBERG, A. HALPERN, D. RUSCH, J. EISEN, C. HUT-

TENHOWER, D. GEVERS, R. KNIGHT, S. ABUBUCKER, J. BADGER, A. CHINWALLA, D. HUSON, A. AUCH, J. QI, S. SCHUSTER, A. BRADY, S. SALZBERG, B. LIU, T. GIBBONS, M. GHODSI, T. TREANGEN, M. POP, N. SEGATA, L. WALDRON, A. BALLARINI, V. NARASIMHAN, O. JOUSSON, C. HUTTENHOWER, G. ROSEN, E. REICHENBERGER, A. ROSENFELD, K. PATIL, P. HAIDER, P. POPE, P. TURNBAUGH, M. MORRISON, T. SCHEFFER, S. AMES, D. HYSOM, S. GARDNER, G. LLOYD, M. GOKHALE, J. ALLEN, D. WOOD, S. SALZBERG, A. BAZINET, M. CUMMINGS, D. KOSLICKI, S. FOUCART, G. ROSEN, S. ALTSCHUL, W. GISH, W. MILLER, E. MYERS, D. LIPMAN, W. KENT, S. VINGA, J. ALMEIDA, K. MAVROMATIS, N. IVANOVA, K. BARRY, H. SHAPIRO, E. GOLTSMAN, A. MCHARDY, T. MAGOC, S. PABINGER, S. CANZAR, X. LIU, Q. SU, D. PUIU, M. ANTONIO, S. HAWES, S. HILLIER, R. HYMAN, M. FUKUSHIMA, L. DIAMOND, J. KUMM, L. GIUDICE, R. DAVIS, J. DOLEŽEL, J. VRÁNA, J. ŠAFÁŘ, J. BARTOŠ, M. KUBALÁKOVÁ, H. ŠIMKOVÁ, S. LONARDI, D. DUMA, M. ALPERT, F. CORDERO, M. BECCUTI, P. BHAT, R. LUO, B. LIU, Y. XIE, Z. LI, W. HUANG, J. YUAN, T. CLOSE, S. WANAMAKER, M. ROOSE, M. LYON, T. CLOSE, P. BHAT, S. LONARDI, Y. WU, N. ROSTOKS, L. RAMSAY, M. MASCHER, G. MUEHLBAUER, D. ROKHSAR, J. CHAPMAN, J. SCHMUTZ, K. BARRY, Q. TU, Z. HE, J. ZHOU, Z. ZHANG, S. SCHWARTZ, L. WAGNER et W. MILLER. 2015, «CLARK : fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers», *BMC Genomics*, vol. 16, nº 1, doi :10.1186/s12864-015-1419-2, p. 236, ISSN 1471-2164. URL http://www.biomedcentral.com/1471-2164/16/236. 57

PAFF, M. L., S. P. STOLTE et J. J. BULL. 2014, «Lethal mutagenesis failure may augment viral adaptation.», *Molecular biology and evolution*, vol. 31, nº 1, doi :10.1093/molbev/mst173, p. 96–105, ISSN 1537-1719. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3879444{&}tool=pmcentrez{&}rendertype=abstract. 71

PEARSON, W. R. et D. J. LIPMAN. 1988, «Improved tools for biological sequence comparison.», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, nº 8, p. 2444–8, ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=280013&tool=pmcentrez&rendertype=abstract. 41

PETERSON, A. T. 2014, «Defining viral species : making taxonomy useful.», *Virology journal*, vol. 11, nº 1, doi :10.1186/1743-422X-11-131, p. 131, ISSN 1743-422X. URL http://www.virologyj.com/content/11/1/131. 32

PRESCOTT, L. M., J. P. HARLEY et D. A. KLEIN. 2003, *Microbiologie*,

ISBN 2804142566, 1137 p.. URL https://books.google.fr/books/about/Microbiologie.html?id=_M4tHiD8VXgC&pgis=1. 15

PRIDE, D. T., R. J. MEINERSMANN, T. M. WASSENAAR et M. J. BLASER. 2003, «Evolutionary implications of microbial genome tetranucleotide frequency biases.», *Genome research*, vol. 13, nº 2, doi :10.1101/gr.335003, p. 145–58, ISSN 1088-9051. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=420360{&}tool=pmcentrez{&}rendertype=abstract. 61

PRIDE, D. T., T. M. WASSENAAR, C. GHOSE et M. J. BLASER. 2006, «Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses.», *BMC genomics*, vol. 7, doi :10.1186/1471-2164-7-8, p. 8, ISSN 1471-2164. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1360066&tool=pmcentrez&rendertype=abstract. 76

RAMENSKY, V. E., MAKEEV VJU, M. A. ROYTBERG et V. G. TUMANYAN. 2000, «DNA segmentation through the Bayesian approach.», *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, nº 1-2, doi : 10.1089/10665270050081487, p. 215–31, ISSN 1066-5277. URL http://www.ncbi.nlm.nih.gov/pubmed/10890398. 83

ROOSSINCK, M. J. 2010, «Lifestyles of plant viruses.», *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 365, nº 1548, doi :10.1098/rstb.2010.0057, p. 1899–905, ISSN 1471-2970. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2880111&tool=pmcentrez&rendertype=abstract. 17

ROOSSINCK, M. J. 2012, «Plant virus metagenomics : biodiversity and ecology.», *Annual review of genetics*, vol. 46, doi :10.1146/annurev-genet-110711-155600, p. 359–69, ISSN 1545-2948. URL http://www.ncbi.nlm.nih.gov/pubmed/22934641. 50

ROSEN, G. L., E. R. REICHENBERGER et A. M. ROSENFELD. 2011, «NBC : the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads.», *Bioinformatics (Oxford, England)*, vol. 27, nº 1, doi :10.1093/bioinformatics/btq619, p. 127–9, ISSN 1367-4811. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3008645&tool=pmcentrez&rendertype=abstract. 62, 63, 101

ROUX, S., F. ENAULT, G. BRONNER et D. DEBROAS. 2011, «Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems.», *FEMS microbiology ecology*,

vol. 78, nº 3, doi :10.1111/j.1574-6941.2011.01190.x, p. 617–28, ISSN 1574-6941. URL http://www.ncbi.nlm.nih.gov/pubmed/22066608. 51

ROUX, S., J. TOURNAYRE, A. MAHUL, D. DEBROAS et F. ENAULT. 2014, «Metavir 2 : new tools for viral metagenome comparison and assembled virome analysis.», *BMC bioinformatics*, vol. 15, doi :10.1186/1471-2105-15-76, p. 76, ISSN 1471-2105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4002922&tool=pmcentrez&rendertype=abstract. 52

SAKO, Y., N. NOMURA, A. UCHIDA, Y. ISHIDA, H. MORII, Y. KOGA, T. HOAKI et T. MARUYAMA. 1996, «Aeropyrum pernix gen. nov., sp. nov., a Novel Aerobic Hyperthermophilic Archaeon Growing at Temperatures up to 100°C», *International Journal of Systematic Bacteriology*, vol. 46, nº 4, doi :10.1099/00207713-46-4-1070, p. 1070–1077, ISSN 0020-7713. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-46-4-1070. 68

SAVORY, F., G. LEONARD et T. A. RICHARDS. 2015, «The role of horizontal gene transfer in the evolution of the oomycetes.», *PLoS pathogens*, vol. 11, nº 5, doi :10.1371/journal.ppat.1004805, p. e1004 805, ISSN 1553-7374. URL http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1004805. 19

SCHOLTHOF, K.-B. G., S. ADKINS, H. CZOSNEK, P. PALUKAITIS, E. JACQUOT, T. HOHN, B. HOHN, K. SAUNDERS, T. CANDRESSE, P. AHLQUIST, C. HEMENWAY et G. D. FOSTER. 2011, «Top 10 plant viruses in molecular plant pathology.», *Molecular plant pathology*, vol. 12, nº 9, doi :10.1111/j.1364-3703.2011.00752.x, p. 938–54, ISSN 1364-3703. URL http://www.ncbi.nlm.nih.gov/pubmed/22017770. 21

SIMMONDS, P. 2015, «Methods for virus classification and the challenge of incorporating metagenomic sequence data», doi :10.1099/jgv.0.000016. URL http://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.000016. 32, 118

SIMMONDS, P., M. J. ADAMS, M. BENKŐ, M. BREITBART, J. R. BRISTER, E. B. CARSTENS, A. J. DAVISON, E. DELWART, A. E. GORBALENYA, B. HARRACH, R. HULL, A. M. KING, E. V. KOONIN, M. KRUPOVIC, J. H. KUHN, E. J. LEFKOWITZ, M. L. NIBERT, R. ORTON, M. J. ROOSSINCK, S. SABANADZOVIC, M. B. SULLIVAN, C. A. SUTTLE, R. B. TESH, R. A. VAN DER VLUGT, A. VARSANI et F. M. ZERBINI. 2017, «Consensus statement : Virus taxonomy in the age of metagenomics», *Nature Reviews Microbiology*, vol. 15, nº 3, doi :10.1038/nrmicro.2016.177, p. 161–168, ISSN 1740-1526. URL http://www.nature.com/doifinder/10.1038/nrmicro.2016.177. 118

SIMON, C. et R. DANIEL. 2011, «Metagenomic analyses : past and future trends.», *Applied and environmental microbiology*, vol. 77, nº 4, doi :10.1128/AEM.02345-10, p. 1153–61, ISSN 1098-5336. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3067235&tool=pmcentrez&rendertype=abstract. 26

SMITH, M. R., T. MARTINEZ et C. GIRAUD-CARRIER. 2013, «An instance level analysis of data complexity», *Machine Learning*, vol. 95, nº 2, doi :10.1007/s10994-013-5422-z, p. 225–256, ISSN 0885-6125. URL http://link.springer.com/10.1007/s10994-013-5422-z. 73

SMITH, T. F. et M. S. WATERMAN. 1981, «Identification of common molecular subsequences.», *Journal of molecular biology*, vol. 147, nº 1, p. 195–7, ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/7265238. 53

SOANES, D. et T. A. RICHARDS. 2014, «Horizontal gene transfer in eukaryotic plant pathogens.», *Annual review of phytopathology*, vol. 52, doi :10.1146/annurev-phyto-102313-050127, p. 583–614, ISSN 0066-4286. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-phyto-102313-050127. 19

SOUEIDAN, H., L.-A. SCHMITT, T. CANDRESSE et M. NIKOLSKI. 2014, «Finding and identifying the viral needle in the metagenomic haystack : trends and challenges.», *Frontiers in microbiology*, vol. 5, doi :10.3389/fmicb.2014.00739, p. 739, ISSN 1664-302X. URL http://journal.frontiersin.org/article/10.3389/fmicb.2014.00739/abstract. 65, 72

STRANNEHEIM, H., M. KÄLLER, T. ALLANDER, B. ANDERSSON, L. ARVESTAD et J. LUNDEBERG. 2010, «Classification of DNA sequences using Bloom filters.», *Bioinformatics (Oxford, England)*, vol. 26, nº 13, doi :10.1093/bioinformatics/btq230, p. 1595–600, ISSN 1367-4811. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2887045&tool=pmcentrez&rendertype=abstract. 49

SUTTLE, C. A. 2007, «Marine viruses–major players in the global ecosystem.», *Nature reviews. Microbiology*, vol. 5, nº 10, doi :10.1038/nrmicro1750, p. 801–12, ISSN 1740-1534. URL http://www.nature.com/nrmicro/journal/v5/n10/execsumm/nrmicro1750.html. 52

SZTUBA-SOLIŃSKA, J., A. URBANOWICZ, M. FIGLEROWICZ et J. J. BUJARSKI. 2011, «RNA-RNA recombination in plant virus replication and evolution.», *Annual review of phytopathology*, vol. 49, doi :10.1146/annurev-phyto-072910-095351, p. 415–43, ISSN 0066-4286. URL http://www.ncbi.nlm.nih.gov/pubmed/21529157. 71

TEELING, H. et F. O. GLÖCKNER. 2012, «Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective.», *Briefings in bioinformatics*, vol. 13, nº 6, doi :10.1093/bib/bbs039, p. 728–42, ISSN 1477-4054. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3504927&tool=pmcentrez&rendertype=abstract. 44, 76

TRIFONOV, V. et R. RABADAN. 2010, «Frequency analysis techniques for identification of viral genetic data.», *mBio*, vol. 1, nº 3, doi :10.1128/mBio.00156-10, p. e00156–10–, ISSN 2150-7511. URL http://mbio.asm.org/content/1/3/e00156-10. 52, 62, 63

VAYSSIER-TAUSSAT, M., E. ALBINA, C. CITTI, J.-F. COSSON, M.-A. JACQUES, M.-H. LEBRUN, Y. LE LOIR, M. OGLIASTRO, M.-A. PETIT, P. ROUMAGNAC et T. CANDRESSE. 2014, «Shifting the paradigm from pathogens to pathobiome : new concepts in the light of meta-omics.», *Frontiers in cellular and infection microbiology*, vol. 4, doi :10.3389/fcimb.2014.00029, p. 29, ISSN 2235-2988. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3942874&tool=pmcentrez&rendertype=abstract. 22

VILLARREAL, L. P., V. R. DEFILIPPIS et K. A. GOTTLIEB. 2000, «Acute and persistent viral life strategies and their relationship to emerging diseases.», *Virology*, vol. 272, nº 1, doi :10.1006/viro.2000.0381, p. 1–6, ISSN 0042-6822. URL http://www.sciencedirect.com/science/article/pii/S0042682200903817. 18

VINGA, S. et J. ALMEIDA. 2003, «Alignment-free sequence comparison–a review», *Bioinformatics*, vol. 19, nº 4, doi :10.1093/bioinformatics/btg005, p. 513–523, ISSN 1367-4803. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg005. 59

WANG, Q., G. M. GARRITY, J. M. TIEDJE et J. R. COLE. 2007, «Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.», *Applied and environmental microbiology*, vol. 73, nº 16, doi :10.1128/AEM.00062-07, p. 5261–7, ISSN 0099-2240. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950982{&}tool=pmcentrez{&}rendertype=abstract. 62

WOESE, C. R. et G. E. FOX. 1977, «Phylogenetic structure of the prokaryotic domain : The primary kingdoms», *Proceedings of the National Academy of Sciences*, vol. 74, nº 11, doi :10.1073/pnas.74.11.5088, p. 5088–5090, ISSN 0027-8424. URL http://www.pnas.org/content/74/11/5088.full. 68

WOMMACK, K. E., J. BHAVSAR et J. RAVEL. 2008a, «Metagenomics : read length matters.», *Applied and environmental microbiology*, vol. 74,

nº 5, doi :10.1128/AEM.02181-07, p. 1453–63, ISSN 1098-5336. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2258652&tool=pmcentrez&rendertype=abstract. 42

WOMMACK, K. E., J. BHAVSAR et J. RAVEL. 2008b, «Metagenomics : read length matters.», *Applied and environmental microbiology*, vol. 74, nº 5, doi :10.1128/AEM.02181-07, p. 1453–63, ISSN 1098-5336. URL http://aem.asm.org/content/74/5/1453.long. 44

WOOD, D. E. et S. L. SALZBERG. 2014, «Kraken : ultrafast metagenomic sequence classification using exact alignments.», *Genome biology*, vol. 15, nº 3, doi :10.1186/gb-2014-15-3-r46, p. R46, ISSN 1465-6914. URL http://genomebiology.com/2014/15/3/R46. 57, 101

WOOLEY, J. C., A. GODZIK et I. FRIEDBERG. 2010, «A primer on metagenomics.», *PLoS computational biology*, vol. 6, nº 2, doi :10.1371/journal.pcbi.1000667, p. e1000667, ISSN 1553-7358. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2829047&tool=pmcentrez&rendertype=abstract. 27, 42, 43

WREN, J. D., M. J. ROOSSINCK, R. S. NELSON, K. SCHEETS, M. W. PALMER et U. MELCHER. 2006, «Plant virus biodiversity and ecology.», *PLoS biology*, vol. 4, nº 3, doi :10.1371/journal.pbio.0040080, p. e80, ISSN 1545-7885. URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040080. 21

WU, M. et J. A. EISEN. 2008, «A simple, fast, and accurate method of phylogenomic inference.», *Genome biology*, vol. 9, nº 10, doi :10.1186/gb-2008-9-10-r151, p. R151, ISSN 1465-6914. URL http://genomebiology.com/2008/9/10/R151. 31

XIAO, C., Y. G. KUZNETSOV, S. SUN, S. L. HAFENSTEIN, V. A. KOSTYUCHENKO, P. R. CHIPMAN, M. SUZAN-MONTI, D. RAOULT, A. MCPHERSON et M. G. ROSSMANN. 2009, «Structural studies of the giant mimivirus.», *PLoS biology*, vol. 7, nº 4, doi :10.1371/journal.pbio.1000092, p. e92, ISSN 1545-7885. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2671561&tool=pmcentrez&rendertype=abstract. 50

YANG, A. C.-C., A. L. GOLDBERGER et C.-K. PENG. 2005, «Genomic classification using an information-based similarity index : application to the SARS coronavirus.», *Journal of computational biology : a journal of computational molecular cell biology*, vol. 12, nº 8, doi :10.1089/cmb.2005.12.1103, p. 1103–16, ISSN 1066-5277. URL http://www.ncbi.nlm.nih.gov/pubmed/16241900. 52

YU, C., T. HERNANDEZ, H. ZHENG, S.-C. YAU, H.-H. HUANG, R. L. HE, J. YANG et S. S.-T. YAU. 2013, «Real time classification of viruses in 12 dimensions.», *PloS one*, vol. 8, nº 5, doi :10.1371/journal.pone.0064328, p. e64 328, ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/23717598http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3661469. 59, 118

ZHANG, Z., S. SCHWARTZ, L. WAGNER et W. MILLER. 2004, «A greedy algorithm for aligning DNA sequences.», *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, nº 1-2, doi :10.1089/10665270050081478, p. 203–14, ISSN 1066-5277. URL http://online.liebertpub.com/doi/abs/10.1089/10665270050081478. 95

Annexes

A.1 Figures annexes

FIGURE A.20 – (3 pages suivantes) Alignement multiple des segments 1 de la grippe humaine A (8486138), B (8486164) et C (52630349). Selon la documentation de T-coffee, le score d'alignement est considéré comme "pretty good" (sic) à partir de 40 et au-delà. Le score d'alignement actuel (29) confirme donc la mauvaise qualité de l'alignement.

```
T-COFFEE, Version_8.93(Thu Aug 5 18:09:23 CEST 2010)
Cedric Notredame
CPU TIME:0 sec.
SCORE=29
*
 BAD  AVG  GOOD
*
gi|52630349|ref    :  29
gi|8486138|ref|    :  30
gi|8486164|ref|    :  28
cons               :  29

gi|52630349|ref    AT---------GTCTTTTCTATTGACAATAGCA
gi|8486138|ref|    AGCGAAAGCAGGTCAATTATATT--CAATATGG
gi|8486164|ref|    AGCAGAAGCGGAGCT-TTAAGAT--GAATATAA
cons               *           *  **    *   ****

gi|52630349|ref    AAGGAATACAAAAGACTATGC----CAAGAT-G
gi|8486138|ref|    AAAGAATAAA-AGAACTAAGAAATCTAATGTCG
gi|8486164|ref|    AT-CCATATTTTCTTTTCATAGATGTACCTATA
cons               *    ***        *        *

gi|52630349|ref    C--TAAGGCAGC--TCAA-ATGATGACAGTAGG
gi|8486138|ref|    CAGTCTCGCACCCGCGAG-ATACTCACAAAAAC
gi|8486164|ref|    CAGGC-AGCAATTTCAACAACATTCCCATACAC
cons               *      ***      *  *   *  **

gi|52630349|ref    AACTGT---ATCAAACTACACTACG-TTCAAGA
gi|8486138|ref|    CACCGTGG-ACCATAT--GGCCATA-ATCAAGA
gi|8486164|ref|    CGGTGTTCCCCCTTAT--TCTCATGGAACGGGA
cons                  **     *  *       *     *  **

gi|52630349|ref    A----ATGGACTACATCAAGGAAGGAAAAGAAT
gi|8486138|ref|    A----GTACACATCAGGAAGACAGGAGAAGAAC
gi|8486164|ref|    ACAGGCTACACAATAGACA-CCGTGATTAGAAC
cons               *     *  **   *   *     **  ****

gi|52630349|ref    CCTTCACTAAGAATGAGATGGGCAATGAGCAGC
gi|8486138|ref|    CCAGCACTTAGGATGAAATGGATGATGGCAATG
gi|8486164|ref|    ACAC----------------------------
cons                *

gi|52630349|ref    AAATTCCC-AATAATAGCTAACAAGAGAATGCT
gi|8486138|ref|    AAATATCC-AATTACAGCAGACAAGAGGATAAC
gi|8486164|ref|    GAGTACTCAAACAAGGGAAAACAATACATTTCT
cons                * *   * **  *  *   **** *   *

gi|52630349|ref    GGAAGAAGCTCAAATTCCTA-AAG---AACACA
gi|8486138|ref|    GGAAA----T--GATTCCTGAGAG---AAATGA
gi|8486164|ref|    G-ATGTTACAGG-ATGTGTA-ATGGTAGATCCA
cons               * *          **   *    *   *   *

gi|52630349|ref    ACAATGTAGCCCTT-TGGGAAGACACAGAA---
gi|8486138|ref|    GCAAGGACAAACTTT-ATGGAGTAAAA----TG
gi|8486164|ref|    ACAAATGGGCCATTACCCGAAGACAATGAACCG
cons                ***        **    * **  *

gi|52630349|ref    GATG---TTTCAAAAAGGGATCATGTTCTTGCA
gi|8486138|ref|    AATG---ATGCCGGATCAGACCGAGTGAT-GGT
gi|8486164|ref|    AGTGCCTATGCACAATTGGATTGTGTTCT-GGA
cons                 **    * *   *   **    **  * *

gi|52630349|ref    AGCGCCTCTTG-TATAAATTATTGGAAT-TTTT
gi|8486138|ref|    ATCACCTCTGGCTGTGACATGGTGGAAT-AGGA
gi|8486164|ref|    GG---CTTT-G-GATAGAATGGATGAAGAACAT
cons                  ** * *   *    *    ***

gi|52630349|ref    GTGGACCTTGTGTCAACAATTCAGAA-GTGATC
gi|8486138|ref|    ATGGACC-AATGACAAA--TACAGTTCATTATC
gi|8486164|ref|    CCAGGTC-TGTTTCAAGCAG---------GGTC
cons                  *  *   *  ***             **

gi|52630349|ref    AAAGAAGTTTATAAATCTA------------GA
gi|8486138|ref|    CAA-AAATCTACAAAACTT------------AT
gi|8486164|ref|    ACAGAATGCCATGGAGGCACTAATGGTCACAAC
cons                 * **    *   *

gi|52630349|ref    TTTGGAAGATTAGAAAGAAGGAAAGA---AATA
gi|8486138|ref|    TTTGAAAGAGTCGAAAGGCTAAAGCATGGAACC
gi|8486164|ref|    AGTGGACAAATTGACTCA-GGGGAGACA-GACC
cons                 ** *  * * **          *    *

gi|52630349|ref    ATGTGG-AAAGAA-----CTTAGATTTACATTA
gi|8486138|ref|    T--TTGGCCCTGTC-CATTTTAGAAACC---AA
gi|8486164|ref|    T--TTG-ATTGGACGGTGTGTAGAAACC---AA
cons                  * *           ****       *

gi|52630349|ref    GTTGATAGACAACGAAGAAGA-GTTGACACTCA
gi|8486138|ref|    G---TCAAAATACGTCGGAGA-GTTGACATAAA
gi|8486164|ref|    CCTGCTG--CAACGGCACTGAACACAACAATAA
cons                        ***     **     ***   *

gi|52630349|ref    GCC--TGTAGAA-CAAAGA-TTGAGAACTGGAG
gi|8486138|ref|    TCC--TGGTCATG-C-AGATCTCAGTGCCAAGG
gi|8486164|ref|    CCTCTTTAGGTTGAATGA-TTTA--AATGGAG
cons                *   *          **  * *        *

gi|52630349|ref    AAATTAAAGACTTGCAAATGTGGACTTTGTT--
gi|8486138|ref|    AGGC-ACAGG-ATGTAATCATGGAAGTTGTTTT
gi|8486164|ref|    CCGAC-AAGG-GTGGATTAGTGCCCTT-TTG--
cons                     **   ** *    **    *  *

gi|52630349|ref    -CGAAGATGAAGCTCCTCTTGCTAGCAAATTTA
gi|8486138|ref|    CCCTA-ACGAAGTGGGA---GCCAGGATACTAA
gi|8486164|ref|    -CCAAGAT-ATCATTGA---TTCATTAGACAAA
cons                *  * *  *           *  * *   *

gi|52630349|ref    T--------TTTAGACAATTATGGTCTAGTCAA
gi|8486138|ref|    CATCGGAATCGCAACTAACGATAACCAAAGAGA
gi|8486164|ref|    C--------CTGAAATGATTTTCTTCACAGTAA
cons                        *   *   *   *     *

gi|52630349|ref    AGAAATGAGAT-------------CAAAGTTTG
gi|8486138|ref|    AGAAAGAAGAACT-----------CCAGGATTG
gi|8486164|ref|    AGAATATAAAGAAAAAATTGCCTGCTAAAAACA
cons               ****   * *             * *

gi|52630349|ref    CAAACAAACCTCTGAATAAAGA---AG------
gi|8486138|ref|    CAAA---ATTTCTCCTTTGA-------------
gi|8486164|ref|    GAAAG---GGTTTCCTTATAAAAAGAATACCTA
cons                ***     * *   *  *

gi|52630349|ref    ---TAGTTGCACACATGTTAG--AAAAACAATT
gi|8486138|ref|    ---TGGTTGCATACATGTTGG--AGAGAGAACT
gi|8486164|ref|    TGAAGGTAAAAGACAGAATAACAAGAGTGGAAT
cons                    **   * ***   *   * *   * *

gi|52630349|ref    CAATCCGG-AAAGTAGATTCTTGCCTGTTTTCG
gi|8486138|ref|    GG-TCCGCAAAACGAGATTCCTCCCAGTGGCTG
gi|8486164|ref|    ACATC----AAAA---------GAGCATTATCA
cons                  **    ***             *

gi|52630349|ref    -GAGCTATAAGG-CCAGAAAGAATGGAATTGAT
gi|8486138|ref|    -GTGGAACAAGC-AGTGTGTACATTGAAGTGTT
gi|8486164|ref|    TTAAACACAATGACTAAAGATGCTGAAA--GAG
cons                     * **            *  **  *

gi|52630349|ref    CCA-TGCATTAGGAGGAGAAACTTGGATACAAG
gi|8486138|ref|    GCA-TTTGACTCAAGGAACATGCTGGGAACA-G
gi|8486164|ref|    GCAAACTAAAAAGAAGAGCAATTG--CCACCGC
cons                **          * **  *       **

gi|52630349|ref    AAGCTAACA-CTGCAG-GGATTTCCAAT-GTTG
gi|8486138|ref|    ATGTATACTCCAGGAG-GGGAAGTGAAGA-ATG
gi|8486164|ref|    TGGGATACA-AATCAGAGGATTTGTATTAGTAG
cons                 *   **       ** **      *      *

gi|52630349|ref    ATCAAA-----GGAAAAATGATATGAGAGCAGT
gi|8486138|ref|    ATGATGT--TGATCAAAGCTTGATTATTGCT--
gi|8486164|ref|    TTGAAAACTTGGCTAAAAATATCTGTGAA-AAT
cons                * *          ***      *

gi|52630349|ref    ATGTAGGAAAGTTTGTCTTGCAGCAAATGCAAG
gi|8486138|ref|    -GCTAGGAACAT-AGTG--AGAAGAGCTGC-AG
```

```
gi|8486164|ref|   CTAGAGCAAAGT-GGTTTACCCGTAGGTGGAAA
cons                  ** **  *  **        *  **  *

gi|52630349|ref   TATAATGAACGCCAAAAGCAAACTGGTTGAGTA
gi|8486138|ref|   TATCAGCAGACCCACTAGCATCTTTATTGGAGA
gi|8486164|ref|   CGAAAAGAAGGCCAAA--CTATCAAATGCAGTG
cons                  *  *   ***    *       *

gi|52630349|ref   TATAAAAGTACAAGTAT-GAG-AATTGG--AG
gi|8486138|ref|   TGTG-----------------CCACAGCACAC
gi|8486164|ref|   GCTAAAATGCTCAGTAATTGTCCACCAGG--AG
cons                *                        *   *

gi|52630349|ref   -AAACAGAAAGAAAGCTTGAAGAACTTATACTT
gi|8486138|ref|   AGATTGGTGGAATTAGGATGGTAGACATCCTTA
gi|8486164|ref|   GGATCAGTATGACTGTGACAGGAGACAATACTA
cons                *   *    *          *        *

gi|52630349|ref   ----GAAACCGAT-GA-TGTCTCACCTGAAGTA
gi|8486138|ref|   AGCAGAACCCAACAGA-AGAGCAAGCCGTG---
gi|8486164|ref|   -----AATGGAAT-GAATGCTTAAATCCAAGAA
cons                   **    *  **  *    *

gi|52630349|ref   ACATTATG--TAAATCTGCTTTAGGAG-GACCA
gi|8486138|ref|   GGTATATG--CAAGGCTGCAATGGGACTGAGAA
gi|8486164|ref|   TCTTTTTGGCTATGACTGAAAGAATA------A
cons                  * **   *   ***       *      *

gi|52630349|ref   TTAGGAAAAACTCTA---TCTTTTGGGCCCATG
gi|8486138|ref|   TTAGCTCATCCTTC----AGTTTTGGTGGATTC
gi|8486164|ref|   CCAGAGACAGCCCAATTTGGTTCCGGGATTTTT
cons                **      *         **  **     *

gi|52630349|ref   CTACTCAAGAAAATTTCTGGTTCCGGAGTAAAA
gi|8486138|ref|   ACATTTAAGAGAACAAGCGGATCATCAGTCAAG
gi|8486164|ref|   GTAGTATAGCACCGGTCTTGTTCTCCAATAAAA
cons                * *  **          * **   * * **

gi|52630349|ref   ---GTTAAAGATACAGTATATA-TCCAAGGTGT
gi|8486138|ref|   AGAGAGGAAGAGGTGC-TTACG-GGCAA--TCT
gi|8486164|ref|   ---TAGCTAGATTGGGAAAAGGGTTCATGATA-
cons                      ***        *     **   *

gi|52630349|ref   CAGAGCAGT-ACAATTTGAATACTGGAGTGAGC
gi|8486138|ref|   TCAAACATTGAAGATAAGAGTGCATGAGGGA-T
gi|8486164|ref|   ---ACAAGT-AAAACAAAAAGACTAAAA-GCTC
cons                 *  * * *  *    *   *   *  *

gi|52630349|ref   AAGAAGAATTCT--ATGGAGAATATAAGTCAGC
gi|8486138|ref|   ATGAAGAGTTCACAATGGTTGGGAGA--AGAGC
gi|8486164|ref|   AAATACCTT-GT-CCCGATCTGTTTAATATACC
cons              *   *   *       *        *    * *

gi|52630349|ref   CACCGCTTTATTCAGCAGAAAGGAAAG-ATCAC
gi|8486138|ref|   AACAGCCATACTCAGA---AAAGCAA---CCAG
gi|8486164|ref|   ATTAGAAAGATATAAT---GAAGAAACAAGGGC
cons                  *    *   *      * * **

gi|52630349|ref   TAGAATGGATTACAA---TAGGAGGAGGAATAA
gi|8486138|ref|   GAGATT-GATTCAGCTGATAGTGAGTGGGAGAG
gi|8486164|ref|   AAAACT-GAAAAAGC---TAAAACCTTTCTTCA
cons               * * * **         **

gi|52630349|ref   ATGAAGACA-GAAAG--AGACTTCTAGCTATGT
gi|8486138|ref|   ACGAACAGTCGATTGCCGAAGCAATAATTGTGG
gi|8486164|ref|   ATGAAGAAG-GAACG--GCATCT----CTTTCG
cons              * *** *   **  *    *        * *

gi|52630349|ref   GCATGATATTTTGCAGAGATGGAGATTATTTTA
gi|8486138|ref|   CCATGGTATTTTC-AC--AAGAGGATTGTATGA
gi|8486164|ref|   CCA-GGAATGAT---GAT--G-GGAATGTTTAA
cons               ** *  **  *        *  ** * * * *

gi|52630349|ref   AAGACGC-CCCTGCA-ACAATAACAATGGCAGA
gi|8486138|ref|   TAAAAGC-AGTTAGAGGTGATCTGAATTTCGTC
gi|8486164|ref|   TA--TGCTATCTACA-GTATTAGGAGTAGCCGC
cons               *   **    *  *     *   * *  *

gi|52630349|ref   TTTAAGTACGAA----G-TTAGGAAGAGAAATT
gi|8486138|ref|   AATAGGGCGAATCAGCGACTGAATCCTATGCAT
gi|8486164|ref|   ACTAGGGATAAAAAACA-TTGGAAACAAAGAAT
cons                ** *    *        *            *

gi|52630349|ref   --CCATATCAATATGTGATGATGAATT-GGAT-
gi|8486138|ref|   CAACTTTTA---------------------AG-
gi|8486164|ref|   --ACTTATGG-GATGGACTGCAGTCTTCGGATG
cons                 * * *                      *

gi|52630349|ref   ACAAAAATCAGAA-GATAATCTCGAAGCCTTAT
gi|8486138|ref|   ACA-TTTTCAGAA-GG--ATGCGAAAGTGCTTT
gi|8486164|ref|   ATTTTGCTCTGTTTGTTAATG-CAAAAGATGAA
cons              *      ** *   *   **    **

gi|52630349|ref   TATACA-GTAGGGGAATTGTAGAAACCAATC-C
gi|8486138|ref|   TTCAAA-ATTGGGGAGTTG----AACCTATC-G
gi|8486164|ref|   GAGACATGTATGG-AA-GGAATAAACGATTTTT
cons                 * *  *  ** *   *    ***   *

gi|52630349|ref   AGGAAA-AATGGGGAGCTCAATGGGA-ATTG--
gi|8486138|ref|   ACAATGTGATGGGAA--TGATTGGGATATTGCC
gi|8486164|ref|   ACCGAA-CATG-TAAGCT-ATTGGGA-ATAA--
cons              *       ***   *  * * ***** **

gi|52630349|ref   ----ATGGTTCCAAAAGAGCAATTA--AAT-CT
gi|8486138|ref|   CGACATGACTCCAA----GCATCGAGATGT-CA
gi|8486164|ref|   ----ACA-------------------TGAGCA
cons                  *                          *

gi|52630349|ref   TTAAGGGCTGTCACAAT----ACAATCAGGAAA
gi|8486138|ref|   ATGAGAGGAGTGAGAATCAGCAAAATGGGTGTA
gi|8486164|ref|   AAAAGAAAAGTTACTGTAATGAAACTGGG-ATG
cons                 **    ** *   *    * * *  *

gi|52630349|ref   GATTGACATGCCAGAATCAAAAGAAAAAATTCA
gi|8486138|ref|   GATGAGTACTCCAGCACGGAGAGGGTA-----G
gi|8486164|ref|   TTTGAATTTACCAGCATGTTT--TACA-----G
cons                *       **** *          *

gi|52630349|ref   CCTTGAGCTCTCTGATAATCTTGAAG-------
gi|8486138|ref|   TGGTGAGCATTGACCGGTTCTTGAGAGTCCGGG
gi|8486164|ref|   AGATGGATTTGTATCTAATTTTGCAA---TGGA
cons                 **             * ***

gi|52630349|ref   --------------------------------C
gi|8486138|ref|   ACCAACGAGGAAATGTACTACTGTCTCCCGAGG
gi|8486164|ref|   AC-TCCCTTCATTTGGA-----GTCGCTGGAGT
cons

gi|52630349|ref   ATTTGATTCATCAGGAA---GAATTGTTGCAAC
gi|8486138|ref|   AGGTCAGTGAA-ACACA-GGGAACAGAGAAACT
gi|8486164|ref|   GAATGAATCAGCAGACATGGCAATAGGAATGAC
cons                 * * * *  *   *    **  *

gi|52630349|ref   A---ATTTTAGACCTTCCTAGTGACAAAAAGGT
gi|8486138|ref|   GACAATAA----CTTACTCATCGTCAATGATGT
gi|8486164|ref|   AATAATAAAGAACAATATGATCAACAATGGGAT
cons                  **      *      *    ***     *

gi|52630349|ref   AACATTTCAGGATGTAAGCTTTCAACATCCT-G
gi|8486138|ref|   G--G------------G-AGATTAATGGTCCTG
gi|8486164|ref|   G--G------------GCCCAGCAACGGCAC-A
cons                                     **

gi|52630349|ref   ATCTGGCA-----GTATTGA-G-----------
gi|8486138|ref|   AAT---CA--GTGTTGGTCAATACCTATCAATG
gi|8486164|ref|   AACAGCCATACAATTATTCA-TAGCTGACTATA
cons              *     **      *  * *

gi|52630349|ref   -----------AGATGAGAAAACGGCCATAACA
gi|8486138|ref|   GAT-CATC-AGAA-----ACTGGGAAACTGTTA
gi|8486164|ref|   GATACACCTACAAATGCCACAGGGGAGATTCCA
cons                         *      *    *    *   *

gi|52630349|ref   AAAGGGTATGAAGCGCTAATCAAAAGGCTAGGA
```


gi|8486138|ref|
gi|8486164|ref|
cons
gi|52630349|ref
gi|8486138|ref|
gi|8486164|ref|
cons

A.2 Tableaux annexes

TABLEAU A.5 – Liste des caractères utilisés lors de la notation de nucléotides ambigus selon la notation définie par l'IUPAC

Notation IUPAC	Signification	Possibilités			
W	**W**eak	A			T
S	**S**trong		C	G	
M	a**M**ino	A	C		
K	**K**eto			G	T
R	pu**R**ine	A		G	
Y	p**Y**rimidine		C		T
B	tout sauf A (**B** suit A)		C	G	T
D	tout sauf C (**D** suit C)	A		G	T
H	tout sauf G (**H** suit G)	A	C		T
V	tout sauf T/U (**V** suit T et U)	A	C	G	
N	n'importe quel **N**ucléotide	A	C	G	T

A.3 Summary of 26 species concepts

Reproduit avec l'aimable permission de John S. Wilkins

There are numerous species concepts at the object-level in the literature. (Mayden 1997) has listed 22 distinct species concepts along with synonyms, and his metataxonomy provides a useful starting point for a review. I have added authors where I can locate them in addition to Mayden's references, and instead of his abbreviations I have tried to give the concepts names, such as biospecies for Biological Species, etc. (following George 1956) except where nothing natural suggests itself. There have also been several additional concepts since Mayden's review, which I have added (the views of Pleijel, Wu, and Hey), and several new revisions presented in (Wheeler and Meier 2000). In addition, I add some "partial" species concepts - the compilospecies concept and the nothospecies concept. I have made a distinction between two phylospecies concepts that go by various names, mostly the names of the authors presenting at the time (as in the Wheeler and Meier volume). To remedy this, I have christened them the Autapomorphic species concept and the Monophyletic species concept.

A.3.1 Agamospecies

Synonyms : Microspecies, paraspecies, pseudospecies, semispecies, quasipecies
Principal authors : Cain 1954, Eigen 1993 (quasispecies)
Specifications : Asexual lineages, uniparental organisms. May be secondarily uniparental from biparental ancestors

A.3.2 Autapomorphic species

See : Phylospecies
Principal authors : Nelson and Platnick 1981 ; Rosen 1979

A.3.3 Biospecies

Synonyms : Syngen, speciationist species concept
Related concepts : Biological species concept, Genetic species, isolation species
Principal authors : Ray, Buffon, Dobzhansky 1935, Mayr 1942
Specifications : Inclusive Mendelian population of sexually reproducing organisms (Dobzhansky 1935 ; Dobzhansky 1937 ; Dobzhansky 1970), interbreeding natural population isolated from other such groups (Mayr 1942 ; Mayr 1963 ; Mayr 1970 ; Mayr and Ashlock 1991). Depends upon endogenous reproductive isolating mechanisms (RIMs).

A.3.4 Cladospecies

Synonyms : Internodal species concept, Hennigian species concept, Hennigian convention
Principal authors : Hennig 1950 ; Hennig 1966 ; Kornet 1993
Specifications : Set of organisms between speciation events or between speciation event and extinction (Ridley 1989), a segment of a phylogenetic lineage between nodes. Upon speciation the ancestral species is extinguished and two new species are named.

A.3.5 Cohesion species

Synonyms : Cohesive individual (in part) (Ghiselin and Hull)
Principal authors : Templeton 1989
Specifications : Most inclusive population with potential for phenotypic cohesion through intrinsic cohesion mechanisms ... having the potential for genetic and/or demographic exchangeability (Templeton 1989).

A.3.6 Compilospecies

Synonyms : None

Related concepts : Introgressive taxa

Principal authors : Aguilar et al. 1999 ; Harlan and De Wet 1963

Specifications : A species pair where one species “plunders” the genetic resources of another via introgressive interbreeding.

A.3.7 Composite Species

Synonyms : Phylospecies (in part), Internodal species (in part), cladospecies (in part)

Principal authors : Kornet and McAllister 1993

Specifications : All organisms belonging to an internodon and its descendents until any subsequent internodon. An internodon is defined as a set of organisms whose parent-child relations are not split (have the INT relation).

A.3.8 Ecospecies

Synonyms : Ecotypes

Related concepts : Evolutionary species sensu Simpson, Ecological mosaics

Principal authors : Simpson 1961 ; Sterelny 1999 ; Turesson 1922 ; Van Valen 1976

Specifications : A lineage (or closely related set of lineages) which occupies an adaptive zone minimally different from that of any other lineage in its range and which evolves separately from all lineages outside its range.

A.3.9 Evolutionary species

Synonyms : Unit of evolution, evolutionary group

Related concepts : Evolutionary significant unit

Principal authors : Simpson 1961 ; Wiley 1978 ; Wiley 1981

Specifications : A lineage (an ancestral-descendent sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies (Simpson).

A.3.10 Evolutionary significant unit

Synonyms : Biospecies (in part) and evolutionary species (in part)

Principal authors : Waples 1991

Specifications : A population (or group of populations) that 1) is substantially reproductively isolated from other conspecific population units, and 2) represents an important component in the evolutionary legacy of the species.

A.3.11 Genealogical concordance species

Synonyms : Biospecies (in part), cladospecies (in part), phylospecies (in part)

Principal authors : Avise and Ball 1990

Specifications : Population subdivisions concordantly identified by multiple independent genetic traits constitute the population units worthy of recognition as phylogenetic taxa

A.3.12 Genic species

Synonyms : none

Related concepts : Genealogical concordance species, genetic species (in part), biospecies (in part), autapomorphic species (in part)

Principal author : Wu 2001a ; 2001b

Specifications : A species formed by the fixation of all isolating genetic traits in the common genome of the entire population.

A.3.13 Genetic species

Synonyms : Gentes (sing. Gens)

Related concepts : Biospecies, phenospecies, morphospecies

Principal authors : Dobzhansky 1950 ; Mayr 1969 ; Simpson 1943

Specifications : Group of organisms that may inherited characters from each other, common gene pool, reproductive community that forms a genetic unit

A.3.14 Genotypic cluster

Synonyms : Polythetic species

Related concepts : Agamospecies, biospecies, genetic species, Hennigian species, morphospecies, non-dimensional species, phenospecies, autapomorphic phylospecies, successional species, taxonomic species

Principal author : Mallet 1995

Specifications : Clusters of monotypic or polytypic biological entities, identified using morphology or genetics, forming groups that have few or no intermediates when in contact.

A.3.15 Hennigian species

Synonyms : Biospecies (in part), cladospecies (in part), phylospecies (in part)

Principal authors : Hennig 1950 ; Hennig 1966 ; Meier and Willman 1997

Specifications : A tokogenetic community that arises when a stem species is dissolved into two new species and ends when it goes extinct or speciates.

A.3.16 Internodal species

Synonyms : Cladospecies and Hennigian species (in part), phylospecies

Principal author : Kornet 1993

Specifications : Organisms are conspecific in virtue of their common membership of a part of a genealogical network between two permanent splitting events or a splitting event and extinction

A.3.17 Least Inclusive Taxonomic Unit (LITUs)

Synonyms : evolutionary group (in part), phylospecies

Principal authors : Pleijel 1999 ; Pleijel and Rouse 2000

Specifications : A taxonomic group that is diagnosable in terms of its autapomorphies, but has no fixed rank or binomial.

A.3.18 Monophyletic species

See : Phylospecies

Principal authors : Cracraft 1983 ; Eldredge and Cracraft 1980 ; Nixon and Wheeler 1990

A.3.19 Morphospecies

Synonyms : Classical species

Related concepts : Linnean species, binoms, phenospecies, monothetic species, monotypes, types, Taxonomic species

Principal authors : Aristotle and Linnaeus, and too many others to name, but including Owen, Agassiz, and recently, Cronquist 1978

Specifications : Species are the smallest groups that are consistently and persistently distinct, and distinguishable by ordinary means (Cronquist 1978).

A.3.20 Non-dimensional species

Synonyms : Folk taxonomical kinds (Atran 1990)

Related concepts : Biospecies, genetic species, morphospecies, paleospecies, successional species, taxonomic species

Principal authors : Mayr 1942 ; Mayr 1963

Specifications : Species delimitation in a non-dimensional system (a system without the dimensions of space and time, Mayr 1963)

A.3.21 Nothospecies

Synonyms : hybrid species, reticulate species

Related concepts : Compilospecies, horizontal or lateral transfer

Principal author : Wagner 1983

Specifications : Species formed from the hybridisation of two distinct parental species, often by polyploidy.

A.3.22 Phylospecies

Synonyms : Autapomorphic phylospecies, monophyletic phylospecies, minimal monophyletic units, monophyletic species, lineages

Related concepts : Similar to internodal species cladospecies, composite species, least inclusive taxonomic units.

Principal authors : Cracraft 1983 ; Eldredge and Cracraft 1980 ; Nelson and Platnick 1981 ; Rosen 1979

Specifications : The smallest unit appropriate for phylogenetic analysis, the smallest biological entities that are diagnosable and monophyletic, unit product of natural selection and descent. A geographically constrained group with one or more unique apomorphies (autapomorphies). There are two versions of this and they are not identical. One derives from Rosen and is what I call the autapomorphic species concept. It is primarily a concept of diagnosis and tends to be favoured by the tradition known as pattern cladism. The other is what I call the monophyletic species concept, and tends to be favoured by process cladists.

A.3.23 Phenospecies

Synonyms : Phena (sing. phenon) (Smith 1994)

Related concepts : Biospecies, genetic concordance species, morphospecies, nondiensional species, phylospecies (in part), phenospecies, successional species, taxonomic species, quasispecies, viral species

Principal authors : Beckner 1959 ; Eigen 1993 ; Sokal and Sneath 1963

Specifications : A cluster of characters that statistically covary, a family resemblance concept in which possession of most characters is required for inclusion in a species, but not all. A class of organisms that share most of a set of characters.

A.3.24 Recognition species

Synonyms : Specific mate recognition system (SMRS)
Related concepts : Biospecies
Principal author : Paterson 1985
Specifications : A species is that most inclusive population of individual, biparental organisms which share a common fertilization system

A.3.25 Reproductive competition species

Synonyms : Hypermodern species concept , Biospecies (in part)
Principal author : Ghiselin 1974
Specifications : The most extensive units in the natural economy such that reproductive competition occurs among their parts.

A.3.26 Successional species

Synonyms : Paleospecies, evolutionary species (in part), chronospecies
Principal authors : George 1956 ; Simpson 1961
Specifications : Arbitrary anagenetic stages in morphological forms, mainly in the paleontological record.

A.3.27 Taxonomic species

Synonyms : Cynical species concept (Kitcher 1984)
Related concepts : Agamospecies, genealogical concordance species, morphospecies, phenospecies, phylospecies
Principal author : Blackwelder 1967, but see (Regan 1926)
Specifications : Specimens considered by a taxonomist to be members of a kind on the evidence or on the assumption they are as alike as their offspring of hereditary relatives within a few generations. Whatever a competent taxonomist chooses to call a species.

A.3.28 References

Aguilar, J. F., J. A. Roselló, and G. N. Feliner. 1999. Molecular evidence for the compilospecies model of reticulate evolution in Armeria (Plumbaginaceae). Systematic Biology 48 :735-754.

Atran, S. 1990. The cognitive foundations of natural history. Cambridge University Press, New York.

Avise, J. C., and R. M. Ball Jr. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Pp. 45-67 in D. Futuyma and J. Atonovics, eds. Oxford Surveys in Evolutionary Biology. Oxford University Press, Oxford.

Beckner, M. 1959. The biological way of thought. Columbia University Press, New York.

Blackwelder, R. E. 1967. Taxonomy : a text and reference book. Wiley, New York.

Cain, A. J. 1954. Animal species and their evolution. Hutchinson University Library, London.

Cracraft, J. 1983. Species concepts and speciation analysis. Current Ornithology 1 :159-187.

Cronquist, A. 1978. Once again, what is a species? Pp. 3-20 in L. Knutson, ed. BioSystematics in Agriculture. Alleheld Osmun, Montclair, NJ.

Dobzhansky, T. 1935. A critique of the species concept in biology. Philosophy of Science 2 :344-355.

Dobzhansky, T. 1937. Genetics and the origin of species. Columbia University Press, New York.

Dobzhansky, T. 1950. Mendelian populations and their evolution. American Naturalist 74 :312-321.

Dobzhansky, T. 1970. Genetics of the evolutionary process. Columbia University Press, New York.

Eigen, M. 1993. Viral quasispecies. Scientific American July 1993

Eldredge, N., and J. Cracraft. 1980. Phylogenetic analysis and the evolutionary process. Columbia University Press, New York.

George, T. 1956. Biospecies, chronospecies and morphospecies. Pp. 123-137 in P. Sylvester-Bradley, ed. The species concept in paleontology. Systematics Association, London.

Ghiselin, M. T. 1974. A radical solution to the species problem. Systematic Zoology 23 :536-544.

Harlan, J. R., and J. M. J. De Wet. 1963. The compilospecies concept. Evolution 17 :497–501.

Hennig, W. 1950. Grundzeuge einer Theorie der Phylogentischen Systematik. Aufbau Verlag, Berlin.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

Kitcher, P. 1984. Species. Philosophy of Science 51 :308-333.

Kornet, D. 1993. Internodal species concept. J Theor Biol 104 :407-435.

Kornet, D., and J. McAllister. 1993. The composite species concept. Reconstructing species : Demarcations in genealogical networks. Unpublished phD dissertation, Institute for Rheoretical Biology, Rijksherbarium, Leiden.

Mallet, J. 1995. The species definition for the modern synthesis. Trends in Ecology and Evolution 10 :294-299.

Mayden, R. L. 1997. A hierarchy of species concepts : the denoument in the saga of the species problem. Pp. 381- 423 in M. F. Claridge, H. A. Dawah and M. R. Wilson, eds. Species : The units of diversity. Chapman and Hall, London.

Mayr, E. 1942. Systematics and the origin of species from the viewpoint of a zoologist. Columbia University Press, New York.

Mayr, E. 1963. Animal species and evolution. The Belknap Press of Harvard University Press, Cambridge MA.

Mayr, E. 1969. Principles of systematic zoology. McGraw-Hill, New York.

Mayr, E. 1970. Populations, species, and evolution : an abridgment of Animal species and evolution. Belknap Press of Harvard University Press, Cambridge, Mass.

Mayr, E., and P. D. Ashlock. 1991. Principles of systematic zoology. McGraw-Hill,, New York.

Meier, R., and R. Willman. 1997. The Hennigian species concept in Q. Wheeler and R. Meier, eds. Species concepts and phylogenetic theory : A debate. Columbia University Press, New York.

Nelson, G. J., and N. I. Platnick. 1981. Systematics and biogeography : cladistics and vicariance. Columbia University Press, New York.

Nixon, K. C., and Q. D. Wheeler. 1990. An amplification of the phylogenetic species concept. Cladistics 6 :211-223.

Paterson, H. 1985. The recognition concept of species. Pp. 21-29 in E. Vrba, ed. Species and speciation. Transvaal Museum, Pretoria.

Pleijel, F. 1999. Phylogenetic taxonomy, a farewell to species, and a revision of Heteropodarke (Hesionidae, Polychaeta, Annelida). Systematic Biology 48 :755-789.

Pleijel, F., and G. W. Rouse. 2000. Least-inclusive taxonomic unit : a new taxonomic concept for biology. Proceedings of the Royal Society of London - Series B : Biological Sciences 267 :627-630.

Regan, C. T. 1926. Organic evolution. Report of the British Association for the Advancement of Science, 1925 :75- 86.

Ridley, M. 1989. The cladistic solution to the species problem. Biology and Philosophy 4 :1-16.

Rosen, D. E. 1979. Fishes from the uplands and intermontane basics of Guatemala : revisionary studies and comparative biogeography. Bulletin of the American Museum of Natural History 162 :267-376.

Simpson, G. 1943. Criteria for genera, species and subspecies in zoology and paleontology. Annals New York Academy of Science 44 :145-178.

Simpson, G. G. 1961. Principles of animal taxonomy. Columbia University Press, New York.

Smith, A. B. 1994. Systematics and the fossil record : documenting evolutionary patterns. Blackwell Science, Oxford, OX; Cambridge, Mass., USA.

Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H.

Freeman, San Francisco,.

Sterelny, K. 1999. Species as evolutionary mosaics. Pp. 119-138 in R. A. Wilson, ed. Species, New interdisciplinary essays. Bradford/MIT Press, Cambridge, MA.

Templeton, A. 1989. The meaning of species and speciation : A genetic perspective. Pp. 3-27 in D. Otte and J. Endler, eds. Speciation and its consequences. Sinauer, Sunderland, MA.

Turesson, G. 1922. The genotypic response of the plant species to the habitat. Hereditas 3 :211-350.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. Taxon 25 :233-239.

Wagner, W. H. 1983. Reticulistics : The recognition of hybrids and their role in cladistics and classification. Pp. 63- 79 in N. I. Platnick and V. A. Funk, eds. Advances in cladistics. Columbia Univ. Press, New York.

Waples, R. S. 1991. Pacific salmon, Oncorhynchus spp., and the definition of 'species' under the Endangered Species Act. Marine Fisheries Review 53 :11-22.

Wheeler, Q. D., and R. Meier. 2000. Species concepts and phylogenetic theory : a debate. Columbia University Press, New York.

Wiley, E. O. 1978. The evolutionary species concept reconsidered. Systematic Zoology 27 :17-26.

Wiley, E. O. 1981. Remarks on Willis' species concept. Systematic Zoology 30 :86-87.

Wu, C.-I. 2001a. Genes and speciation. Journal of Evolutionary Biology 14 :889-891.

Wu, C.-I. 2001b. The genic view of the process of speciation. Journal of Evolutionary Biology 14 :851-865. Summary of 26 species concepts [Copyright © 2002 John S. Wilkins]