THESIS / THÈSE

MASTER EN SCIENCES INFORMATIQUES

En quoi consiste la pertinence liée à l'utilisation d'un pipeline d'extraction de données génétiques dans le milieu médical

Dartevelle, Jason

Award date:
2016

Awarding institution:
Universite de Namur

Université de Namur
Faculté d'informatique
Année académique 2015–2016

En quoi consiste la pertinence liée à l'utilisation d'un pipeline d'extraction de données génétiques dans le milieu médical ?

Jason Dartevelle

Maître de stage : Federico Divina

Promoteur : _____ (Signature pour approbation du dépôt - REE art. 40)
Wim Vanhoof

Mémoire présenté en vue de l'obtention du grade de
Master en Sciences Informatiques.

Résumé

Dans un contexte de technologies et d'intérêts concernant l'étude de maladies génétiques, en constante évolution, il peut se révéler être pertinent de se demander en quoi l'élaboration d'un outil bio-informatique pourrait être utile dans l'élaboration d'études et de recherches pouvant aboutir à un moyen de remédier à ces maladies.

Dès lors, nous nous penchons, dans ce travail, sur la création d'un outil, un « pipeline d'extraction de données à partir de séquences d'ARN », afin de pouvoir y observer la pertinence de son application dans le milieu médical.

Pour ce faire, ce travail se concentre, dans un premier temps, sur les concepts et notions biologiques et bio-informatiques permettant la compréhension liée à l'utilisation d'un tel outil. Dans un deuxième temps, nous élaborons, de manière conceptuelle, sa création. Un exemple d'utilisation sera effectué dans le cadre d'une étude dans le milieu agricole. Ces résultats permettent d'illustrer le type de résultats pouvant être obtenus.

Après cela, nous discutons de l'utilisation de l'outil dans deux cas de maladies génétiques faisant l'objet d'un grand nombre d'études, à savoir le diabète et le cancer. Ces observations et discussions permettent de savoir en quoi cet outil peut se révéler être pertinent dans ce contexte.

Abstract

In a context of constant evolution of technologies abilities and interest for the medical research, about genetical diseases, it could be interesting to ask how the technology can be applied to this matter of research.

To this aim, we look, in this document, in the construction of a tool, the “data extraction pipeline”, to see what in what consists the application of this tool in the medical area.

The work focuses, firstly, on the biological and technical knowledge needed to understand the construction of the tool. An example is presented in view to observe the kind of results can be observed from this use.

The use of the tool, in the context of a genetic study in the agricol domain, is explained as an example. The results obtained can be used to help the reader to figure out about the kind of results that can be produced by this tool.

After that, we will discuss about the ability of this tool to be used in the medical domain by thinking about the case of the diabetes and cancer diseases. The observations and the conclusions, that we obtain, give us the ability find out if the tool can be really applied in this matter.

Avant-propos

Avant de commencer ce travail, nous souhaiterions, tout d'abord, remercier notre promoteur, le Professeur Wim Vanhoof, pour l'aide qu'il a apporté à l'élaboration de ce travail.

Nous souhaiterions, également, remercier mon maître de stage, le Professeur Federico Divina (Universidad de Pablo de Olavide, Séville, Espagne), ainsi que l'équipe de biologistes, les Professeurs Stephania Pilati et Claudio Moser (Fondazione Edmund Mach, Italie), ayant offert leur aide pendant le stage.

Nous remercions, finalement, les personnes proches, amis et membres de la famille, nous ayant soutenu tout au long de cette épreuve.

Table des matières

Glossaire

Nucléotide Elément de base de la molécule d'ADN ou d'ARN

ARN Ensemble de nucléotides permettant la création de protéine

ADN Ensemble de nucléotides constituant le matériel génétique, situé dans le noyau de chaque cellule d'une être vivant.

Gène Séquence d'ADN, situées sur les chromosomes, permettant de représenter un trait de caractère physiologique.

Chromosome Support physique sur lequel viennent se placer les gènes.

Génome Carte d'identité génétique d'un individu, composé de tous les chromosomes.

Génome de référence Génome d'une espèce vivante contenant tous les gènes répertoriés à ce jour.

Allèle Valeur définie à un gène.

Locis Emplacement sur le gène.

Microarrays Technique d'étude génétique par extraction d'ADN basé sur des réactions chimiques.

Exons Partie codantes inclues dans les séquences d'ARN.

Introns Partie intermédiaires, permettant de faire le lien entre deux exons.

Partie codante Sous-séquence d'une séquence d'ARN permettant de synthétiser une protéine.

Pipeline Processus technique permettant de réaliser des études génétiques.

Read séquence d'ARN seulement composée d'exons (partie codante).

Alignement Etape du pipeline permettant de placer les reads d'un échantillon sur un génome de référence

Expressivité génétique

Etape permettant de déterminer le niveau de représentation du génome de référence dans un échantillon donné.

FASTQ Format de fichier contenant les séquences d'ARN ou d'ADN.

SAM Format de fichier contenant les résultats de l'étape d'alignement.

BAM	Format de fichier contenant les résultats de l'étape d'alignement en représentation binaire.
Tumeur	Ensemble de cellules atteint d'un disfonctionnement de reproduction (cellules cacéreuses)
Cancer	Maladie génétique liée à un disfonctionnement de reproduction au niveau des cellules.
Diabète	Maladie liée à une présence insuffisance d'insuline en réponse à un fort taux de glycémie.
Glycémie	Taux de glucose dans le sang

Motivations

De nos jours, le domaine, en pleine expansion, que constitue celui de l'étude des gènes constitue un centre d'intérêt particulièrement utile. En effet, ces dernières années ont donné lieu à l'élaboration d'une multitude d'entre elles. Cependant, dû au lourd traitement de données que ces études impliquent, il devient, alors, naturel d'observer l'apparition d'une limitation dans leur accomplissement.

De son côté, les sciences informatiques, émanant de l'avènement des technologies, ont permis aux activités liées à la recherche scientifique, nécessitant des fonctionnalités complexes, d'acquérir une certaine aisance dans l'accomplissement de leurs tâches et, ainsi, contribué à l'obtention de résultats nécessaires à l'aboutissement de conclusions.
Sur base de ces faits, nous comprenons aisément qu'il devient, alors, pertinent d'essayer de fournir des moyens technologiques capables d'apporter une réelle contribution dans l'aboutissement de ces projets de recherche.

Ayant déjà travaillé dans le domaine de la bio-informatique, lors de la création d'un pipeline d'extraction de données à partir de séquences d'ARN dans le secteur vinicole, et dû à la menace que certaines maladies génétiques représentent, nous avons décidé de nous pencher sur la contribution que pourrait offrir ce genre d'outil au milieu médical. Notre question de recherche concernera l'utilisation de l'outil dans le milieu de médical et sera formulée comme suit :

« En quoi consiste la pertinence liée à l'utilisation d'un pipeline d'extraction de données génétiques dans le milieu médical ? ».

D'après notre expérience dans le domaine, nous sommes en mesure d'affirmer l'apparition de difficultés lors de l'étude du domaine. Ces difficultés sont principalement liées à une mauvaise connaissance du domaine biologique et bio-informatique, impliquant une remise à niveau non négligeable concernant ces notions ; ainsi qu'à la création de l'outil, dont la construction nécessite une attention particulière.

Le but de ce travail sera, donc, de fournir toutes les informations nécessaires à la création d'un raisonnement pouvant aboutir à une réponse quant à la question de recherche, basée sur les observations faites à partir de l'application de l'outil au cas du traitement de maladies génétiques et

ce, tout en essayant de contourner les diverses difficultés pouvant être rencontrées lors de la création d'un tel outil. Le travail concernera, donc, une discussion sur la création théorique d'un outil, à savoir, le pipeline, et sa probable application au milieu médical.

Nous espérons que le point de vue général, que ce travail offrira, d'une part, la preuve de la pertinence de l'utilisation de cet outil dans le milieu médical et, par conséquent, la motivation du corps scientifique à utiliser ce genre d'outil dans le milieu médical; et d'autre part, fournir la base de connaissances pouvant être utilisée par un informaticien, novice dans le milieu bio-informatique, lors de la réalisation de recherches similaires.

Introduction

Afin de pouvoir atteindre nos objectifs, nous avons décidé d'articuler notre travail sur trois parties, l'une étant dédiée à l'état de l'art ; la deuxième, à la construction d'un pipeline d'extraction de données et la troisième étant consacrée à l'application de cet outil au milieu médical. Une conclusion reprenant les faits importants sera, également, proposée.

L'état de l'art permettra de poser les fondements de base concernant les notions biologiques et bio-informatiques. Ces principes sont importants puisqu'ils permettent la compréhension de concepts plus élaborés dispensés plus loin dans le travail. Cette partie a pour but de résoudre les problèmes, pouvant être rencontrés, quant au manque de connaissances concernant les concepts généraux de ce domaine de recherches.

La deuxième partie, quant à elle, permettra de donner les informations relatives à la construction du pipeline d'extraction de données. Dans un premier temps, tous les composants le constituant seront explicités de manière abstraite. Ensuite, nous compléterons ces informations en ajoutant un cas concret d'utilisation servant d'exemple. Celui-ci concernera une étude biologique ayant été menée dans le secteur vinicole. Plus précisément, celle-ci portera sur les conséquences que pouvait avoir l'utilisation de l'hormone de Gibbérelline sur les vignes de raisins Sauvignon blanc et Pinot gris. Cet exemple permettra au lecteur d'observer le type de résultats pouvant être obtenus grâce à l'utilisation de cet outil. Le but de cette partie sera, donc, de sensibiliser le lecteur sur la construction de l'outil ainsi que sur le type de résultats pouvant être obtenus.

Après cela, dans le but de pouvoir fournir une réponse pertinente à la question de recherche, nous allons faire le lien avec le milieu médical. Concrètement, nous allons expliquer en quoi cet outil pourrait être appliqué aux différentes études portant sur les maladies génétiques. Pour ce faire, nous allons, tout d'abord, donner les explications de base concernant le génome humain. Ensuite, nous traiterons de l'application de l'outil à deux cas de maladies génétiques, à savoir : le diabète et le cancer. Pour chacune d'entre elles, nous étudierons leur fonctionnement, les méthodes existant concernant leur diagnostic ainsi que celles concernant leur traitement. Après cela, nous tenterons d'expliquer comment nous pourrions utiliser l'outil afin de pouvoir apporter une contribution relativement pertinente à leur contexte.

A terme de ce travail, nous espérons pouvoir être en mesure de fournir une conclusion pouvant répondre à la question de recherche de manière pertinente. Celle-ci sera, également, l'occasion pour nous d'exprimer notre ressenti par rapport à ce travail afin de pouvoir discuter d'éventuelles améliorations possibles.

Histoire

Pour pouvoir situer notre domaine de recherche et donner un certain cadre facilitant la compréhension du lecteur, nous proposons de commencer par l'introduction de l'histoire concernant l'étude des gènes. Celle-ci, basée sur les sources [Jork, Carey, Bamshad 2010 & Pevsner 2011], permettra au lecteur de comprendre l'intérêt de ce domaine et de se rendre compte de ce qui a déjà été réalisé dans celui-ci.

Depuis des millénaires, l'étude des traits physiques a suscité chez l'Homme un intérêt sans précédent. Les peuples de l'Antiquité élaboraient déjà des théories sur ces phénomènes de la vie. La plupart d'entre elles se sont avérés être fausses. Cependant, celles-ci nous permettent de comprendre l'origine de l'engouement que représente ce domaine de recherche.

Au fil du temps, les théories se sont améliorées jusqu'à donner, à notre époque, les théories contemporaines.

Malgré les théories existantes, une contribution majeure dans le domaine est due à un scientifique autrichien, du nom de Gregor Mendel, qui a permis de définir la base concernant la théorie de la transmission de caractères héréditaires. En effet, pendant plusieurs années, il a étudié le comportement génétique qui pouvait apparaître dans une culture de petit-pois. Ces expériences lui ont permis de découvrir les fondements du caractère héréditaire d'un individu.

Cette découverte lui permettra, en 1865, de constituer, en publiant ces résultats, une théorie de base dans le domaine de la génétique et d'être considéré comme étant le fondateur de ce domaine.

A partir de ces résultats, d'autres découvertes ont vu le jour. Charles Darwin publiera au même moment sa théorie de l'évolution qui défend l'hypothèse selon laquelle toutes les espèces vivantes auraient évoluées à partir d'un ancêtre commun. Cette hypothèse

Le 20ème siècle a constitué un tournant majeur dans les de ce domaine. En effet, de nombreux travaux sur la génétique ont permis d'enrichir les travaux de Mendel. Par exemple, Landsteiner découvre la présence des groupes sanguins, en 1900 ; Archibald Garrod découvre, en 1902, la présence d'erreur interne au métabolisme et Johanson définit, en 1909, le gène comme étant l'unité de base pour l'hérédité.

A partir de là, d'autres travaux théoriques et expérimentaux ont été réalisés lors des décennies suivantes :

Ronald Fisher, JBS Haldone et Sewall Wright étudient la génétique en termes de population (étude par groupes ethniques) et découvrent que certaines maladies sont plus favorables à certains milieux environnementaux et à certains groupes ethniques.

En 1944, Oswald Avery, médecin américain, découvre l'existence d'ADN (acide désoxyribonucléique) et que les gènes en sont composés.

En 1953, James Watson et Francis Crick définissent la structure moléculaire de l'ADN.

En 1956, le nombre exact de chromosomes que contient un être humain a été découvert et est au nombre de 46 par cellule, ce qui revient à 23 paires de chromosomes. Cette découverte permet d'instaurer une notion de normalité en ce qui concerne la constitution humaine. En d'autres termes, elle nous permet, de nos jours, de savoir si un être humain est normalement constitué.

En 1959, élaboration de la cytogénétique permet de définir certaines maladies génétiques en se basant sur les caractères génétiques. Par exemple, le syndrome de Down, généralement appelé « trisomie 21 », est caractérisé par un triplement du 21ème chromosome.

Dans les années soixante, le progrès scientifique a permis de faire progresser le domaine génétique moléculaire en découvrant la localisation des gènes sur les chromosomes.

Dans les années nonante, des projets de grandes envergure ont été menés afin de pouvoir enrichir la base de connaissances génétique. Un des projets les plus importants est le « Human Genome Project » qui consiste à fournir l'ensemble de l'ADN pouvant être rencontré dans le corps humain. Ce projet a abouti avec succès en 2003. Il aura donc fallu près de 15 ans pour pouvoir définir entièrement l'information génétique d'un être humain.

D'autres découvertes, concernant le déclenchement des maladies génétiques et leur compréhension, ont également été identifiées pendant cette période.

Ce bref historique nous aura permis de comprendre que la génétique est un centre d'intérêt prometteur, dont l'origine remonte à l'Antiquité, et que les théories les plus importantes, notamment celles relatives au domaine de l'étude du génotype humain, sont relativement récentes.

Première partie : Fondements

Cette première partie a pour but de fournir au lecteur toutes les informations nécessaires à la bonne compréhension du sujet. Même si certains concepts ne semblent pas être d'une importance capitale, nous les expliquerons, tout de même, afin de pouvoir y faire référence plus tard dans le travail. Cette parte est donc nécessaire et doit donc être considérée avec beaucoup d'attention.

Cette partie nous permettra de définir, d'une part, les concepts liés au domaine biologique : nous y définirons la notion de cellule, de matériel génétique identifiant un individu et des actions pouvant être performées à ce stade; et d'autre part, de poser les concepts liés au domaine de la bio-informatique, à savoir: l'obtention des données et le traitement de celles-ci.

A son terme, le lecteur possèdera toutes les informations biologiques et technologiques permettant l'assemblage des outils, présentés dans la deuxième partie.

Chapitre 1 : Fondements biologiques

Cette partie constituera notre base de connaissance concernant les concepts biologiques. Elle est essentiellement basée sur les concepts de nos sources [Jork, Carey, Bamshad, 2010 & Karp, 2002 & Raven, Jonhson, Losos et Singer, 2007]. Nous y verrons la structure cellulaire, et les liens qu'elle possède avec son environnement. Pour ces explications, nous avons préféré utiliser une approche « top-down », ce qui nous permet de partir d'une explication générale concernant l'organisme d'un être vivant et d'aller de plus en plus loin en fournissant des détails concernant les cellules et matériel génétique.

Organisme

Un organisme est un système vivant complexe, composé de plusieurs organes possédant une fonctionnalité bien particulière. Ces organes sont, à leur tour, composés de tissus, eux-mêmes composés de composants, appelés cellules, composés d'éléments cellulaires pouvant être différent d'une espèce à une autre.
Etant donné l'aspect génétique de notre travail, nous nous limiterons à l'étude les cellules.
Par exemple, l'être humain peut être considéré comme un organisme possédant plusieurs organes composés de différentes cellules.

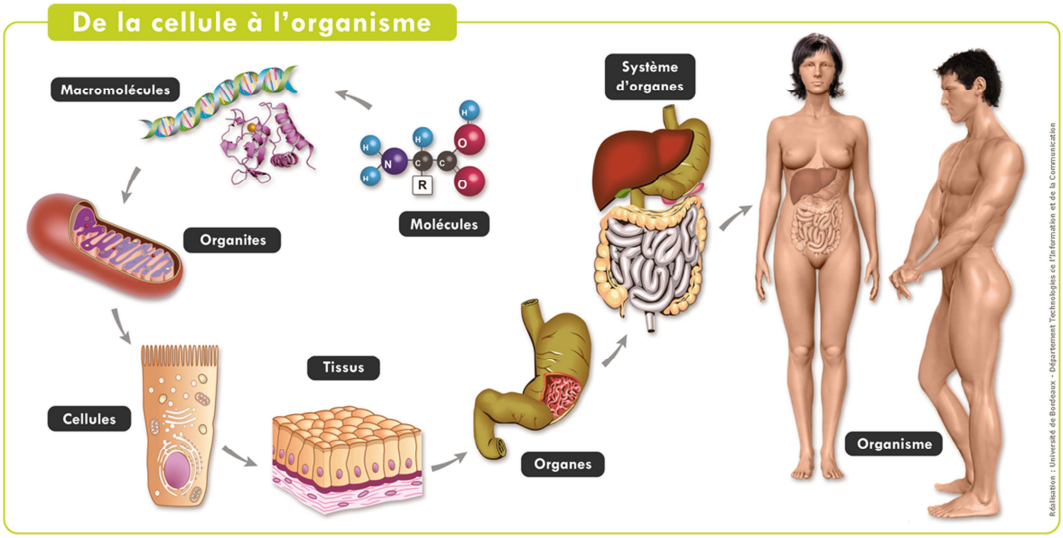

Figure 1.1: Organisme de l'être humain[1]

[1] Source image: http://ressources.unisciel.fr/biocell/chap1/co/module_Chap1_2.html

La figure 1.1, présentée ci-dessus, présente les différents composants étant pris en considération dans le concept de composition d'organismes, tel que l'être humain. Il est important de noter que les autres organismes sont également composés de manière similaire. Seule la composition des cellules est susceptible d'être modifier.

D'après de nombreux ouvrages portant sur la biologie cellulaire, il existe deux types caractéristiques de cellules : les procaryotes et eucaryotes. Le premier type, cellules procaryotes, ne sont composés que de composants cellulaires tout en étant dépourvue de structure membranaire, il n'y a donc pas de présence de membrane nucléaire (noyau). Ce genre de cellules représente des organismes simples tels que les bactéries, qui ne sont composés que d'une seule cellule. La figure 1.2 montre la composition d'une telle cellule.

A l'opposé, les cellules eucaryotes présentent une structure complexe contenant un noyau bien séparé du reste des composants. Ce genre de cellules est celui couramment rencontré dans les organismes humains, végétaux, animaux ainsi que celui des champignons. La figure 1.3 représente ce genre de cellules. La section suivant permet d'étudier de manière plus profonde ce genre de cellules. En effet, étant donné que notre travail porte sur les organismes végétaux et humains, nous ne nous limiterons qu'à l'étude de celle-ci.

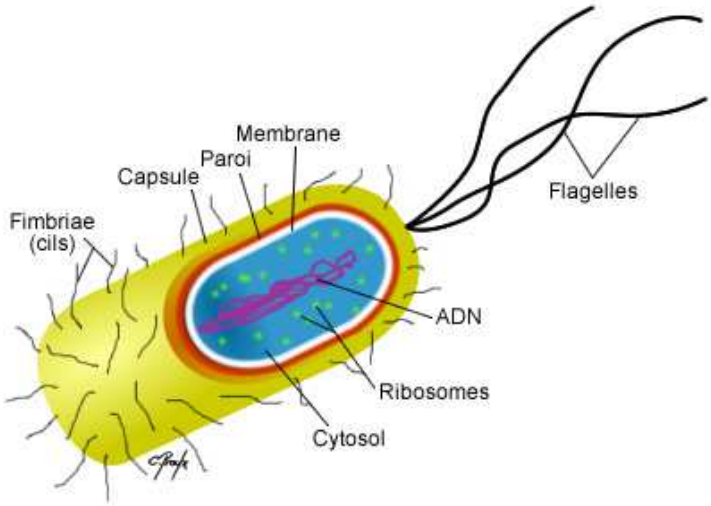

Figure 1.2: Cellule procaryote[2]

[2] Image obtenue à l'adresse suivante : Mathieu Simon, *Cellules procaryotes et cellules eucaryotes*, http://www.cours-pharmacie.com/biologie-cellulaire/cellules-procaryotes-et-cellules-eucaryotes.html, 2009, Consultation : 07/2016

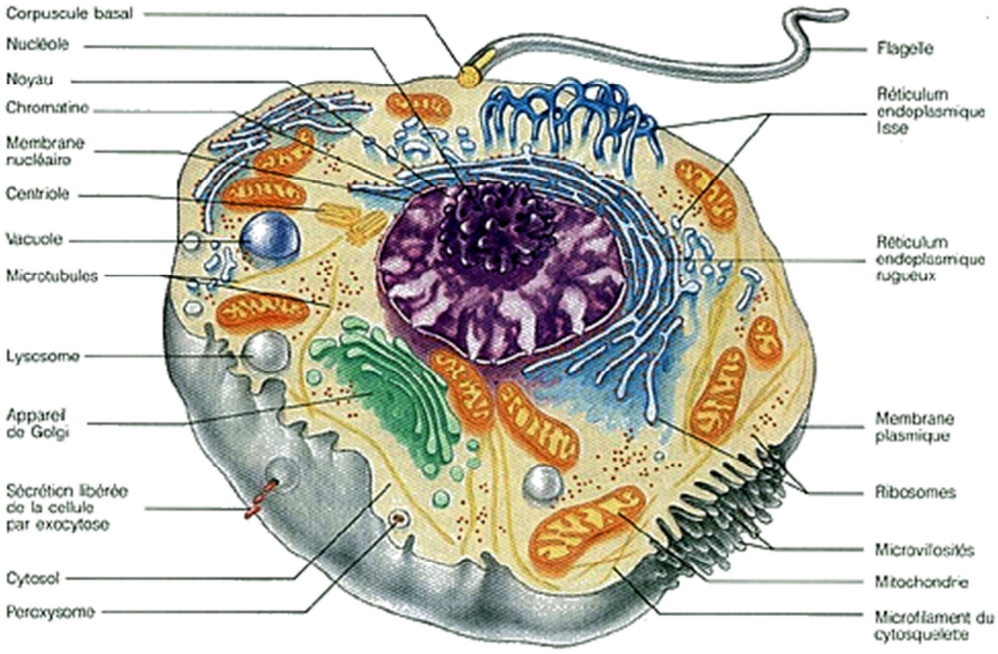

Figure 1.3: Cellule eucaryote[3]

N.B. : Etant donné la diversité d'organismes d'espèces vivantes, qui requièrent des analyses expérimentales différentes d'une espèce à une autre, les biologistes ont concentré leurs efforts sur faisant des recherches que sur un nombre restreint d'organismes modèles. Ceux-ci sont considérés comme étant assez représentatifs de l'ensemble des organismes vivants et permettent donc de ne perdre de ressources. A l'heure actuelle, il existe six organismes modèles dont la souris. C'est la raison pour laquelle il est normal de considérer le fait que certaines expériences biologiques concernant les recherches de traitements de maladies humaines soient faites, dans un premier temps, sur des souris.

[3] Image obtenue à l'adresse suivante : http://www.astrosurf.com/luxorion/bio-fonctionnement-cellules4.htm, Consultation : 07/2016

La cellule

Comme dit précédemment, chaque espèce d'être vivant possède ses types de cellules ayant un aspect qui leur est propre. Cette section permet de donner une définition ainsi que les principes majeurs étant liés à la notion de « cellule eucaryote ». Nous allons, pour ce faire, présenter les cellules animales et végétales.

Tout d'abord, nous présentons la figure suivante, construite par nos soins, qui permet de voir les liens entre ces catégories de cellules :

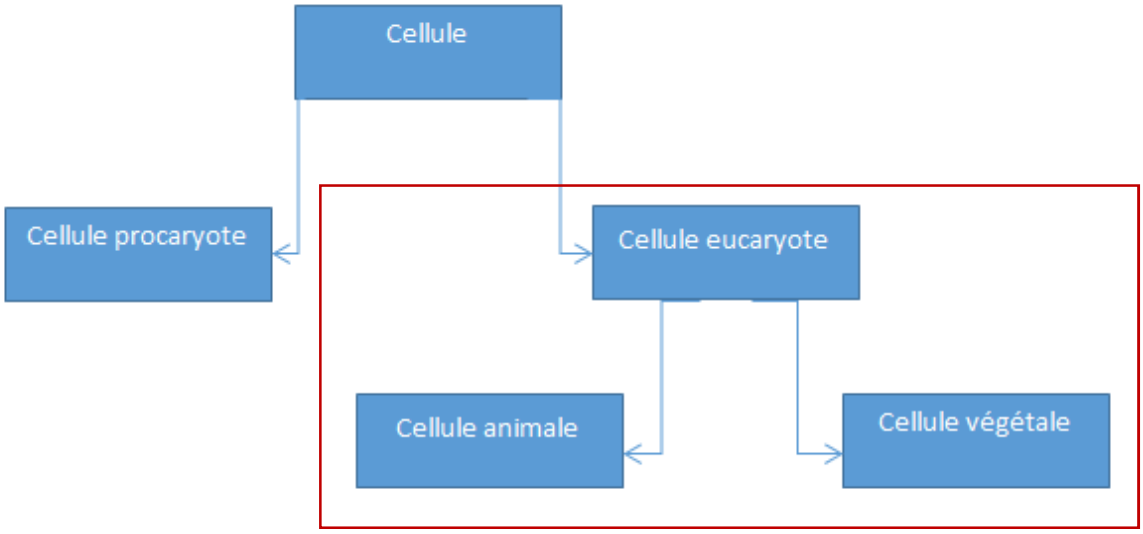

Figure 1.4 : Organisation des cellules

Etant donné notre domaine de recherche, nous ne nous limiterons à l'étude des cellules eucaryotes, et donc des cellules animales et végétales.

Définition

La cellule peut être vue comme étant la plus petite structure d'un organisme. Celle-ci contient toute le matériel génétique d'un individu et permet de constituer la composition d'un tissu qui, une fois assemblés, représentent un organe. [4]

[4] D'après le Larousse, 2016

Fonctionnement

Le fonctionnement d'une cellule peut être vulgarisé comme correspondant à celui d'une usine. En effet, une cellule est composée de plusieurs composants et son but principal est de produire des protéines utilisées, soit, pour la cellule elle-même, soit, pour le tissu que celle-ci compose. La protéine produite devra, donc, correspondre au type de tissu contenant la cellule. Notons, dès lors qu'un disfonctionnement au niveau de cette production peut avoir des conséquences en cascades et, par conséquent, aboutir à des problèmes à l'échelle de l'organe, puis, à celle de l'organisme en entier. Par exemple, le cancer provient d'un disfonctionnement lié au processus de reproduction cellulaire entraînant une reproduction infinie d'une cellule.

Composants

Les éléments des cellules sont diverses et ont une prédisposition à accomplir une fonctionnalité particulière qui permettra, à long terme, d'assurer le bon fonctionnement de la cellule. Les cellules ont besoin, en moyenne, d'un dizaine de composants pour pouvoir accomplir leur tâche au sein de l'organisme. Par exemple, certains composants, tels que le réticulum endoplasmique lisse, permet la fabrication de protéines et de stocker les lipides nécessaires.

Etant donné que nous traitons deux types de cellules, nous pensons qu'il est nécessaire de présenter leur composition afin de déterminer leurs ressemblances et, par conséquents, leurs distinctions particulières.

Nous nous limitons, pour le moment, à une simple présentation de leur structure. Un complément d'informations concernant leurs fonctions (ou rôle), dans l'organisme, sera présenté plus tard.

Les deux figures suivantes, 1.5 et 1.6, représentent, respectivement, les deux types de cellules eucaryotes animale et végétale :

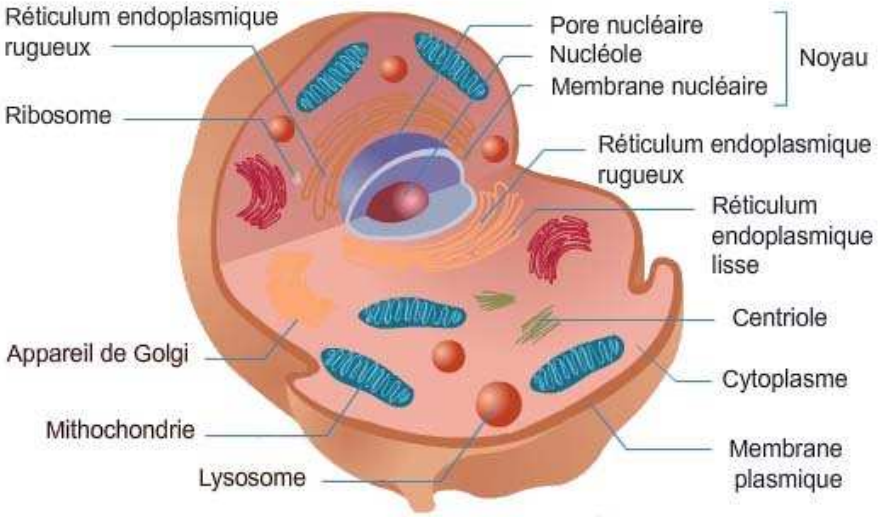

Figure 1. 5 : Cellule animale [5]

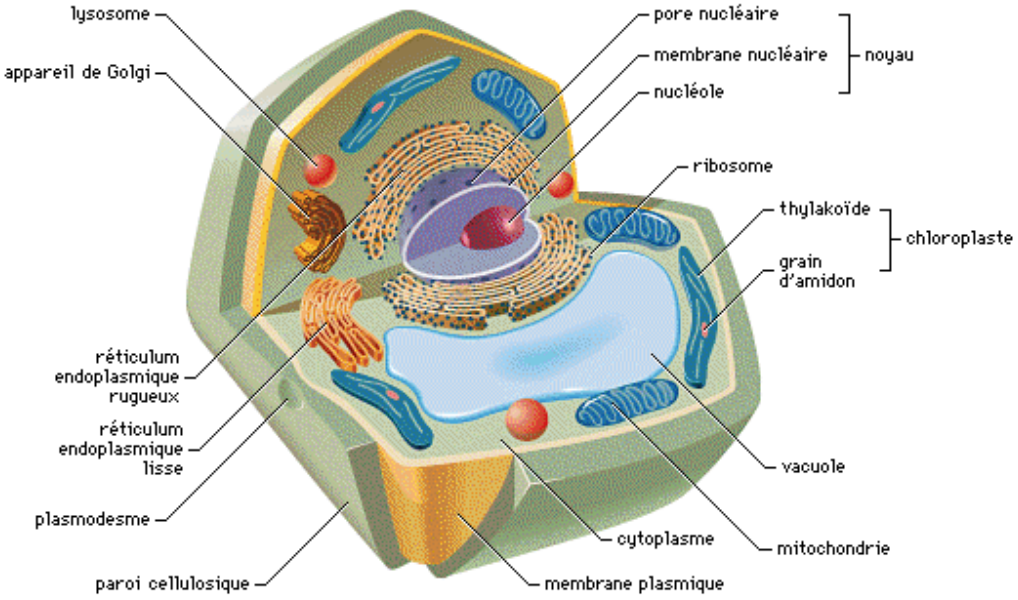

Figure 1.6 : Cellule végétale [5]

[5] Image obtenue à l'adresse suivante : *La Terre dans l'Univers, la vie et l'évolution du vivant : une planète habitée*, http://svt.seconde.free.fr/Vocabulaire/Voca1.html, Consultation : 07/2016

Comme nous pouvons le constater, celles-ci possèdent des composants différents. Par exemple, la cellule végétale possède un chloroplaste que l'animale ne possède pas. Ceci est dû au fait que la ce type de cellule doit supporter une autre fonctionnalité nécessitant ce composant, à savoir la photosynthèse.

Cependant, ces deux cellules possèdent un ensemble de composants de base constitué, entre autre, du noyau et de ribosomes jouant un rôle primordial dans le cycle de vie de la cellule.

Nous pensons, également, qu'il est nécessaire de présenter les éléments de cet ensemble de base, à savoir:

- Noyau : contient toutes les formations liées au bagage génétique d'un individu
- Ribosomes : permet, comme nous le verrons plus tard, à la cellule de produire des protéines
- Membranes : constitue le caractère structural qui est spécifique aux cellules eucaryotes

Le rôle de ces éléments sera expliqué en temps voulu.

N.B. : Nous notons que chaque composant joue un rôle important dans le cycle de vie de la cellule, mais ne rentre pas dans notre champ d'intérêt lié à l'élaboration de ce travail. C'est pourquoi, nous avons décidé de ne pas en tenir compte.

Expressivité cellulaire

Un des rôles fondamentaux d'une cellule réside dans sa capacité à pouvoir représenter, grâce à son matériel génétique, les caractéristiques de l'individu en entier. De nombreuses expériences concernant ces capacités ont été menées durant le siècle dernier et continuent de susciter l'attention du domaine de la recherche scientifique.

Cette section permet au lecteur, novice ou moins novice, de se familiariser avec les résultats et théories obtenues jusqu'à présent lors de ces recherches.

Pour ce faire, nous commencerons par expliquer la notion de « génome », ainsi que les structures physiques sous-jacentes. La section suivante aura pour but de compléter les connaissances avec des informations concernant les fonctionnalités, ou rôle de participation, que la cellule peut entreprendre au cours de son cycle de vie.

Structure

La structure du matériel génétique est composée comme suit : la cellule contient un noyau possédant toute l'information génétique de l'individu ; ces informations génétiques sont concentrées dans un génome (considéré comme étant la « carte d'identité » d'un individu) contenant, dans le cas de l'être humain, 46 chromosomes placés par paires et se partageant l'ensemble de l'ADN. C'est sur ces chromosomes que sont localisés les gènes de l'individu permettant de définir ses traits caractéristiques.

La figure 1.7 représente les relations entre les différents composants.

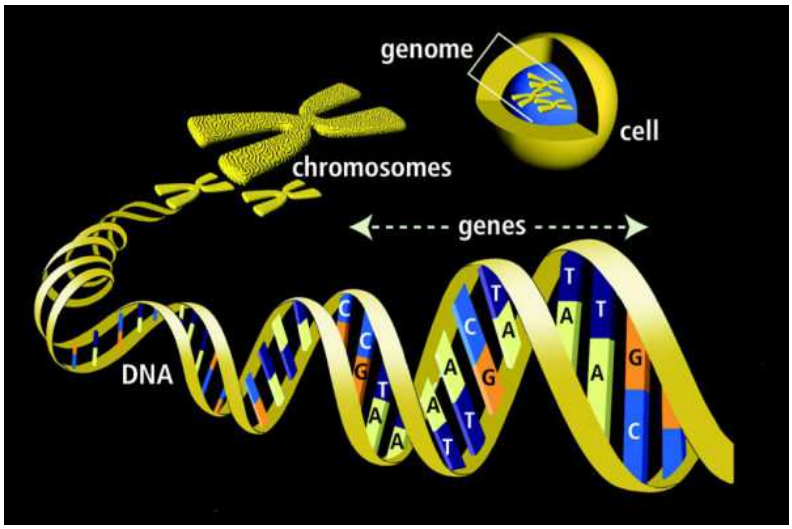

Figure 1.7 : Vue générale des gènes[6]

Génome

Comme dit précédemment, le génome représente l'ensemble des informations génétique d'un individu. Il se situe dans le noyau des cellules. Il est représenté par l'ensemble des chromosomes.

La figure 1.8 montre sa représentation.

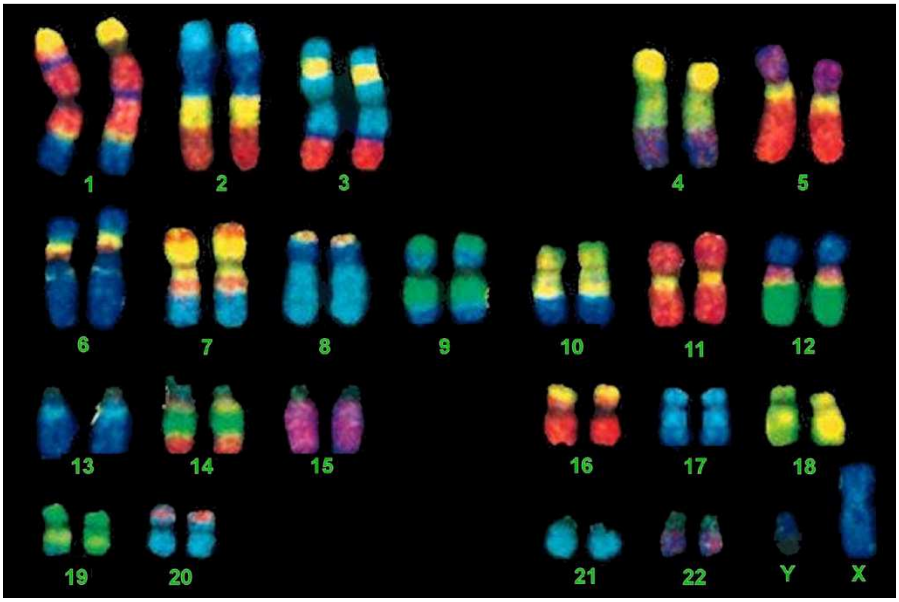

Figure 1.8: Génome d'un être humain[7]

[6] Source image : Philippe Pognonec, *Fête de la Science*, http://slideplayer.fr/slide/513519/, slide 29, 2007, Consultation : 07/2016

Comme nous le suggère cette figure, le génome permet de représenter les chromosomes par paire. Ceci nous permettra de comparer les allèles (valeur du gène sur chaque chromosome) entre eux afin de pouvoir ressortir les caractéristiques génétiques (voir section suivante pour la détermination de caractères et traits individuels).

Les sections suivantes offrent des compléments d'informations concernant les caractéristiques du chromosome, composant de base de ce composant.

N.B. : Dans le jargon, nous faisons la distinction entre deux sous-catégories de cellules. En effet, nous parlerons de cellules « diploïdes », pour des cellules dont le génome contient des paires de chromosomes, et de cellules « haploïdes » pour des cellules dont le génome ne contient que des chromosomes « célibataires » (sans homologues), comme, par exemple, les cellules germinales.

[7] Source image : Charles Coutton, *Nouveaux outils de cytogénétique moléculaire utilisés en constitutionnel*, http://www-sante.ujf-grenoble.fr/SANTE/cms/sites/medatice/dcem1/dcem1/docs/20120126140538/COUTTON_Cours_EC_G_n_tique_12022012.pdf , 2012

Chromosome

Le chromosome est la structure physique du gène. Il est composé essentiellement de protéines autour desquelles s'enroulent les séquences d'ADN. Il se situe dans le noyau et est facilement décelable à partir de ses deux chromatines (branches) qui lui donnent sa forme particulière. La figure 1.9 nous présente la structure d'une paire de chromosomes.

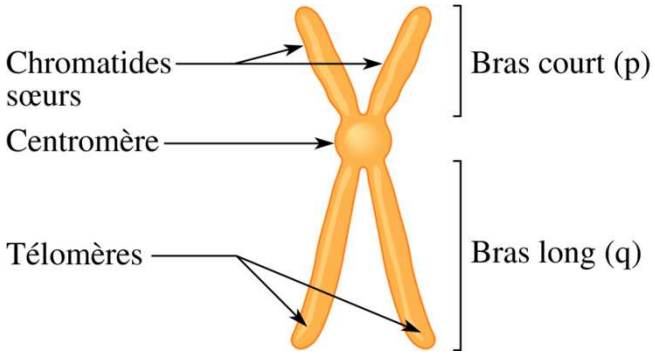

Figure 1.9: Structure paire chromosomes[8]

Un chromosome possède un certain nombre de gènes situés sur son corps à différents endroits, généralement appelés « loci ».

Le fait qu'un chromosome se mette en relation avec un homologue, permet à l'individu de posséder deux « allèles » (valeur génétique) correspondant à un trait. La figure 1.10 présente une paire de chromosomes possédant deux allèles pour un même trait.

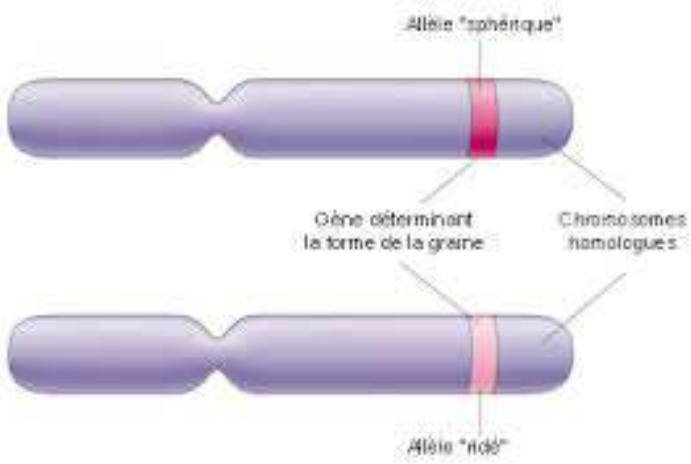

Figure 1.10: Positions des allèles [9]

[8] Source image : *Le chromosome au cours du cycle cellulaire*, http://www.assistancescolaire.com/eleve/TST2S/biologie/reviser-le-cours/le-chromosome-au-cours-du-cycle-cellulaire-tst2s_bio_08, Consultation : 07/2016

Cette figure montre que les traits de l'individu sont donc, en réalité définis par plusieurs valeurs. Or, étant donné qu'un seul de ces allèles sera utilisé, la cellule doit instaurer une politique afin d'en « choisir » une des deux. Cette politique est basée sur le caractère dominant de certains allèles. En effet, certains allèles sont « dominants » et d'autres « récessifs » lorsqu'ils sont mis en relation ensemble. Nous obtenons, par exemple, pour le trait de la couleur des yeux, si nous mettons l'allèle « yeux bleus » et « yeux bruns », un individu possédant des yeux bruns.

ADN

L'ADN, ou acide désoxyribonucléique, est le composant principal du chromosome, qui contient l'ensemble du matériel génétique. Il se compose de trois éléments chimiques de base :

- Le pentose
- Le désoxyribose
- Quatre bases nitrogènes : Cytosine, Thymine, Adénine et Guanine (C, T, A, G)

Les quatre bases nitrogènes sont appelées nucléotides dans ce contexte-ci. Ces nucléotides, mis bout-à-bout, forment une séquence d'ADN qui permettra, à long terme, jouera un rôle dans la vie de l'individu.

Sa structure se présente sous forme de deux brins, ou séquences, d'ADN mis en concordances dans le sens opposé sous forme d'une hélice. La mise en correspondance des deux séquences est permise grâce à une liaison d'hydrogène présente entre les couples de nucléotides mis en relation. Notons que ces couples ne se forment pas aléatoirement, mais en suivant une règle de concordance. En effet, La cytosine ne peut se mettre en relation qu'avec la guanine (et inversement) et la thymine ne peut se mettre en relation qu'avec l'adénine (et inversement).

La figure 1.11 représente la structure d'une séquence d'ADN normale.

[9] Source image : Génétique, http://coursde-medecine.blogspot.be/2013/01/genetique.html, Consultation : 07/2016

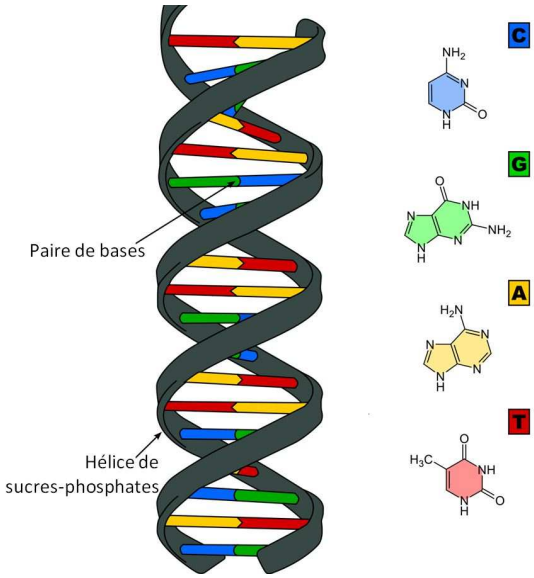

Figure 1.11: Structure ADN[10]

Une des fonctionnalités majeures que la cellule doit pouvoir gérer concernant l'ADN est sa réplication. Celle-ci est utilisée lors du processus de copie de cellule. En effet, le fait de copier la cellule implique, également, une copie de tous ses composants, dont le noyau ainsi que l'ensemble des séquences d'ADN composant les chromosomes.

Cette procédure se réalise de la manière suivante : les deux brins se séparent en brisant leur liaison d'hydrogène. Nous obtenons donc, deux brins « célibataires », à partir desquels nous pouvons faire correspondre, à leurs nucléotides, les nucléotides homologues, satisfaisants la politique de concordance précédemment établie. Ce procédé nous permet d'obtenir à partir d'une séquence d'ADN, deux séquences identiques.

[10] Source image : *ADN*, http://sites.crdp-aquitaine.fr/stl/lexique/adn/, Consultation : 07/2016

La figure 1.12 représente ce procédé.

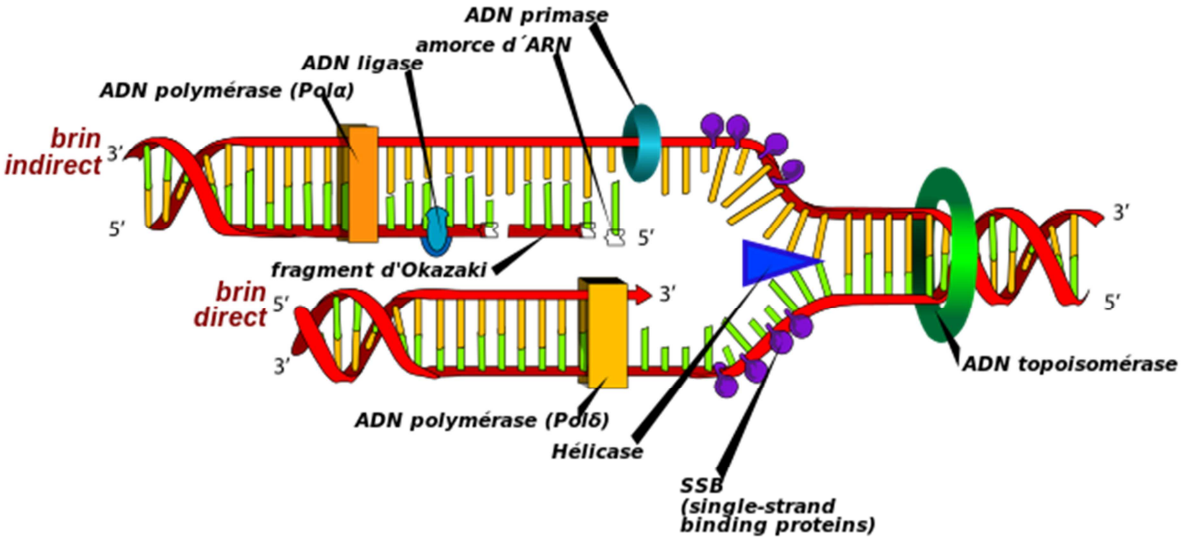

Figure 1.12: réplication ADN[11]

Notons que dans cette figure, nous remarquons que les deux brins sont bien juxtaposés dans le sens inverse et que la réplication se fait dans le sens « downstream », ou à contre-courant qui part de la troisième boucle jusqu'à la cinquième.

Un autre composant est souvent couplé à la manipulation des brins. Il s'agit des enzymes. Celles-ci sont utilisées par la cellule afin d'accomplir différentes actions, telles que la réplication des brins, ou encore, la détermination des nucléotides manquants (appelée « polymérase »). Des enzymes permettent, par exemple, de briser la juxtaposition des brins, de les maintenir ceux-ci et d'effectuer la polymérase de l'ADN.

N.B. : Notons l'existence de certains faits, le taux de réplication de l'ADN chez l'être humain est de l'ordre de 40 à 50 nucléotides par secondes. Certains chromosomes comportent, jusqu'à, 250 000 000 nucléotides. De ce fait, la réplication d'ADN de ces chromosomes devenant trop longue, dans une optique de processus linéaire (partant du début et s'arrêtant à la fin), la cellule permet de contourner ce problème en exécutant ce processus à partir de plusieurs points différents. Ce qui revient, en quelques sortes, à paralléliser la réplication.

[11] Source image : *Le séquençage de l'ADN et ses applications*, http://fr.slideshare.net/Olivez/le-squenage-de-ladn-et-ses-applications, 2015, Consultation : 07/2016

ARN

L'ARN, ou acide ribonucléique, est un composant chimiquement similaire à l'ADN, composé de sucres ribose, et non pas désoxyribose (comme l'ADN), de groupes phosphate et de quatre bases nitrogènes (Cytosine, Thymine, Adénine et Guanine). La différence majeure entre ces deux composants réside dans le fait que l'ARN dispose d'une base d'uracile à la place de celle de thymine, présente dans l'ADN. Malgré cette différence, il est important de noter que la thymine et l'uracile se comportent de la même manière et, par conséquent, peuvent, tous deux, être mis en relation avec un nucléotide d'adénine lors de la polymérase.

Concernant sa structure, celle-ci se présente également sous forme d'hélice, mais ce n'est pas toujours le cas. En effet, comme nous e verrons dans les sections suivantes, l'ARN peut parfois se composer que d'une seule séquence.

L'ARN se construit sur base de l'ADN. En effet, dans le contexte de création de protéine notamment, l'enzyme de polymérase permet de parcourir d'un des deux brins d'ADN afin de compléter la séquence d'ARN sur base de la politique utilisée lors de la polymérase d'ADN à l'exception, que nous faisons correspondre un nucléotide d'adénine à un nucléotide d'uracile à la place de celui de thymine. La séquence obtenue est appelée « ARN messager », ou « ARNm ».

Notons qu'il existe d'autres forment d'ARN. En effet, lors de la production de protéines, la cellule utilise également une séquence d'ARN dite de « transfert », qui mise en correspondance avec la séquence messagère et étant couplée avec le ribosome, permet de produire une protéine.

La figure 1.13 présente les séquences d'ADN et d'ARN.

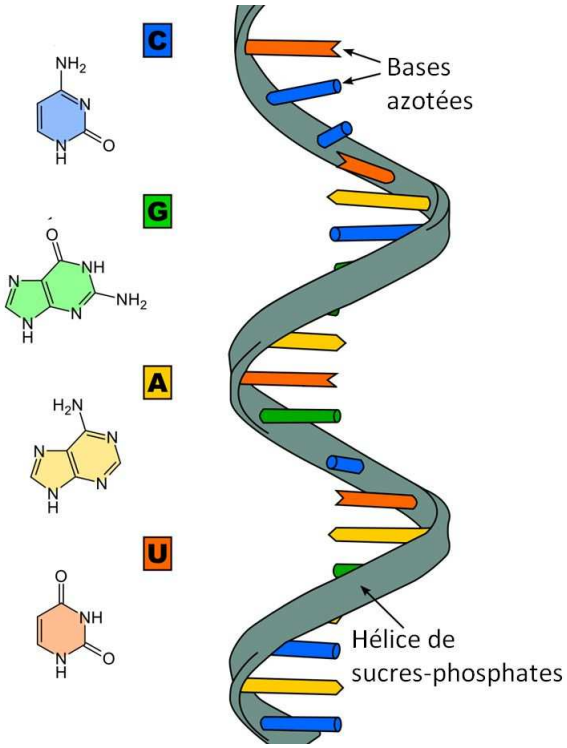

Figure 1.13: Structure ARN[12]

N.B. : Le but étant de donner des explications brèves concernant ce concept, nous nous limitons, pour le moment, à de simples faits. Ce processus, décrit brièvement ci-dessus, sera donc décortiqué par la suite.

[12] Source image : *ARN*, http://sites.crdp-aquitaine.fr/stl/lexique/arn/, Consultation : 07/2016

Fonctions

La cellule, composée des éléments structurels précédemment élicités, est également couplée à un ensemble de fonctionnalités vitales à la survie et au bon fonctionnement de son organisme hôte. Parmi celle-ci, nous notons les principes de reproduction cellulaire et de production de protéines.

Reproduction cellulaire

Afin de pouvoir assurer la prospérité d'un individu, les cellules doivent être munies de capacités reproductrices. Ces capacités sont utilisées dès la formation de l'embryon, initialement composé d'une unique cellule, et ne s'arrêtent qu'au moment où l'hôte décède. Ce fait permet d'expliquer la transition d'un œuf (embryon) unicellulaire vers un organisme multicellulaire. Il est important de spécifier que seules certaines cellules persisteront de la création jusqu'à la mort d'un individu. Nous pouvons, donc, comprendre la nécessité d'assurer aux cellules des capacités d'autoreproduction.

Celles-ci, comme le proposent les deux sections suivantes, sont principalement définies par la mitose, qui concerne principalement reproduction cellulaire, et la méiose, concernant plus généralement la reproduction de l'individu. Ces mécanismes sont utilisés dans divers processus cellulaire durant le cycle de vie de la cellule.

Mitose : la division cellulaire

La mitose constitue la base du principe de reproduction. En effet, elle constitue une des deux partie du cycle de vie de la cellule, qui permet de construire, à l'identique, cette cellule. La figure 1.14 représente le cycle de reproduction cellulaire.

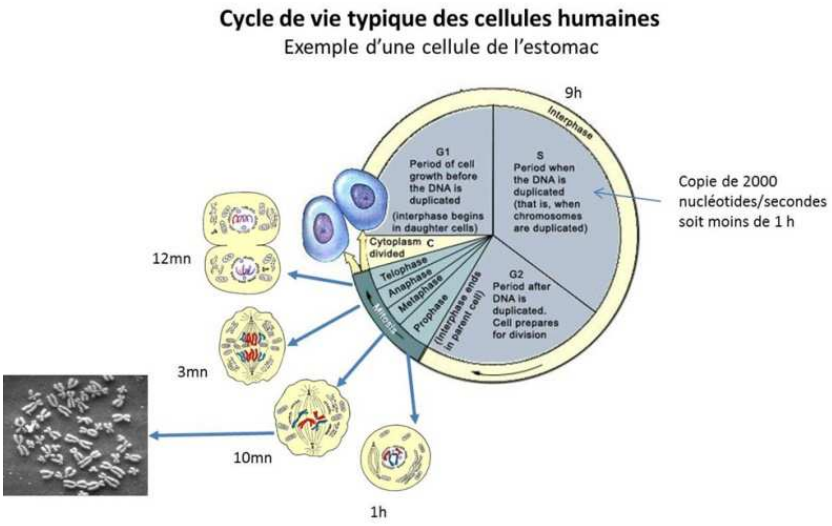

Figure 1.14: cycle cellulaire[13]

Ce cycle de vie est composée, comme le suggère cette figure, de deux phases générales : l'interphase (partie préparatoire) et la mitose (partie concernant la division). Cette première partie est constituée des étapes suivantes :

- GAP 1 : permettant la synthèse de séquence d'ARN et de protéines
- Synthèse : permettant la réplication d'ADN
- GAP 2 : permettant de réparer l'ADN et la préparation des cellules à la mitose
- Mitose : reproduction cellulaire

Quant à la deuxième partie, la mitose, celle-ci concerne purement la copie cellulaire et est composée de trois phases :

- Prophase :
    - mise en correspondance des chromosomes homologues (autour d'un point central appelé « centromère »)
    - disparition de la membrane nucléaire
    - apparition de fibres de fuseau (« spindle fibers ») à partir de chaque centriole (situés aux pôles opposés)

[13] Source image : Olivier Ezratty, *Les technologies de séquençage du génome humain*, http://www.oezratty.net/wordpress/2012/technologies-sequencage-genome-humain-1/?output=pdf, consultation : 07/2016

    - Liens produits entre les centromères et ces fibres et séparent les chromosomes homologues afin d'en tirer un à chaque pôle
- Métaphase :
    - Chromosomes atteignent leur état le plus condensé
    - Les centromères sont rejetées l'une de l'autre, ce qui laisse apparaître deux sous-cellules
- Anaphase
    - Eclatement et éjection des centromères de chaque chromosome
    - Nous obtenons donc deux cellules filles à partir de la cellule mère, contenat toute deux, le jeu de 46 chromosomes
- Télophase
    - Constitue l'étape finale
    - Apparition, dans les deux cellules filles, d'une membrane nucléaire autour des 46 chromosomes

Ce procédé permet donc de crée, à partir d'une cellule mère diploïde, deux cellules filles diploïdes.

La figure 1.15 représente ce processus.

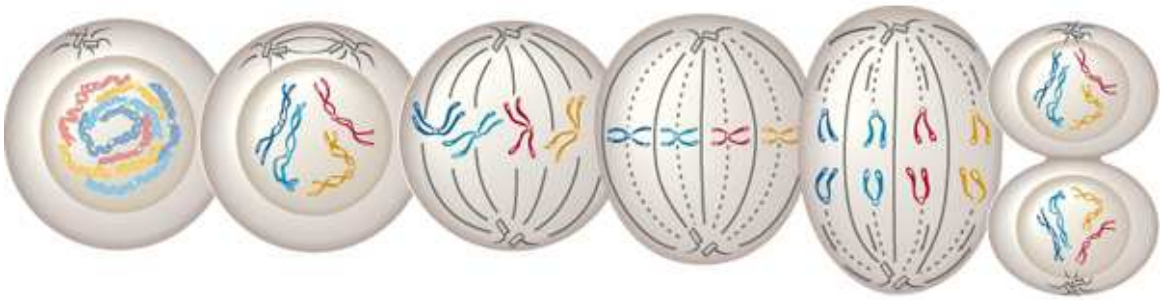

Figure 1.15: La mitose[14]

N.B. : Notons, également, que le temps de ce cycle de vie dépend du type de la cellule et de l'état de maturation de l'individu. Il faudrait, en effet, une dizaine d'heures pour une cellule intestinale et les cellules construisant les fibres musculaires ne se renouvèlent de moins en moins bien au fil du temps de la vie de l'individu.

[14] Source image : Phases de la mitose, http://www.larousse.fr/encyclopedie/images/Phases_de_la_mitose/1009782, Consultation : 07/2016

Méiose : la reproduction cellulaire (pour reproduction)

La méiose entre en action en réponse à un problème rencontré lors de la reproduction de l'organisme. En effet, lors de la reproduction deux organismes secrètent une cellule contenant leur matériel génétique. Habituellement, ce genre de cellules sont diploïdes, cela signifie, par définition, qu'elles possèdent, dans le cas de l'être humain, 23 paires de chromosomes, donc 46 chromosomes au total. Ce fait pose problème puisqu'étant donné que deux cellules fusionnent, nous obtiendrions un embryon contenant 92 chromosomes.

Pour pallier ce problème, la cellule va, avant de sécréter la cellule contenant le matériel génétique, créer à partir de cette cellule quatre cellules filles haploïdes (ne contenant que des chromosomes célibataire, pour un total de 23 chromosomes par parent). Ce procédé de création cellulaire est appelée « méiose ». La figure 1.16 permet de fournir une vue globale de ce processus.

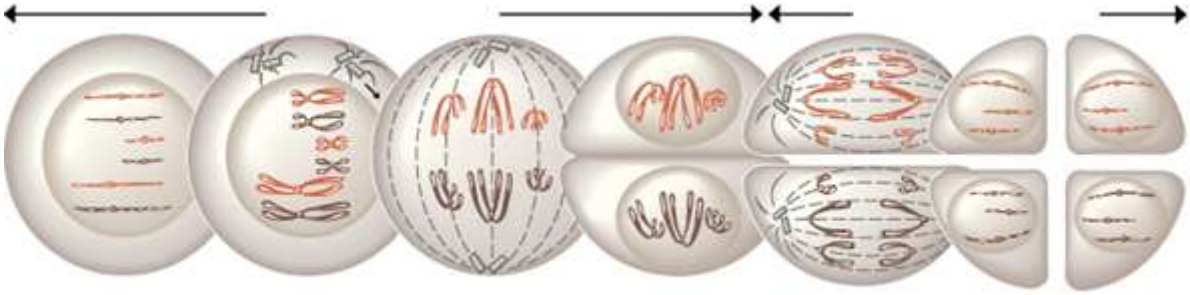

Figure 1.16: La méiose[15]

Notons que la méiose n'est rien de plus que deux divisions cellulaires, contenant les mêmes étapes que la mitose, appliquées l'une après l'autre.

La méiose est, également, le processus qui permet, comme nous le verrons ultérieurement, l'introduction d'un brassage génétique et donc, d'une grande variation de production de cellules haploïdes à partir du génome de l'individu.

[15] Source image : *méiose*, http://www.larousse.fr/encyclopedie/divers/m%C3%A9iose/69066, Consultation : 07/2016

Production de protéines

Une des fonctionnalités clefs dans la maintenance et développement des organismes, est la production des protéines.

Celles-ci consiste, à partir des brins d'ADN, à créer les séquences d'ARNm, essentielles à cette étape, y correspondantes, à partir desquelles, la cellule pourra produire les protéines.

Ce processus est composé de trois étapes :

- Transcription de l'ADN en ARNm
- Migration de l'ARNm du noyau vers le cytoplasme
- Elaboration de la protéine correspondante à la séquence d'ARNm dans le cytoplasme (étape de traduction)

Les sous-parties suivantes reprennent les explications concernant chaque étape.

La transcription

La transcription a pour but de produire une séquence d'ARNm correspondante, suivant le principe de la polymérisation et ce, comme le suggère la figure 1.17, dans le sens « upstream » (vers 5').

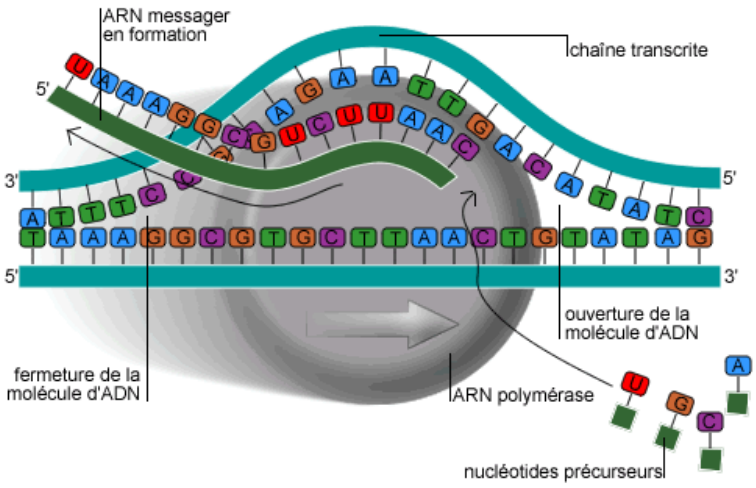

Figure 1.17: La transcription[16]

L'enzyme utilisée, appelée « RNA polymérase II », permet de trouver l'emplacement adéquat, d'écarter les deux brins d'ADN, et d'en sélectionner un pour le parcourir et de créer, à partir de celui-ci, la séquence d'ARNm.

Notons que la séquence d'ARN se complète dans le sens opposé à celui du parcours du brin d'ADN.

Notons, également, que certaines variantes peuvent exister. En effet, la séquence, ainsi produite, peut attirer un nucléotide de guanine chimiquement modifié, ayant pour but de protéger la séquence et d'y reconnaître la position de départ. De plus, il se peut, également, que 100 à 200 bases d'adénine soient ajoutées après la séquence de terminaison. Nous appelons cette dernière structure « poly A-tail ».

La séquence d'ARN finalement obtenue, avec ou sans l'application des légères modifications possibles, est appelée « primary transcript ».

[16] Source image : Transcription du matériel génétique, http://ressources.unisciel.fr/DAEU-biologie/P3/co/P3_chap1_c09.html, Consultation : 07/2016

Migration vers le cytoplasme

Concernant la deuxième étape consistant à permettre le passage de la séquence obtenue du noyau, où elle a été construite, vers le cytoplasme et plus particulièrement, vers les ribosomes où elle sera traitée. Nous avons décidé, par soucis de simplicité, de ne pas rentrer dans les détails de cette étape.

La translation

L'étape de translation, quant à elle, concerne la partie de production de protéine. Avant de rentrer dans le vif du sujet, nous devons expliquer certaines caractéristiques préliminaires, à savoir : l'extraction des sous-séquences codantes et des notions de code génétique.

La première d'entre elles, concerne l'extraction des séquences codantes à partir de la première transcription d'ARNm. En effet, cette séquence est, en réalité, composée de deux sortes de sous-séquences d'ARN : les codantes, appelées « extrons » et les non-codantes, appelées « introns ». Le but de cette étape étant d'extraire que les extrons afin d'obtenir une séquence d'ARNm « purifiée ».

La deuxième notion concerne le code génétique. En effet, cette partie nous permet de faire le lien entre la séquence d'ARN (composée que d'extrons) et la protéine. Pour ce faire, nous devons définir une protéine comme étant un ensemble de polypeptides eux-mêmes composés pas un ensemble d'acides aminés. Un acide aminé, quant à lui, est représenté par un groupement de trois nucléotides, appelés « codons ».

La figure 1.18 représente tous les acides aminés qu'il est possibles d'obtenir sur base des quatre nucléotides (A, U, C et G).

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA Stop UAG Stop	UGU, UGC } Cys UGA Stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile AUG Met	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

Figure 1.18: Ensemble des acides aminés[17]

Remarquons le fait qu'il n'existe que 64 possibilités (puisque quatre nucléotides : 4^3) de codons dont trois représentant la terminaison de la protéine et qu'un acide aminé peut être représenté par plusieurs codons (l'inverse n'étant pas possible).

Maintenant que nous avons posé les notions de bases, nous allons nous pencher sur l'étape de translation en tant que telle. En effet, celle-ci ne sert que de séquence de transmission permettant de rassembler les éléments utiles en temps voulu.

La séquence formée des exons de la première transcription d'ARNm obtenue précédemment ne peut être directement utilisée pour la synthèse de la protéine. En effet, les ribosomes, chargés de la mission de la synthétisation, utilisent un dérivé appelé « ARN de transfert » (ARNt) qui est le complément de l'ARNm : A traduit en U (et inversement) et C traduit en G (et inversement). Ce composant prend la forme d'un trèfle dont la partie « haute » est composée de l'anti-codon (complément du codon de l'ARNm). C'est cet anticodon qui sera utilisé par le ribsome pour synthétiser la protéine.

La figure 1.19 présente cette transition.

[17] Source image : Querioz, Emmechec, El-Hani, *Information and Semiosis in Living Systems: a Semiotic Approach*, http://see.library.utoronto.ca/SEED/Vol5-1/Queiroz_Emmeche_El-Hani.htm, Consultation: 07/2016

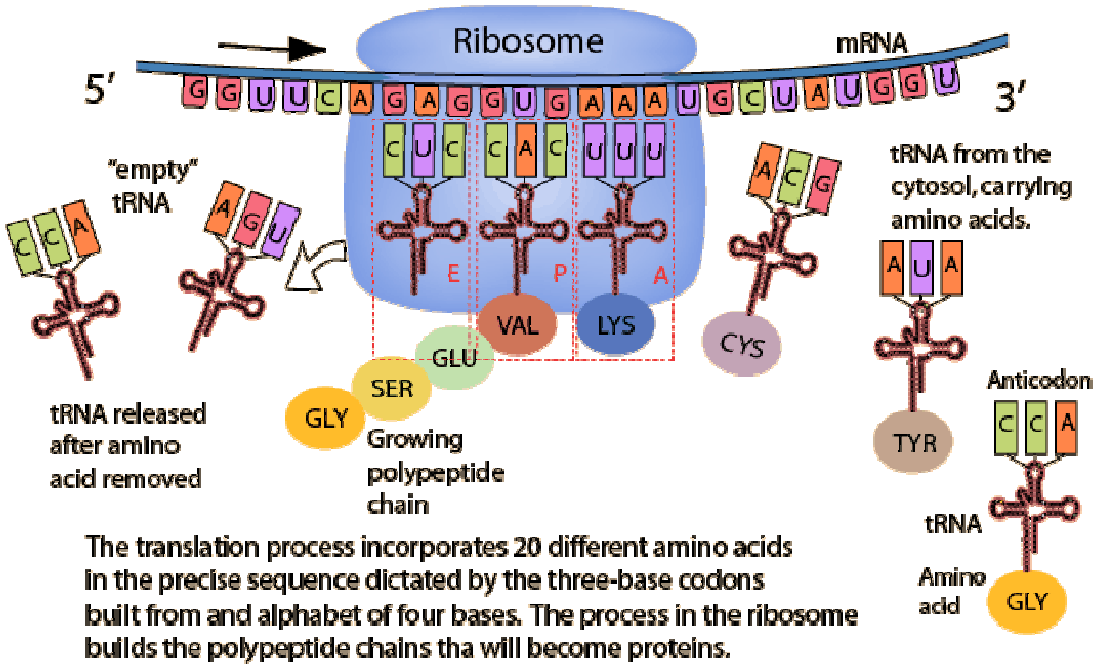

Figure 1.19: Translation[18]

N.B. : Notons que si l'ARNt est le complément de l'ARNm, étant lui-même le complément de l'ADN, nous pouvons dire que la synthétisation à partir de l'ADN (à l'exception des nucléotides de thymine qui est remplacés par les nucléotides d'uracile). L'ARNm ne servant, donc, qu'à transmettre la séquence aux ribosomes.

Nous pouvons, donc, affirmer que la translation a pour but à traduire chaque codon en acide aminé, ce qui permettra à la cellule de produire les polypeptides, et donc, la protéine.

Variations génétiques

Les explications précédentes nous ont permis de comprendre les principes fondamentaux de la génétique. Cependant, ils ne nous permettent pas de comprendre la raison d'une telle variété de caractères héréditaires présents dans la nature.

Cette partie, consacrée aux variations génétiques, se penchera sur les principes de mutation génétique ainsi que celui de l'enchâssement (crossing-over en anglais) qui permettent l'apparition du concept de brassage génétique.

[18] Source image : *Como se expresa el ADN*, http://slideplayer.es/slide/9863646/, Consultation : 07/2016

La mutation génétique

Le principe de mutation consiste en une modification du code génétique (ADN), soit dans les cellules somatiques (composant un organisme), lors d'apparition de cancers, ou dans les cellules germinales, lors de mutations héréditaires.

Il existe plusieurs types (ou causes) de mutations :

- Mutations de base
    - Substitution de paires de bases

      Consiste en un remplacement de paires de bases par une autre. Ceci a pour conséquence d'altérer les séquences d'ARN produites et, donc, à long terme, de modifier la synthétisation des acides-aminés. Il existe deux sortes de modifications possibles : la « nonsense » (« misens » en anglais), lors d'apparition d'un autre codon que celui attendu ; et « nonsense », lors de l'apparition de codons de terminaison.

    - Ajout/Suppression de paires de bases

      Consiste en une modification du code génétique d'un individu par insertion ou suppression de nucléotides. Ce type de mutation est réellement nuisible à partir du moment où le nombre de modification est un multiple de trois.

      La figure suivante (1.20) présente ces deux premiers types de mutations.

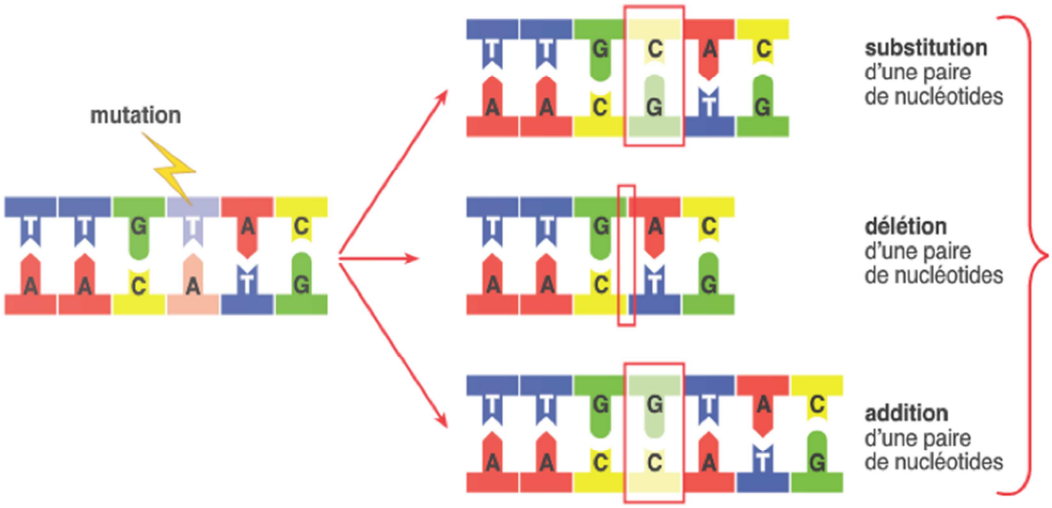

Figure 1.20: Mutations subs + add/del[19]

- Mutations spécifiques
  - Mauvaises identification introns/exons

    Consiste en une mauvaise délimitation des frontières des parties codantes et non-codantes (introns et exons) de la séquence d'ARNm. Ce genre de mutation engendre une mutation de base présentée ci-dessus.

  - Mauvais choix de position de départ

    Consiste en un mauvais choix concernant le point de départ pour la formation de la séquence d'ARNm, ce qui se traduit par une diminution de la qualité de la protéine résultante. Ce genre de mutation engendre une mutation de base présentée ci-dessus.

Toutes les mutations aboutissent à un disfonctionnement du procédé de synthétisation de protéines et peuvent, donc, avoir d'importantes répercussions conséquentes sur la santé de l'individu (apparition de maladies génétiques).

[19] Source image : http://antibiotique-utile-remplacable.revolublog.com/-a114297936, 2015, Consultation : 07/2016

Crossing-over (Lien avec Méiose) plutôt dans variété génétique

L'enchâssement est un principe qui permet d'instaurer de la variation génétique au sein des caractéristiques héritables d'un individu.

Elle apparait pendant la méiose, lors de la réplication chromosomique. En effet, il se peut que deux allèles, situés sur deux chromosomes homologues, s'intervertissent. Dans ce cas, il se pourrait qu'un chromosome ne soit pas entièrement « légué » à un héritier, mais une composition des deux chromosomes homologues. Ce qui a pour but d'instaurer une grande diversité génétique au sein d'une seule personne.

La figure 1.21 représente ce processus.

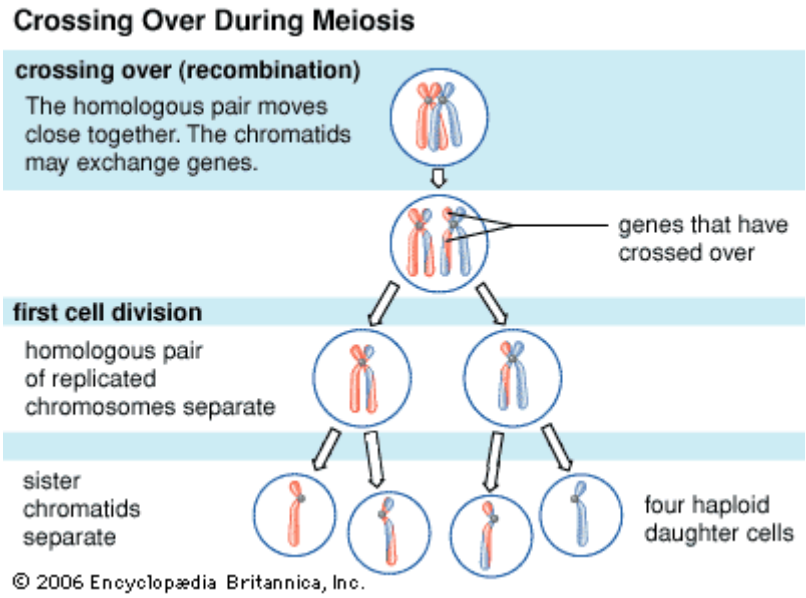

Figure 1.21: Crossing-over[20]

Notons que ce processus que si les locis (emplacements) des gènes sur le chromosome sont suffisamment éloignés l'un de l'autre. En effet, au plus la distance augment entre ces deux locis et au plus la probabilité de chance d'observer un enchâssement est probable.

Par exemple, supposons que les gènes définissants la couleur des cheveux et celles des yeux se situent sur le même chromosome, à des locis suffisamment éloignés que pour permettre l'enchâssement. Supposons, également, que :

[20] Source image : *Meiosis*, https://www.britannica.com/science/meiosis-cytology/images-videos/During-meiosis-an-event-known-as-chromosomal-crossing-over-sometimes/106567, Consultation : 07/2016

- le père d'un individu possède des yeux bruns et des cheveux noirs, il se pourrait qu'il possède les chromosomes suivants

    - chromosome1 : yeux bruns et cheveux blonds
    - chromosome2 : yeux bleus et cheveux noirs

- la mère possède des cheveux blonds et des yeux bleus, il se pourrait qu'elle possède les chromosomes suivants :

    - chromosome1 : yeux bleus et cheveux blonds
    - chromosome2 : yeux bleus et cheveux blonds

Imaginons maintenant, que l'enfant possède des yeux bleus et des cheveux blonds. A première vue, cela semble impossible puisque pour cela, il faudrait que le père lègue à son enfant l'allèle correspondant aux cheveux blonds, situé sur le deuxième chromosome, et celui correspondant aux yeux bleus, situé sur le deuxième chromosome.

Or, étant donné que, par hypothèse, les deux gènes responsables des couleurs des yeux et des cheveux sont suffisamment éloignés, il se pourrait que, par le processus d'enchâssement, les deux allèles nécessaires se retrouvent sur le même chromosome.

Cet exemple prouve l'importance de ce brasage génétique et pourrait susciter quelques curiosités concernant la prédiction des variations.

Mort de la cellule

Les explications précédemment présentées, impliquent que les cellules sont obligées de mourir. En effet, étant donné qu'elles se reproduisent, elles sont obligées de mourir afin que leur nombre ne dépasse pas un certain seuil maximal. La mort des cellules est, en quelque sorte, un mal nécessaire pour assurer le bon fonctionnement de l'organisme et, donc, le bien être de l'individu.

Cependant, il se peut que certains composants, comme les bactéries ou les virus, ne déclenchent la mort prématurée de la cellule. La section suivante a pour but d'expliquer les conséquences que l'instauration d'un virus, source de nombreuses maladies génétiques, provoque dans une cellule.

Les virus

Les virus sont, généralement, constitué que d'une seule cellule, contenant le code génétique viral, protégée par une capsule protéique.

Les virus constituent un moyen d'impliquer la mutation génétique chez un individu. En effet, lors de son instauration dans une cellule, le virus va modifier le code génétique, de cette cellule hôte, en y insérant son propre matériel génétique. De là s'en suit l'apparition d'une mutation impliquant la présence d'une infection.

Concernant les infections en découlant, il en exits deux sortes :

- Infections impliquant l'arrêt du fonctionnement normal de la cellule hôte pour que celle-ci produise des protéines renforçant le virus. La cellule infectée peut se rompre, on dit alors, qu'elle est « lysée », en libérant des particules virales dans son environnement, ce qui causera, l'infection d'autres cellules.
- Infections impliquant l'instauration du matériel génétique du virus dans celui de la cellule hôte. Celle-ci peut, soit, se comporter normalement jusqu'à ce qu'un élément externe ne déclenche une réaction négative ; soit, créer d'autres virus ; soit, perdre le contrôle de son fonctionnement interne et, par conséquent, dans le cas d'une reproduction, générer un cancer.

Il est important de noter, que malgré l'aspect négatif que ces virus représentent, ils possèdent un aspect positif dans le sens où ils permettent la mutation génétique.

Chapitre 2 : Fondements bio-informatiques

La bio-informatique, discipline apparue dans les années nonante, est le domaine d'expertise visant à offrir, à la recherche biologique, une facilité de développement en lui offrant des capacités importantes d'évaluations et de visualisations de données. En effet, cette approche permet, comme dans les autres domaines d'activités adaptés à l'informatique, non seulement, de décharger les utilisateurs des difficultés liées à l'accomplissement de leurs tâches, mais également, d'aller plus loin dans les possibilités liées à leur domaine. Ce qui, étant donné le volume de données, devenu de plus en plus important, s'est révélé être une nécessité.

Généralement, plusieurs professions entrent en considération dans ce domaine. En effet, nous retrouvons, dans les projets liés à ce domaine, des informaticiens, biologistes, mathématiciens (statisticiens) et bien d'autres métiers. Ceci est dû au fait que ce domaine peut être décomposé en plusieurs aspects liés à différents domaines, à savoir : le domaine biologique et génétique, les statistiques, les visualisations informatiques, et bien d'autres encore.

Chaque année, le progrès dans ce milieu se fait ressentir [Liang, 2013]. Notamment, grâce au développement technologique de l'informatique permettant, à long terme, le progrès dans l'élaboration de bases de connaissances relatives aux données biologiques.

Tout comme la section précédente, cette partie aura pour but de donner une vue générale du domaine bio-informatique afin d'offrir au lecteur une compréhension globale du sujet.

Pour ce faire, celle-ci se composera d'un bref contexte historique, suivi d'un condensé d'informations relatif aux données largement utilisées, telles que les séquences ADN et d'ARN, les génomes, etc. ; ainsi que leur obtention. Ensuite, nous traiterons de leur traitement en proposant une présentation des fonctionnalités ainsi que des outils disponibles.

Extraction des données

Dans le but d'acquérir des informations génétiques relatives à des individus, nous devons, dans un premier temps, obtenir le matériel génétique les concernant. Pour ce faire, deux moyens s'offrent à nous : nous pouvons, soit, extraire les séquences d'ADN à partir de composants organiques, tels que le sang, la salive, ou encore, les tissus; soit, en ayant recourt à des outils de partage de connaissances disponibles sur le web.

Cette section donnera au lecteur une présentation générale de ces méthodes ainsi que des types de données.

Extraction à partir d'échantillons

Les scientifiques peuvent extraire le génome (séquences d'ADN) à partir d'une cellule de l'individu. Celle-ci peut provenir d'un de ses composants organiques, sang ou salive, ou d'un tissu. Cette section a pour but d'expliquer les méthodes d'extraction de séquences d'ADN et d'ARN composant le matériel génétique d'un individu. Nous présentons, d'abord, les explications relatives à l'extraction des séquences d'ADN, suivies par celles relatives à l'extraction des séquences d'ARN.

Concernant l'acquisition des séquences d'ADN, nos sources [Fan, Gulley 2001] prouvent l'existence d'une procédure standard, pouvant être utilisée à partir de tissus. Celle–ci se base sur le fait que l'ADN et les lipides se décomposent, respectivement dans l'eau et dans le phénol. A partir de ce fait, des réactions biochimiques peuvent être utilisées.

Concernant le processus en lui-même, nous pouvons suivre les étapes suivantes afin de l'accomplir.

Les tissus sont, dans un premier temps, désagrégés et traités sous l'action de détergents. Les cellules sont, ensuite, « lissées », ce qui implique la dissolution de la membrane cellulaire. Après cela, l'étape de protéinase est nécessaire afin de digérer la protéine. Après toutes ces étapes, il ne reste plus que, donc, l'ADN et le phénol, qui sera extrait grâce à du chloroforme.

Notons bien, qu'il existe différentes manière d'obtenir ce genre de données. La figure 2.1 présente un processus d'extraction en entier.

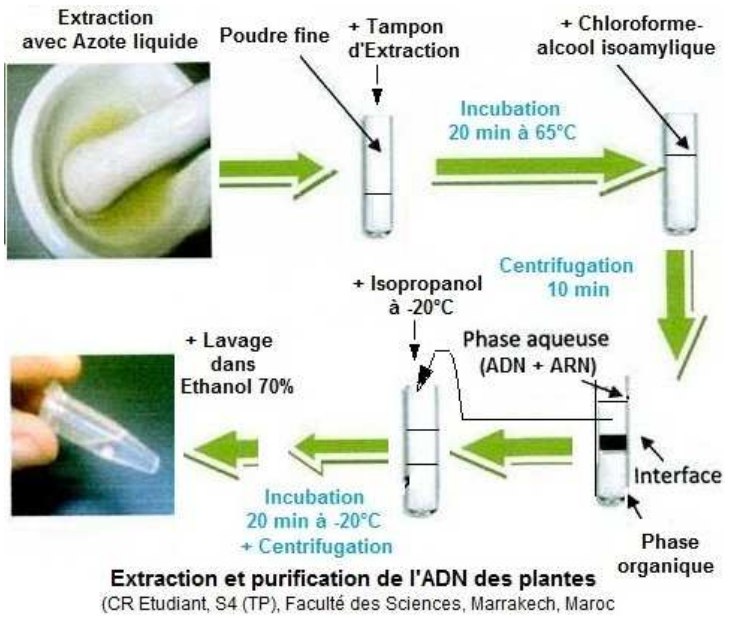

Figure 2.1: Cycle d'extraction ADN[21]

Au cours du temps, d'autres méthodes ont vu le jour. Celles-ci ont l'avantage d'être plus rapide et d'éviter la contamination des séquences d'ADN au phénol. Par exemple, une autre source [Aljanabi , Martinez, 1997] propose la présentation d'une méthode universelle pouvant gérer plusieurs types d'échantillons.

Concernant l'extraction des séquences d'ARN, la méthode décrite précédemment pour l'extraction des séquences d'ADN permet également d'obtenir les séquences d'ARN. Il nous suffit donc juste d'appliquer à ce résultat un autre composant chimique. Il existe, également, d'autres méthodes plus évoluées.

Comme nous pouvons le constater, ces méthodes requièrent des compétences particulières dans le domaine ainsi que du matériel adapté et, ne sont, par conséquent, pas accessible à tous les publics.

N.B. : Notons que, malgré le fait que l'extraction d'ARN puisse être considérée comme étant proche de celle de l'ADN, ces séquences possèdent des avantages qui leurs sont propres et sont, par conséquent, plus utilisées dans certaines situations. En effet, les séquences d'ARN sont plus sensibles et permettent donc d'obtenir de meilleures mesures pour l'expression des génétiques.

[21] Source image : *Extraction de l'ADN*, http://www.takween.com/techniques/ADN-purification-protocole.html, Consultation : 07/2016

Extraction à partir de bases de connaissances

Une alternative à l'extraction de données à partir d'échantillon existe et consiste à simplement télécharger le contenu approprié à partir du web. Par contenu, nous entendons toutes les séquences d'ADN ou ARN ainsi que les génomes et les protéines.

A l'heure actuelle, il est disponible de les obtenir grâce à trois bases de données publiques principales:

- « GenBank » au Centre National d'Information Biotechnologique (NCBI)
- Base de données d'ADN du Japon (DDBJ)
- Laboratoire Européen de Biologie Moléculaire (EMBL)

Chaque source possède une large base de connaissances qui lui est propre ce qui leurs permet de couvrir une large partie des espèces existantes.

D'autres sources existent, mais celles présentées ci-dessus représentent la plupart des données disponibles et continuent à s'enrichir quotidiennement dans le but de pouvoir offrir des génomes de plus en plus complets grâce aux méthodes d'extractions de séquences précédemment présentées.

La constitution de ces sources, s'effectue en extrayant directement les échantillons, de manière physique, comme explicité dans la section précédente, à partir des échantillons de tissus (ou autres constituants organiques) d'un individu de l'espèce étudiée.

La figure 2.2 permet au lecteur de se rendre des progrès qui sont fait chaque année dans la constitution de base de connaissances relatives aux génomes des espèces.

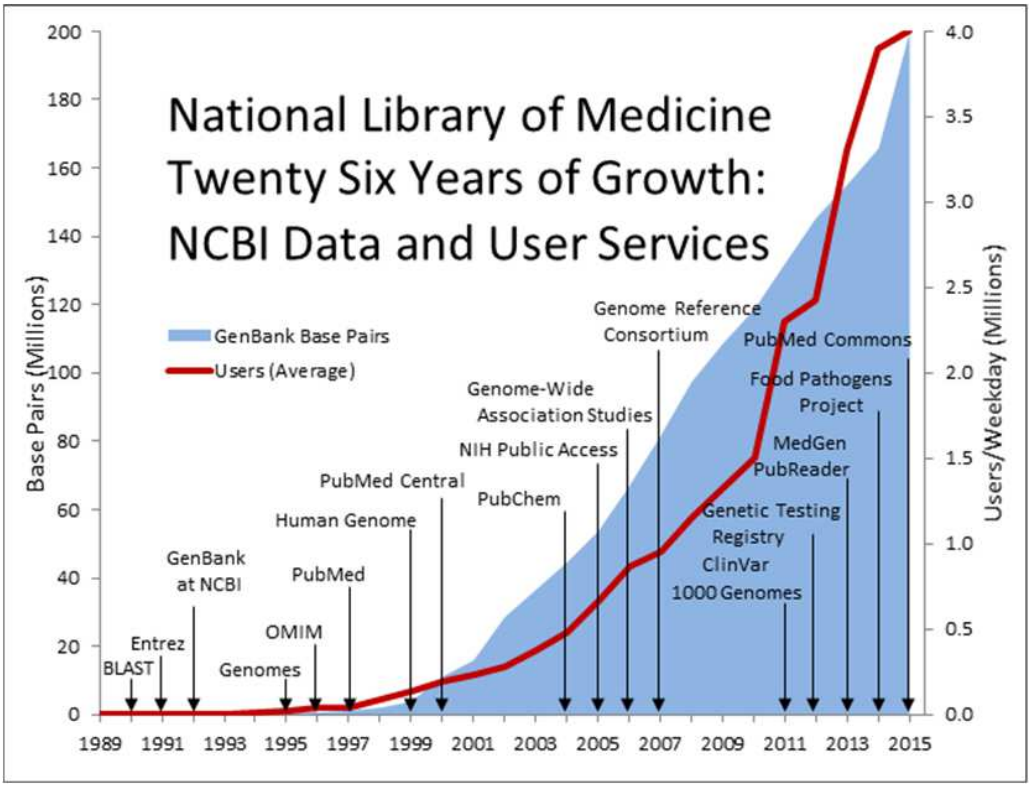

Figure 2.2: Evolution de la connaissance de GenBank[22]

Chacune de ces sources offrent un ensemble d'outils permettant l'accessibilité aux différentes informations, qu'elles concernent les séquences ou les articles scientifiques. Parmi ces outils, nous avons les browsers, visualisateurs de protéines, gènes et génomes.

Ceux-ci nous permettent d'obtenir non-seulement les génomes relatifs à des situations spécifiques, mais également le génome de référence qui est considéré comme étant le génome le plus représentatif de l'espèce étudiée. Ce génome de référence est construit sur base de plusieurs individus de l'espèce et est composé de tous les gènes de l'espèce ayant été rencontré à ce jour.

Il peut être téléchargé dans le format « GFF » ou « fastqc », qui sont des formats assez répandus et utilisé par les outils dans le domaine, à partir des dépôts mis à disposition par les fournisseurs d'informations (NCBI, etc).

N.B. : Comme dit plus haut, il existe d'autres librairies constituant une source de données. C'est le cas d'Illumina, qui est à l'origine des sources d'échantillons du projet mené dans la partie deux, ou encore de « Entrez ».

[22] Source image : https://www.nlm.nih.gov/about/2017CJ.html, 2016, Consultation : 07/2016

Types de données utilisées

Les données, pouvant être obtenues sur le web, en ayant recourt aux moyens précédemment exposés, ou directement extraites des composants organiques, doivent être représentées dans un format électronique afin de pouvoir être traité électroniquement.

Pour ce faire, plusieurs formats de données ont été développés au cours du temps par les différentes institutions à travers le monde. Les formats que nous présenterons dans cette section, seront limités à l'ensemble suivant : FASTA, GFF, SAM et BAM. Ce choix est justifié par le fait que ces formats sont largement utilisés dans ce domaine et par le but de ce travail, qui est de fournir une vue générale du domaine bio-informatique.

Pour des raisons de clarté, nous allons séparer ces formats en deux catégories(en fonction de leur utilité). La première catégorie concerne les formats utilisés pour stocker des séquences.(FASTA, FASTQ) et la seconde pour enregistrer les résultats de certaines fonctionnalités (SAM, BAM).

Formats de stockage de séquences

Cette section a pour but de présenter les différents formats de données couramment rencontrés lors de manipulation d'outils tels que le pipeline. Ceux-ci servent à présenter les séquences des échantillons ainsi que le génome de référence. Il est important de noter que ces formats sont équivalents et représentent le même type de données. Nous avons décidé d'en parler afin de pouvoir fournir au lecteur un éventail relativement complet concernant les types de fichiers pouvant être rencontrés dans le domaine.

N.B. : D'autres formats, tels que « BED », « ENCODE », etc., existent, mais ne seront pas pris en compte dans ce travail.

FASTA

Ce format de fichier est largement utilisé pour représenter les séquences de données (ADN/ARN) par la plupart des outils traitant ce genre de données. La figure 2.3 propose un exemple de fichier FASTA.

```
>1|chr12|64798729-64798930|354.27082|-1.0|1127
CTGGCTGGGCGGACCGGGTGGGGGTGGGTACGAGCCGGGGCCGCCGCCGAGGAGCGCGT
TTGGTGTTTCATCACCCGAATTGCCACGAGGCTTCCTTTAGGGGAGGGATCGGGGGAGG
GGGTCGGCATCGCCTGTGGTTCCGAAGCCCGTTAG
>2|chr12|57848784-57848985|336.02993|-1.0|635
CCACCTGGCTCATAAGGCGTTCCCTCCCCCAAGTCCCAGACCTTGGGGACTGAGCATGT
TGCGTGTCCACATTGCACCCCCCCACCCCCATACCCCTACTTCAGGCCCAGTCACCATG
TGGGGAGGAGGACCTCCACCCCCTGCAGGGGCCTG
```

Figure 2.3: Exemple fichier FASTA[23]

Cet exemple va permettre de décrire plus facilement la structure de ce type de format. Grâce à celle-ci, nous pouvons remarquer que chaque séquence commence par le symbole « > », la première ligne contient les informations concernant l'identification et autres caractéristiques de la séquence, alors que les autres lignes contiennent la séquence. Le tout étant d'une taille inférieure à 80 caractères.[24]

FASTQ

Ce format est un dérivé du précédent. Il possède les mêmes caractéristiques concernant les informations qu'il peut stocker. Cependant, des différences peuvent, tout de même, être observées. En effet, ce format permet, par exmple,d'ajouter une chaine de caractères représentant la qualité de la séquence de nucléotides.

[23] Source image : http://couger.oit.duke.edu/documentation/, 2015, Consultation : 07/2016
[24] Sur base de la documentation consultable à l'adresse suivante : *Blast documentation*, http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml , 2007, date de consultation : 23/07/2016

La figure 2.4 présent e un exemple de ce genre fichiers. Cette figure, provenant des données que nous avons utilisées, permet d'illustrer ce genre de fichiers.

```
@HISEQ1:102:D099AACXX:1:1207:4630:182147 1:N:0:TGACCA
TACAGACATAGGGAACTTCTCATCTTGGTTTCCTCCGGCAAAC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJIJJIJJJIIJJJJ
```

Figure 2.4: fichier FASTQ

N.B. : Il est important de noter que pour représenter la double chaine de nucléotides, nous utilisons deux fichiers. En d'autres termes, nous ne représentons les deux chaines dans le même fichier.

Formats de stockage de résultats

Cette section a pour but de présenter les différents formats de données couramment rencontrés lors de manipulation d'outils tels que le pipeline. Ceux-ci permettent de présenter les différentes entrées et résultats des différents composants de l'outil.

GFF/GTF

En nous basant sur notre expérience, nous pouvons affirmer que ces deux formats, GFF, pour (« General Feature Format »), et GTF, pour (« General Transfert Format »), sont deux formats largement utilisés dans le domaine pour représenter les annotations. Ces annotations permettent de faire le lien entre les chromosomes et les séquences.

Ces deux formats sont équivalents et leur syntaxe est fort proche. Ils possèdent, tous deux, une structure similaire basée sur neuf colonnes délimitées par des tabulations.[25]

Cependant, le format GFF est considéré comme étant plus général que le GTF.

Concernant la conversion entre les deux, nous savons, d'après notre expérience dans le domaine, qu'il n'existe pas de script tout fait. En effet, il existe des solutions devant être adaptées aux noms utilisés pour les champs du format GFF. Notons que la conversion de GTF vers GFF est plus aisée que celle permettant l'opération inverse.

[25] D'après la documentation disponible à l'adresse suivante: *GFF/GTF Format – Definition and supported options*, http://www.ensembl.org/info/website/upload/gff.html, 2016, Consultation : 07/2016

La figure 2.5 propose un exemple de fichier GFF.

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr12	unknown	exon	87984	88017	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88257	88392	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88570	88771	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88860	89018	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	89675	89827	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90387	90655	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90796	91263	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	147946	148309	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	148612	148814	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	149052	149412	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	CDS	176049	176602	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	176049	176602	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	start_codon	176049	176051	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	186542	186878	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	CDS	208312	208380	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	234799	235078	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	234799	235078	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	246577	246793	.	-	.	gene_id "LOC574538"; gene_name "LOC574538"; transcript_id "NR_033859"; tss_id "TSS17153";
chr12	unknown	CDS	247433	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	247439	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	start_codon	247439	247441	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";

Figure 2.5: Exemple fichier GTF[26]

SAM

SAM (« Sequence Alignment/ Map Format) est un format permettant de représenter les résultats obtenus lors de l'accomplissement de l'étape d'alignement (voir section suivante). Ceux-ci peuvent être obtenus grâce à de divers outils permettant la réalisation de l'alignement. En effet, il nous permet de représenter les résultats liés à l'étape d'alignement, etc.

La figure 2.6 propose un exemple de fichier SAM.

[26] Source image : *Aligning SE RNA-Seq Reads to a Reference Genome*, http://bioinformatics.ucdavis.edu/docs/2014-june-workshop/Thursday_MB_Tophat_Model_SE.html, 2014, Consultation: 07/2016

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJHHHHHFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Figure 2.6: Exemple fichier SAM[27]

Comme le suggère cette figure, les fichiers SAM sont des fichiers, délimités par des tabulations, organisés en deux parties. La première, est constituée de la section, optionnelle, 'header', contenant les informations sur le fichier (fichier trié, etc); tandis que la deuxième concerne la partie renseignant tous les alignements.[28]

BAM

D'après notre expérience, nous comprenons que ce format de fichier correspond à la transposition binaire du format SAM, précédemment exposé. Il peut être utile, dans certains cas, de directement traiter ce genre de données binaires. En effet, certains outils ont été créés dans le but de réaliser l'étude d'expressivité génétique en utilisant ce genre de fichiers.

Des outils, tels que « SAMTools », proposent des méthodes de transpositions entre ces deux formats.

[27] Source image: *SAM File Format*, https://insidedna.me/tutorials/view/samtools-commands-tutorial-working-sam-bam-files, 2015, Consultation : 07/2016

[28] D'après la documentation : *Sequence Alignment/Map Format Specification*, The SAM/BAM Format Speci cation Working Group, https://samtools.github.io/hts-specs/SAMv1.pdf, Novembre, 2015, date de consultation: 08/2016

Traitement des données

L'atout principal du domaine d'activité, que représente la bio-informatique, réside dans la capacité de pouvoir offrir au domaine de la recherche biologique toute la puissance computationnelle offerte par les technologies informatiques actuelles.

C'est pourquoi, nous avons décidé de dédier une partie de cette section aux différentes fonctionnalités d'analyse pouvant y être associées.

Analyse

Les différentes étapes du processus d'analyse de données dépendent du but final des recherches. Cependant, un ensemble de fonctionnalités semblent constituer un ensemble de base étant généralement utilisé lors d'expériences et étant, par conséquent, commun à de multiples expériences. Celui-ci est constitué des étapes d'alignement, d'expressivité génétique (étape de comptage) et de visualisation.

Cette sous-section permet de fournir une présentation globale de ces fonctionnalités. Le but étant de pouvoir les détailler ultérieurement dans les deux parties suivantes consacrées aux expériences.

N.B. : Avant d'aller plus loin, notons que ces fonctionnalités peuvent nécessiter l'accomplissement de certaines étapes préliminaires visant, sur base des procédés biologiques concernant la constitution des séquences d'ARN, à vérifier la qualité et corriger ou adapter, si besoin, ces séquences. Par exemple, certaines séquences d'ARN, comme exposé précédemment, peuvent être complétées par des nucléotides (« poly A-tail »).

Alignement

Cette étape permet de trouver la localisation de séquences, introduites par un utilisateur, sur un génome de référence. Concrètement, ce processus permet de définir le gène du génome de référence qui sera utilisé dans la séquence [Oshlack, Robinson, Young, 2010].

En plus de pouvoir définir les gènes impliqués dans un échantillon de données, cette fonctionnalité permet d'obtenir le pourcentage de similarité entre un génome de référence et un échantillon,

prélevé à partir d'un organisme ayant été transformé grâce à des hormones ou autres composants similaires. Ce qui, dans certaines analyses, permet d'obtenir de premières conclusions. En effet, étant donné que l'étape d'alignement permet déplacer les reads d'un échantillon sur le génome de référence, au plus, votre pourcentage de reads alignés est élevé, au plus, votre échantillon correspond au génome de référence.

Il est important de noter que deux sortes de types d'alignement existent : l'alignement local et l'alignement global. Dans le premier cas, le génome de référence est décomposé en plusieurs parties à partir desquelles, nous alignerons la séquence étudiée. Le second type, quant à lui, permet d'aligner la séquence étudiée à l'entièreté du génome de référence, ce qui, nécessite des algorithmes plus performants et sophistiqués étant donné la taille des donnés plus importantes à manipuler [Oshlack, Robinson, Young, 2010].

D'un point de vue plus technique, comme nous le verrons dans la partie suivante, cette étape peut être réalisée grâce à l'outil « Subread » ou par la libraire homologue pour le langage statistique R « Rsubread ». Notons, que d'autres outils existent (Blast, etc), mais qu'ils ne seront pas présentés dans ce travail.

Il est, cependant, important de noter que ces outils se basent essentiellement sur deux algorithmes, à savoir : « la transformé de Burrows Wheeler (BWT) » et les hash tables. Sans rentrer dans les détails, le premier consistant à représenter les données sous une forme plus compacte; tandis que le deuxième, permet de construire une table de hash contenant la localisation de chaque sous-séquence du génome de référence. Ces deux algorithmes possèdent des avantages et inconvénients. En effet, le premier, peut accomplir l'alignement efficacement, d'un point de vue des ressources, mais devient très lent lorsqu'il s'agit de cas plus complexes. Le second algorithme, quant à lui, est plus extensible que le second, et est, par conséquent, plus performant pour détecter les différences les plus compliquées entre les reads et le génome de référence mais impose des contraintes liées à l'utilisation des ressources. [Oshlack, Robinson, Young, 2010 & Fenwick, 2007]

Concrètement, l'utilisateur peut utiliser les outils, cités ci-dessus, en fournissant un fichier au format « fastqc » (voir précédemment) et un fichier contenant l'index du génome de référence. Ce dernier est créé en raison de la taille importante du génome de référence. En effet, celui-ci recueille tous les gènes recensés à ce jour et a, par conséquent, une taille importante. Ce fichier peut être créé grâce à « Samtools », « Subread » e bien d'autres..

Les résultats seront, quant à eux, représentés sous la forme d'un fichier « sam », ou dans sa version binaire « bam ».

Les résultats, obtenus à partir de ces outils, permettront de tirer des conclusions sur l'étude en cours de réalisation. Plus tard, nous ferons le lien entre l'alignement et la localisation d'un gène sur un chromosome, ce qui nous sera utile pour obtenir des conclusions relatives à des études.

N.B. : Le terme « mapping », ou encore « mapper », peuvent être rencontrés dans la littérature et font référence au processus d'alignement.

Détermination de l'expressivité génétique

Après avoir obtenu les localisations des séquences sur le génome de référence, nous pouvons définir l'expressivité des gènes de ce génome de référence. Cette opération, pouvant être appelée « opération de comptage », permettra de définir, pour chaque gène du génome de référence, le nombre de reads, provenant des échantillons étudiés, y étant alignés. Un pourcentage d'utilisation des gènes du génome de référence en fonction d'un échantillon donné pourra, selon les outils, être obtenu. Notons que cette information ne représente qu'une information secondaire, la principale étant le comptage.

Cette opération est très importante dans le domaine de la recherche puisque celle-ci permet de savoir quels gènes seront le plus utilisés. Par exemple, si nous étudions l'influence que possède une hormone, ou autre agent, sur le fonctionnement génétique d'un individu, nous serons en mesure de définir les gènes qui réagissent le plus sous l'effet.

Les résultats peuvent être perçus comme étant une distribution, représentant le nombre de reads alignés sur chaque gène du génome de référence, pouvant être statistiquement étudiée. Ce fait est important puisqu'il permet de comprendre le genre de graphiques qui sont utilisés pour représenter les résultats

En ce qui concerne l'aspect technique, nous pouvons utiliser l'outil « Rsubread » qui possède une fonction « featureCounts », ou encore, l'outil « Samtools » pour obtenir les résultats.

Dans la pratique, nous utilisons cet outil avec les résultats obtenus précédemment (sous forme de fichier « sam ») afin de pouvoir obtenir de nouveaux résultats concernant l'expressivité génétique. Il est à noter qu'il se peut qu'un fichier d'annotation « GFF » soit nécessaire avec certains outils, par exemple « Subread ». Cependant, cette contrainte peut, facilement, être contourner grâce à l'utilisation de « Samtools »

Visualisation

Etant donné le large nombre de données étant traitées, une étape permettant de rendre les données visuellement accessibles aux êtres-humains est nécessaire afin permettre l'élaboration de conclusions. C'est la raison pour laquelle, une des branches de la bio-informatique traite de la visualisation des données.

Dans cette section, nous verrons trois types de techniques permettant de fournir un état visuel aux scientifiques. Celles-ci sont : les arbres génétiques, les différents graphiques et les rendus visuels en trois dimensions.

Graphiques

Etant donné que nous travaillons dans un domaine fortement liés au domaine des statistiques, il est alors logique que certains de ces graphiques, généralement utilisés dans le cadre de projets statistiques, soient aussi utilisés dans ce domaine.

Parmi les graphiques les plus utilisés, nous pouvons citer : boxplot, barplots, PCA, etc.

Chacun d'entre eux permet de mettre en avant une caractéristique génétique observée dans les échantillons fournis. Par exemple, nous pouvons la dispersion des gènes utilisés, ou encore, former des groupes d'échantillons possédant des similitudes génétiques similaires.

Les figures suivantes permettent fournissent au lecteur des exemples concernant ces graphiques. En effet, la figure 2.7 un boxplots, permettant de représenter les caractéristiques statistiques de la distribution des résultats, et la figure 2.8 un PCA qui permet de représenter le comportement de plusieurs échantillons.

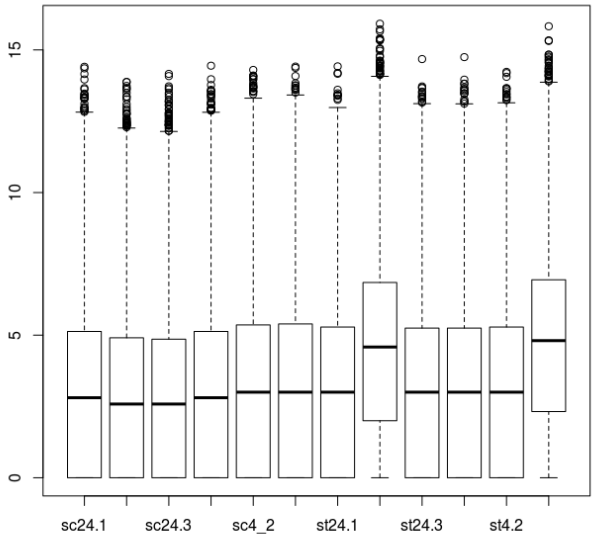

Figure 2.7: Figure sous forme boxplots

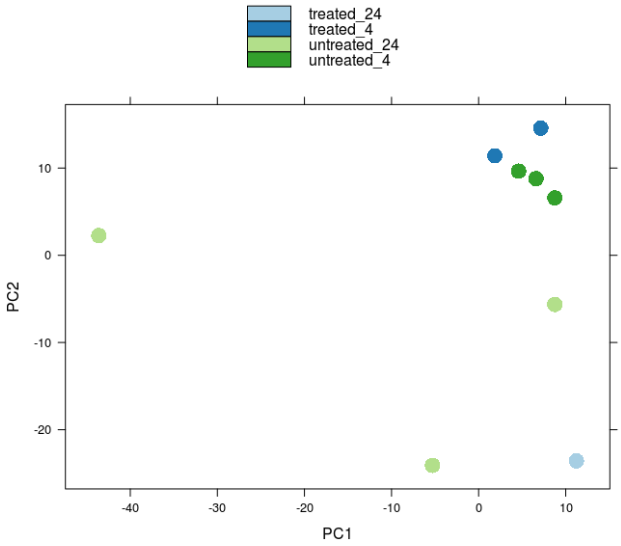

Figure2.8: figure sous forme de PCA

Certains de ces graphiques nécessitent l'emploi de machine learning notamment lorsque nous voulons diminuer le nombre de dimensions des caractéristiques afin de mieux représenter les groupes d'échantillons étudiés.

En effet, nous pourrions, aisément, imaginer l'exemple suivant : une expérience portant sur l'utilisation de plusieurs antidotes ayant pour but de soigner des patients atteints d'une certaine maladie génétique. L'expérience pourrait se dérouler en prélevant des échantillons des patients ayant pu bénéficier, chacun, d'un des différents traitements. Après cela, nous pourrions continuer cette expérience, en réalisant l'étape d'alignement afin de mapper les différents échantillons sur le génome humain de référence (sain). Les résultats obtenus nous permettraient de comparer les différents antidotes, afin de définir le plus adéquat.

Arbres phylogénétiques

Les arbres phylogénétiques sont des structures représentées en arbres. Donc, partant d'une racine, commune à toutes les branches, toutes les bifurcations représentent des variations génétiques. Celles-ci aboutissent, en bout de chaine, à la définition d'une sous-espèce.

L'intérêt de ces structures réside dans le fait qu'elles permettent de représenter visuellement l'évolution et les similarités génétiques (et donc similarités) entre les espèces. Pour cela, cette structure utilise les nœuds, pour représenter un sous-ensemble d'espèces communes, et de longueur de branches afin de représenter les rapports de similitudes entre les nœuds mère et fille.

La figure 2.9 représente une de ces structures.

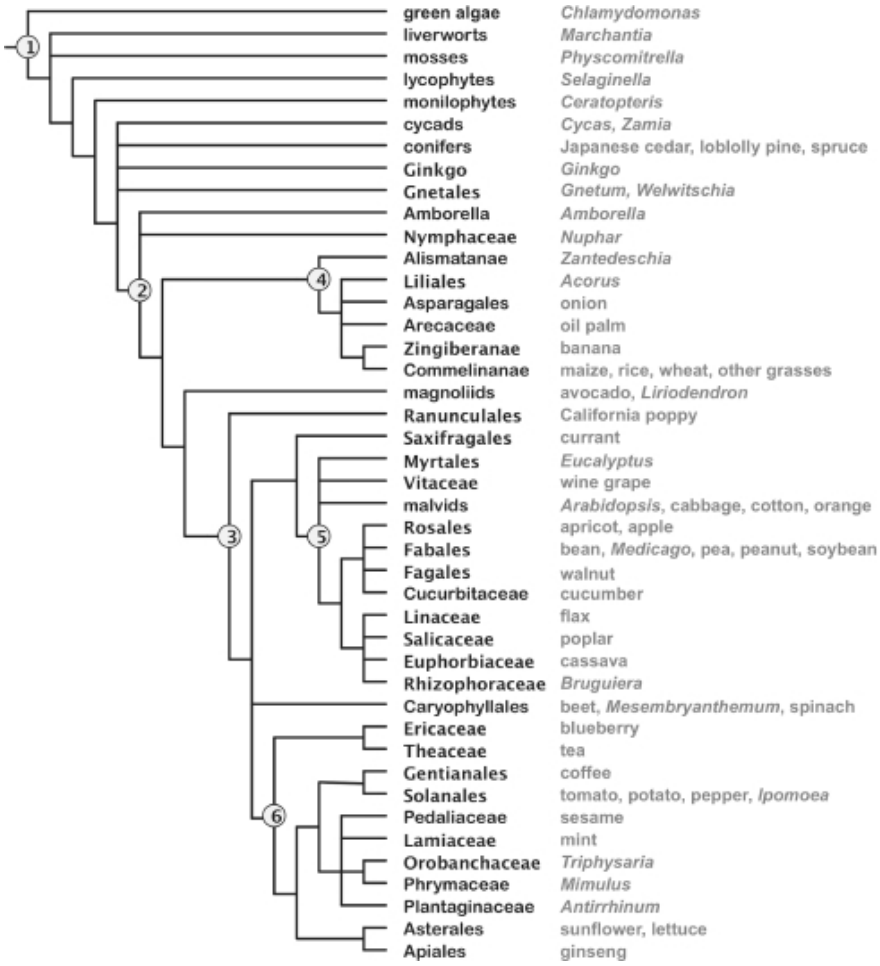

Figure 2.9: Arbre phylogénétique[29]

Notons que, comme ces figures représentent les similarités génétiques entre espèces, avec une notion quantitative de distance, nous pouvons directement le générer grâce à l'étape d'alignement qui permet de définir les similitudes entre espèces.

N.B. : Cette structure peut également être utilisée dans le cas de représentation de familles de protéines.

[29] Source image : Hartmann S., Phillips J, Vision TJ, *Phytome : a platform for plant comparative genomics,* https://openi.nlm.nih.gov/detailedresult.php?img=PMC1347408_gkj045f1&req=4, 2006, Consultation: 07/2016

Rendus visuels

Le dernier type de figure présenté ici est celui des figures en trois dimensions. Celles-ci peuvent représenter visuellement des composants biologiques, tels que des protéines, séquences d'ADN ou ARN et bien d'autres.

Elles sont souvent utilisées soit, comme résultats, à la suite d'observations et de recherches; soit, pour simplement visualiser des composants biologiques, hors de projets de recherche.

Le but de ces visualisations est, encore une fois, de permettre aux chercheurs et scientifiques d'établir plus facilement des conclusions à leurs recherches.

La figure suivante (2.10) représente le rendu visuel d'une protéine.

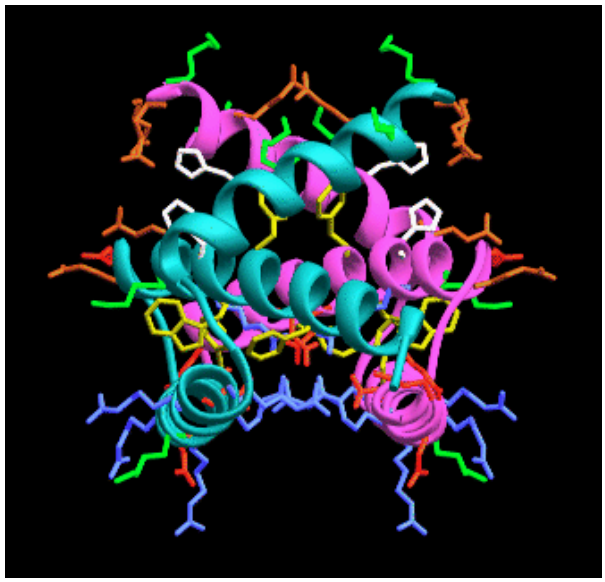

Figure 2.10: Représentation 3D protéine[30]

N.B. : Ce genre de représentation est utilisé à tous les niveaux de la biologie.

[30] Source image : *Protein NMR Structure Gallery*, http://www-nmr.cabm.rutgers.edu/photogallery/structures/, Consultation : 07/2016

Outils disponibles

Afin de pouvoir réaliser les traitements, exposés ci-dessus, plusieurs outils ont été développés. Ceux-ci constituent soit des logiciels, outils disponibles sur le web ou des librairies de programmation, pouvant être utilisées, par exemple, avec le logiciel statistique R.

Ces différents outils réalisent un ensemble de fonctionnalités bien particulières. Nous sommes, donc, dans l'obligation d'utiliser plusieurs d'entre eux afin de tirer, à partir de codes génétiques, des informations plus complètes.

Toujours dans l'optique de fournir une vue générale du domaine bio-informatique, nous n'aborderons, dans cette section, que certains outils, à savoir : BLAST de NCBI, R, bioconductor, Subread, SAMTools, FastQC, FQTrim et DESeq.

Ceux-ci sont des logiciels (ou librairies) libres, souvent développés par des centres de recherche spécifiés dans le domaine ou dans un autre.

Cette section a pour but de décrire ces outils de manière externe à tout contexte particulier.

FastQC

Cet outil permet de donner des informations concernant les séquences de données. Etant donné le caractère instable de la création des séquences, dû aux mécanismes biologiques expliqués précédemment, il nous est obligatoire de vérifier certains critères de fiabilités les concernant. En effet, il se pourrait, par exemple, que ces séquences contiennent des nucléotides tronqués (sans valeur).

Cet outil, nous permet, donc, de vérifier la pertinence des données en se basant sur l'étude de plusieurs critères de fiabilité et permet de créer un rapport visuel, contenant des graphiques, afin de mieux représenter au mieux l'état de ces critères de fiabilité. Parmi ceux-ci, nous avons : le taux de qualité lié à une base particulière, le taux de nucléotides non déterminés dans les séquences et bien d'autres.

N.B. : Ce programme fournit, en plus des résultats, une appréciation permettant à l'utilisateur de savoir si sa séquence est pertinente ou non, et ainsi, l'aider dans sa prise de décision, de manière implicite.

FQTrim

Sur base d'évaluations des séquences, cet outil nous permet de les corriger adéquatement. Il est disponible à partir de son dépôt. Son utilisation ne requiert aucun environnement de programmation, mis à part, l'utilisation d'un terminal Unix. Des options spécifiques permettent à l'outil de mieux corriger les séquences en fonction de leurs caractéristiques biologiques (taille minimum des séquences, etc). Le site web du fournisseur permet de prendre connaissance des options à utiliser.[31]

Il est à noter qu'une des forces de cet outil réside dans le fait que celui-ci permet, également, de prendre en considération les deux séquences ensembles. De ce fait, il permet de corriger les paires de nucléotides correspondantes avec plus de précision. Il permet, par exemple, de définir les nucléotides indéfinis.

BLAST

BLAST est un outil, proposé par NCBI, permettant l'accomplissement de certaines fonctionnalités directement sur le web. En effet, cet outil peut être utilisé pour obtenir des résultats relatifs aux alignements génétiques.

L'avantage, que possède cet outil, est le fait que l'utilisateur novice puisse réaliser l'étape d'alignement sans nécessairement installer tous les outils dont l'utilisabilité n'est pas toujours aisée.

N.B. : Notons, que nous n'avons pas utilisé cet outil dans nos recherches, mais que nous avons décidé de, tout de même, le citer parce qu'il est largement connu dans le milieu.

R

R est un logiciel statistique qui est très utilisé dans différents milieux scientifiques. R peut, également, décrit être comme étant un langage de programmation fonctionnel utilisé à des fins

[31] D'après la documentation : *fqtrim: trimming & filtering of next gen reads*, http://ccb.jhu.edu/software/fqtrim/, date de consultation: 17/08/2016

statistiques pouvant se prêter correctement dans le traitement d'un grand nombre de données [Torgo 2011 & R Development Core Team, 2008].

Un avantage que possède ce langage est le fait qu'il permette la visualisation des données. Dû à le caractère lié aux probabilités de la bio-informatique, nous comprenons, dès lors la pertinence de son utilisation ce le milieu.

Outre cet avantage, R est, également, régulièrement mis à jour. En effet, de plus en plus de librairies, correspondantes à divers milieux scientifiques (biologie, génétique, etc), existent tout en étant « open source ».

N.B. : Certains IDE, comme « RStudios », existent et permettent d'utiliser le langage plus facilement (ajout de librairies facilité, présence de manuels d'utilisation et de tutoriels, …).

Bioconductor

Bioconductor est un projet open-source proposant une multitude d'outils et de librairie pouvant être utilisées dans le contexte de recherches biologiques.[32] Ceux-ci permettent le traitement d'un large nombre de données biologiques. Le nombre de ces outils ne cesse d'augmenter au fil des années [Morgan, 2016].

Dans le cadre de nos recherches, celui-ci nous a permis d'obtenir les librairies que nous avons utilisées dans R, telles que les librairies graphiques, ou encore les librairies biologiques : DESeq.

Subread

Cet outil permet d'accomplir bon nombre des fonctionnalités de bases, telles que l'alignement, la détermination de l'expressivité génétique, etc. Cet outil utilise, généralement, les types de fichiers FASTA et produisent des fichiers de type SAM ou BAM.

Pour effectuer l'alignement, il utilise un algorithme basé sur la stratégie « seed-and-vote » [Liao Young, Smyth GK and Shi W., 2013], qui est largement connue dans le milieu, et qui consiste à

[32] D'après les sources : *Bioconductor open source software for bioinformatics*, https://www.bioconductor.org/about/, 2016, date de consultation : 20/07/2016

subdiviser la séquence devant être mappée en plusieurs sous-séquences qui sont mappées, individuellement, au le génome de référence à une localisation particulière. La localisation ayant été le plus référencé devient la localisation de la séquence principale sur le génome de référence.

La figure 2.11 représente ce processus.

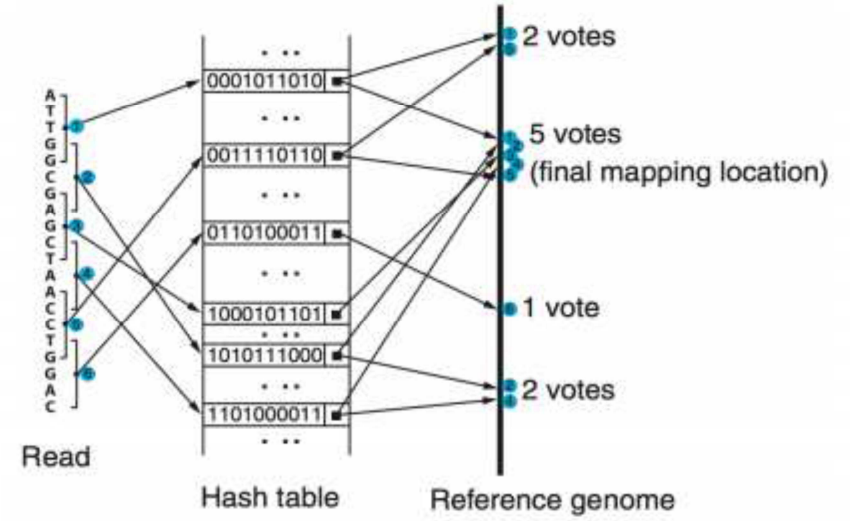

Figure 2.11: Stratégie 'seed-and-vote' [Liao Young, Smyth GK and Shi W., 2013]

Notons que dans le projet, décrit dans la deuxième partie, nous n'avons pas utilisé cet outil pour la réalisation de l'étape de comptage (détermination de l'expressivité génétique). En effet, celui-ci nécessite des connaissances biologiques plus poussées.

N.B. : Une librairie, pour le langage R, équivalente existe (« Rsubread ») qui peut s'avérer être plus rapide dans certains cas.

SAMTools

Cet outil permet d'extraire des informations à partir, comme l'indique son nom, des fichiers de type SAM. Il permet d'utiliser une large gamme de fonctionnalités pouvant être appliquées sur les résultats de l'étape d'alignement.[33]

Il est disponible à partir du web. Son utilisation se fait à partir d'un terminal Unix et requiert, par conséquent, une connaissance de l'utilisation de ce genre de système.

Ses qualités résident dans sa haute flexibilité, utilisable avec des résultats de l'étape d'alignements provenant de multiples programmes, il est également de taille compacte.

Dans le cas de nos recherches, nous l'avons utilisé afin de pouvoir extraire les données à partir des résultats de l'alignement qui seront, ensuite, utilisés.

DESeq

Ce package R est généralement utilisé pour analyser, de manière statistique, les résultats obtenus lors de la détermination de l'expressivité des gènes du génome [Anders, 2016]. Nous pouvons le télécharger librement à partir du dépôt de paquets biologiques R : Bioconductor.

Il a été utilisé afin d'obtenir les résultats obtenus suite à l'étude menée (voir section suivante). En effet, grâce à lui, nous avons pu utiliser les résultats obtenus lors de la réalisation de graphiques permettant aux scientifiques d'aboutir à des conclusions.

[33] D'après les informations proposées par le développeur : *SAMtools*, http://samtools.sourceforge.net/, 2012, date de publication : 20/07/2016

Conclusion

Cette partie aura permis au lecteur de pouvoir poser les fondements de base du domaine de la biologie et de la bio-informatique, tant au point de vue historique, scientifique que technique de l'intérêt lié à ce domaine. Nous espérons, ainsi, pouvoir aider le lecteur n'ayant aucune connaissance en la matière. En effet, cette partie a été créée en vue de résoudre un problème que nous avons rencontré lors de la réalisation du stage.

Notons que les explications, relatives aux notions biologiques et aux outils, constituent, selon nous, la base nécessaire pour débuter dans le milieu.

Notons que ces outils sont basés sur des concepts plus élaborés de machine learning. Etant donné l'objectif du travail consistant à offrir une vision générale des concepts de base et celles-ci suffisent à la compréhension, nous avons préféré nous limiter aux explications générales. Toutefois, nous pouvons recommander, aux lecteurs souhaitant s'intéresser d'avantage à ces outils, d'étudier les concepts généraux liés au machine learning. Cela leurs permettra de mieux comprendre le fonctionnement des algorithmes de ces outils.

D'un point de vue global, ces fondements nous permettront d'énoncer les faits qui constitueront la base de notre réponse que nous proposerons afin de répondre à la question de recherche. En effet, les deux sections suivantes auront pour but de démontrer la pertinence quant à l'utilisation d'un pipeline d'extraction de données en se basant sur ces faits théoriques.

Deuxième Partie : Le pipeline d'extraction de données

Cette partie, a pour but, de présenter l'outil d'un point de vue conceptuel. Ceci est important dans le cheminement du raisonnement puisque nous devons, comme présenté dans l'introduction, présenter l'outil afin de pouvoir, ensuite, vérifier sa pertinence dans le milieu médical.

Un exemple d'étude pouvant être menée en utilisant cet outil sera, ensuite, présentée. Celle-ci permettra au lecteur de prendre un premier contact avec les résultats pouvant être obtenus grâce à l'utilisation de ce genre d'outils.

A terme de cette partie, nous serons, donc, en mesure de traiter du sujet principal de ce travail, à savoir l'application de cet outil au domaine médical.

Chapitre 3 : Présentation du pipeline

Le contexte, décrit dans la partie précédente, permet de faire apparaître le besoin de l'acquisition de résultats en vue d'aboutir à des conclusions de manière rapide et efficace. Pour ce faire, l'élaboration d'un outil, un « pipeline d'extraction de données », sera étudié.

Cette section a donc pour but de décrire sa structure, en étudiant chaque composant, pour ensuite, dans la seconde partie, démontrer sa nécessité et sa pertinence grâce à un cas concret d'utilisation.

Notons que le but de cette section, malgré la présentation d'un cas concret, reste principalement axé sur l'outil en lui-même et non pas sur l'exemple d'utilisation et des résultats y émanant. Nous n'assurons, par conséquent, la véracité des résultats obtenus

Mise en contexte

Dans un contexte d'étude, au niveau transgénique, des interactions et activités génétiques, nous jugeons utile l'existence d'un outil permettant l'extraction de données à partir d'un problème biologique concret. Ces données permettant l'élaboration de conclusions.

Les études transgéniques se font de plus en plus connaître dans différents milieux. En effet, plusieurs études, portant sur ce domaine, ont déjà été menées lors des dernières années. Celles-ci portent souvent sur les conséquences génétiques que possède une hormone sur un individu.

Ces études nous permettent de comprendre l'intérêt que constitue ce domaine d'activités, ainsi que celui concernant l'élaboration d'un outil capable de générer les résultats attendus.

Cette section décrira, la structure d'un pipeline, pour ensuite en présenter un exemple simple d'utilisation.

Le pipeline

Dans le but de pouvoir aider les scientifiques dans leurs recherches, nous pouvons élaborer un outil, appelé « pipeline d'extraction de données », afin de pouvoir extraire des informations statistiques à partir des séquences, ou données génétiques, d'un individu. Son but principal est donc de fournir des informations complémentaires pouvant aider les utilisateurs (scientifiques) dans l'élaboration de conclusions.

Un tel outil se compose généralement de plusieurs composants, entre lesquels, nous ferons transiter les données et différents résultats. Ces composants apparaissent dans le besoin de répondre aux exigences et problèmes biologiques contenus dans les séquences de données. Ces composants sont les suivants : évaluation des séquences de donnés, ajustement de celles-ci, l'étape d'alignement sur le génome de référence, l'évaluation de l'expressivité du génome, la normalisation et la production de graphique.

La figure 3.1, ci-dessous, représente de manière schématique, les différentes étapes et processus de cet outil.

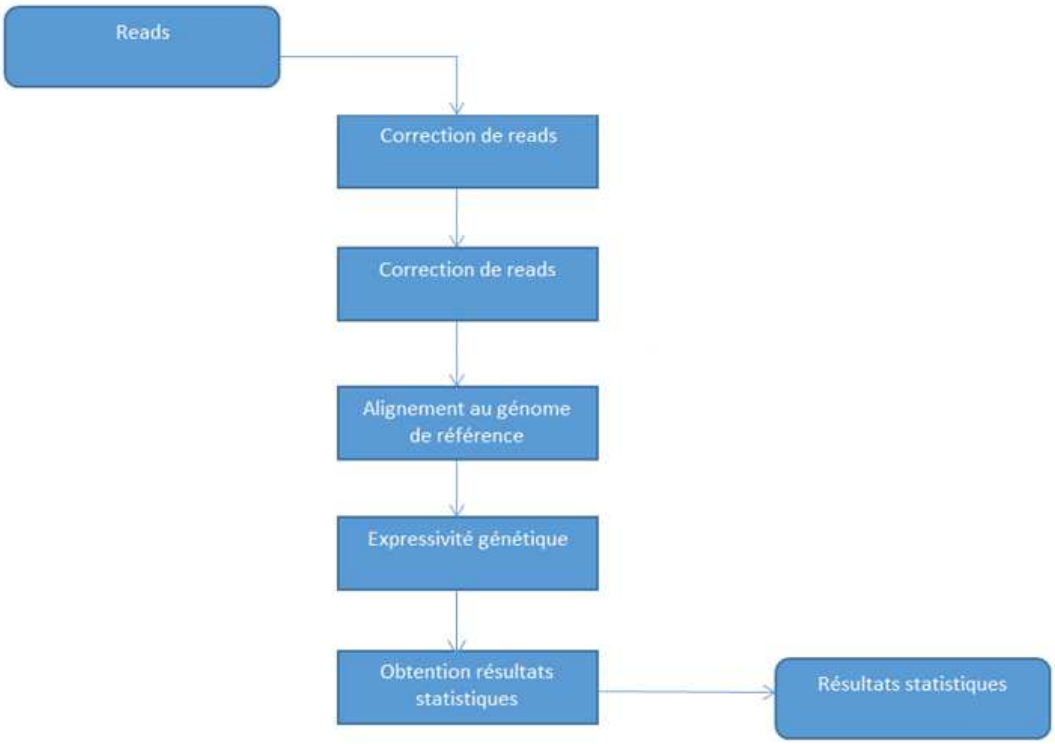

Figure 3.1: Pipeline workflow

Cette figure s'inspire fortement de celle qui est fournie par A. Oshlack, M. D. Robinson, M. D. Young[34].

Cette section aura, donc, pour but de décortiquer, de manière théorique, chacune de ces étapes. La section suivante présentera, ensuite, les résultats obtenus lors de l'étude d'un projet de recherche.

[34] Alicia Oshlack, Mark D Robinson, Matthew D Young, *From RNA-seq reads to differential expression results*, Genome Biology 2010, p. 2, BioMed Central, 2010

Evaluation des séquences de données

Comme dit, précédemment, les séquences de données utilisées pour l'expérience peuvent, soit, être extraites à partir des organismes, soit, prises à partir de librairies des centres de recherches. Ces séquences, provenant des deux types de sources, peuvent contenir des erreurs. Par exemple, il se peut que certains nucléotides ne contiennent de valeur (nucléotide « N »). Notons qu'il existe également d'autres points suscitant notre intérêt.

La figure 3.2 représente un extrait de séquence d'ARN contenant des nucléotides indéfinis.

```
@HISEQ1:102:D099AACXX:1:1207:4630:182147 1:N:0:TGACCA
TACAGNCATAGGGAACTTCTCANCTTGGTTTNCTCCNGCNAAC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJIJJIJJJIIJJJJ
```

Figure 3.2: séquence ARN avec nucléotides indéfinis

Afin de résoudre ce problème, nous sommes dans l'obligation d'utiliser des outils d'évaluation. Par exemple, nous pouvons utiliser l'outil « FastQC » qui nous permet de vérifier la pertinence des séquences fournies en entrée en proposant des informations sur les séquences.

Cet outil permet, grâce à cette panoplie d'informations, aux scientifiques de juger la pertinence des séquences afin de leurs permettre de les corriger de manière adéquate.

Ajustement des données

Sur base des résultats obtenus lors de l'évaluation des séquences, nous pouvons réajuster celles-ci grâce à des outils correctifs. Dans le cas de notre pipeline, nous avons considéré l'utilisation de l'outil « FQTrim » qui permet de corriger les séquences présentées par paires. Cet outil est utilisable à partir d'un terminal.

Concrètement, nous utilisons cet outil en lui fournissant deux fichiers de séquences en entrée en y activant certaines options. Les résultats obtenus sont deux autres fichiers de séquences contenant les éventuels ajustements.

Alignement sur le génome de référence

L'alignement constitue, généralement, la première étape du traitement des données. En effet, jusqu'à maintenant, les étapes précédentes n'auront servies qu'à vérifier et corriger les données. L'alignement, comme expliqué précédemment, a pour but de définir les régions, du génome de référence, représentées par les échantillons. Ceci permettra de faire ressortir les gènes mis en évidence plus tard dans le processus global.

Avant toute chose, nous devons nous occuper du génome de référence. En ce qui concerne ce génome, nous pouvons, le télécharger à partir d »une banque de données. Celui-ci sera représenté au format « Fasta » (ou « faa »). A partir de celui-ci, nous devons créer un index afin de pouvoir améliorer les performances des algorithmes suivants.

Cette étape est réalisée grâce à l'outil « Subread » depuis un terminal. Nous pouvons, toutefois, utiliser la librairie « Rsubread » de R pour réaliser cette étape, si nous préférons utiliser les résultats dans l'environnement du langage R. En effet, l'option « build-index » de cet outil permet la création de cet index.

En ce qui concerne les performances, nous conseillons d'utiliser « Rsubread » à la place de « Subread ». En effet, la version adaptée au logiciel R réalise les opérations plus rapidement.

Ensuite, nous pouvons nous préoccuper de l'étape d'alignement, en utilisant l'option « align » de cet outil. Notons que nous pouvons activer des options permettant, par exemple, de gérer deux fichiers contenants les deux séquences de nucléotides, ainsi que celle qui indique à l'outil que l'on manipule des séquences d'ARN. Les options peuvent être consultées dans le guide de l'utilisateur.[35]

Les résultats obtenus sont sous la forme de fichiers « SAM » et permettent de faire référence entre les sections.

Pour plus de détails, nous encourageons le lecteur à relire la partie consacrée aux traitements des données présentée dans la deuxième partie.

N.B : En plus des résultats obtenus (au le format « SAM »), nous pouvons, également obtenir un pourcentage d'alignement.

[35] Publication: http://bioinf.wehi.edu.au/subread-package/SubreadUsersGuide.pdf

Evaluation de l'expressivité du génome

Une fois l'alignement fait, nous pouvons réutiliser les fichiers résultats produits afin d'en extraire les données relatives à l'expressivité du génome de référence. Ce processus permettra de savoir combien de gènes du génome de références sont utilisés. En d'autres termes, cette étape consiste à aller plus loin que l'étape d'alignement, mais en nous focalisant sur les gènes du génome de référence à la place des séquences de l'échantillon traité.

En effet, cette étape permet de vérifier les gènes du génome de référence utilisé. Ce fait nous permet, sur base des propriétés biologiques expliquées dans la première partie, d'analyser le comportement que pourrait obtenir la cellule.

Nous pensons, donc, que cette étape est primordiale dans notre cas puisque le pipeline sert, dans de nombreux cas, à comparer le comportement génétique de plusieurs échantillons.

Concernant l'aspect technique, dans le cas de notre pipeline, nous avons utilisé l'outil « Samtools ». Cet outil peut facilement être utilisé à partir d'un terminal.

Pour plus d'informations concernant cette étape, nous invitons le lecteur à consulter la section concernant le traitement des données présentée dans la première partie du travail.

Les résultats obtenus sont présentés à la section suivante.

La normalisation

Arrivé à ce stade du processus, il nous est de pouvoir comparer les différents résultats obtenus et ce, dans le but d'apporter les premières conclusions à l'étude en cours.

Du à la grande diversité de ceux-ci, il peut être souhaitable, voir recommandé, d'instaurer une étape de normalisation dans le processus afin d'améliorer l'adaptation des données. En effet, il se peut que les données ne possèdent pas les mêmes caractéristiques et ne soient, par conséquent, aptes à la comparaison. Une préparation des données est alors nécessaire.

Des graphiques contenant les différentes données normalisées est présentée dans la section suivante et permettent de prendre conscience de l'impact que cette opération a sur les données.

La production de graphiques

A partir des résultats obtenus précédemment, il nous est possible de construire des graphiques afin de pouvoir augmenter le caractère visuel des données, aidant, ainsi, l'élaboration de conclusions.

Parmi ces techniques de visualisations, nous avons, dans un premier temps, les graphiques plus communs, tels que les boxplots, et dans un second temps, les graphiques plus spécifiques au milieu biologique, comme les graphiques phylogénétiques.

Les premiers graphiques peuvent être obtenus grâce à l'acquisition de résultats quantitatifs, lors de l'exécution des étapes du pipeline, et grâce à leur utilisation dans R. En effet, le logiciel R, disposant, déjà de librairie permettant la représentation graphique des données, peut directement être utilisé à cette fin.

En ce qui concerne l'autre type de graphiques, les outils utilisés sont uniquement utilisables dans R et ne se limitent qu'à des librairies de ce langage, telles que : « DESeq », « RColorBrewr », « gplots », etc.

Ces librairies nous permettront d'obtenir les graphiques biologiques liés à l'expérience et nous permettront, ainsi, de nous rendre compte des produits pouvant être obtenus grâce au pipeline.

Exemple

Un exemple d'étude transgénique pourrait être l'observation des comportements, au niveau génétique, des plantes lors de leur exposition à la lumière du soleil et à celles provenant des lampes UV. Nous pourrions, dès lors, observer les interactions transgéniques pendant la photosynthèse, par exemple. En effet, nous disposerions de plusieurs échantillons par type d'exposition de lumière. Chaque échantillon serait analysé d'un point de vue génétique grâce au pipeline. Ces analyses nous permettraient, par exemple, de vérifier la production de protéines créées lors de la photosynthèse. Selon nous, le taux de ces protéines serait plus important dans le cas d'une exposition à la lumière naturelle.

Cet exemple permet au lecteur, dès lors, de comprendre quel genre d'études peut être lié à ce domaine.

Chapitre 4 : Application (étude de cas) : Exemple de l'étude liée à l'application de l'hormone de Gibbérelline sur le raisin

Maintenant que nous avons donné des explications théoriques concernant l'outil, nous allons compléter ces informations en présentant une expérience scientifique ayant réellement été menée. Celle-ci nous permettra d'exposer le genre de résultats pouvant être obtenus à partir de séquences de données en y appliquant le pipeline.

Nous commencerons donc, par une présentation du projet, pour ensuite décrire, la manière dont nous avons adapté les différentes étapes du pipeline au contexte, pour ensuite, présenter les résultats obtenus à chaque étape du pipeline.

Notons que les outils utilisés pour la réalisation des différentes étapes, présentées à la figure 3.1, ce pipeline sont dépendants de notre choix. En effet de nombreux outils existent et pourraient remplacer ceux utilisés ci-après [Oshlack, Robinson, Young, 2010]. Nous justifions notre choix par le fait que ces outils nous ont été recommandés par les experts.

Présentation du projet

Ce projet, conduit par les Professeurs C. Moser et S. Pilati, a été créé afin d'essayer de répondre à un des problèmes actuels rencontré dans la production agricole, à savoir : l'optimisation du rendement des récoltes. En effet, ce projet permet à des biologistes d'étudier le comportement génétique des plants de raisins (L. Vinifiera) traités avec l'hormone de Gibbérelline afin de pouvoir déterminer une utilisation permettant d'obtenir un rendement optimal dans les récoltes de raisins.

Afin de mener l'étude bio-informatique à bien, le génome de référence du raisin ainsi que les différentes séquences d'ARN correspondantes aux échantillons correspondantes à des situations nous ont été mis à disposition par l'équipe de biologistes. Parmi celles-ci, nous avons considéré l'étude de deux types de vignes de raisins (le Pinot gris et le Sauvignon blanc), traités de manière hormonale ou non, différents moments de récolte des fleurs (4 ou 24 heures après utilisation de l'hormone). Ces situations sont importantes puisque celles-ci seront étudiées, d'un point de vue génétique, afin d'en tirer des conclusions.

Concrètement, nous avons reçu ces échantillons d'ARN sous forme de fichiers « fastq » ne contenant qu'un seul brin chacun. Des fichiers devaient, donc, être mis en relation afin de reformer les paires de séquences d'ARN.

Nous commencerons, donc, par une description des informations génétiques liées à l'hormone de Gibbérelline ainsi que celles de l'hormone. Ensuite, nous décrirons la manière dont nous avons utilisé le pipeline, pour finir par les résultats obtenus.

Présentation génétiques

Comme nous l'aurons compris, le concept d'hormone constitue la base du raisonnement dispensé dans l'explication de l'étude d'exemple proposée dans cette section. Celle-ci aura, donc, pour but de permettre au lecteur d'en savoir plus sur la notion d'hormones ainsi que sur les aspects génétiques de celle de Gibbérelline ainsi que des plants de raisins utilisés dans l'expérience. De brèves explications relatives au génome du raisin, « Vitis viniefera », viendront compléter la connaissance du lecteur expliquant, ainsi, la structure du génome ainsi quelques caractéristiques propres.

Tout d'abord, une hormone peut être définie, d'après le Larousse, comme étant, une substance secrétée par une glande permettant de modifier le fonctionnement d'un organe cible [36]. Cette substance est créée par l'individu ou la plante hôte, par des organes, en fonction du taux déjà présent dans l'organisme [Tostain, Rossi, Martin, 2004], ce qui aura une répercussion sur le développement de l'individu hôte [Granell, Carbonell, 1996].

Par exemple, prenons le cas de la testostérone, l'hormone masculine humaine, est produite par les testicules, sous le contrôle de l'hypophyse, tout au long de la vie de l'individu et se régule automatiquement en fonction du taux présent dans l'organisme [Tostain, Rossi, Martin, 2004]. En effet, au plus ce taux est élève, au plus la production de cette hormone sera réduite. Cette hormone agit sur différents organes du corps humain en appliquant ses effets sont liés au phénotype masculin de l'individu : apparition de barbe, musculature, etc. Le principe présenté dans cet exemple est valable pour n'importe quelle hormone.

[36] Définition obtenue à l'adresse suivante: http://www.larousse.fr/encyclopedie/divers/hormone/185887

Maintenant que nous en connaissons plus sur le concept d'hormone, penchons-nous sur l'hormone de Gibbérelline constituant une famille d'hormones généralement produite dans les graines des organismes végétaux.

Cette famille hormonale fait l'objet d'une multitude d'études, notamment concernant ses bienfaits et méfaits liés à son utilisation[37][38][39]. Les quelques études parcourues, dans le cadre de ce travail, nous ont permis de comprendre que cette hormone possèdent des effets sur le développement et l'apparence du fruit, l'aspect physiologique de la plante ainsi que d'autres aspects.

La figure 4.1 suivante présente la disposition phylogénétique de celle-ci.

---

[37] Chan Jin Jung, Youn Young Hur, Sung-Min Jung, Jung-Ho Noh, Gyung-Ran Do, Seo-June Park, Jong-Chul Nam, Kyo-Sun Park, Hae-Sung Hwang, Doil Choi, Hee Jae Lee, *Transcriptional changes of gibberellin oxidase genes in grapevines with or without gibberellin application during inflorescence development,* The Botanical Society of Japan and Springer Japan, p.359-371, 2013

[38] Atiako Kwame Acheampong, Jianhong Hu, Ariel Rotman, Chuanlin Zheng, Tamar Halaly, Yumiko Takebayashi, Yusuke Jikumaru, Yuji Kamiya, Amnon Lichter, Tai-Ping Sun and Etti Or, *Functional characterization and developmental expression profiling of gibberellin signalling components in Vitis vinifera*, Journal of Experimental Botany, 2014

[39] Lisa Giacomelli, Omar Rota-Stabelli, Domenico Masuero, Atiako Kwame Acheampong, Marco Moretto, Lorenzo Caputi, Urska Vrhovsek and Claudio Moser, *Gibberellin metabolism in Vitis vinifera L. during bloom and fruit-set: functional characterization and evolution of grapevine gibberellin oxidase*, Journal of Experimental Botany, 2013

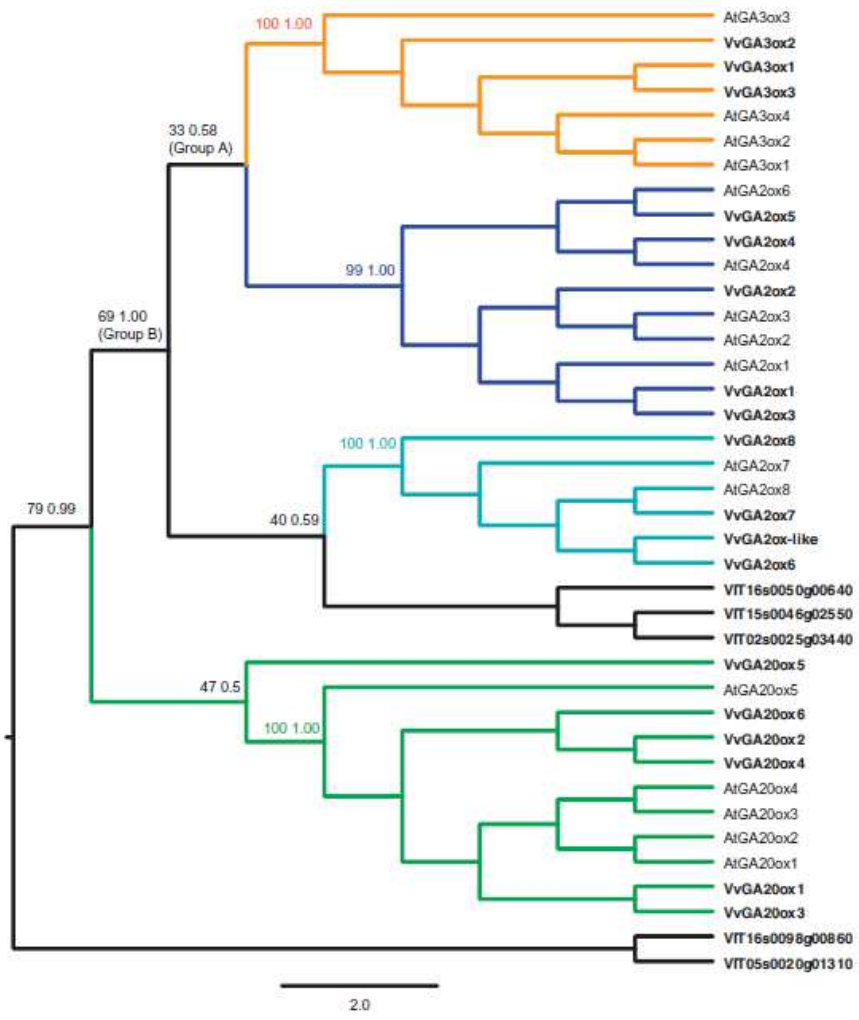

Figure 4.1 : diagramme phylogénétique de l'hormone de Gibbérelline[40]

Sans rentrer dans les détails, sur base de cette figure, nous constatons que l'hormone de Gibbérelline se dérive en protéines utilisées par les plants de raisins (« Vitis Vinifera »). Nous comprenons, dès lors, en quoi cette hormone est utile pour le développement végétal (protéines dont le nom commence par « VIT »).

[40] Lisa Giacomelli, Omar Rota-Stabelli, Domenico Masuero, Atiako Kwame Acheampong, Marco Moretto, Lorenzo Caputi, Urska Vrhovsek and Claudio Moser, *Gibberellin metabolism in Vitis vinifera L. during bloom and fruit-set: functional characterization and evolution of grapevine gibberellin oxidase*, p 5, Journal of Experimental Botany, 2013

Nous notons, à titre d'information, qu'il n'existe pas moins de 20 sortes de molécules de Gibbérelline. Cependant, seuls certaines de ces molécules sont couramment rencontrées et produites.[8]

Ces explications permettent au lecteur de comprendre l'importance que possède cette hormone dans ce domaine d'activité et donc la pertinence que possèdent les études.

En ce qui concerne le plant de raisin, le génome de référence est le celui de la vigne, appelé « Vitis Vinifera ». Son matériel génétique est codé sur 19 chromosomes qui permettent, comme dit précédemment, de générer ses traits particuliers.

Utilisation du pipeline

Cette section a pour but d'expliquer l'application du pipeline, décrit de manière théorique dans la section précédente, dans le contexte du projet actuel.

Nous allons, donc, expliquer concrètement, comment les étapes ont été performées. En effet, nous donnerons des détails concernant les données insérées en entrées et celles obtenues en résultats. Ceci permettra, nous l'espérons, au lecteur de pouvoir, grâce à un exemple concret, de mieux comprendre l'utilisation de l'outil en pouvoir observer le genre de résultats pouvant être obtenu à chaque étape.

Notons de notifier que les échantillons ont été séparés en fonction de leur sorte de raisin. Nous avons, donc, exécuté chaque étape du pipeline en considérant la sorte, et donc, par conséquent, groupé tous les échantillons. Ce fait explique la raison de la présence de deux ensembles de résultats, une concernant le Pinot Gris, et l'autre, le Sauvignon Blanc.

Evaluation des séquences de donnés

FastQC est utilisé en y insérant les séquences de données. Notons que, dans cet outil, il n'est pas possible de prendre en considération les séquences par paires. Par exemple, nous utilisons les deux séquences de données suivante : "pc24-1_TGACCA_L001_R1_001.fastq.gz"; afin d'obtenir les informations relatives à son niveau de qualité de séquences. Précisons que le format utilisé est « fastq » et qu'il permet de représentes la séquence « brutes » (sans modifications).

Comme dit précédemment, plusieurs critères sont à prendre en compte afin d'évaluer la pertinence des séquences. En effet, nous devons, par exemple, tenir compte du taux de présence de nucléotides indéfinis, ou encore, de taux de pertinence des résultats. Un bilan récapitulatif est fourni par cet outil.

La figure 4.2 présente le bilan concernant les résultats de l'analyse en tenant compte de caractéristiques générales pouvant être étudiées avec l'outil.

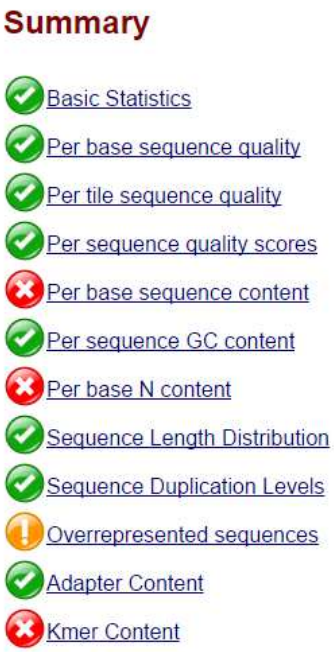

Figure 4.2: Bilan général

Comme le suggère cette figure, le programme permet d'indiquer et de conseiller les scientifiques sur les éléments nécessitant une attention particulière (pictogrammes d'indications).

La figure 4.3, quant à elle, représente un exemple de graphique pouvant être obtenu. Cette figure présente, donc, le taux de présence de nucléotides indéfinis, tandis que la figure 4.4 représente la pertinence des résultats contenus dans ces séquences.

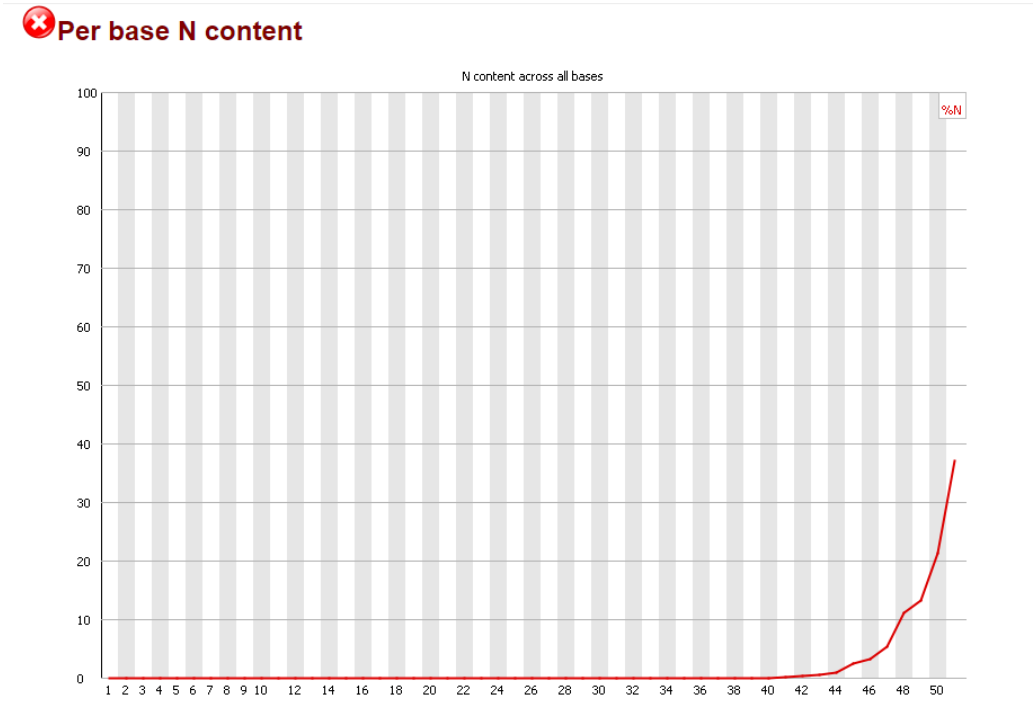

Figure 4.3: Taux de présence N

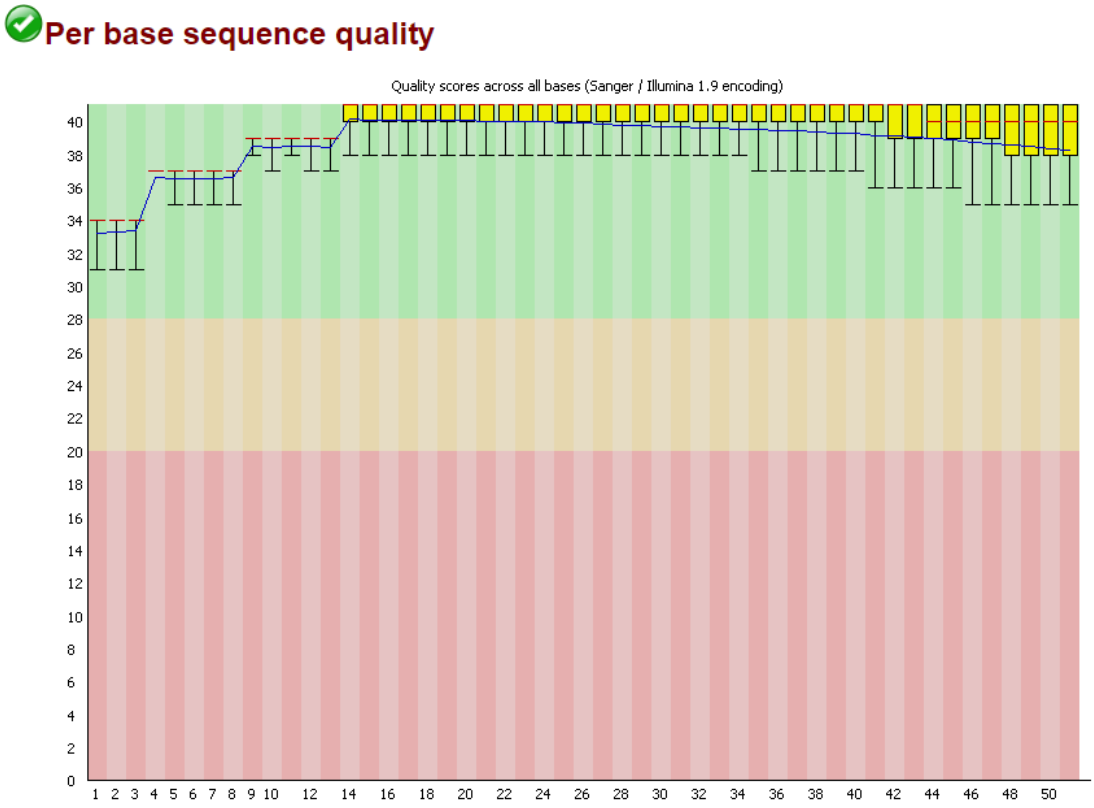

Figure 4.4: Niveau qualité des séquences

Par soucis de synthèse, nous avons décidé de présenter, dans cette section, que les résultats obtenus avec les deux échantillons cités plus haut et en ne prenant en compte que ces deux critères. Les autres échantillons présentent des résultats similaires, leur étude est, alors, jugée comme n'étant pas pertinente.

Ajustement des séquences

Comme dit précédemment dans la présentation théorique, nous avons utilisé l'outil de correction de séquences « FQTrim » en y insérant les séquences par paires ("pc24-1_TGACCA_L001_R1_001.fastq.gz" et "pc24-1_TGACCA_L001_R2_001.fastq.gz"). Nous avons, également, activé les options associées à la longueur des séquences (25 bases) et permis la gestion de séquences contenant une suite de bases d'adénine (« poly-A tail », voir première partie).

Les résultats obtenus étaient sous la forme de fichiers de séquences adaptées pouvant être utilisés pour la suite du pipeline.

Par exemple, les figures 4.5 et 4.6 présentent les corrections pouvant être apportées à des séquences. Celles-ci proviennent des échantillons mis à disposition.

```
@HISEQ1:102:D099AACXX:1:1207:4630:182147 1:N:0:TGACCA
TACAGNCATAGGGAACTTCTCANCTTGGTTTNCTCCNGCNAAC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJIJJIJJJIIJJJJ
```

Figure 4.5 : séquences initiales

```
@HISEQ1:102:D099AACXX:1:1207:4630:182147 1:N:0:TGACCA
TACAGACATAGGGAACTTCTCATCTTGGTTTCCTCCGGCAAAC
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJIJJIJJJIIJJJJ
```

Figure 4.6 : Séquences corrigées

Grâce à ces deux figures, nous pouvons constater que certaines paires de nucléotides ont été adaptées, par rapport à la séquence mise en relation. Par exemple, certaines paires de nucléotides ont été adaptées sur base des règles biologiques de correspondance entre nucléotides et des nucléotides non définis, « N », le sont devenu.

Alignement sur le génome de référence

Comme dit précédemment, afin de pouvoir mener à bien l'expérience, nous avons dû faire la distinction entre les échantillons sur base de leur origine. En effet, nous les avons dû séparer en deux groupes respectivement en fonction de leur sorte de raisin. Le but étant de comparer les deux espèces de raisin, nous avons aligné les échantillons de chaque ensemble sur le génome de référence.

Avant de nous concentrer sur l'étape d'alignement, nous avons, en premier lieu, créer un index. Pour ce faire faire, nous avons utilisé

La figure 4.7 présente, quant à elle, les données quantitatives présentant les taux d'alignement possible par échantillons de chaque ensemble.

Samples	PC24-1	Pc24-2	Pc24-3	Pc4-1	Pc4-2	Pc4-3	Pt24-3	Pt4-2	Pt4-3
Average percentage of reads mapped	71.6	66.56	61.02	49.82	60.3	59.275	58.8	60.4	60.3

Samples	Sc24-1	Sc24-2	Sc24-3	Sc4-1	Sc4-2	Sc4-3	St24-1	St24-2	St24-3	St4-1	St4-2	St4-3
Average percentage of reads mapped	61,4	47,7	62	61,9	62,3	62,2	60,2	52,1	61,1	61,2	59,2	60,3

Figure 4.7: Résultats moyens alignement

Ces tableaux permettent de présenter, pour chaque échantillon, le pourcentage de « reads », séquences d'ARN uniquement composées d'exons, alignés au génome de référence. Les échantillons sont triés sur base de leur origine (Pinot gris : premier tableau ; Sauvignon blanc second tableau), du temps écoulé entre la pulvérisation et la récolte des feuilles ainsi que d'une position sur le génome (appelée « réplication biologique »).

Ces résultats nous permettent de comprendre que le taux d'alignement atteint son minimum dans le cas du Sauvignon blanc, non traité et récolté 24haprès le traitement général ; tandis que, le pourcentage de reads alignés atteint son maximum dans les échantillons de Pinot gris. Les trois

premiers échantillons du Sauvignon (les trois réplications) nous permettent de remarquer une variation plus importante dans les résultats du raisin étant en provenance de cette origine.

N.B. : un « read » est une sous-séquence d'ARN composée que d'extron (voir première partie sur la composition de l'ARN pour plus d'informations).

Evaluation de l'expressivité du génome

Dans cette section, nous allons nous pencher sur le problème de l'expressivité génétique du gène de référence. Pour ce faire, nous allons utiliser les résultats obtenus précédemment pour pouvoir en extraire les informations nécessaires.

Comme expliqué, dans la section précédente, cette étape permet de nous concentrer sur le génome de référence en obtenant tous les gènes étant référencés lors de l'étape d'alignement d'un échantillon sur le génome de référence. Ceci nous permet, au final, de connaître les gènes représentés par l'échantillon ainsi que leur « degré » d'expressivité.

D'un point de vue technique, nous avons utilisé le Samtool à partir des résultats obtenus à l'étape d'alignement.

Il est à noter qu'un autre outil, comme Subread, avec son outil « featureCounts », aurait, également, pu être utilisé. Cependant, cet outil nous a causé quelques problèmes. En effet, afin de pouvoir réaliser cette étape, nous aurions dû fournir un fichier d'annotations, au format « GTF », permettant de faire le lien entre le nom des chromosomes dans les résultats, obtenus lors de l'étape d'alignement, ainsi que le génome de référence. Or, par manque de connaissances dans le milieu, il nous était compliqué de réaliser cette étape de cette manière. C'est la raison pour laquelle, nous avons utilisé l'outil « Samtools » qui ne nécessité pas ce genre de fichiers.

Face à cette difficulté, l'utilisation d'un autre outil était nécessaire. Celui-ci est, comme vous l'aurez compris, « Samtool » qui nous a permis d'extraire les résultats relatifs à l'expressivité génétique directement à partir des résultats de la première étape.

Concrètement, nous avons repris les résultats de la première étape (alignement) sous forme de fichier « SAM ». Or, Samtools, ne sachant utiliser que les fichiers « BAM », nous a imposé une contrainte. Pour y remédier, nous avons dû convertir ce fichier « SAM » en « BAM » à partir duquel, nous avons pu extraire les données concernant les comptes.

Concernant ces résultats, la figure 4.8 présente le genre de données que nous pouvons obtenir grâce à cette étape, dans le contexte des échantillons relatifs au Pinot gris.

gene	pc24-1	pc24-2	pc24-3	pc4-1	pc4_2	pc4_3	pt24-3	pt4-2	pt4-3
VIT_201s0011g00010.1	41	24	30	26	32	28	38	35	27
VIT_201s0011g00030.1	77	40	70	68	60	46	82	101	85
VIT_201s0011g00040.1	4	7	3	7	3	8	2	9	6
VIT_201s0011g00050.3	1	0	0	0	0	0	3	0	0
VIT_201s0011g00050.1	7	14	15	13	13	9	76	12	13
VIT_201s0011g00050.2	45	33	10	12	13	9	29	10	6
VIT_201s0011g00060.1	16	14	26	19	17	21	39	21	19
VIT_201s0011g00070.5	1	5	9	7	6	6	16	12	3
VIT_201s0011g00070.4	16	13	1	0	2	0	0	1	0
VIT_201s0011g00070.1	148	79	66	57	56	43	91	59	53
VIT_201s0011g00070.2	2	0	0	1	1	2	0	0	0
VIT_201s0011g00070.3	28	9	5	8	6	5	20	4	6
VIT_201s0011g00080.1	2	1	5	2	2	8	4	1	8
VIT_201s0011g00080.2	0	0	0	2	0	0	1	0	0
VIT_201s0011g00080.3	4	0	0	1	1	1	0	0	0
VIT_201s0011g00090.1	1	0	1	0	0	1	0	0	0
VIT_201s0011g00100.1	228	284	363	252	215	190	457	344	200
VIT_201s0011g00110.1	74	111	134	146	116	119	81	178	148
VIT_201s0011g00110.2	38	22	16	14	12	10	45	14	14
VIT_201s0011g00110.3	89	47	31	35	26	24	62	31	31
VIT_201s0011g00120.1	48	146	121	132	93	86	197	150	94
VIT_201s0011g00130.1	6	26	35	23	24	26	55	31	25
VIT_201s0011g00130.2	65	53	35	36	34	28	47	35	25
VIT_201s0011g00140.1	176	31	69	83	82	93	52	134	111
VIT_201s0011g00150.1	30	28	54	45	26	37	109	36	40
VIT_201s0011g00160.1	835	1003	1157	988	856	939	1077	1479	1057
VIT_201s0011g00170.1	3	5	7	7	4	3	2	9	3
VIT_201s0011g00190.1	1	9	0	0	3	1	0	0	1
VIT_201s0011g00210.1	88	119	152	98	90	103	249	125	86
VIT_201s0011g00230.5	4	5	5	4	1	7	10	3	4
VIT_201s0011g00230.6	17	11	11	14	9	10	10	7	10
VIT_201s0011g00230.3	26	10	13	11	5	6	10	7	11
VIT_201s0011g00230.4	1	0	0	0	0	0	2	0	0
VIT_201s0011g00230.1	1	1	1	0	1	0	0	0	0
VIT_201s0011g00230.2	0	0	0	0	0	0	1	0	0

Figure 4.8: Comptes expressivité génétique Pinot gris

Ces résultats sont obtenus en appliquant la deuxième étape du pipeline aux différents échantillons disponibles. Comme nous pouvons le remarquer, cette disposition permet incite la comparaison de ces échantillons. Outra l'aspect de comparaison, cette étape permet, également, de vérifier quels gènes rentrent en compte dans les différents échantillons. Rappelons, sur base des concepts biologiques décrits dans la première partie, qu'un ensemble de gènes peuvent produire l'apparition d'hormones.

Nous comprenons, dès lors, l'importance que possède cette étape dans le processus d'extraction de données.

Afin de rendre les valeurs, obtenues précédemment, plus intuitives, nous pouvons, également, réaliser des graphiques afin de les présenter d'une manière plus visuelle. Ceci peut être réalisé grâce à l'obtention de graphiques de types boxplots qui permettent de présenter plusieurs types de données, tels que : la moyenne, la variance et bien d'autres informations pertinentes [Potter, 2006].

Les graphiques 4.9 et 4.10 représentent le graphique boxplots correspondant au Pinot et Sauvignon respectivement.

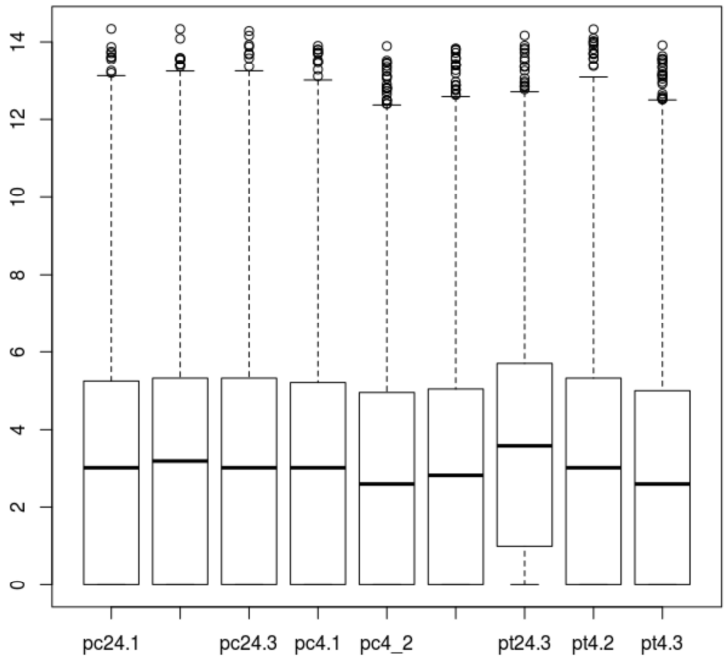

Figure 4.9: Graphique boxplots Pinot gris

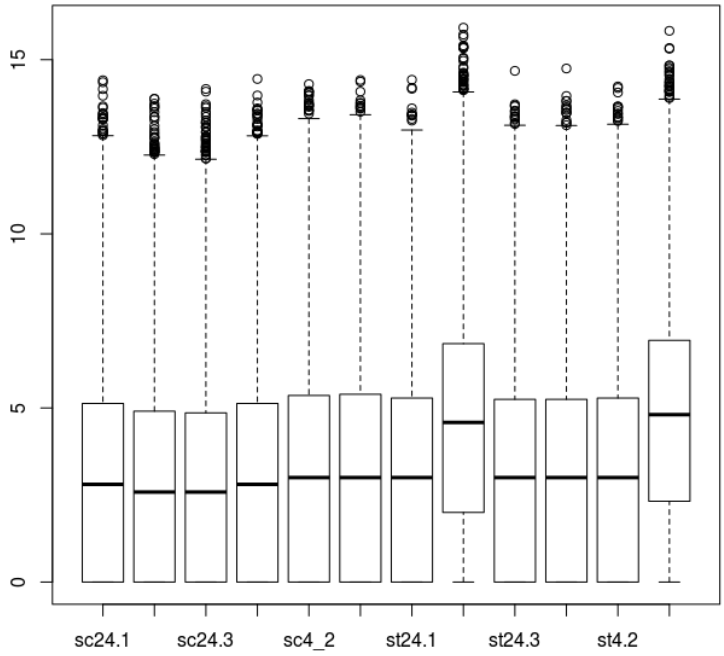

Figure 4.10: Graphique boxplots Sauvignon blanc

Ces résultats peuvent être expliqués de la manière suivante : sur l'axe de l'abscisse représente les échantillons et l'axe des ordonnées, le nombre de reads. Plus précisément, la première lettre désigne l'origine du raisin, la seconde, le fait que l'échantillon ait été traité ou non avec l'hormone, le nombre suivant désigne, quant à lui, le nombre d'heures entre le moment de la pulvérisation et celui où on a prélevé les échantillons. Concernant les données représentées, nous constatons que type de graphiques nous permet de tirer des conclusions sur la distribution des données relatives à l'étape d'alignement par rapport à chaque gène. que représente l'échantillon par rapport au génome de référence. Par exemple, nous pouvons constater que, dans certains échantillons, la médiane est supérieure à celle des autres échantillons.

Notons que nous n'avons pas les connaissances nécessaires pour élaborer des conclusions pertinentes. Cependant, grâce à ces premiers résultats nous pouvons déjà remarquer et, peut-être, confirmer, à ce stade, l'hypothèse selon laquelle l'hormone de Gibbérelline serait plus facilement assimilée par les échantillons provenant de plants de Sauvignon blanc.
Ceci pourrait être réellement observé puisque les échantillons de ce type de raisin, ayant été traités, par l'hormone présentent de meilleurs résultats d'alignement. En effet, sur base des figures 3.12 et

3.13, nous constatons que les différences entre les échantillons et non traités sont plus significatives dans le cas du Sauvignon que dans celui du Pinot (la distribution varie plus fortement). Ce fait implique que l'hormone permet de mapper plus de reads (séquences d'ARN) sur le génome de référence.
Malgré le fait que cette conclusion ne soit pas approuvée par des experts, nous pouvons, tout de même, observer le genre de raisonnement pouvant être réalisé sur base de ces résultats.

Ces premiers résultats permettent, à ce stade, d'illustrer le genre de données et de conclusions que nous pouvons extraire à partir des échantillons. Malgré le fait que nous ne disposons pas des connaissances nécessaires à l'élaboration de conclusions constructives, nous pouvons, néanmoins, affirmer que ce pipeline fournit des résultats pertinents. Ceux-ci, étant considérés comme « pertinent » par les experts, nous permettent de donner plus de crédibilité à l'utilisation de l'outil dans ce domaine.

N.B. : un « read » est une sous-séquence d'ARN composée que d'extrons (voir première partie sur la composition de l'ARN pour plus d'informations).

La normalisation

Les résultats ayant été obtenus à ce stade ne suffisent pas aux scientifiques d'apporter une réponse pertinente à leurs recherches. En effet, comme dit précédemment, la large présence de caractéristiques différentes entre les échantillons ne permet pas de pouvoir les comparer correctement. Par exemple, il se peut que la taille des transcripts représentant les gènes dans le génome de référence, ne soient pas de la même taille. Ce qui pourrait induire le jugement des experts en erreurs puisque plus de reads, provenant des échantillons, pourraient être alignés sur ces transcripts, dû à leur taille plus importante.

Nous devions, par conséquent, normaliser les résultats obtenus précédemment afin d'obtenir des graphiques de conclusions finales.

Pour ce faire, nous avons utilisé les options de la librairie « DESeq » dans R qui permettent automatiquement de normaliser les données, en activant les options nécessaires, avant de produire les graphiques pertinents. En effet, l'étape de normalisation ne requiert, en utilisant cet outil, de spécifier l'option de normalisation dans les fonctions permettant de créer les figures.

Nous avons, donc, comme dans les parties précédentes, fait la distinction entre les deux sortes de raisin. Les deux graphiques 4.11 et 4.12 présentent, respectivement, le graphique correspondant aux échantillons de Pinot gris et le second, le graphique correspondant aux échantillons de Sauvignon blanc.

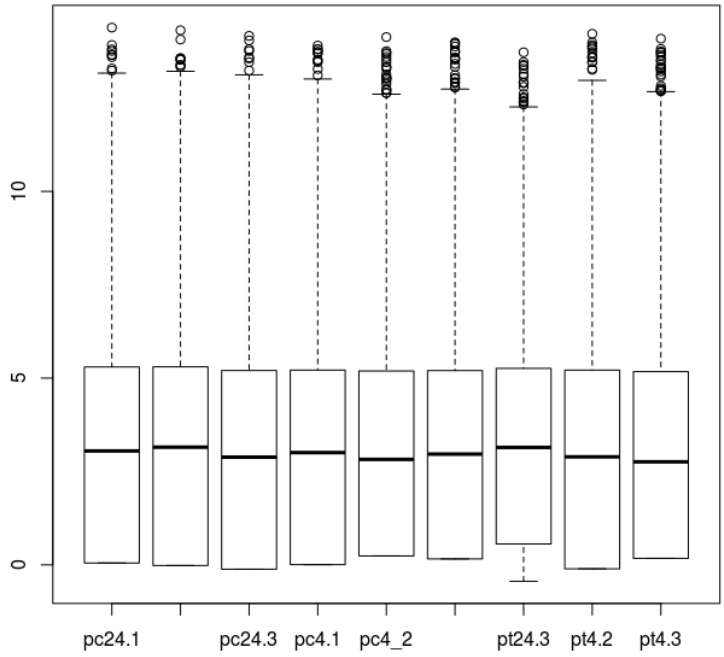

Figure 4.11: Boxplots données normalisées Pinot gris

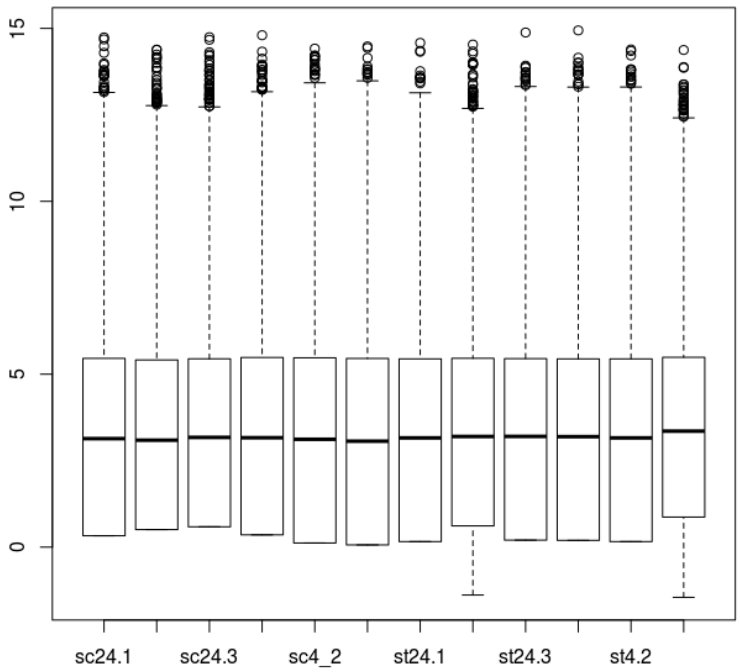

Figure 4.12: Boxplots données normalisées Sauvginon blanc

Comme nous pouvons le voir, la normalisation permet d'aligner les médianes des différents échantillons, en comparaison avec les graphiques homologues de la section précédente. Ces données et résultats, ainsi obtenus, seront utilisés plus tard dans le processus afin d'obtenir des graphiques plus pertinents.

Malgré le fait que nous ne sommes pas en mesure de fournir une conclusion sr base de ces résultats, ceux-ci nous permettent, malgré tout, de nous rendre compte, en comparant les deux séries de résultats (des étapes deux et trois), de l'impact correcteur que possède l'étape de normalisation sur ceux-ci. Par conséquent, nous pouvons la juger comme étant pertinente.

La production de graphiques spécifiques

Malgré le fait que nous avons déjà produit des graphiques, nous ressentons toujours la nécessité d'en obtenir de nouveaux, permettant de représenter autrement les résultats, obtenus lors de l'étape de comptage, dans le but d'aboutir à des conclusions plus pertinentes. Par exemple, nous jugeons qu'il est intéressant d'obtenir des graphiques qui permettent de représenter les données sous forme de groupes, grâce à un mécanisme de visualisation (PCA) de données complexes ; ainsi que de graphiques phylogénétiques, grâce auxquels, nous pourrons examiner les différences de productions protéiques entre les différents échantillons.

Cette section a, donc, pour but d'expliquer les résultats obtenus. Une fois encore, nous avons dû faire la distinction entre les deux sortes de raisin.

En ce qui concerne les outils utilisés, la librairie « DESeq » suffit amplement à l'élaboration de ces graphiques.

Les graphiques PCA sont intéressant dans le sens où ils permettent d'observer et tirer des conclusions à partir d'observations faites à partir du comportement des échantillons.

En ce qui concerne l'outil, cet outil peut être utilisé en utilisant les données de la deuxième étape (comptes) dans le cas des données normales et dans le cas des données normalisées. Pour le deuxième cas, nous devons seulement d'activer une option « normalized ».

Les graphiques 4.13 et 4.14 présentent la répartition des échantillons en groupes d'échantillons correspondant aux données des données normalisées des deux sortes de raisin.

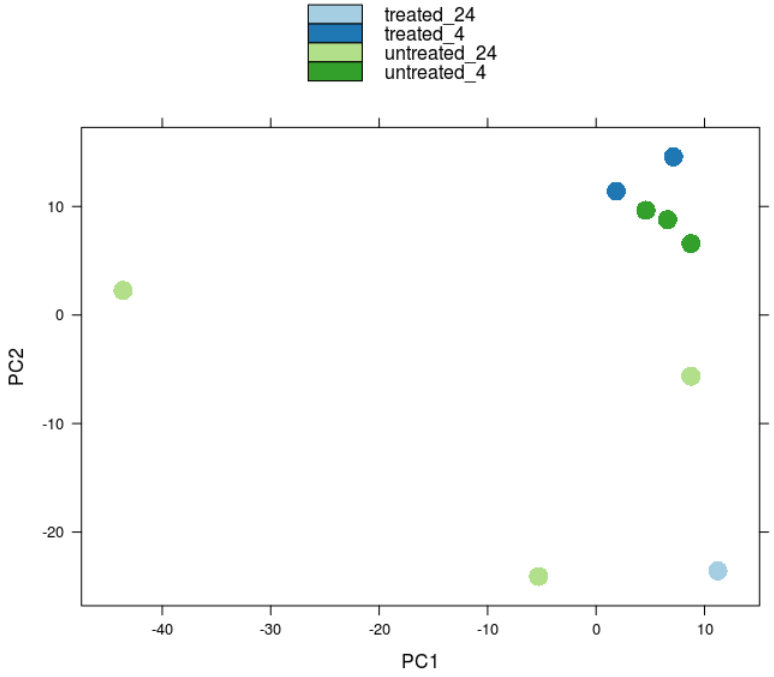

Figure 4.13: graphique PCA Pinot gris

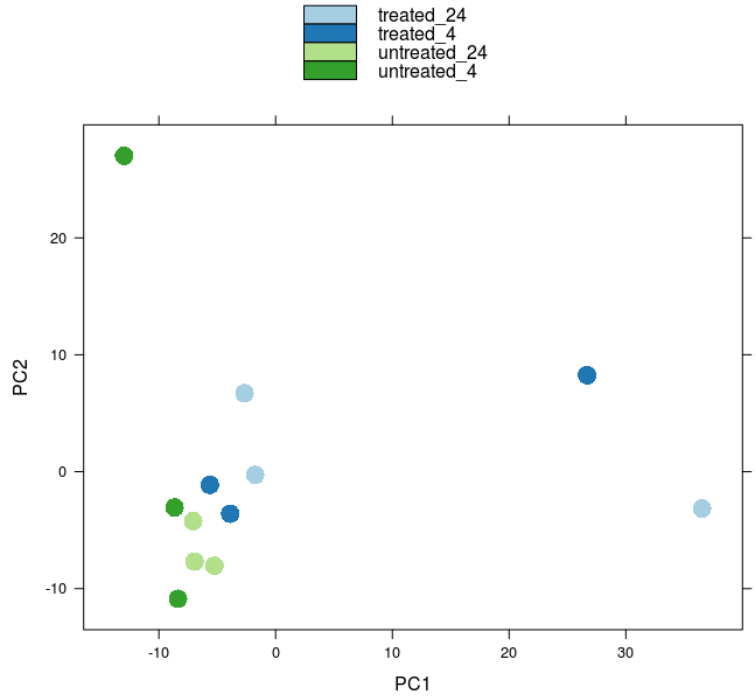

Figure 4.14: graphique PCA Sauvignon blanc

Ces résultats nous permettent de notifier, qu'une fois de plus, le traitement hormonal se fait plus ressentir dans la sous-espèce de raisin Sauvignon blanc. En effet, nous pouvons constater, dans le second graphique, une répartition plus aléatoire concernant les échantillons qui ont fait l'objet de traitement hormonal. De plus, lorsque nous analysons les deux graphiques, nous remarquons que ce fait est plus observé dans le second.

Ce qui est intéressant avec ce genre de graphique est le fait de pouvoir associer des échantillons possédant des caractéristiques semblables. Cependant, nous notons que nous n'en connaissons pas suffisamment sur le sujet que pour pouvoir fournir une explication détaillée de ce genre de graphique. Cependant, nous pouvons expliquer que ce graphique permet de présenter des données en fonction d'un grand nombre de caractéristiques, la répartition est donc difficilement explicable lorsque l'on ne connait pas vraiment le domaine d'expertise dans lequel l'expérience est menée.

Un autre type, utile à l'élaboration de conclusions biologiques, peut être utilisé dans notre situation. En effet, nous pouvons construire, grâce aux données normalisées, des graphiques permettant d'observer la production de protéines. Ce genre de graphiques, appelé « heatmap », permet, pour chaque échantillon, de fournir le taux de protéines produites, et ce pour chaque protéine importante.

Par exemple, pour les échantillons provenant de plants de Sauvignon blanc, la figure 4.15 présente les résultats obtenus.

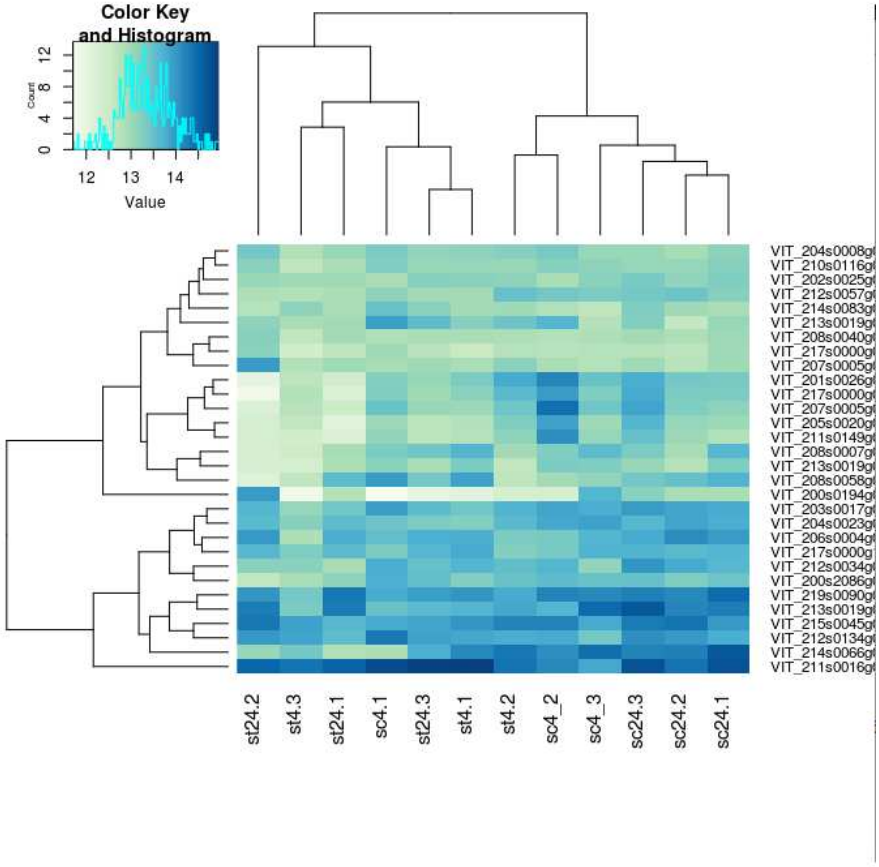

Figure 4.15: Graphique « HeatMap »

Les échantillons sont présentés, dans ce graphique, de manière similaire que dans les précédents. C'est-à-dire que la première lettre correspond à l'origine du plant de raisin, le nombre au délai d'attente entre la pulvérisation et les différents moments de récolte.

Au niveau de l'interprétation, nous pouvons remarquons, sur ce graphique, que la protéine « VIT_211s0016g0 » est plus produite lorsque les plants de raisins sont traités avec l'hormone. Inversement, nous remarquons que la protéine « VIT201s0026g0 » est, quant à elle, moins produite par les plants traités avec l'hormone de Gibbérelline.

N.B. : D'autres types de graphiques pourraient, également, être utilisés, mais ne l'ont pas été dans le cadre de ces expériences et ne seront, par conséquent, pas abordés dans ce travail.

Limites des résultats

Nous insistons sur le fait que nous ne sommes pas biologistes et que, par conséquent, il se pourrait que la véracité de nos résultats et conclusions ne puissent pas être vérifiées. Malgré ce problème, cette section, ayant pour but de présenter les différentes étapes du pipeline ce projet, atteint bien son objectif puisque ce projet ne constitue qu'un exemple assez pertinent qui permet, au lecteur, d'observer les différents résultats pouvant être obtenus.

De ce fait, même si les résultats sont faux, cette section reste valable dans la mesure où elle permet de présenter au lecteur un cas concret d'utilisation du pipeline.

Cependant, nous jugeons pertinent de notifier que le type de résultats, ainsi obtenus, a été jugé comme étant intéressant par les experts. Ce fait renforce la pertinence liée à l'utilisation de cet outil dans ce genre de recherche.

Discussion

Ces résultats et recherches nous ont permis de nous rendre compte de l'importance et aisance d'utilisation que constitue cet outil. Si nous voulons aller plus loin, nous pouvons discuter la pertinence de l'automatisation de cet outil. En effet, le fait qu'il soit composé de composants utilisables dans un terminal et le fait que leur distribution soit libre, nous permettrait d'écrire un script permettant de la génération automatique des résultats. Malgré le fait que cette avancée technologique pourrait contribuer à l'épanouissement de la recherche scientifique, elle pourrait peut-être aboutir, à long terme, à des problèmes moraux et éthiques.

En effet, comme l'Histoire l'a déjà démontré, la science a pu être utilisée, à de multiples reprises, dans le but d'appuyer le raisonnement et l'idéologie de certains organismes politiques et religieux. Le fait de rendre ce genre d'outils plus accessible à un plus large public, pourrait, donc, rendre les conclusions scientifiques plus accessibles et avoir des répercussions conséquentes.

L'éthique, tout comme dans les autres domaines technologiques et scientifiques, n'est, par conséquent, pas à écarter de l'évolution de ce domaine d'expertise.

Conclusion

Cette partie nous a permis de prendre conscience de la pertinence de l'apport majeur que constitue un pipeline d'extraction de données. En effet, nous avons pu remarquer que le processus d'extraction de données et de création de graphiques, permettant l'élaboration de conclusions, pouvaient être générés.

Pour ce faire, nous avons démontré l'utilisabilité d'un tel outil dans le cadre d'un exemple de projet scientifique auquel nous avons réellement pris part. Celui-ci avait pour but de déterminer les conséquences qu'avait l'utilisation de l'hormone de Gibbérelline sur deux sortes de raisins, à savoir : le Pinot gris et le Sauvignon blanc. Le pipeline nous aura permis, à terme, d'obtenir des résultats intéressants permettant d'illustrer le type de résultats pouvant être extraits et servant à l'aboutissement de conclusions.

Ces premières observations permettent, à ce stade, d'illustrer le genre de résultats que nous pouvons obtenir à partir des échantillons. Malgré le fait que nous ne disposons pas des connaissances nécessaires à l'élaboration de conclusions constructives, nous pouvons, néanmoins, affirmer que ce pipeline fournit des résultats pertinents. Ceux-ci, étant considérés comme « pertinent » par les experts, nous permettent de vérifier la crédibilité liée à l'utilisation de l'outil dans ce domaine.

Notons, toutefois, que les brèves conclusions, présentées précédemment, sont purement personnelles et n'ont pas été vérifiées par des experts du domaine biologique ou génétique. Elles ne servent, uniquement, à illustrer le genre de conclusions pouvant être élaborées à partir des résultats.

Cet exemple nous permet, donc, de répondre partiellement à la question de recherche posée en affirmant que cet outil est d'une utilité majeure dans la recherche scientifique ayant trait au milieu biologique puisqu'il permet, comme nous venons de le voir, d'obtenir des résultats pertinents, selon les experts, lors d'études portant sur l'influence que possède une hormone ou tout autre modification génétique sur un corps.

A ce stade, nous nous rendons, dès lors, compte d'un lien pouvant être fait entre le milieu génétique et l'outil. Nous nous demandons, donc, comment cet outil pourrait être applicable au milieu médical, lors d'étude portant sur les maladies génétiques. La section suivante aura comme but de vérifier ce rapprochement en se concentrant sur l'étude de deux maladies génétiques.

La section suivante nous permettra, donc, à terme, de répondre à la question de recherche en démontrant que l'outil peut aller plus loin et qu'il nous permet, par conséquent, d'affirmer qu'il n'est pas seulement utile dans le milieu de la recherche biologique, mais également dans celui de la recherche médicale.

Troisième Partie: Application eu milieu médical

Dans cette partie nous allons découvrir la manière dont le pipeline d'extraction de données à partir de séquences d'ARN, présenté dans la partie précédente, pourrait être appliqué au domaine médical. Le but sera, donc, de tenter d'élaborer la réponse à la réponse de recherche, en essayant d'expliquer en quoi l'utilisation de l'outil serait pertinente dans ce milieu.

Cette partie sera, donc, articulée autour de deux points clefs : dans un premier temps, une partie destinée à présenter le génome humain sera proposée, afin de fournir au lecteur la base de connaissances nécessaires à la compréhension des concepts plus élaborés ; et dans un deuxième temps, nous étudierons deux cas de maladies génétiques afin d'en savoir plus sur celles-ci d'un point de vue génétique. Pour chacune d'entre elles, une brève définition, ainsi que des explications concernant les méthodes de détection et de guérison seront fournies au lecteur. Ensuite, nous expliquerons comment l'outil pourrait être utilisé et les conséquences que cette utilisation aurait dans le milieu.

Deux sections portant sur des discussions éthique et technologique seront également proposées. Le but étant de motiver une réflexion concernant notre point de vue sur la question de recherche.

Au terme de cette partie, nous espérons être en mesure de répondre pertinemment à la question de recherche.

Chapitre 5 : Génome humain

Cette partie sera consacrée à la définition et aux explications concernant le génome humain. Ceci permettra au lecteur d'obtenir les connaissances initiales nécessaires à la compréhension des concepts plus élaborés exposés plus loin dans cette partie.

Tout d'abord, un être humain peut être considéré, d'un point de vue biologique, comme étant un ensemble de cellules animales incluant les différentes propriétés présentées dans la première partie. Ce fait nous permet, dès lors, de comprendre les fondements principaux des cellules humaines.

Son génome est composé de 23 paires de chromosomes possédant tout le matériel génétique nécessaire à sa survie. D'après T. A. Brown [Brown, 2004], ce génome est le produit obtenu du projet de recherche, appelé « Human Genome project », démarré en 1990 et qui a abouti en 2001. Malgré le fait que ce résultat ne contienne que 83-84% de l'entièreté du génome humain, celui-ci possède toutes les séquences générales.

Selon R.Lewis [Lewis, 2007], les traits génétiques d'un individu sont généralement hérités des parents et incluent toutes les caractéristiques physiques d'un individu, telles que : la taille, le poids, la couleur des yeux, de la peau et des cheveux, la pression artérielle, etc. ; mais également, des traits psychologiques liés aux comportements, telles que : l'insomnie, schizophrénie, la bipolarité, etc.

Concernant les études génétiques, portant sur le génome humain, nous pouvons affirmer que les études utilisant des techniques de manipulations génétiques ont apporté une contribution importante dans le domaine de la biologie et de la médecine [Primrose, Twyman, Old 2004].

Une des premières contributions observées dans le domaine médical est le fait d'avoir pu rendre l'hormone de croissance humaine (HGH, « Human Growth Hormone ») disponible en plus grande quantités chez les adolescents souffrants de problèmes hormonaux permettant l'arrivée de la puberté [Primrose, Twyman, Old 2004].

Un autre avantage que le domaine médical a pu retirer de cette technique concerne les vaccins. En effet, de nouvelles techniques ont ainsi été découverte et utilisées, notamment dans le cas du vaccin actuel de l'hépatite B [Primrose, Twyman, Old 2004].

Ce fait nous permet de comprendre que l'étude des gènes n'est pas nouvelle dans le milieu médical et qu'elle fera encore l'objet de nombreuses études dans le futur. Nous savons, également, que ces maladies génétiques ne concernent pas seulement les maladies physiques, mais également mentales, ce qui laisse supposer que ce genres de maladies peuvent également être étudié de cette manière.

Grâce à ces explications, nous comprenons directement le lien existant entre le génome humain et l'outil que nous proposons. En effet, l'outil, dont le but étant de pouvoir obtenir des informations (statistiques ou autres), permettrait à des scientifiques d'aboutir à des conclusions relatives aux problèmes causés par les disfonctionnements génétiques.

Ce lien entre les deux concepts sera explicité plus en profondeur dans les sections à venir.

Chapitre 6 : Etudes de cas

Afin de pouvoir fournir le plus d'informations concernant l'utilisation d'un tel outil dans le domaine médical, nous allons étudier certaines maladies génétiques. Ceci nous permettra à terme de pouvoir tirer des conclusions utiles à la réponse de la question de recherche.

Parmi ces maladies, nous avons décidé, dans un souci de synthétisation, de nous limiter à des maladies fréquentes. Nous ne traiterons, donc, que du diabète et du cancer. Pour chacune de ces maladies, nous allons fournir une brève description ainsi que les méthodes employées par les médecins et scientifiques pour les déceler dans l'organisme et les moyens mis en œuvres pour les traiter. Ensuite, nous compléterons ces explications par celles relatives à l'emploi de cet outil à ce domaine ainsi que les informations relatives aux caractéristiques génétiques de ces maladies.

Notons que, selon nos sources [Lewis, 2007], les maladies génétiques sont rarement causées par un seul gène particulier, mais par un ensemble, ainsi que le comportement génétique (déclenchement gènes, maladies) ne dépend pas seulement de l'héritage parental, mais également, de l'environnement. Par exemple, un enfant ayant une prédisposition à être diabétique pourra retarder ou atténuer la maladie en vivant dans un environnement qui lui sera plus favorable. Ce fait, nous permet de comprendre qu'il est peut être possible de contourner les prédispositions génétiques à faire apparaitre une maladie génétique. Nous reparlerons de ce fait plus tard dans ce travail.

N.B. : les explications présentées ci-dessous feront des références aux notions biologiques présentées dans la première partie et à celles, relatives au pipeline, présentées dans la deuxième partie.

Diabète

Dans un contexte de mode de vie dans lequel bon nombre de personnes possèdent une mauvaise hygiène alimentaire, plusieurs maladies liées à ce mode de vie ont pu facilement émerger. Parmi celles-ci, le diabète est une maladie considérée comme étant l'une des plus importantes dans les sociétés actuelles. En effet, d'après nos sources médicales [Grimaldi, 2000], on estime les nombre de personnes souffrantes de cette maladie à 30 millions rien qu'en Europe et à 100 millions dans le monde entier.

A. Grimaldi [Grimaldi, 2000] nous apprend également qu'il existe en réalité deux sortes de diabètes : le diabète insulino-dépendant, couramment appelé « type un », et le non-insulino-dépendant, couramment appelé « type deux ». Le premier survient avant l'âge de 20 ans et représente que 10-15% des diabètes, tandis que le second, quant à lui, ne survient que vers 50 ans et représente 85-90% des diabètes. Parmi ces deux maladies, c'est la seconde qui est la plus dangereuse. En effet, les effets secondaires pouvant en découler peuvent avoir de lourdes conséquences sur le mode vie du patient. En effet, d'après la même source, le diabète de « type deux » serait à l'origine de près 50% des cas de cécité et d'amputations chez les personnes de plus de quarante ans.

De plus, d'après [Jork, Carey, Bamshad, 2010], le diabète peut être transmis aisément vers les membres de la descendance. Ce qui nous encourage à étudier cette maladie d'un point de vue génétique et, par conséquent, voir comment l'outil pourrait s'y adapter.

Ces faits nous permettent, dès lors, de comprendre l'importance que possèdent les découvertes concernant le traitement de ce genre de maladie.

Explications

De manière plus concrète, nous pouvons, sur base des explications fournies par les sources recueillies [Grimaldi, 2000 & Loghmani, Stang, Story 2005], définir le diabète comme étant une maladie liée à un excès de glycémie dans le sang, couramment appelée « hyperglycémie ».

Plus précisément, ce haut taux de glycémie engendrerait une défectuosité dans la sécrétion d'insuline. L'insuline étant l'hormone, produite par le pancréas, permettant d'utiliser le glucose comme une source d'énergie. Concrètement, cette hormone permet de stocker le glucose sous

forme de réserves nutritionnelles, appelées « glycogènes » [P. Ferré 2005]. Cette réserve est utilisée lors de a journée.

Sous l'action d'un haut taux de glycémie, la production d'insuline diminue et, par conséquent, la création de réserve nutritionnelle est plus difficile à créer. L'absorption des glucides dans les tissus sanguins et musculaires se révèle, par conséquent, être de plus en plus difficile à réaliser.

Nous comprenons, sur base des explications relatives à la génération de protéines et hormones (de la première partie), qu'il s'agit d'un problème lié à une mauvaise production hormonale qui, comme toute production hormonale et protéiques, se réalise à partir du matériel génétique de l'individu. Nous comprenons, également, que la cause du disfonctionnement génétique est dû à l'environnement [Ferré, 2005].

Diagnostic

Le diagnostic du diabète se fait généralement sur base d'une prise de sang lorsque le patient est à jeun. A titre d'information, cette prise de sang permet de vérifier que le taux de glycémie ne dépasse pas un certain seuil, à savoir, 26 mg/dL [Loghmani, Stang, Story 2005].

Pour être plus précis, selon la même source, la synthétisation de l'insuline se produit sur base de l'information génétique contenue sur le chromosome 11 de l'individu. Nous remarquons, dès lors, le caractère génétique de la cause de la maladie ainsi que la source de la déficience de la production protéique.

Nous pouvons, sur base de ces données, comprendre que le diagnostic de cette maladie dans l'organisme des patients se fait grâce à la présence d'un taux glycémique ou par une production réduite d'insuline dans le sang. Sachant que l'hormone d'insuline est constituée de certaines protéines, nous pouvons, donc, déceler un disfonctionnement hormonal sur base d'une production protéique trop faible.

Traitement

D'après nos sources [Loghmani, Stang, Story 2005], les traitements « communs » relatifs à cette maladie consistent à l'absorption par le patient d'une dose d'insuline, par voir orale ou

intraveineuse, afin de pouvoir contrôler le taux de glycémie. Cette absorption va permettre au patient de réguler son taux de glycémie. Le médicament servira donc, en quelques sortes, de complément protéique ou hormonal devant être ingéré afin de répondre à l'absence de certaines protéines, normalement générées à partir de l'ADN du patient, permettant une régulation normale de la glycémie.

Il est important de noter que ce genre de traitement, ne constitue en rien une solution définitive dans le sens où, le patient devra prendre ce genre de substances tout au long de sa vie. Une seule utilisation ne constitue, donc, qu'une solution à court terme.

Outre le problème de la récurrence de ce traitement, ceux-ci possèdent, comme tout médicament, des effets secondaires liés à sa composition.

Par exemple, le « Glucophage » est un médicament connu pour dans ce contexte. Lorsque nous analysons la notice d'un peu plus près, nous remarquons que … est essentiellement présente dans sa composition. Ce médicament possède des effets secondaires liés à des troubles digestifs (nausée, etc), troubles liés aux sens, sanguins et cutanés[41].

Nous comprenons, dès lors, qu'une solution à long terme, qui serait plus souhaitable, consisterait à modifier le matériel génétique afin de pouvoir réinstaurer une production de protéines capables de réguler le taux d'insuline. Ce genre de traitement consiste en la technique de thérapie génique.

En effet, de nos jours, certaines sources supposent que les scientifiques tentent de trouver un moyen de modifier la production d'insuline soit en instaurant de l'insuline synthétisée d'une source externe [Baeshen, Baeshen, Sheikh, Bora, Ahmed, Hassan A I Ramadan, Kulvinder Singh Saini, Elrashdy M Redwan, 2014]. Ce fait prouve bien l'intérêt que représente le domaine de la thérapie génique dans la détermination de remède contre le diabète.

[41] D'après la monographie du produit pouvant être obtenue à l'adresse suivante : http://products.sanofi.ca/fr/glucophage.pdf

Cancer

Le cancer est l'une des maladies les plus répandue dans le monde. En effet, d'après des statistiques sur le cancer [Siegel, Miller, Jemal, 2015], cette maladie occuperait, de nos jours, la deuxième place du podium des maladies les plus mortelles. Il est attendu qu'elle dépasse le nombre de morts causées par des arrêts cardiaques d'ici quelques années.

D'après les chiffres, pouvant être obtenus à partir de la plateforme du service « Belgian Cancer Registry »[42], nous remarquons que le nombre de cancers chez les femmes en régions flamande entre 1999 et 2009 a augmenté de 4000 cas.

Cependant, malgré ces chiffres alarmants, les technologies actuelles permettent à la recherche scientifique de pousser les ses limites. En effet, d'après nos sources [Jorde, Carey et Bamshad, 2010], les scientifiques sont capables de comprendre les mécanismes internes du développement de cette maladie et ce, tant d'un point de vue cellulaire, que génétique. De plus, les conclusions proposées sur le taux de mortalité des cancers auraient diminué de 22% entre 1991 et 2001. Ce qui prouve, d'une certaine manière, que le progrès dans ce centre d'intérêt, se fait ressentir au fil du temps, en permettant de fournir des remèdes adéquats.

Nous comprenons, dès lors, la pertinence du traitement de cette maladie dans cette section. Les sections suivantes auront pour but de permettre au lecteur d'en savoir plus sur cette maladie en fournissant des explications générales, les méthodes relatives au diagnostic et la présentation d'un traitement général.

Explications

Cette section a pour but de donner des explications concernant le cancer en fournissant une définition, les méthodes de dépistage ainsi que la thérapie génétique qui est un traitement général.

D'après nos sources [Lewis 2007], le cancer est une maladie génétique apparaissant lorsque le phénomène de reproduction cellulaire se produit de manière anormale. En effet, comme expliqué précédemment, le nombre de cellules générées doit être proportionnel à celui des cellules mourant. Dans le cas contraire, les tissus risquent de contenir trop de cellules et, par conséquent, faire augmenter la taille de l'organe. Le fait que cette maladie apparaisse dans un contexte de

[42] *Cancer Burden in Belgium 2004-2013*, Belgian Cancer Registry, Bruxelles, 2015

reproduction cellulaire implique l'existence du lien avec la génétique (voir explications de la première partie).

Pour être plus précis, selon nos sources [Jork, Carey, Bamshad, 2010 & Lewis, 2007 & Weinberg, 2014], le cancer aurait une source génétique et environnementale. La première cause (génétique) peut être expliquée comme suit : l'altération génétique due à des mutations peut entraîner un disfonctionnement dans la production de protéines et, par conséquent, dans la reproduction cellulaire. Ce qui peut entraîner l'apparition de cancer. Il est important de préciser que ce genre de mutations n'apparait que dans les cellules somatiques et non pas dans les cellules germinales. Il en résulte, donc, que les cancers ne se transmettent pas de générations en générations. Alors, lorsque nous disons qu'ils sont génétiques, nous voulons dire qu'ils proviennent d'une source génétique, mais la maladie, quant à elle, n'est pas héritée.

La cause environnementale, quant à elle, peut être expliquée par le fait que, malgré le fait que le cancer ne semble n'être lié qu'au caractère génétique, celui-ci ne déclenche un processus de mauvaise reproduction cellulaire qu'après un changement particulier lié à l'environnement. Jork, Carey et Bamshad [2010] affirment que ces changements concernent des mutations génétiques (voir première partie du travail pour plus d'informations), dues à des utilisations d'agents chimiques, ou à des facteurs externes influençant, alors, la production de protéines sans, pour autant, provoquer de mutations. Par exemple, le tabagisme favoriserait l'apparition d'un cancer des poumons plus rapidement.

Précisons que ces mutations génétiques ne prennent part que dans les tumeurs. En effet, lorsqu'une mutation est observée chez un individu dans le cas d'un cancer, seules les cellules ne permettant plus de se reproduire correctement possèdent une erreur de code génétique. Ce fait implique que le cancer ne soit pas héréditaire puisque les cellules haploïdes, servant à la reproduction de l'individu, ne seront pas affectées par ces mutations et transmettent, par conséquent, le matériel génétique d'origine de l'individu.

Le cancer peut, donc, être définit comme étant une maladie génétique, mais pas directement héréditaire, liée à un disfonctionnement lié à la reproduction cellulaire, déclenchée suite à un stimulus environnemental apparaissant dans le long terme.

Notons que, d'après nos sources [Jork, Carey et Bamshad 2010], chaque organe produit son type de cellule. En effet, en fonction de l'organe dans lequel la cellule, qui se reproduit, se trouve, un certaines protéines vont être synthétisées. En d'autres termes, nous pouvons affirmer que la production de protéines est directement liée à l'organe hôte. Donc, nous pouvons faire le lien entre ne production anormale de certaines protéines et l'apparition d'une tumeur cancéreuse dans un

organe. Donc, si on peut déceler une production anormale de certaines protéines, on peut les relier à une présence possible d'un cancer.

D'après nos sources (), nous pouvons, également, noter que les cancers possèdent plusieurs stades de développement. En effet, si nous prenons le cas du cancer de la prostate (article []), il existe quatre stades du cancer. Ce fait est important dans le sens où il nous permet de comprendre qu'une « simple » détection du cancer ne suffit plus, nous devons être capables de définir l'état d'avancement afin de pouvoir fournir le traitement adéquat.

Diagnostic

Une multitude de cancers existent, par conséquent, les méthodes de dépistages se diversifient également. Le but n'étant pas de décrire en détails ce qui se fait à l'heure actuelle dans ce domaine, mais de fournir, au lecteur, une vue générale de ce sujet. Nous allons, donc, dans cette section expliquer les méthodes de dépistage les plus populaires liés à ces maladies.

D'après les sources que nous disposons [Jork, Carey, Bamshad, 2010 & Rousseau, P. Bohet, J. Merlière, H. Treppoz, B. Heules-Bernin, R. Ancelle-Park, 2002 & Belgian Cancer Registry, 2015], nous pouvons aisément classer les techniques en deux catégories. Celles-ci sont les extractions de données à partir d'échantillons organiques, à savoir : la biopsie, la technique du frotti, la prise de sang, etc. ; et celles qui consistent à analyser visuelles des zones : radiologie, échographie, mammographie, IRM, etc. Certaines de ces techniques sont plus utilisées dans certains cas de maladies.

Par exemple, la technique du frotti, consistant à prélever un échantillon de peau de la zone à examiner, est largement utiliser pour détecter le cancer du col de l'utérus. Tandis qu'en ce qui concerne le cancer du poumon ou du sein, les techniques de radiologies ou de mammographie seront plus utilisées.

Notons, toutefois, que tous les cancers peuvent être décelés par la première catégorie de techniques, à savoir par prélèvements de matière organiques. En effet, étant donné que le cancer est génétique, celui-ci peut être détecté à partir des séquences d'ADN des tumeurs (trouver une source). Pour ce faire, nos sources nous permettent de comprendre qu'une des techniques les plus utilisées, par les biologistes à l'heure actuelle, est l'analyse de « mico-arrays » qui permet d'étudier l'expressivité génétique des échantillons formés par les tissus des tumeurs cancéreuses. Concrètement, cette technique utilisant des lamelles de verres dans lesquelles se situent une

multitude de d'encoches. Ces encochent, pouvant atteindre le nombre de plusieurs milliers, permettent d'insérer des molécules d'ADN de l'individu et peuvent contenir plusieurs copies d'ADN correspondant à un gène en particulier. Ensuite, nous pouvons utiliser un laser afin d'exciter les différentes molécules d'ADN avec un laser. Ce procédé permet de pouvoir remarquer les différents comportements que produisent ces molécules.

A ce stade, plusieurs applications peuvent y être appliquées pour mesurer l'expressivité génétique du sujet. Cependant, une application semble être plus couramment utilisée que les autres. En effet, cette technologie est souvent utilisée pour comparer l'expressivité génétique d'une cellule soumise à différentes situations.

La figure 6.1 suivante permet au lecteur de visualiser ce procédé.

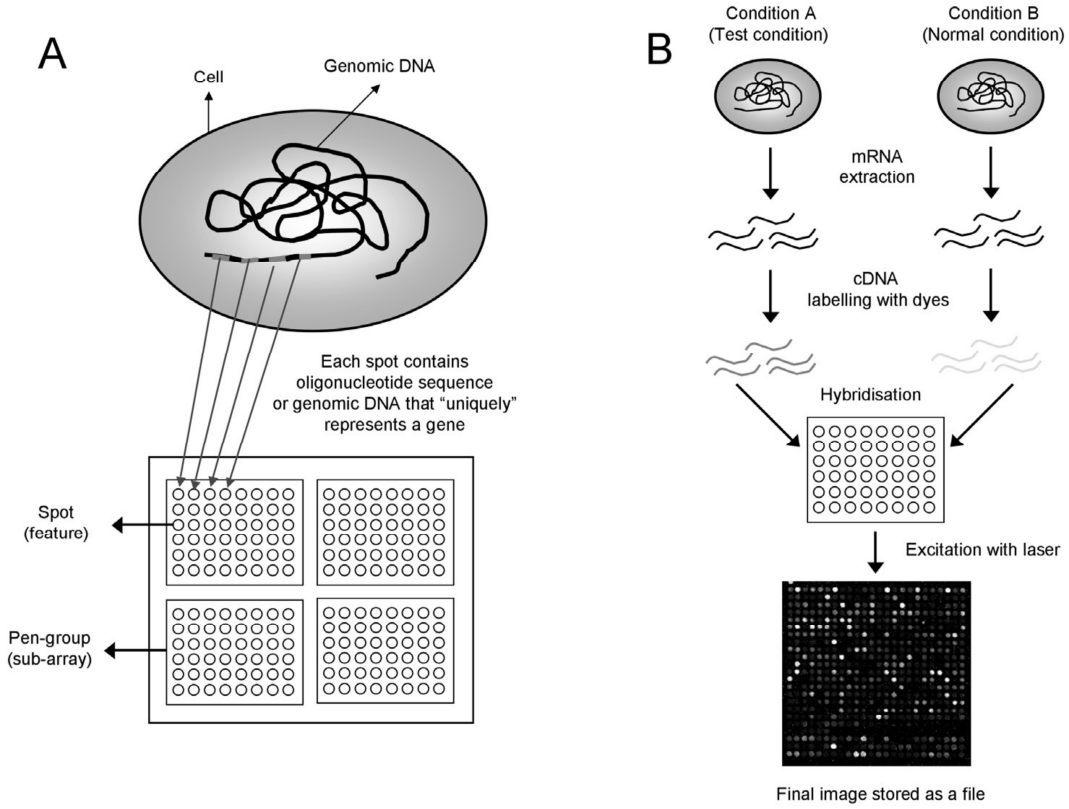

Figure 6.1: Processus de la technique "Micr-arrays" [Babu, 2004]

Nous pouvons y remarquer, grâce au sigle « A », la constitution de l'étude ; tandis que la partie « B », représente l'étude du génotype dans deux situations particulières, dans des conditions particulières et dans des conditions normales. [Babu, 2004]

Il est à noter que les résultats, obtenus par cette technique, peuvent être interprétés grâce à des techniques informatiques à partir de l'image des spots produite. Nous pouvons, alors, procéder à une

normalisation, création de clusters ainsi qu'une classification supervisée sur base des résultats extraits.

En ce qui concerne les différents stades d'avancement du cancer, certaines techniques moins ordinaires utilisent le machine learning afin de pouvoir, d'une part, catégoriser les radiographies, ou autres images médicales, afin de pouvoir détecter les tumeurs à partir des images médicales ; et d'autre part, déterminer leur niveau d'agressivité, lorsqu'il s'agit d'une tumeur cancéreuse [43] [44] [45]

Comme nous pouvons le remarquer, les techniques principales actuelles peuvent être qualifiées de « manuelles » dans le sens où elles ne sont pas directement liées aux technologies informatiques qui pourraient rendre certaines tâches automatiques. Nous remarquons, également, qu'étant donné que les cancers peuvent être décelés à partir de matières organiques, nous pouvons en récupérer les séquences d'ADN ou d'ARN y étant associées. Ces remarques seront utiles dans les explications à venir.

Traitement

De multiples traitements pour lutter contre le cancer existent. Parmi ceux-ci, nous pouvons citer : la chimiothérapie, qui consiste à détruire les zones d'infections suite à une prise de médicament ; la chirurgie, qui consiste à extraire la partie infectée ; la radiothérapie, consistant à irradier la zone infectée ; l'hormonothérapie, consistant agir sur les productions des hormones permettant l'épanouissement des cellules infectées ; et l'immunothérapie, qui permet de faire intervenir le système immunitaire. Chaque cancer, nécessitant un traitement approprié, bénéficiera d'un traitement précédemment cité[46].

---

[43] Jimmy C. Azar, Martin Simonsson, Ewert Bengtsson, and Anders Hast, *Automated Classification of Glandular Tissue by Statistical Proximity Sampling*, Hindawi Publishing Corporation International Journal of Biomedical Imaging Volume 2015, Centre for Image Analysis, Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden, 2014

[44] Scott Doyle, Michael D Feldman, Natalie Shih, John Tomaszewski and Anant Madabhushi, Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer, BioMed Central (BMC) Bioinformatics, 2012

[45] Ali Tabesh, Vinay P. Kumar, Ho-Yuen Pang, David Verbel, Angeliki Kotsianti, Mikhail Teverovskiy, and Olivier Saidi, *Automated Prostate Cancer Diagnosis and Gleason Grading of Tissue Microarrays*, Proc. of SPIE Vol. 5747, p. 58-70

[46] *Comprendre la chimiothérapie*, Institut National du Cancer, Belgique, 2009

Malgré les qualités curatives que ces traitements possèdent, ils possèdent également des effets secondaires. Par exemple, la chimiothérapie, par exemple, ne se limite pas à agir sur les cellules cancéreuses, mais agit, également, sur les autres cellules du corps saines. Ce qui peut entraîner des complications sur la santé de l'individu, telles que : la stérilité, problèmes rénaux, etc.[16]. Ce fait nous prouve que ces techniques ne sont pas parfaites. En d'autres termes, d'autres techniques, plus élaborées, pourraient les remplacer.

Parallèlement à l'utilisation de ces traitements, les chercheurs travaillent sur l'élaboration d'un autre traitement, semblant être plus innovant et plus pertinent dans notre étude. Il s'agit de la technique de thérapie génique.

Malgré l'abondance des traitements, et dû à la cause génétique de l'apparition des cancers, nous avons décidé de nous limiter, dans ce travail, qu'au traitement directement lié aux gènes : la thérapie génique. En effet, celle-ci offre la possibilité, selon nos sources [Jork, Carey et Bamshad 2010], de pouvoir modifier le matériel de l'individu dans le but que d'autres protéines puissent être générées. Nous voyons, donc, que celle-ci s'adapte parfaitement au sujet traité puisqu'elle consiste en une méthode principalement axée sur l'analyse génétique.

Concrètement, d'après R.Lewis [Lewis, 2007], la thérapie génique, contrairement à une thérapie médicamenteuse, consiste à altérer le génotype d'un individu afin qu'il ne produise plus le même ensemble de protéines pouvant aboutir à l'apparition d'une maladie. Ce traitement constituerait, donc, un remède définitif, ce qui n'est pas le cas de tous les autres.

Dans son ouvrage, l'auteur [Lewis, 2007], présente l'exemple d'Ashanti DeSilva. Cette petite fille était atteinte d'une maladie génétique qui l'empêchait son génotype de produire des nucléotides d'adénosine. Cette carence empêche certains mécanismes biochimiques de se produire et entraîne des défectuosités immunitaires. Ashanti a, alors, été la première personne à pouvoir bénéficier d'un traitement médical lié à la thérapie génique. En effet, après avoir planifié l'intervention pendant des années, les médecins ont finalement pu modifier son génotype afin que celui-ci puisse avoir un comportement normal. D'après l'auteur, Ashanti serait toujours en vie au moment de la rédaction en 2007.

Cet exemple nous prouve donc la pertinence de ce genre de traitement dans ce contexte.

Généralités sur les maladies génétiques

Avant d'aller plus, essayons, à ce stade, de faire une conclusion relative aux maladies génétiques précédemment décrites dans notre de cas. En effet, nous allons décrire nos conclusions concernant le moyen de détection, les sources et les traitements qu'il est possible d'effectuer dans le cas des maladies génétiques.

En ce qui concerne la détection de ces maladies, nous pouvons affirmer qu'il existe des moyens biologiques consistant à vérifier le taux de certaines substances à partir de prises de sang ou de tissus infectés. En d'autres termes, ce fait nous laisse supposer qu'il est possible de déceler les maladies génétiques à partir de matières organiques.

D'un point de vue des causes, nous pouvons affirmer que les maladies génétiques proviennent d'un disfonctionnement dans la production de protéines, ou autre composants biologiques, dont la cause serait génétique.

Concernant les remèdes, certains d'entre eux permettent de combler le manque du au disfonctionnement génétique. Ceux-ci semble être efficaces, mais seulement à court terme. Cependant, nous pouvons constater qu'un autre type de remède, consistant à modifier directement le matériel génétique en vue de gérer le problème directement à la source du problème, existe. Il s'agit de la thérapie génique. D'après une source, recueillie sur le web[47], cette méthode serait basée sur le transfert de gènes. Concrètement, cette technique permettrait de modifier le matériel génétique d'une cellule en introduisant une séquence d'ADN contenue cellule cible. Cette manière de procéder ressemble fortement au mécanisme de mutation génétique liée à l'introduction d'un virus dans l'individu hôte. Ce procédé permet, comme nous le comprendrons, à instaurer des capacités nouvelles ou perdues dans ce génome hôte.

Selon la même source, cette technique ne peut être menée à terme que sous certaines conditions morales et éthiques. En effet, ce domaine d'activités semble être lié à des problèmes éthiques. En effet, ce fait semble évident puisqu'il consiste à modifier la nature humaine. Cependant, malgré ces contraintes et limites, la thérapie génique possèderait, comme domaine de prédilection, les maladies génétiques et plus particulièrement le cancer. Ce fait nous permet de renforcer le lien existant entre thérapie génique et le traitement des maladies génétiques.

---

[47] *Chapitre G: Thérapie génique : principes et applications en urologie*, http://urofrance.org/fileadmin/documents/data/PU/2000/PU-2000-00101021/TEXF-PU-2000-00101021.PDF, 2010, date de consultation : 17/08/2016

A titre d'informations, l'aboutissement à un processus de thérapie génique passe, en premier lieu, sur un modèle animal qui, étant capable de reproduire la maladie humaine étudiée, permet de vérifier les conséquences des modifications génétiques apportées.

Cette section nous aura, donc, permis de faire le point sur les premières conclusions que nous pouvions faire sur base des informations récoltées concernant les maladies génétiques. Elle nous aura, également, permis d'en savoir plus sur la technique de thérapie génique et, dès lors, de pouvoir faire le lien entre cette technique et les mécanisme induit dans les mécanismes de mutations causées par un virus détaillées à la première partie.

Notons, également, que cette dernière description, concernant la thérapie génique, nous permet de faire le lien entre cette notion et l'étude menée dans la deuxième partie. En effet, nous réaliser que l'étude génétique menée n'était, en réalité, qu'une étude génétique pouvant être utilisée lors de la thérapie génique. Ce fait nous semble important et nous permettra d'aboutir à des conclusions.

Application de l'outil

Dans cette section, nous allons traiter de la problématique liée à l'application de l'outil au milieu médical en tentant d'expliquer comment l'utiliser dans le cas des deux maladies génétiques précédemment explicitées. En effet, nous discuterons de la possible utilisation du pipeline, dans le but d'alléger le travail des scientifiques et des médecins, tant au point de vue du dépistage, qu'au point de vue du traitement.

Afin de prouver la pertinence de cet outil nous allons, d'une part, vérifier que les données s'y adaptent pour, ensuite, vérifier que l'ensemble des fonctionnalités, relatives au diagnostic ainsi qu'aux traitements, à appliquer peuvent être accomplies grâce à l'utilisation du pipeline.

Avant toute chose, notons que cette section ne se base que sur notre avis personnel. Nous n'en assurons, donc, nullement la pertinence et la cohérence. Cependant, ayant déjà travaillé dans l'aboutissement d'un projet génétique dans lequel nous avons utilisé cet outil, nous estimons avoir assez de connaissances pour pouvoir avancer nos propos.

D'une part, les maladies génétiques sont analysables, comme indique leur nom, à partir des gènes de l'individu. En effet, comme nous l'avons vu, dans la partie précédente, les scientifiques les décèlent sur base d'un taux de production de substances, une hormone par exemple. Or étant donné que la détection génétique de ces maladies peut se faire à partir de matières organiques, prises de sang ou échantillons organique, nous pouvons affirmer que nous pouvons obtenir les séquences relatives à ces maladies nécessaires aux différentes analyses. En effet, nous avons vu que le diabète pouvait se diagnostiquer à partir d'une prise de sang et que les cancers, quant à eux, pouvaient être décelés à partir d'échantillons des tumeurs à partir desquelles nous pourrions extraire le matériel génétique modifié.
Ce fait implique qu'il est, dès lors, possible traiter ces maladies génétiques grâce à notre outil. En d'autres termes, l'outil proposé est pertinent pour la détection de maladies génétiques.

D'autre part, les traitements de la plupart des maladies génétiques, comme vu précédemment, pourraient être élaborés en suivant les procédés de la thérapie génique. Ce qui nous prouve, de manière évidente, une fois de plus, que la pertinence de l'utilisation de l'outil.
En effet, les différents résultats pouvant être obtenus à partir du pipeline permettraient de vérifier les protéines produites par le génotype du patient et, par conséquent, en dériver un traitement

adapté. Ensuite, une fois le traitement administré, le pipeline pourrait être utilisé afin de nous assurer que le traitement permet au génotype du patient de produire les protéines adéquates. En d'autres termes, le pipeline servirait, en quelques sortes, de mécanisme de monitoring.

De plus, l'outil pourrait, également, être utile lors de recherches liées à l'obtention d'un remède (vaccins, etc.). En effet, des scientifiques pourraient tester des remèdes différents en les administrant à des cobayes, recevant chacun un remède différent. L'outil pourrait, ensuite, sur base des échantillons de sang prélevés des cobayes, classer les différents remèdes en utilisant les différentes étapes du pipeline sur les différents échantillons prélevés.

Tous ces cas d'utilisations utiliseraient les différentes étapes du pipeline. Celles-ci permettraient de créer différents graphiques permettant l'aboutissement de premières conclusions. Nous avons décidé, dans une volonté d'augmenter le caractère concret de notre raisonnement, nous avons décidé de traiter une expérience que l'on pourrait exécuter avec cet outil en expliquant chaque étape du processus. Ceci permettra au lecteur de mieux comprendre le but de chaque étape et, d'autre part, de prouver une fois de plus que le pipeline est pertinent dans ce domaine.
Pour ce faire, nous allons essayer de vérifier le fonctionnement d'un médicament pouvant être prescrit par un médecin dans le cas d'un traitement de patient souffrant de diabète. Le fonctionnement de ce médicament serait le suivant : le médicament fonctionnerait comme un traitement hormonal permettant aux cellules de produire des protéines permettant la synthétisation de l'hormone d'insuline, essentielle pour pouvoir stocker les glucides, mécanisme impossible à réaliser dans le cas du diabète. Le but de l'expérience serait de vérifie, d'un point de vue génétique, que le médicament fonctionne correctement.
Pour commencer notre expérience, nous prendrions deux prises de sang du patient : une dans son état naturel, appelée « A », et l'autre après la prise du médicament, appelée « B ». Ces échantillons sanguins nous permettraient d'obtenir les séquences d'ADN et d'ARN construisant le matériel génétique de l'individu (voir première partie sur l'extraction d matériel génétique à partir d'échantillon sanguin).
Une étape de pré-processing est alors nécessaire pour vérifier et corriger, si nécessaire, les séquences, afin d'obtenir, par la suite, des résultats plus pertinent. Le lecteur est invité à consulter la deuxième partie du travail pour en savoir plus sur les outils utilisés ainsi que sur les mécanismes de corrections.

En ce qui concerne le processing, nous allons commencer par l'étape d'alignement qui nous permettra d'aligner, sur le génome de référence humain, ce qui nous permettra de relocaliser, sur celui-ci, les séquences provenant des échantillons A et B. Ce qui, permettra de comprendre les gènes entrant en jeu lors de la prise du médicament.
La deuxième étape, permettant de définir l'expressivité génétique, permet de donner le nombre de reads, provenant des deux échantillons, alignés sur chaque gène du génome de référence. Ceci, nous permettra de comparer les deux échantillons d'un point de vue génétique. En effet, nous supposons, que plus de reads seront alignés sur les gènes permettant la création de protéines permettant de synthétiser l'hormone d'insuline.
A ce stade, déjà, nous serions dans la capacité de donner de premières conclusions quant au bon fonctionnement du médicament.

La troisième étape, étape de normalisation permettrait, comme expliqué dans la deuxième partie permettrait de nous assurer le moyen de mieux comparer les échantillons.
Après, cela, nous pourrions, dans le but d'améliorer la prise de décision, obtenir des graphiques permettant de visualiser les conséquences génétiques qu'implique ce genre de médicaments.

Comme nous pouvons le remarquer cette expérience fictive ressemble fortement à celle portant sur les conséquences de l'utilisation de l'hormone de Gibbérelline sur les plants de vigne présentée dans la deuxième partie. Ce qui prouve, en quelques sortes, la faisabilité de celle-ci. Malgré ce point positif, nous devons nous limiter qu'à une expérience fictive. En effet, pour pouvoir la mener à bien, nous aurions dû disposer du matériel nécessaire à son bon fonctionnement.

Pour conclure, ces explications permettent au lecteur de comprendre l'utilité de ce genre d'outil dans le domaine médical. En effet, celui-ci pourrait aider les médecins et scientifiques dans l'élaboration de conclusions relatives aux génotypes de patients.
Par exemple, l'obtention de résultats pourrait se produire de manière automatique. Ce qui permettrait de faciliter l'accomplissement des tâches. Nous allons donc prouver que son utilisation est pertinente.
Rappelons que cette section ne se base que sur notre avis personnel. Nous n'en assurons, donc, nullement la pertinence et la cohérence. En effet, cette partie a été rédigée sans aucune intervention biologique ou génétique de la part d'experts. Il va donc de soi que ceci est un exemple fictif ayant pour but de l'outil selon notre point de vue. Nous ne garantissons, dès lors, aucune garantie sur sa pertinence.

Afin de pouvoir donner plus de détails pertinents à notre raisonnement, nous avons décidé de passer en revue chaque étape du pipeline. En effet, nous jugeons pertinent d'expliquer les différents résultats que pourrait nous fournir chacune de ses étapes.

Discussion éthique

Tout comme dans le cas de l'utilisation dans le secteur agricole, ce domaine d'activités pourrait, selon nous, être sujet à de diverses controverses éthiques. L'application de ce genre d'outil à ce milieu possède, comme nous l'avons vu, des avantages sans précédents : aide pour les tâches de recherche, automatisation possibles, etc. ; mais également, des inconvénients majeurs liés à l'éthique.

En effet, nous pourrions aisément imaginer que, d'une part, l'automatisation de ce pipeline devienne utilisable par un plus grand nombre de personnes ; et d'autre part, que, comme le suggère Mme Gevers concernant l'évolution des inventions [Gevers d'Udekem-d'Acoz, année académique 2011-2012], ce pipeline pourrait servir de base à l'élaboration d'outils futurs plus élaborés qui permettraient d'obtenir des informations sur les défaillances génétiques des individus. Par défaillances génétiques, nous parlons des maladies génétiques au sens large, incluant, ainsi, les différents troubles de la personnalité héréditaires, tels que la schizophrénie et la bipolarité. Nous pouvons, aisément, comprendre le genre de discrimination que ce genre de données impliquerait dans le milieu professionnel par exemple.

De plus, étant donné que l'outil pourrait être fortement lié à des études traitant de la thérapie génique, nous devons tenir compte des problèmes éthiques liés à cette technique. En effet, la thérapie génique fait l'objet de multiples discussions éthiques, notamment sur les étdues incluant les manipulations génétiques embryonnaires.

Ceci nous permet, selon nous, de comprendre l'enjeu éthique dont ce genre de technologie pourrait faire l'objet ainsi que des probables précautions à envisager.

Discussion technologique

Suite aux explications de base, nous pouvons imaginer une certaine évolution fonctionnelle de l'outil proposé. En effet, nous pouvons, pour aller plus loin, utiliser cet outil dans un contexte plus novateur en lui intégrant des capacités de prédiction. Selon nous, ce genre d'améliorations serait innovant et apporterai une réelle plus-value à l'outil. Nous pourrions imaginer, par exemple, le cas dans lequel le machine learning serait couplé à ce pipeline dans le but de trouver, de manière automatique, une succession de modifications génétiques à appliquer au génotype du patient malade afin d'obtenir, en bout de course, un génotype sain. Pour ce faire, nous pourrions une technologie semblable à celle de la programmation par contrainte proposée par Prolog, par exemple. En effet, ce type de programmation permet, au moteur d'inférence, de trouver une combinaison respectant les règles introduites dans la base de connaissances. Nous pouvons, alors, imaginer le cas dans lequel ce moteur d'inférence contiendrait, dans sa base de connaissance les propriétés génétiques pouvant y être appliquées, le génome malade du patient, le génome de référence de l'être humain ainsi que le pipeline afin de déduire si une situation est saine ou non.

Cet exemple nous permet de comprendre que l'outil pourrait, par conséquent, être utile dans le milieu de la recherche médicale en aidant les scientifiques à trouver la meilleure manière d'appliquer la technique de thérapie génique au patient malade.

Une autre amélioration possible liée à la prédiction concernerait l'aspect préventif que présenterait l'outil. Par exemple, celui-ci pourrait donner fournir des informations concernant l'individu en le mettant en garde concernant les défaillances probables de son génotype. En effet, sur base des explications concernant la méiose (voir première partie) et la capacité que possède l'être humaine à produire une variété génétique lors de la reproduction (voir « crossing-over », première partie), à utiliser le machine learning afin de prédire la probabilité qu'un parent ne lègue un caractère néfaste. En effet, nous savons, grâce aux explications de la première partie, qu'un chromosome, dû au mécanisme de crossing-over, ne se lègue rarement entièrement. Ce fait implique qu'un trait néfaste, provenant d'un des deux parents, ne soit pas transmis alors qu'un autre gène, se situant sur le même chromosome, l'est. De plus, nous savons qu'au plus les locis (emplacements sur le chromosome) des traits sont éloignés sur le chromosome, plus la probabilité que ces deux traits ne soient pas transmis ensemble est grande.

Plus concrètement, la position d'un gène pourrait être obtenue en effectuant l'étape d'alignement. A partir de ce stade, nous pourrions utiliser ces données afin de pouvoir définir la probabilité de transmission d'un gène néfaste.

Sur base de ces explications, nous voyons clairement apparaître une utilisation possible de l'outil afin de définir la probabilité de chance qu'un des parents ne transmette une prédisposition génétique non souhaitée.

Cet exemple, prouve, encore une fois, l'avantage que possède l'utilisation d'un tel outil. Il pourrait donc être utilisé dans un but préventif afin de prévenir les dommages liés à la génétique d'un individu en diminuant la cause d'apparition de certains facteurs.

Par exemple, si nous savons qu'un enfant, dont le père est diabétique, possède 80% de chance de recevoir la prédisposition génétique favorable au développement du diabète, il nous est, alors, possible, en prévenant les parents, de conditionner son environnement pour que la maladie ne se manifeste le plus tard possible.

Nous pouvons également discuter de la capacité à pouvoir coupler le pipeline à d'autres outils utiles dans la détection de maladies. Par exemple, certains outil existent déjà afin de pouvoir classer les différentes tumeurs entre celles qui sont bénignes et malignes, ainsi que déterminer le niveau d'avancement d'un cancer. Nous imaginons donc qu'il serait intéressant, dans le but d'obtenir des résultats fiables, de pouvoir combiner ces outils.

En effet, selon nous, nous pourrions imaginer que celui-ci soit doté d'un composant y étant couplé afin de pouvoir extraire des informations relatives à des prédictions. Ces capacités pourraient, notamment, être utilisées dans le secteur de la recherche afin d'éviter d'effectuer tout traitement sur des cobayes.

Ces différents cas d'utilisations et exemples permettent de prouver la pertinence de l'utilisation d'un tel outil. En effet, les améliorations imaginées permettent d'ajouter des capacités supplémentaires à l'outil, ce qui rend son utilisation encore plus pertinente dans le domaine.

Conclusion

Cette partie nous aura permis de pouvoir obtenir des éléments permettant de constituer une réponse à la question de recherche posée en présentant l'utilisation du pipeline d'extraction de données dans le cas de deux maladies génétiques.

Pour ce faire, nous avons, dans un premier temps, défini les concepts de maladies génétiques en fournissant deux cas concrets, à savoir, celui du diabète et du cancer. Notons que le fonctionnement des autres maladies génétiques est fortement semblable. Pour chacune d'entre elles, nous avons donné des explications relatives à leur mode de fonctionnement, leurs méthodes de diagnostic ainsi que les traitements étant liés.

Notons que la transmission génétique n'est pas le seul facteur prise en compte lors de a détermination de traits de caractère d'un individu. En effet, l'environnement est aussi à prendre en compte. Cela nous laisse une once d'espoir quant à l'apparition de celles-ci, en nous faisant comprendre que ces maladies peuvent être évitées, du moins retardé. Nous comprenons, dès lors l'importance que possèderait un outil capable de prédire les maladies pouvant émerger de notre matériel génétique. C'est la raison pour laquelle, l'aspect de prévention a été élaboré dans cette section.

Ensuite, nous avons eu l'opportunité de décrire brièvement en quoi la technique de thérapie génique consistait. Ces informations et exemples, que nous avons traités, nous ont permis de constater que cette technique était innovante et que celle-ci avait pris place au centre de l'intérêt scientifique. Cette technique nous semble, donc, selon nous, être l'avenir en ce qui concerne la lutte contre les maladies génétiques. Ce qui nous a permis de renforcer le lien entre cette technique et les maladies génétiques.

Cette section aura, également, présenté l'opportunité de discuter d'éventuelles amélioration à apporter à l'outil. En effet, outre l'aspect de calculabilité fiable proposé par l'outil informatique, nous avons imaginé quelques cas d'utilisations dans lesquels l'outil permettrait de présenter d'autres qualités tout aussi importantes. Par exemple, nous avons traité de la capacité de prédiction concernant les réactions génétiques du génotype d'un individu, atteint d'une maladie génétique, ainsi que de la probabilité de chance, que possèderait un parent, de transmettre une prédisposition génétique défavorable à un enfant.

Certains cas de couplage avec d'autres outils ont également été étudiés. Citons, par exemple, le cas concernant le couplage avec un autre outil, permettant de détecter les tumeurs cancéreuses, pourrait s'avérer être d'une grande utilité.

Ce genre de réflexions nous aurons permis de réfléchir à des solutions rendant, ainsi, l'accès, aux résultats et conclusions, facilité.

Le lecteur aura pu prendre conscience du fait que cette partie prouve, d'une part, que l'outil est utilisable dans le domaine médical ainsi que, d'autre part, l'existence des multiples améliorations possibles permettant d'appuyer notre point de vue lié à l'utilisation de l'outil dans ce secteur, et ce, tant au point de vue de la prévention, que du traitement et que de la technique.

Conclusion générale

Ce travail nous aura permis d'en savoir plus sur le domaine de la bio-informatique liée au domaine médical. En effet, il nous permet, à ce stade, de donner une réponse à la question de recherche : « En quoi consiste la pertinence liée à l'utilisation d'un pipeline d'extraction de données génétiques dans le milieu médical ? ». Le but de ce travail était, donc, de définir la pertinence que possédait l'utilisation d'un tel outil dans le milieu médical.

Pour ce faire, nous avons, tout d'abord, dû donner les explications relatives aux notions biologiques et bio-informatiques de base. Cette compréhension nous aura permis, ensuite, de passer aux explications relatives à la construction d'un pipeline d'extraction de données à partir de séquences d'ARN en le présentant dans un cas concret d'utilisation, à savoir les conséquences liées à l'utilisation de l'hormone de Gibbérelline sur les plants de raisins. Toutes ces explications de bases nous ont permis, dans la troisième partie, de présenter l'application de cet outil au domaine médical en présentant le génome humain, quelques maladies génétiques, le diabète et le cancer, ainsi que les avantages que possèderait l'utilisation de ce genre d'outil à ce domaine.

Une des premières conclusions pouvant être présentée concerne le fait que ce genre d'outil puisse directement être utilisé dans le cas d'étude portant sur la thérapie génique. En effet, comme dit précédemment dans la conclusion de la deuxième partie, l'étude présentée consistait en une étude incluant la technique de thérapie génique. Or, nous avons pu observer la pertinence de ce genre d'outils dans cette étude, nous aboutissons au fait que l'utilisation de l'outil soit pertinente dans ce contexte.
D'autre part, nous avons découvert, grâce à la troisième partie que la thérapie génique était pertinente dans l'étude des maladies génétiques.
Nous obtenons, par conséquent, le fait que l'outil puisse être utilisé dans le cas de l'étude de maladies génétiques.
Ce fait répond, partiellement, à la question de recherche, mais nécessite les conclusions, liées aux fonctionnalités, proposées ci-après.

En plus de cette première réponse, nous pouvons ajouter que ce travail nous aura permis d'obtenir un cheminement de pensées permettant l'élaboration d'une réponse plus élaborée tenant compte des éventuelles évolutions. Celui-ci s'articule autour des différents avantages que pourrait

posséder l'utilisation de ce genre d'outils dans ce milieu. Parmi ceux-ci nous proposons : l'accès aux conclusions facilité, l'aide de prises de décision, l'aspect technique ainsi que l'aspect préventif.

Le premier point, concerne l'aspect pratique et réside dans le fait que l'outil permet de rendre l'accès aux conclusions plus facilement. En effet, le pipeline peut être utilisé plus facilement que d'autres outils et processus permettant l'analyse génétique puisque, pour l'exécuter, nous ne devons disposer que du matériel génétique ainsi que d'un simple ordinateur.
Nous pouvons, dès lors, imaginer qu'un médecin pourrait, s'il existait un dispositif permettant de d'extraire directement le matériel génétique d'un patient, analyser le sang d'un patient directement dans son cabinet sans devoir attendre le délai de temps nécessaire à l'analyse sanguine par laboratoire. Ce qui permettrait au médecin de directement traiter le patient de manière adéquate en le redirigeant, par exemple, chez un spécialiste.

Le second point, quant à lui, concerne l'aspect de support d'aide à la prise de décision et consistant à aider les scientifiques et médecins dans l'élaboration de conclusions concernant le bilan sanguin d'un patient. En effet, l'outil pourrait être couplé à des algorithmes de machine learning permettant de définir les situations dangereuses ou non. Ce qui pourrait avertir les médecins sur des situations nécessitant un intérêt particulier.

Le troisième point concerne l'aspect purement technique. En effet, l'outil est déployable sur du matériel disposant de la technologie informatique. Ce qui implique, alors, que l'outil dispose de toute l'expressivité et capacités calculatoires inclues les technologies informatiques actuelles. Ce fait rend, donc, le pipeline capable d'utiliser des ressources importantes nécessaires au traitement du grand nombre de données qu'il est obligatoire de manipuler afin d'obtenir des résultats.

Le quatrième aspect concerne l'aspect préventif que possède cet outil. En effet, comme nous l'avons vu, certains cas d'utilisation pourraient informer l'apparition de maladies génétiques. Or, sur base du fait que l'apparition des maladies génétiques peut être également contourner grâce au contexte. Ce fait nous rassure et nous motive à apporter une contribution concernant l'aspect préventif que représente l'outil.

Par exemple, nous pourrions étudier, sur base des principes de méioses et d'enchâssement (voir première partie), la probabilité de chance qu'un individu puisse hériter d'une prédisposition génétique favorable à la venue de problème médicaux provenant d'un de ses parents. Permettant, ainsi, aux proches de l'enfant de le confiner dans un environnement ne favorisant pas le développement de la maladie.

Nous remarquons, par cet exemple, de l'impact positif qu'aurait cette caractéristique le dans le domaine de la détection de maladies génétiques, et, par conséquent, sur le confort de vie des individus.

Grâce à ces aspects positifs, nous pouvons, donc, selon nous affirmer que cet outil est utile dans le milieu médicale. Selon nous, ceci pourrait réellement être utile puisque dans certains cas de maladies, ne sont pas facilement décelables. En effet, étant donné que certaines maladies possèdent les mêmes symptômes, les médecins ont parfois du mal à pouvoir les diagnostiques sans d'examen approfondis. Outre la haute puissance de calculs, due à la nature informatique de l'outil, pouvant être mis à disposition, l'accès à des conclusions facilités ainsi que l'aide à la prise de décision permettraient à pipeline de (constituer une aide majeure dans) contribuer, de manière non-négligeable, à l'épanouissement du domaine médical. Ces faits traitent d'améliorations importantes dans le domaine, mais sont tout à fait concevables.

Ce travail nous aura, également, permis de discuter d'autres technologies et outils disponibles ayant pour but d'améliorer l'accomplissement de tâches liées aux maladies génétiques. En effet, nous avons vu que des outils permettent de trier les images médicales représentant les tumeurs en tumeurs cancéreuses et non-cancéreuses.

Outre l'aspect purement technologique, nous avons également étudié l'aspect éthique de ce genre d'outil. Comme toute avancée technologique, nous ne pouvons pas écarter l'aspect éthique. En effet, ce genre d'outil pourrait aider, dans des cas, des personnes malveillantes à appuyer leurs idéologies et permettrait de contribuer à une certaine catégorisation de personnes. Par exemple, comme dit précédemment, certains comportements, tels que la schizophrénie et la bipolarité, seraient d'origine génétique. Nous comprenons, dès lors, qu'un tel outil pourrait aider les médecins à définir la nature d'une personne en découvrant si elle est bipolaire ou non, per exemple. Ce genre

d'informations pourrait, alors être réutilisé lors d'entretiens d'embauche. Nous voyons, donc, clairement le problème que cela pourrait poser.

En ce qui concerne les points critiques de notre travail, nous pensons qu'il aurait été intéressant de pouvoir mener une réelle expérience sur le diabète à la place de la traiter de manière conceptuelle. En effet, pour améliorer ce travail, nous pensons qu'il aurait été intéressant de créer une brève étude, sur le diabète par exemple, afin de pouvoir utiliser le pipeline et d'obtenir des résultats concrets dans le domaine. Ces expériences n'auraient servis qu'à consolider la connaissance déjà connue, à ce jour, dans le but de pouvoir démontrer que le pipeline puisse, également, aboutir à des résultats déjà vérifiés. Ceci, nous aurait, donc, permis de renforcer la pertinence ainsi que la crédibilité de l'outil et de son utilisation dans ce domaine. Malheureusement, l'accomplissement de ce genre d'études nécessite l'utilisation de matériel adéquat, notamment pour extraire les données génétiques, et également l'obtention de ressources humaines, sujets atteints de maladies génétiques. Nous comprenons, dès lors, qu'il nous est impossible de réaliser ce genre d'expériences.

Pour conclure, nous pouvons affirmer que ce genre d'outils est bel et pertinent dans le milieu médical, puisqu'il permet l'amélioration de certains facteurs importants, tels que : l'accès aux conclusions, l'aide de prises de décision, l'aspect technique ainsi que la prévention.

Nous espérons avoir été assez clairs, précis et complets dans les explications et raisonnements qui nous ont permis d'aboutir à l'élaboration de notre réponse à la question de recherche. Nous espérons, en effet, avoir traité la problématique avec un rayon d'action suffisamment large. En plus de l'élaboration de la réponse, nous avons eu l'opportunité de produire un état de l'art concernant les notions biologiques et bio-informatiques utiles à la compréhension du domaine. Nous espérons, ainsi, pouvoir apporter une contribution dans l'étude de sujets de recherches à venir dans le domaine en proposant un état de l'art pouvant aider.

Nous terminerons, donc, en notant, qu'une fois de plus, l'intégration de l'outil informatique dans un domaine d'activité s'est révélée être bénéfique pour l'épanouissement de celui-ci. Prouvant, alors, l'existence de points bénéfiques, inclues dans cette discipline, et, par conséquent, l'importance que possède le développement de ces technologies.

Références

Jork, Carey, Bamshad, *Medical genetics* (Fourth Edition), StudentConsult, USA, 2010

Karp, *Biologie Cellulaire et moléculaire* (3ième édition), deBoeck, 2002

Jonathan Pevsner, *Bioinformatics and functional genomics* (2nd edition), Whiley-Blackwell, 2009

Raven, Jonhson, Losos et Singer, *Biologie* (7ième edition) , de Boeck, 2007

King-Hao Liang, *Bioinformatics for biomedical science and clinical application*, WoodHead Publishing, 2013

Hongxin Fan, Margaret L. Gulley, *DNA Extraction from Fresh or Frozen Tissues*, in *Methods in Molecular Medicine*, *vol. 49: Molecular Pathology Protocols*, A.A. Killeen Humana Press Inc., Totowa, NJ, USA, 2001

Salah M. Aljanabi , Iciar Martinez , *Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques*, in *Nucleic Acids Research*, pages Vol. 25, No.22 - Oxford University Press, 1997

Luis Torgo, *Data Mining with R: Learning with Case Studies*, p.3, CRC Press, 2011

Liao Y, Smyth GK and Shi W., *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote*, Nucleic Acids Research, 2013, URLhttp://www.ncbi.nlm.nih.gov/pubmed/23558742

R Development Core Team (2008), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Kristin Potter, *Methods for Presenting Statistical Information: The Box Plot*, University of Utah School of Computing Salt Lake City, UT, 2006

Simon Anders, Wolfgang Huber, *Differential expression of RNA-Seq data at the gene level – the DESeq package*, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, 2016

Antonio Granell, Juan Carbonell, *Les hormones végétales*, Pour la science, n°228, octobre 1996

J. TOSTAIN, D. ROSSI, P.M MARTIN, *Physiologie des androgènes chez l'homme adulte*, Physiologie des Androgènes : de l'Homme Adulte à l'Homme Vieillissant, p. 639-660, Marseille, 2004

Chan Jin Jung, Youn Young Hur, Sung-Min Jung, Jung-Ho Noh, Gyung-Ran Do, Seo-June Park, Jong-Chul Nam, Kyo-Sun Park, Hae-Sung Hwang, Doil Choi, Hee Jae Lee, *Transcriptional changes of gibberellin oxidase genes in grapevines with or without gibberellin application during inflorescence development,* The Botanical Society of Japan and Springer Japan, p.359-371, 2013

Atiako Kwame Acheampong, Jianhong Hu, Ariel Rotman, Chuanlin Zheng, Tamar Halaly, Yumiko Takebayashi, Yusuke Jikumaru, Yuji Kamiya, Amnon Lichter, Tai-Ping Sun and Etti Or, *Functional characterization and developmental expression profiling of gibberellin signalling components in Vitis vinifera*, Journal of Experimental Botany, 2014

Lisa Giacomelli, Omar Rota-Stabelli, Domenico Masuero, Atiako Kwame Acheampong, Marco Moretto, Lorenzo Caputi, Urska Vrhovsek and Claudio Moser, *Gibberellin metabolism in Vitis vinifera L. during bloom and fruit-set: functional characterization and evolution of grapevine gibberellin oxidase*, Journal of Experimental Botany, 2013

Marie Gervers-d'Udekem d'Acoz, *Anthropologie de l'informatique*, Presses universitaires, FUNDP (UNamur), Namur, 2011-2012

Robert A. Weinberg, *The biology of cancer (second edition)*, Garland Science, USA, 2014

T. A. Brown, *Génomes*, Médecine-Sciences, Flammarion, 2004

Ricki Lewis, *Human Genetics : Concepts and Applications (7ème édition)*, McGraw-Hill International Edition, 2007

Rebecca L. Siegel, Kimberly D. Miller, Ahmedin Jemal, *Cancer Statistics, 2015*, A cancer Journal for Clinicians, USA, 2015

A. Rousseau, P. Bohet, J. Merlière, H. Treppoz, B. Heules-Bernin, R. Ancelle-Park, *Evaluation du dépistage organisé et du dépistage individuel du cancer du col de l'utérus : utilité des données de l'Assurance maladie*, Ministère de l'Emploi et de la Solidarité, INSTITUT DE VEILLE SANITAIRE, France, 2002

Pr. A. Grimaldi, *Diabétologie : Questions d'internat*, Université Pierre et Marie Curie, Faculté de Médecine, France, 2000

Emily Loghmani, Jamie Stang, Mary Story, Chapitre 14: *DIABETES MELLITIS: TYPE 1 AND TYPE 2*, in *GUIDELINES FOR ADOLESCENT NUTRITION SERVICES* , p. 167-182, School of Public Health University of Minnesota, Mineapolis, 2005

Cancer Burden in Belgium 2004-2013, Belgian Cancer Registry, Bruxelles, 2015

Alicia Oshlack, Mark D Robinson, Matthew D Young, *From RNA-seq reads to differential expression results*, Genome Biology 2010, BioMed Central, 2010

Jimmy C. Azar, Martin Simonsson, Ewert Bengtsson, and Anders Hast, *Automated Classification of Glandular Tissue by Statistical Proximity Sampling*, Hindawi Publishing Corporation International Journal of Biomedical Imaging Volume 2015, Centre for Image Analysis, Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden, 2014

Scott Doyle, Michael D Feldman, Natalie Shih, John Tomaszewski and Anant Madabhushi, Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer, BioMed Central (BMC) Bioinformatics, 2012

Ali Tabesh, Vinay P. Kumar, Ho-Yuen Pang, David Verbel, Angeliki Kotsianti, Mikhail Teverovskiy, and Olivier Saidi, *Automated Prostate Cancer Diagnosis and Gleason Grading of Tissue Microarrays*, Proc. of SPIE Vol. 5747, p. 58-70

Nabih A Baeshen, Mohammed N Baeshen, Abdullah Sheikh, Roop S Bora, Mohamed Morsi M Ahmed, Hassan A I Ramadan, Kulvinder Singh Saini, Elrashdy M Redwan, *Cell factories for insulin production*, BioMed Central, Microbial Cell Factories, 2014

Madan Babu, Chapitre 11 : *An Introduction to Microarray Data Analysis*, in *Microarray Data Analysis*, 2004

Pascal Ferré, *Action et sécrétion de l'insuline : Double jeu pour les canaux potassiques*, Centre de Recherches Biomedicales des Cordeliers, Université Pierre et Marie Curie, 2005

Peter Fenwick, *Burrows Wheeler Compression: Principles and Reflection*, in Theoritical Computer Sciences, vol. 387, p200-219, Department of Computer Science, The University of Auckland, Private Bag 92019, Auckland, New Zealand, 2007

Chapitre G: Thérapie génique : principes et applications en urologie, http://urofrance.org/fileadmin/documents/data/PU/2000/PU-2000-00101021/TEXF-PU-2000-00101021.PDF, 2010, date de consultation : 17/08/2016

Comprendre la chimiothérapie, Institut National du Cancer, Belgique, 2009

Monique Le Gen, *La boîte à moustaches de TUKEY : un outil pour initier à la Statistique*, CNRS-MATISSE

Blast documentation, http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml , 2007, date de consultation : 23/07/2016

fqtrim: trimming & filtering of next gen reads, http://ccb.jhu.edu/software/fqtrim/, date de consultation: 17/08/2016

Bioconductor open source software for bioinformatics, https://www.bioconductor.org/about/, 2016, date de consultation : 20/07/2016

SAMtools, http://samtools.sourceforge.net/, 2012, date de publication : 20/07/2016

Simon Anders, *Package 'DESeq'*, EMBL Heidelberg, Allemagne, 2016

Sequence Alignment/Map Format Specification, The SAM/BAM Format Speci cation Working Group, https://samtools.github.io/hts-specs/SAMv1.pdf, Novembre, 2015, date de consultation: 17/08/2016

Primrose, Twyman, Old, *Principes de génie génétique*, de boeck, 2004

Monographie de produit : GLUCOPHAGE®(chlorhydrate de metformine), http://products.sanofi.ca/fr/glucophage.pdf, 2009, date de consultation : 17/08/2016

Martin Morgan, *Bioconductor Annual Report*, Roswell Park Cancer Institute, 2016