Projet de Génie Logiciel :

Élaboration d'une interface permettant l'utilisation centralisée de multiples outils issus des nouvelles générations de séquençage.

Cahier des charges

Auteurs :

Denet Lola
Desquerre Émilie
Hui Tongyuxuan
Meguerditchian Caroline
Niu Wenli
Pratx Julie
Vu Thao Uyen

Étudiants en Master 2 de Bio-informatique
Université de Bordeaux

15 Octobre 2020

Table des matières

Introduction

Ce projet s'inspire du programme "Les sentinelles du climat" porté en Nouvelle Aquitaine [1]. Ce programme a pour but l'étude des changements climatiques et de la biodiversité dans différents milieux. Pour chaque milieu, des espèces sont étudiées durant 6 ans afin d'évaluer l'impact des changements climatiques sur la faune et la flore. Ces espèces sont appelées les sentinelles du climat. Leurs comportements et répartitions dans les différentes zones, permettra l'identification des espèces susceptibles de s'adapter, de migrer ou de disparaître et ainsi d'en mesurer l'impact sur l'environnement et la biodiversité de la région [2].

La Nouvelle Aquitaine est particulièrement impactée par le réchauffement climatique, en grande partie à cause des kilomètres de côtes qui la composent. D'autre part, la région bénéficie d'une diversité importante de milieux : les dunes atlantiques, les hêtraies de plaine, les zones humides, les pelouses calcioles, les torrents des montagnes, le pelouses et rocailles des montagnes. Cette diversité de milieux implique une diversité d'espèces permettant une étude approfondie de l'impact des changements climatiques sur la biodiversité.

Les zones humides [3] ont une importance toute particulière car elles abritent une plus grande diversité de sentinelles, c'est-à-dire une plus grande quantité d'espèces susceptibles d'être impactées par les changements climatiques car ces zones sont plus fraîches. Ainsi, l'augmentation des températures dans ces zones pourrait avoir des effets catastrophiques sur les espèces qui y vivent car elles pourraient avoir plus de difficultés à s'adapter ou à trouver un nouvel habitat propice à leur développement et leur survie.

Six espèces ont été identifiées comme sentinelles du climat dans les zones humides dont une qui a été ajoutée très récemment : la vipère péliade. La libellule, la rainette ibérique et rainette verte, l'azurée des mouillères ainsi que le lézard vivipare sont les autres sentinelles identifiées dans ce milieu [4].

Ce projet s'inclut dans ce contexte de recherche et propose d'étudier la vipère péliade. Comme indiqué précédemment, cette espèce a été identifiée très récemment en tant que sentinelle des zones humides car elle est difficile à trouver. Elle est susceptible d'être particulièrement impactée par le réchauffement climatique car sa capacité à s'adapter aux chaleurs en raison de sa physiologie serait limitée. De plus, elle ne se déplace que peu chaque année ce qui pourrait être un frein pour une éventuelle migration. C'est une espèce protégée susceptible de s'éteindre dans la région. L'étude menée par le programme "Les sentinelles du climat" va permettre de vérifier ces hypothèses et de prévoir leur impact sur la biodiversité.

Son inclusion récente dans l'étude et les aspects présentés précédemment en font un sujet d'étude pertinent pour ce projet. En effet, peu d'informations ont été recueillies jusqu'à ce jour dans le cadre de cette étude. Il paraît alors intéressant de collecter des informations sur cette espèce et de créer un outil qui permettra d'en faciliter le traitement.

Ce travail a pour but de créer une interface qui permettrait de centraliser plusieurs outils issus des nouvelles générations de séquençage. En effet, les recherches menées sur la vipère péliade pourront permettre d'identifier un groupe de gènes particulièrement étudié par la communauté scientifique. Un autre aspect de ce travail consistera en l'identification d'outils pertinents pour l'étude de ce groupe de gènes afin d'aboutir à l'élaboration d'un arbre phylogénétique.

La bioinformatique sera un atout primordial dans ce projet car elle permettra de faciliter le travail des chercheurs par la centralisation d'outils complexes sur une même plate-forme. En effet, l'une des difficultés de la recherche actuelle est d'identifier les outils les plus pertinents pour un domaine particulier. La multiplicité des plate-formes qui ralentit l'aboutissement au résultat est un autre exemple de difficulté rencontrée par les chercheurs. Il paraît essentiel de pouvoir centraliser tout un *pipeline* en un même outil afin de rendre plus efficace l'étude du groupe de gêne identifié. De plus, la création de cette interface pourra permettre dans le même temps de collecter des informations sur l'une des espèces sentinelle du

1. `https://www.sentinelles-climat.org`
2. `https://www.sentinelles-climat.org/a-propos/`
3. `https://www.sentinelles-climat.org/milieu/zones-humides/`
4. `https://technique.sentinelles-climat.org/developpement-dindicateurs-sentinelles-climat/`

climat : la vipère péliade.

Dans ce cahier des charges, une analyse du sujet sera tout d'abord réalisée à travers un état de l'art biologique et bioinformatique ce qui permettra d'aboutir à l'identification précise des objectifs de ce travail. Dans un second temps, une analyse des besoins fonctionnels et non fonctionnels sera réalisée afin de préciser les attentes potentielles concernant cette problématique. Enfin, des propositions de réalisation seront présentées. Des engagements et des améliorations potentielles pourront être proposés afin de produire un outil efficace et adapté.

Chapitre 1

Analyse du sujet

1.1 État de l'art

1.1.1 État de l'art biologique

Présentation de l'espèce

Vipera berus, la vipère péliade représentée dans la figure 1.1 [1], est une espèce de serpent venimeux de la famille des *Viperidae*. Son habitat se trouve en Europe et en Asie, dans les zones humides, les tourbières et les forêts. Selon la zone dans laquelle se trouve l'individu, la coloration varie du gris, bleu-gris, marron, vert-brun au noir.

FIGURE 1.1 – Vipère péliade

Cette espèce se nourrit de grenouilles ainsi que de petits rongeurs. Les plus jeunes mangent aussi des lézards. Ces derniers ainsi que les grenouilles sont également des espèces sentinelles. La période d'activité de la vipère diffère en fonction de l'altitude et du climat (de mars-mai à septembre-novembre)[Losange, 2008].

Un bocage dégradé, une biologie exigeante, le réchauffement climatique, l'impopularité, tout cela fait que la vipère péliade est, aujourd'hui, une espèce menacée vulnérable (selon l'union internationale pour la conservation de la nature) [CREUX, 2019].

Il est intéressant d'étudier cette espèce pour les différentes propriétés médicales de son venin mais aussi pour sa capacité à réguler les espèces sentinelles déjà introduites. En effet, le venin de *Vipera berus* a principalement une activité hémotoxique. Les hémotoxines peuvent être classées en fonction de leur effet sur l'activation de facteurs de coagulation sanguine, sur des agents anticoagulants, sur des inhibiteurs ou activateurs de plaquettes, sur des agents affectant la fibrinolyse ainsi que sur les hémorragines [Phillips *et al.*, 2010].

La vipère péliade est un serpent des plus dangereux à cause de son venin mortel chez l'Homme si la prise en charge n'est pas rapide. Cependant, celui-ci est composé de nombreux composants protéiques qui sont actuellement testés pour leur utilité dans le traitement de nombreuses maladies (neurologiques, cardiovasculaires, cancers) [Bocian *et al.*, 2016] ainsi que pour la production de sérum antivenimeux pour

1. `https://biodiversite.parc-naturel-pilat.fr/espece/78141`

le traitement des morsures [Netgen, 2011].

D'après Al-Shekhadat et al., *Vipera berus* est une espèce médicalement importante [Al-Shekhadat *et al.*, 2019]. L'équipe scientifique de Malina et al. s'est intéressée aux types de dégradation des protéines et ceci pour des individus d'âge et de sexe différents, mais aussi, sur l'effet paralysant du venin. Les échantillons de venin avaient des effets neuromusculaires paralysants variables [Malina *et al.*, 2017].

Une étude réalisée par l'équipe de Czajka et al., traite des situations dans lesquelles l'administration d'antivenin est nécessaire pour neutraliser les propriétés toxiques du venin et ses effets indésirables [Czajka *et al.*, 2013]. L'antivenin ViperaTAb a été utilisé par l'équipe de Hamilton et al. car les morsures sont une urgence médicale rare au Royaume-Uni, avec 20 à 50% des 50 à 200 cas estimés par an, nécessitant un traitement avec un antivenin. Ainsi ils ont démontré que lorsque les effets sont systémiques et locaux, ils nécessitent une réanimation hors hôpital, un soutien vasopresseur et une rééducation prolongée [Hamilton *et al.*, 2019].

L'étude de Hermansen et al. a pour but de présenter une vaste série de cas consécutifs de patients mordus par *Vipera berus*, et d'identifier les signes et symptômes indiquant une maladie compliquée[Hermansen *et al.*, 2019]. Malina et al. ont étudié les effets que provoque une morsure chez l'Homme. Le patient a développé des troubles nerveux crâniens sans ambiguïté, se manifestant par une atteinte bilatérale caractérisée par une paralysie oculomotrice avec ptose partielle (paupière qui tombe), parésie du regard et diplopie. La somnolence et la photophobie étaient ses symptômes supplémentaires [Malina *et al.*, 2013].

Composition protéique du venin

L'étude, qui a été réalisée par Bocian et al., a permis de recueillir du venin à partir d'individus mâles et femelles adultes [Bocian *et al.*, 2016]. Tandis que celle de Al-Shekhadat et al. a étudié les activités toxiques et enzymatiques et a déterminé la composition protéomique de son venin. Cette étude a également été conçue pour évaluer l'efficacité préclinique *in vivo* et *in vitro* de l'antivenin russe Microgen pour neutraliser les principaux effets du venin [Al-Shekhadat *et al.*, 2019].

Guillemin et al. ont utilisé une méthode basée sur la PCR pour déterminer les séquences d'ADN génomique codant pour les phospholipases A2 (présentent à 59% [Bocian *et al.*, 2016]) à partir des venins de différentes espèces de vipère dont *Vipera berus*.

Ce travail portera sur ces 3 séquences : AY158636 (phospholipase A2), AY158639 (ammodytin I2), AY159811 (ammodytin I1) où l'ammodytin est une protéine variante de la phospholipase A2 [Guillemin, 2003].

1.1.2 État de l'art bio-informatique

Bases de données biologiques

Bien que le séquençage de l'ADN et des protéines ait lieu dans le monde entier, presque toutes les données collectées sont stockées et partagées dans un certain nombre de grandes banques généralistes tels que :

- NCBI (*National Center for Biotechnology Information*[2]) : est un référentiel de ressources médicales et biotechnologiques. Les principales bases de données du NCBI comprennent Genbank (des informations sur les séquences d'ADN) et PubMed (une base de données bibliographiques en biologie et en médecine). Structurellement, Genbank se compose de deux grandes sections distinctes : la base de données de protéines et la base de données de nucléotides, dans laquelle la base de données de nucléotides est utilisée comme un chemin d'accès aux données protéiques respectives.
- EBI (*European Bioinformatics Institute*[3]) : Cette base de données est organisée et gérée dans environ 80 domaines différents, dont le plus important se concentre sur 3 domaines : la base de données des structures d'ADN, la base de données des structures protéiques (TrEMBL et SWISS-PROT) et la base de données des structures macromoléculaires (EBI-MSD).

2. `https://www.ncbi.nlm.nih.gov/`
3. `https://www.ebi.ac.uk/`

— DDBJ (*DNA Data Bank of Japan*[4]) : est une base de données sous la gestion de JNIG (*Japan National Institute of Genetics*). Comme Genbank et EMBL, DDBJ recueille des données sur les séquences de nucléotides et fournit des données de séquences de nucléotides et un système de supercalculateur disponibles gratuitement pour soutenir les activités de recherche en sciences de la vie.

Les trois bases de données Genbank (NCBI), EMBL et DDBJ effectuent une connexion directe et échangent quotidiennement des informations, de sorte qu'elles possèdent toutes les informations l'une de l'autre. La collaboration entre ces trois banques les aide par conséquent à se développer et devenir les plus grandes bases de données du monde.

Avantages et Inconvénients des banques généralistes

1. Avantages :

En général, les banques généralistes offrent de nombreux avantages à la recherche scientifique, en particulier dans le cadre de l'analyse des séquences. Ce sont des outils indispensables à la diffusion rapide des résultats scientifiques. Elles présentent complètement toutes les données, en double publication depuis 1988 mais centralisées en un seul ensemble de données depuis le début des années 2000, cela est important pour la recherche de similitudes avec une nouvelle séquence. D'autre part, la représentation d'une grande diversité d'organismes dans les bases de données généralistes rend possible des analyses phylogénétiques. Un autre bonus de ces bases réside dans l'information qui accompagne les séquences (annotations, expertise, bibliographie) qui peuvent parfois constituer les rares annotations disponibles sur certaines séquences. Enfin, l'inter-référence à d'autres bases de données permet d'avoir accès à d'autres informations non répertoriées.

2. Inconvénients :

Contrairement aux avantages ci-dessus, les bases de données généralistes ont aussi des inconvénients. D'abord, c'est le manque de contrôle des données soumises ou saisies surtout pour les séquences anciennes (pas d'expertise). Ensuite, la variabilité de l'état de connaissances sur les séquences est aussi une lacune des bases de données. La détermination des caractéristiques biologiques et des fonctions des séquences demande un travail expérimental et une analyse qui doivent se surajouter à l'étape automatisée et systématique du séquençage. Un autre défaut vient des erreurs dans les séquences. La cause est dérivée de la contamination du fragment original due à la technologie ou à la méthodologie. En outre, le dernier désavantage qu'on ne peut pas ignorer est le biais d'échantillonnage : le biais d'échantillonnage taxonomique (les séquences ont été extraites à partir des organismes qui ont une quantité inégale), le biais d'échantillonnage des séquences (les gènes des génomes étudiés sont inégalement représentés dans chacun d'eux) et la redondance des données (certains gènes extrêmement similaires correspondent à des entrées différentes dans la banque et c'est difficile de savoir si cela concerne au polymorphisme génétique ou à la duplication des gènes ou tout simplement aux erreurs lors de la détermination des séquences).

Alignement des séquences

Après avoir rassemblé les séquences choisies, il est essentiel de les traiter de façon à pouvoir déterminer leur relation. Pour cela, il faut réaliser un alignement qui servira par la suite à créer des arbres phylogénétiques.

L'alignement de séquences a pour but d'arranger les séquences de façon à mettre en valeur leurs similarités. Cela permet notamment de détecter des relations fonctionnelles ou évolutives entre ces séquences. De nombreux outils existent permettant de générer un alignement de séquences à partir d'un certain nombre de séquences d'ADN.

Il existe deux types de mesure de similarité : l'alignement global et l'alignement local. L'alignement global permet de regarder l'alignement des séquences dans leur entièreté ; il est conçu pour comparer des séquences homologues. Par conséquent il est possible d'y retrouver des zones où une similarité faible sera observée. L'alignement local lui, s'intéresse seulement aux régions relativement bien conservée. Il compare uniquement les régions similaires d'une séquence. Il existe plusieurs outils permettant d'effectuer des alignements locaux.

— Blast (*Basic Local Alignement Search Tool*) est une méthode de recherche heuristique qui se base sur l'alignement local de séquence par paires. Il permet la recherche dans une banque de données de séquences similaires. Son fonctionnement est le suivant : il décompose tout d'abord la séquence en segments (appelé k-uplets) qui se chevauchent. Puis chacun de ces k-uplets est comparé aux séquences cibles définies par l'utilisateur. Quand une homologie est détectée pour un segment, Blast étend l'alignement en amont et en aval pour chercher une zone de similarité plus étendue. Une fois

4. `https://www.ddbj.nig.ac.jp/index-e.html`

cette étape effectuée, il évalue la probabilité que la similarité soit due au hasard (*E-value*)[Altschul *et al.*, 1990].

— Il existe également des outils tels que Water et SSearch, qui se basent sur l'algorithme Smith-Waterman. Cet algorithme repose sur l'utilisation de matrice de similarité. Il cherche à optimiser l'alignement de deux séquences en insérant des indel (insertion ou délétion) afin de maximiser le nombre de nucléotides positionnés de manière identique dans les deux séquences. Un score permet ensuite de quantifier le résultat selon le nombre de nucléotides identiques et le nombre de substitution[Olsen *et al.*, 1999].

Il existe également de nombreux outils permettant d'effectuer des alignements globaux.

— Certains d'entre eux se basent sur la méthode de l'alignement progressif, qui utilisent le regroupement d'alignements deux à deux pour construire un alignement multiple. C'est le cas par exemple de T-Coffee, MAFFT ou Pileup. Le programme le plus populaire utilisant cette méthode est Clustal Omega. Son algorithme possède 5 étapes : tout d'abord, les séquences sont alignées deux à deux en les segmentant comme vu précédemment avec l'algorithme Blast. Ensuite, les distances entre ces séquences sont calculées. Puis un algorithme de clustering (*k-means*) est utilisé afin de regrouper les séquences. Une matrice de distance est calculé pour chaque *cluster*. Grâce à cet matrice, un arbre guide est construit en utilisant la méthode UPGMA : à chaque étape les deux *clusters* les plus proches sont combinés jusqu'à ce qu'on obtienne l'arbre final. Cet arbre déterminera l'ordre dans lequel les séquences sont alignées. Enfin, les séquences sont alignées grâce au *package* HHAlign.[Sievers et Higgins, 2014].

— Il existe également un outil appelé Muscle qui se base sur un algorithme itératif. A la première étape, un alignement progressif est construit en utilisant un arbre guide, construit en mesurant la similarité des séquences deux à deux puis en estimant leur distance. Ensuite, la deuxième étape améliore l'arbre construit en réalisant un nouvel alignement progressif basé sur l'arbre. Enfin, dans une troisième étape, l'alignement est à nouveau affiné en réalisant encore un nouvel alignement qui sera gardé si son score est supérieur au précédent. La troisième étape est répétée plusieurs fois.[Edgar, 2004].

Dans le contexte de ce projet, il sera nécessaire de réaliser des alignements d'espèces proches, afin de situer la vipère péliade par rapport à d'autre espèces d'intérêt. Il est donc préférable de choisir une méthode d'alignement global. Des alignements multiples devront être réalisés puisqu'il s'agit de comparer plusieurs espèces entre elles. Ainsi, Clustal Omega ou Muscle, des programmes d'alignement multiple global semblent les plus adaptés pour réaliser cette tâche.

Phylogénie

Après avoir sélectionné, nettoyé puis aligné les séquences choisies, une étude phylogénétique approfondie d'un gène cible sera effectuée. Celle-ci se fera sur 2 niveaux différents : une plus générale qui servira à situer l'espèce dans le monde du vivant et une plus spécifique sur le plan génique. Cette étude phylogénétique aboutira donc à la réalisation de minimum 2 arbres phylogénétiques :

— Un arbre des espèces : celui-ci représente les relations évolutives entre les espèces en général, celui-ci peut-être déduit à partir de molécules, toutefois cela reste risqué, car de nombreux biais existent. Il y a en effet plus de risques d'obtenir des paralogies (duplication d'un gène entre 2 espèces différentes ou un transfert horizontal de gènes). C'est pour cela que des données génomiques plus larges seront utilisées pour les construire. Celui-ci permettra ainsi de situer la vipère péliade par rapport à d'autres espèces d'intérêts notamment les 6 espèces sélectionnées dans le programme « Les sentinelles du climat » et donner une idée générale de ses relations évolutives.

— Un arbre des gènes : celui-ci représente l'histoire évolutive des molécules apparentées (gènes, protéines etc ... par exemple). Cette représentation peut différer de manière plus ou moins importantes de l'arbre des espèces. Pour cela une sélection sur des gênes d'intérêt sera effectuée

Un arbre phylogénétique est composé de plusieurs éléments : des OTUs aussi appelés feuillets qui représentent les unités taxonomiques opérationnelles (les espèces/molécules choisies), des HTUs ou nœuds internes qui représentent les unités taxonomiques hypothétiques (un ancêtre hypothétique commun) et finalement des branches représentant les liens de parenté entre les unités. Tous ces éléments sont organisés de façon à former des branchements ce qui correspond à la topologie de l'arbre. Celui-ci peut aussi être enraciné ou non, même si un arbre non-enraciné ne correspond pas réellement à un arbre phylogénétique

puisqu'il ne met pas en avant les relations entre les OTUs. La racine, elle, symbolise le dernier ancêtre commun de toutes les OTUs.

Actuellement, de nombreux outils bio-informatiques ont été développés afin de faciliter la création de ces arbres. Des *softwares* gratuits ou encore de nombreux *webserver* sont actuellement disponibles pour faire ces tâches. Chacun de ces outils se basent sur des algorithmes tirés de méthodes de construction pré-existantes :

— méthodes cladistiques : elles se basent sur une étude des états de caractères avec par exemple le maximum de parcimonie.
— méthodes de distances : basées sur des mesures de distances pour cela l'alignement multiple est utilisé conjointement avec le calcul d'une matrice de distance entre chaque paire de séquences. Ici, l'horloge évolutive est prise en compte et l'arbre peut donc être enraciné ou non. On a plusieurs méthodes très utilisées telles que : UPGMA et le NJ (*neighboor joining*).
— méthodes statistiques : elles comprennent à la fois l'étude des états de caractères et les distances. Il y a par exemple : le maximum de vraisemblance.

Maximum de parcimonie :
C'est donc une méthode dite cladistique, elle se base sur l'étude des états de caractère. Elle consiste principalement à l'identification de la topologie d'arbres impliquant le plus petit nombre de changements évolutifs pour rendre compte des différences que l'on peut observer entre les OTUs étudiés. Pour cela, il va falloir construire tous les arbres possibles, pour chaque caractère le nombre l'arbre qui en demande le moins est conservé.

Avec cette méthode, il n'y a pas de bonnes solutions à proprement parler, puisque plusieurs arbres sont possibles avec un nombre de substitutions identiques. De plus, la longueur des branches ne tient pas compte de la distance évolutive. L'arbre obtenu est non-enraciné.

Celle-ci présente donc plusieurs désavantages :

— Le nombre d'arbres augmentent de manière exponentielle avec le nombre d'OTUs.
— Pas de prise en compte de l'horloge moléculaire = pas de données évolutives.
— Ne fonctionne qu'avec des régions/protéines très conservées.

Cette méthode ne sera donc pas utilisée pour la construction de l'arbre des gènes, mais pourra être utilisée pour celui des espèces.
Outils bio-informatique disponibles :

UPGMA (*Unweighted Pair-Group method by arithmetic averaging.*)
Cette technique se base sur un principe de distance, il va y avoir un regroupement des séquences par ordre de similarité. L'arbre obtenu sera enraciné et tiendra donc compte de l'horloge évolutive. IL possédera un ancêtre hypothétique commun à toutes les espèces. Cette technique nécessite un alignement multiple des séquences, il y aura ensuite construction d'une matrice de distance par itérations successives et celle-ci sera utilisée pour construire l'arbre correspondant. Sa construction se fait des feuilles vers les noeuds.

Cette technique possède des avantages :
— Rapide à mettre en place et rapidité dans l'obtention des résultats.
— Simple à comprendre et à mettre en œuvre.
Mais aussi de nombreux désavantages :

— Égalité des taux d'évolution entre les lignées (= horloge moléculaire biaisée).
— Distance évolutive sous-estimée.
— Les branches longues avec une évolution rapide sont considérées comme des groupes extérieurs.

C'est pour cela que cette technique est assez peu utilisée actuellement.

Neighboor Joining.
Cette méthode se base sur le même principe que la méthode UPGMA et se construit à partir de matrices de distance. Cependant, contrairement à celle-ci, la méthode NJ tient compte du biais des différentes vitesses d'évolution entre les différentes branches de l'arbre en conservant l'additivité des distances. L'arbre obtenu sera, alors non-enraciné. Il ne mettra donc pas en avant les similarités globales entre les différentes espèces. Ce n'est pas une méthode dite phénéticiste : celui-ci reflétera leurs relations de parenté. Cette technique est particulièrement utilisée pour la construction d'arbre des gènes, toutefois il requiert de connaître la distance entre chaque OTUs.

Avantages :

— rapide.

— simple à mettre en place.

— permet de travailler sur un grand nombre de taxa en même temps.

Désavantages :

— Manque de précision.

— Ne fonctionne qu'avec des séquences dont la distance évolutive est connue.

Cette méthode pourrait être utilisée lors du projet.
Les outils bio-informatiques disponibles sont SYLVA, T-Rex.

Maximum de vraisemblance.
Cette méthode est dite statistique et va permettre de calculer à partir d'un échantillon observé, la ou les meilleures valeurs d'un paramètre en suivant une loi de probabilité. Ici, il va y avoir une sélection de l'arbre qui maximise la vraisemblance (celui avec la plus forte probabilité de retrouver les données utilisées). C'est principalement utilisé lorsque les taux de changement sont très élevés entre 2 séquences. Chaque base ou AA (acides aminés) des séquences va être considéré séparément. Un *log fit* de vraisemblance est alors calculé pour une topologie donnée en utilisant un modèle de probabilité. Ce *log* est alors cumulé sur tous les sites et la somme maximisée pour estimer la longueur de branche de l'arbre. Ce processus est utilisé pour toutes les topologies possibles et seule celle ayant la plus haute vraisemblance est conservée.

Cette méthode est particulièrement efficace et possède de nombreux avantages :

— estimation de la longueur des branches.

— permet de différencier les transitions et transversions.

Inconvénients :

— calculs très longs.

Les outils bio-informatiques disponibles sont PhyML, RAxML.

1.2 Objectifs

Compte tenu des recherches exposées précédemment, il paraît essentiel de rappeler le but de ce projet et d'en préciser les objectifs.

Plusieurs difficultés rencontrées par les chercheurs ont été mises en avant et relèvent de la multiplicité des outils et plate-forme de NGS et de phylogénétique. L'état de l'art a permis la sélection d'outils pertinents pour ce projet. Il a également permis l'identification de séquences particulièrement étudiées par la communauté scientifique. Il sera alors nécessaire de fournir une plate-forme qui permettra aux utilisateurs d'accéder à ces séquences, de les traiter, de les analyser et de les modéliser sous forme d'arbres.

Le premier objectif est donc de concevoir une interface web qui permettra l'intégration de divers outils issus de *webservers*.
Le second objectif consiste en cette intégration. Dans un premier temps, l'interface devra permettre à l'utilisateur d'accéder aux séquences d'intérêt par l'intégration d'une base de données. Ensuite, il sera

nécessaire d'intégrer l'outil d'alignement afin de traiter ces séquences afin de les préparer à l'étude phylogénétique. Cette dernière sera également possible depuis l'interface par l'intégration d'outils dédiés.

Le dernier objectif est de permettre à l'utilisateur la sauvegarde des résultats à chaque étape dans des fichiers conformes aux formats standards.

L'intégration est un aspect central dans ce projet. Cela sous-entend un travail sur la compatibilité à différents niveaux : celui de l'interface, des outils, des formats, etc. Il sera donc capital de tester et de vérifier à chaque étape que ces notions sont respectées.

Chapitre 2

Analyse des besoins

2.1 Besoins fonctionnels

L'analyse réalisée précédemment permet de comprendre l'intérêt d'étudier le venin de la vipère péliade et principalement les séquences impliquées dans la production de phospholipase A2 (PLA2) (principal constituant du venin). La phylogénie en ressort comme un élément important à étudier.

Compte tenu de ces constats, ce projet devra permettre à l'utilisateur d'obtenir les séquences, de les traiter et de produire un arbre des espèces ainsi qu'un arbre phylogénétique. Ces éléments constitueront la base de l'analyse des besoins fonctionnels.

L'utilisateur devra avoir accès aux différents outils depuis un seul environnement facilement utilisable. Une interface web semble être la méthode la plus appropriée pour faire le lien entre les outils issus de *webservers*.

Cette interface devra permettre à l'utilisateur :

— d'accéder à une page principale qui présentera le contexte et expliquera les différentes étape du *pipeline* ;

— d'accéder aux différents outils par les biais d'onglets dédiés sur l'interface web ;

— de récupérer des séquences identifiées depuis une base de données en permettant le choix des espèces d'intérêt ;

— d'obtenir des données et des résultats dans des formats standardisés mais aussi de les afficher directement dans l'interface ;

— d'accéder à différents outils d'alignement et d'en sélectionner les options ;

— d'accéder à un outils de construction d'arbres laissant le choix du type d'arbre et de la méthode à utiliser.

2.2 Besoins non fonctionnels

Afin de satisfaire les besoins fonctionnels de l'utilisateur, il est nécessaire d'identifier les aspects techniques par l'analyse des besoins fonctionnels :

— L'interface web utilisera une association de langages *front-end* : HTML, CSS et JavaScript et devra permettre de récupérer les requêtes/choix de l'utilisateur.

— Un *web framework* devra être utilisé afin de construire une interaction entre la partie *front-end* et *back-end* sur la base de Flask ou Django qui utilisent le langage de programmation Python.

— Un programme devra être conçu à l'aide du langage de programmation Python et de *frameworks* pour permettre la transmission des requêtes aux différents *webservers* détaillés dans l'état de l'art et ainsi obtenir les résultats et les transmettre à l'utilisateur par le biais de l'interface.

— Les différentes interactions décrites précédemment devront être transparentes pour l'utilisateur.

— L'utilisation d'un gestionnaire de versions tel que GitHub sera également nécessaire et permettra d'accéder à la plate-forme en ligne.

Chapitre 3

Réalisation

3.1 Engagements

Le but principal de ce projet est de créer une plate-forme permettant de réaliser une étude phylogénique d'une espèce précise, du début jusqu'à la fin. Il va donc falloir rendre accessible divers outils bio-informatiques de phylogénie et de NGS. L'espèce cible est la vipère péliade qui est particulièrement répandue en Gironde. Pour permettre un accès simple et rapide à ces divers outils, il faudra mettre à disposition un site web comportant au minimum 2 onglets :

— Une page d'accueil générale contenant des informations sur le sujet, le projet auquel celui-ci est rattaché et les sources bio-informatiques utilisées.
— Un second onglet donnant accès aux outils incorporés au site permettant donc d'effectuer l'alignement et la construction de l'arbre.

Le contenu de la page peut être divisé en deux parties distinctes : une plus générale et descriptive axée sur l'espèce cible et une partie analytique.

Dans cette partie générale diverses informations seront disponibles. Il y aura dans un premier temps, une description de la vipère péliade et quelques données phylogénétiques. Des informations génériques seront fournies telles que son nom vernaculaire et scientifique, une introduction à sa phylogénie sera faite à l'aide d'un arbre des espèces qui sera réalisé en se focalisant sur des espèces d'intérêt. Sa taxonomie sera aussi présentée. Pour compléter la présentation, les informations issues de l'état de l'art biologique telles que des informations morpho-physiologiques, éthologiques et géographiques seront également présentes sur la page d'accueil.

La partie analytique permettra dans un deuxième temps une analyse de séquences et la construction d'un arbre phylogénétique. L'analyse phylogénétique moléculaire est divisée en trois étapes :
— l'acquisition et la sélection de séquences ;
— l'alignement multi-séquence d'un groupe de séquences homologues ;
— la construction d'arbres.

Les sources des outils bio-informatiques utilisées seront fournies lors des différentes étapes analytiques comme des banques de données servant à obtenir des séquences (NCBI, EMBL, DDBJ), des sites d'alignement multiple global les plus courants pour comparer plusieurs espèces entre elles (Clustal Omega ou Muscle) ainsi que des sites permettant de construire un arbre de gène à partir des séquences choisies (SILVA pour des arbres de gènes et PhyML pour des arbres d'espèces).

Par ailleurs, la plate-forme sera accessible en ligne par le biais de GitHub et un manuel utilisateur sera fourni.

3.2 Améliorations potentielles

La première amélioration qui pourra être proposée est l'augmentation du nombres d'onglets permettant de faciliter la navigation de l'utilisateur. Il serait pertinent de proposer un onglet par outil en plus de l'onglet d'accueil comme présenté dans la figure B.1.

L'onglet d'accueil contiendrait toujours les informations descriptives sur l'espèce et le projet. Il y aurait ensuite un onglet pour l'accès à la base de données puis un autre pour l'outil d'alignement et enfin, un dernier pour la construction d'arbres.

Une autre amélioration potentielle pourrait être d'élargir la base de données permettant ainsi d'accéder à d'autres séquences identifiées précédemment pour la vipère péliade.

3.3 Points de questionnements

En l'état actuel, il est envisagé que la plate-forme accède directement aux bases de données ainsi qu'aux autres outils et permette à l'utilisateur de récupérer des fichiers de sauvegarde au fur et à mesure du *pipeline*. En parallèle, la plate-forme devra afficher les résultats en temps réel. Il s'agirait alors principalement d'intégrer des outils existants et reconnus.

Cependant, il se peut que cela ne soit pas possible ce qui imposerait de revoir la méthodologie. Étant donné que le sujet traité est très ciblé, il serait envisageable de réaliser toutes les analyses en amont, de stocker les résultats et d'en réaliser l'affichage sur la plate-forme. Cependant, cette solution présenterait plusieurs inconvénients comme la mise à jour régulière des données et résultats qui supposerait de refaire toutes les analyses. Cela provoquerait un perte de temps de considérable et un risque d'erreur non négligeable. De plus, il serait nécessaire de réduire le nombre d'options proposées à l'utilisateur.

Une autre possibilité serait de récupérer les données cibles sur les bases de données identifiées précédemment et de les stocker. Ensuite, il serait alors nécessaire de concevoir les outils de bout en bout afin de les proposer à l'utilisateur en se basant sur les algorithmes d'analyses présentés par la communauté scientifiques. Cependant, il ne s'agirait plus ici d'intégration mais de conception. D'autre part, cela supposerait également un suivi important pour assurer la mise à jour des données et la maintenance logicielle. De plus, il serait nécessaire de mener une étude chargée de comparer les résultats obtenus par les outils conçus et les outils reconnus par la communauté scientifique afin d'en évaluer la fiabilité.

Conclusion

Ce cahier des charges a permis de présenter le contexte et les objectifs du projet dont il est issu. Un état de l'art biologique a permis de préciser les aspects biologiques et bio-informatiques de ce projets en identifiant des séquences d'intérêt ainsi que les outils nécessaires à leur étude. Une démarche comparative a permis de sélectionner les outils les plus pertinents au traitement analytique de ces données.

Ce projet a donc pour ambition de concevoir une plate-forme chargée de rassembler plusieurs outils en un seul site web. De cette façon, il serait possible de réaliser tout un *pipeline* phylogénétique en évitant à l'utilisateur de faire face à la multiplicité des plate-formes.

La plate-forme pourrait alors permettre d'interroger des bases de données génomiques, de réaliser l'alignement des séquences qui en sont issues et de réaliser des arbres mettant en évidence la phylogénie de gènes codant la protéine PLA2 du venin de la vipère péliade.

Ce travail permettra d'expérimenter et d'approfondir la notion d'intégration dans le domaine du génie logiciel mais également de faciliter le travail des chercheurs intéressés par le domaine présenté ici. Enfin, un calendrier prévisionnel a été établi pour la réalisation de ce projet et est présenté dans la figure A.1.

Bibliographie

[Al-Shekhadat *et al.*, 2019] Al-Shekhadat, R., Lopushanskaya, K., Segura, , Gutiérrez, J., Calvete, J. et Pla, D. (2019). Vipera berus berus venom from russia : Venomics, bioactivities and preclinical assessment of microgen antivenom. 11.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. et Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

[Bocian *et al.*, 2016] Bocian, A., Urbanik, M., Hus, K., Łyskowski, A., Petrilla, V., Andrejčáková, Z., Petrillová, M. et Legath, J. (2016). Proteome and peptidome of vipera berus berus venom. *Molecules*, 21(10):1398.

[CREUX, 2019] CREUX, T. (2019). Les vipères filent vers l'extinction, dans le silence le plus complet.

[Czajka *et al.*, 2013] Czajka, U., Wiatrzyk, A. et Lutyńska, A. (2013). Mechanism of vipera berus venom activity and the principles of antivenom administration in treatment. 67:641–646, 729–733.

[Edgar, 2004] Edgar, R. (2004). Muscle : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

[Guillemin, 2003] Guillemin, Bouchier, G. W. C. (2003). Sequences and structural organization of phospholipase a2 genes from vipera aspis aspis, v. aspis zinnikeri and vipera berus berus venom. 270:2697–2706.

[Hamilton *et al.*, 2019] Hamilton, J., Kause, J. et Lamb, T. (2019). Severe systemic envenomation following vipera berus bite managed with viperatab antivenom. 30:56–58.

[Hermansen *et al.*, 2019] Hermansen, M., Krug, A., Tjønnfjord, E. et Brabrand, M. (2019). Envenomation by the common european adder (vipera berus) : a case series of 219 patients. 26:362–365.

[Losange, 2008] Losange (2008). *Amphibiens et reptiles*. Editions Artemis.

[Malina *et al.*, 2013] Malina, T., Babocsay, G., Krecsák, L. et Erdész, C. (2013). Further clinical evidence for the existence of neurotoxicity in a population of the european adder (vipera berus berus) in eastern hungary : second authenticated case. 24:378–383.

[Malina *et al.*, 2017] Malina, T., Krecsák, L., Westerström, A., Szemán-Nagy, G., Gyémánt, G., M-Hamvas, M., Rowan, E., Harvey, A., Warrell, D., Pál, B., Rusznák, Z. et Vasas, G. (2017). Individual variability of venom from the european adder (vipera berus berus) from one locality in eastern hungary. 135:59–70.

[Netgen, 2011] Netgen (2011). Une morsure de serpent venimeux : une mort sûre ?

[Olsen *et al.*, 1999] Olsen, R., Hwa, T. et Lässig, M. (1999). Optimizing smith-waterman alignments. *Pacific Symposium on Biocomputing*, page 302–313.

[Phillips *et al.*, 2010] Phillips, D. J., Swenson, S. D., Francis, S., Markland, J. et Mackessy, S. (2010). Thrombin-like snake venom serine proteinases. *Handbook of Venoms and Toxins of Reptiles*, 139:154.

[Sievers et Higgins, 2014] Sievers, F. et Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology*, 1079:105–116.

Appendices

Annexe A

Calendrier (Diagramme de Gantt)

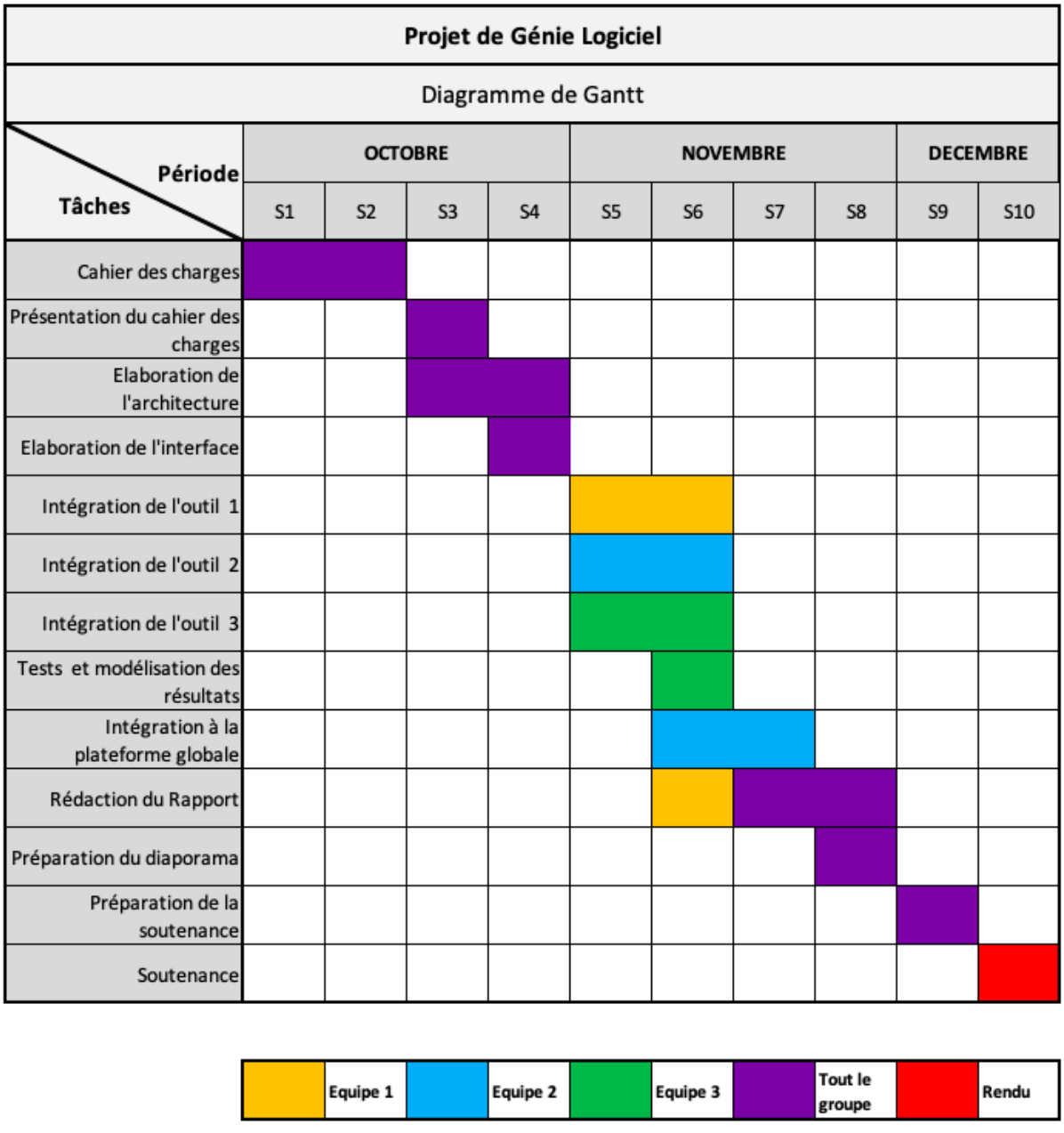

FIGURE A.1 – Diagramme de Gantt

Ce diagramme représente la répartition des tâches nécessaires à l'élaboration de ce travail sur une période donnée. Les couleurs permettent d'identifier les différentes équipes qui travailleront sur chaque tâche.

Annexe B

Prototype

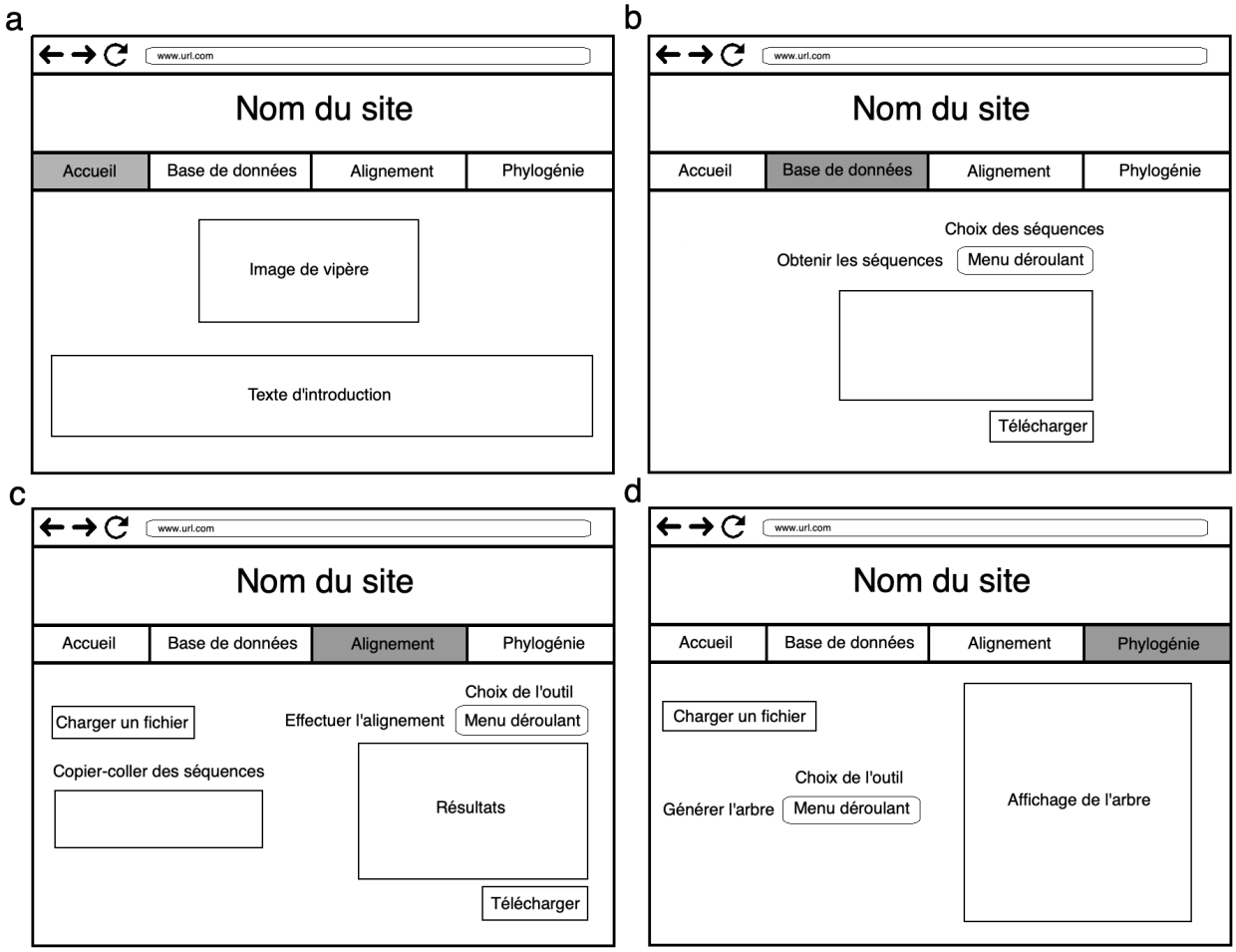

FIGURE B.1 – Maquette du site

Cette maquette représente les différents onglets qu'il sera possible de trouver sur le site.
L'image **a** représente la page d'accueil qui permet d'introduire la thématique.
Sur l'image **b** se trouve le second onglet "Base de données", qui permet de récupérer des séquences de gènes de l'espèce choisie.
Sur l'image **c** se trouve le troisième onglet "Alignement", qui permettra d'effectuer l'alignement des séquences.
Enfin, sur l'image **d** se trouve le dernier onglet "Phylogénie" qui permettra de construire un arbre phylogénétique.