Rapport du projet

Projet de Programmation

L2D1

Illustration 1: ***Image de présentation***

Boutajar Manal
Bodnar Adelin
Hamiche Mehdi

Groupe L2D1

Les informations d'identification du document :

Référence du document :
Version du document : 2
Date du document : 04/05/2021
Auteurs : **Mehdi Hamiche** **Manal Boutajar** **Adelin Bodnar**

Les éléments de vérification du document :

Validé par :
Validé le :
Soumis le :
Type de diffusion :
Confidentialité :

Les éléments d'authentification :

Maître d'ouvrage :	**Chef de projet :**
Date / Signature :	**Date / Signature :**

Sommaire

1. Introduction

Cette deuxième année de licence, un projet informatique devait être réalisé dans le cadre de l'UE Projet. Le choix de ce projet s'est fait par l'envie d'approfondir la découverte des langages de programmation, mais notamment pouvoir concevoir par nos propres moyens une application web qui nous correspond réellement en termes de critère et d'utilisation.

Cette UE a donc pour but de nous faire découvrir les enjeux en tant que développeurs par réaliser un projet complet dans un temps limité, tout en respectant certaines contraintes et règles.

Grâce à ce projet, nous avons eu l'opportunité de nous mettre à la place d'un développeur et de découvrir en profondeur les manières de travailler, d'apprendre et de nous mettre en pratique vis-à-vis du développement, à l'aide des encadrants et de nos efforts personnels.

Le projet que nous avons choisi est le L2D qui consiste à créer un programme réalisant des alignements globaux et multiples des séquences protéiques et nucléotidiques, ainsi récupérer le logo d'un alignement multiple.

2. Guide de lecture

2.1. Maîtrise d'œuvre

La maîtrise d'œuvre présente l'équipe du développement chargé du bon suivi du rapport du projet et des besoins dont le maître d'ouvrage fait commande.

Elle représente l'équipe du développement :

- Adelin Bodnar
- Manal Boutajar
- Mehdi Hamiche

Cette équipe veillera au bon suivi du rapport du projet représentant les besoins des enseignants encadrants.

2.1.1. Responsable

Il est conseillé pour le responsable de la maîtrise d'œuvre de lire le document dans sa totalité afin de prendre conscience de l'ensemble des éléments.

2.1.2. Personnel administratif

Il est conseillé pour le personnel administratif de lire la partie des connaissances, les fonctionnalités principales ainsi que les différentes phases de projet.

2.1.3. Personnel technique

Il est conseillé pour le personnel technique de prendre en compte la partie sur les différentes phases du projet ainsi que les langages, logiciels et environnements.

2.2. Maîtrise d'ouvrage

La maîtrise d'ouvrage représente dans notre cas le client du projet, c'est-à-dire les personnes dont les besoins permettent la conception du projet.

La maîtrise d'ouvrage est assistée par l'équipe de la maîtrise d'œuvre et donc ce rôle sera assuré par les enseignants encadrants Dragutin Jastrebic et Koviljka Lukic Jastrebic.

2.2.1. Responsable

Il est conseillé pour le responsable de la maîtrise d'ouvrage de lire le document dans toute sa totalité afin de prendre conscience de l'ensemble des documents.

2.2.2. Personnel administratif

Il est conseillé pour le personnel administratif de lire la partie des connaissances, les fonctionnalités principales ainsi que les différentes phases de projet.

2.2.3. Personnel technique

Il est conseillé pour le personnel technique de prendre en compte la partie sur les différentes phases du projet ainsi que les langages, logiciels et environnements.

3. Connaissances

- ***Données biologiques***

Une séquence aux formats différents (.fasta - format le plus utilisé dans BIONFO / autres formats utilisés)

- ***Internet***
  - Eclipse IDE un environnent de développement intégré libre à télécharger à partir de l'URL : "https://www.eclipse.org/downloads/"
  - Serveur web (Apache 2.4.9) WampServer est une plate-forme de développement Web sous Windows pour des applications Web dynamiques à l'aide du serveur Apache2, du langage de scripts PHP et d'une base de données MySQL. Il possède également PHPMyAdmin pour gérer plus facilement les bases de données
  - Serveurs de BDDs

- ***Donnée - Extraction / Récupération***

Pour récupérer des données biologiques il y a plusieurs moyens (EMBL-UniProt / NCBI - Protein, Nucleotide, Gene / DDBJ et PDB) comme :

- à la main
  - se rendre sur le site précis et mettre l'identifiant de la séquence
- à partir de l'URL
  - par exemple : "https://www.uniprot.org/uniprot/P00015.3.gff" et ensuite récupérer les données du lien et les mettre dans un fichier .txt ou autres
- à partir de TOGO WEB SERVICES
  - ici : "TogoWS (dbcls.jp)" et ensuite choisir entre Uniprot / DDBJ etc…,

- mettre le nom de identifiants de la séquence,
- le type,
- extension du format.

- ***Stockage***

Voici les différents fichiers qui peuvent être compatibles avec notre projet :

- .fasta
- .txt
- .json
- .gff

Pour les bases de données :

- table proteine
- sequences_nucleiques
- bnlearn
- gene

- ***Utilisation pendant le Développement / Production / Développement de l'application BIOINFO***
  - Alignement global
    - Résultats en forme de matrice / score / alignement
  - Alignement multiple
    - Résultats en forme de logo / alignement
  - Weblogo
    - Pendant le développement, il y a juste à exécuter le code TestWeblogo.java et de bien choisir le fichier dans File dans lequel nous voulons le logo. Le code va d'abord lire le contenu du fichier ensuite se connecter sur le site Weblogo3, transformer la connexion

de l'URL en byte [] et transformer ce byte en image, création de l'image et enfin affichage.

- Dans l'application, lors de la comparaison de deux séquences nucléiques ou protéiques d'un alignement multiple, on peut afficher le logo à partir d'un bouton en bas.

- Interface graphique
  - Permet l'affichage de nos différents modules que ce soit l'alignement global, multiple et weblogo
    - alignement global - afficher l'alignement de deux séquences protéiques ou nucléotidiques en indiquant les valeurs de gap
    - alignement multiple - affichage l'alignement de plusieurs séquences (au moins 3 séquences) en insérant manuellement les séquences ou les récupérer automatiquement en parcourant un fichier fasta
    - logo - récupérer le logo de chaque alignement multiple
  - Résultats unifiés

4. Langages – Logiciels - Environnements

- Comprendre l'utilité / l'importance de tous ces langages durant le développement en Java et la production de modules de BIOIFO

JAVA - ECLIPSE

Eclipse gére de nombreux types de projets, en particulier des projets complexes avec plusieurs sous-projets contenant de nombreux fichiers sources. Eclipse simplifie le développement Java, en particulier pour les gros projets dans un contexte professionnel.

Grace à Eclipse, cela a permis de créer nos modules et à les unifier.

R – RSTUDIO (4.0.4)

R est une suite logicielle dédiée à la manipulation de données, aux calculs statistiques et à leur présentation graphique. C'est aussi un langage de programmation. Cela permet une bonne visualisation de nos différents modules.

Exemple avce msa et psa en R qui nous a permis de voir les alignements multiples et globaux et une visualisation des fichiers aux formats .fasta.

PYTHON (version 3.8.5) - ANACONDA

Anaconda est un utilitaire pour Python offrant de nombreuses fonctionnalités. Il offre par exemple la possibilité d'installer des librairies et de les utiliser dans ses programmes, mais aussi propose des logiciels pouvant aider à développer.

Anaconda nous a permis de lire de lire des fichiers de différents formats depuis différents sites (EMBL-UniProt / NCBI-Protein, Nucleotide, Gene / DDBJ et PDB)

HTML (HTML5 ou autre)

HTML est un langage de description de document utilisé sur Internet pour faire des pages Web. Le balisage HTML est incorporé dans le texte du document et est interprété par un navigateur Web.

CSS (version 3)

Le CSS permet à l'utilisateur de personnaliser une page web

Aujourd'hui, de nombreux sites web permettent à l'utilisateur de changer la mise en page d'un site sans modifier le contenu. Les feuilles de styles qui sont stockées en externe permettent à l'utilisateur d'effectuer les changements requis par eux-mêmes.

SQL

Base de données (Select / Create / Joint...)

- Rencontre avec le monde scientifique (informations relatives aux séquences - formats gb,gff,txt,xml)

Chaque format de fichiers a sa particularité :

- .gff - décrire les gènes et d'autres éléments de séquences d'ADN, d'ARN et de protéines. et le type de contenu qui leur est associé est text / gff3.
- Les serveurs qui génèrent ce format : Uniprot. ("https://www.uniprot.org/uniprot/P99999.gff")
- .txt - permet d'avoir un détail complet de la séquence ("https://www.uniprot.org/uniprot/P99999.txt")
- .xml
- .gb

- Ouverture au monde réel dans la situation sanitaire actuelle - COVID

Futures améliorations envisagées:

Nous avons envisagé d'améliorer notre programme dans le futur pour qu'elle soit plus complète et qu'elle puisse séduire les clients à venir. Par exemple, améliorer l'affichage des différents modules…

5. Différentes phases du projet

5.1. Phase de la documentation

Le début du projet a commencé par la phase de réflexion, de la manière dont le projet allait se dérouler semaine après semaine et avoir une idée sur la conception de celui-ci.

Pour cela nous avons dans un premier temps conçu le cahier des charges pour nous fixer des objectifs et avoir une idée générale de la structure de notre application web.

Par la suite vient le cahier des recettes qui sert à lister les fonctionnalités attendues du site et qui devra réaliser après les tests.

Des documents tels que le manuel d'utilisation (conception générale) est là pour guider l'utilisateur sur l'utilisation du site.

5.2. Phase de développement

Arrivé à notre phase de développement, nous voulions tout d'abord partir comme nous l'avions prévu sur le cahier des charges. Du coup, nous avons commencé réellement à avancer et donc le principal problème que nous avions à partir de ce moment était la compatibilité du code des uns des autres mais aussi avec le code déjà écrit.

Cependant nous avons dû se référer aux méthodes de fonctionnement de certaines méthodes avec Java pour comprendre comment réaliser le programme.

5.3. Phase de test

Nous avons testé l'application à chaque ajout de fonctionnalité dans le but de vérifier si celle-ci est compatible avec ce qui est déjà fonctionnel.

Cela nous a permis de voir ce qui n'allait pas et d'en apporter la solution.

Les tests réalisés sont :

- les résultats des alignements globaux et multiples,
- le bon résultat du logo de l'alignement multiple de chaque séquence.

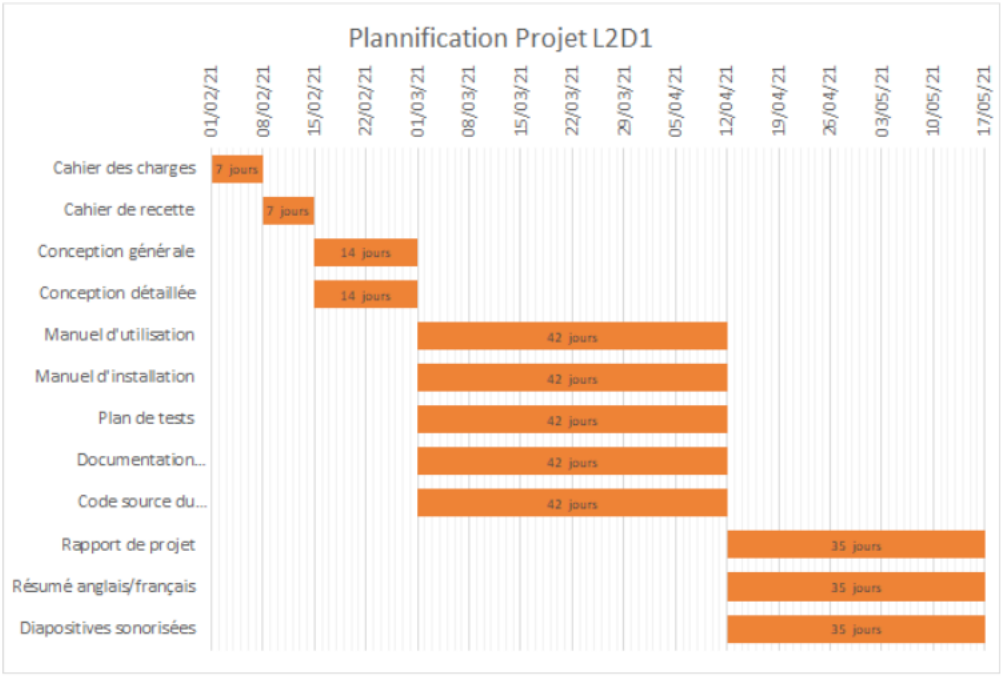

Illustration 2: ***Diagramme de Gantt***

5.4. Difficultés rencontrées

Parmi les difficultés rencontrées durant notre projet :

- Interface graphique : problèmes de threads - heap space, reconnaissance des résultats de l'alignement multiple pour récupérer le logo.
  (Solution : optimisation du code de l'alignement multiple, créer un fichier à partir du résultat pour récupérer le logo;
- Logo : problèmes de reconnaissance de l'URL, faire attention de récupérer la connexion de l'URL, ne pas utiliser un Reader, mais un InputStream, puis transformer cela en byte[] et ensuite en image.
- Alignement : affichage des résultats sur la Frame de l'interface
  (Solution : toString)

6. Fonctionnalités principales

Les différents points forts de notre application :

- **Alignement Global :**
  - lors de la comparaison de deux séquences protéiques ou nucléiques, on obtient les résultats en forme
    - matrice
    - score
    - alignement
  - alignement global complet
- **Alignement Multiple :**
  - choix de fichier .fasta de séquences protéiques et nucléiques pour comparer les séquences
  - ou écrire les séquences manuellement
  - donne l'alignement multiple des deux séquences complet
- **Logo :**
  - afficher le logo de l'alignement multiple
  - rendre la qualité d'image meilleure depuis le code TestWeblogo en choisissant dans les paramètres "png_print" pour que l'image ne soit pas floue
- **Interface Graphique :**
  - permettre une utilisation simple des 3 modules avec une page d'accueil
  - choix de deux modules alignement global et multiple (multiple intègre le logo)

7. Conclusion

Ce projet nous a permis de nous rendre compte de ce que représente un projet informatique et de ce qui est important lors d'un tel projet. Nous avons pu constater l'importance d'avoir une direction précise pour avancer efficacement.

Nous avons également pris conscience de l'importance de bien communiquer car cela pouvait amener à des incompatibilités qui risquerait de causer du retard par la suite. Pour chacun des membres du groupe il s'agit d'un de nos premier "vrai" projet informatique et nous sommes heureux d'avoir pu profiter de cette expérience.

En effet, ce projet est une première expérience pseudo-professionnelle qui nous a permis de nous faire une meilleure idée de la gestion de projet informatique dans le monde du travail. Cela nous donne également une base pour pouvoir continuer chacun de notre côté à travailler sur ce site et pouvoir toujours plus l'améliorer et développer nos compétences.

Nous remercions Madame Koviljka Jastrebic et David Jastrebic qui ont pris en charge d'encadrer notre projet et qui nous ont conseillé pendant son intégralité.

8. Annexes

Illustration 4: ***Page d'accueil de l'application***

Illustration 3: ***Alignement Global produit via l'application***

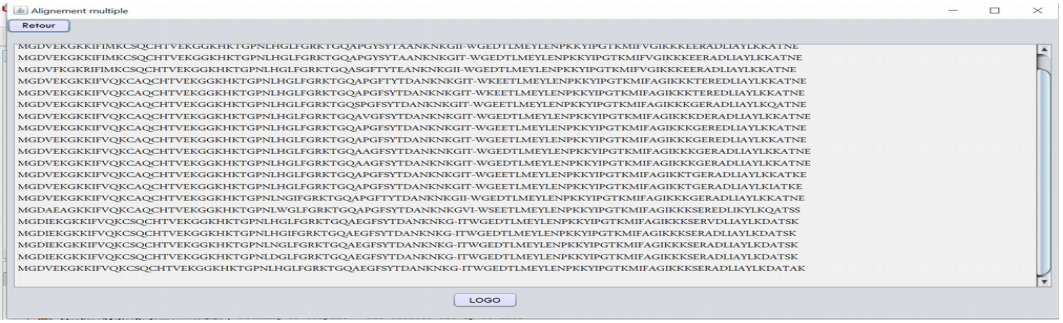

Illustration 6: ***Alignement Multiple produit via l'application***

Illustration 5: ***Logo produit via l'application***

9. Glossaire

★ ***BIO-INFORMATIQUE :*** discipline permettant d'analyser de l'information biologique qu'il y a dans les séquences nucléotidiques et protéiques

★ ***PROGRAMMATION ORIENTÉE OBJET :*** met en œuvre différents objets (instances de classes). Chaque objet associe des données et des méthodes (fonctions) agissant exclusivement sur les données

★ ***ALGORITHME NEEDLEMAN - WUNSCH :*** effectue un alignement global maximal de deux chaînes de caractères

★ ***ALIGNEMENT GLOBAL :*** recouvre les séquences alignées (par exemple 2 séquences) sur l'ensemble de leur longueur

★ ***ALIGNEMENT MULTIPLE :*** aligner un ensemble de séquences homologues, comme des séquences protéiques ou nucléiques qui assure des fonctions similaires dans différentes espèces vivantes

★ ***WEBLOGO :*** application web conçue pour rendre la génération de logos de séquence aussi simple que possible

★ ***INTERFACE GRAPHIQUE :*** relie les 3 modules et les présente de manière simple et efficace

★ ***GAP, MATCH, MISMATCH :*** valeurs pour alignement global

★ ***FASTA :*** format de fichier texte dans lesquels les séquences sont affichées par une suite de lettres

★ ***NCBI :*** National Centre for Biotechnology Information

★ ***DNA :*** Data Bank Japan

★ ***EMBL :*** European Molecular Biology Laboratory

★ ***PROTEIN :*** Base de données pour la récupération des séquences protéiques en format fasta

★ ***GENE :*** Base de données pour récupération des séquences nucléiques en format fasta

★ ***NUCLEOTIDE :*** Base de données pour la récupération des séquences nucléiques en format fasta

★ ***UNIPROT :*** Base de données pour la récupération des séquences protéiques en format fasta

★ ***DDBJ :*** Base de données pour la récupération des séquences nucléiques en format fasta

★ ***EMBOSS :*** outil bioinformatique pour l'alignement global – NEEDLE

★ ***BLAST :*** outil bioinformatique pour l'alignement local

★ ***MULTALIN :*** outil bioinformatique pour l'alignement multiple

10. Références

Alignement de séquences

Alignement global

Alignement multiple

JAVA - Main (site 1)

JAVA - Main (site 2)

Chaînes de caractères

Tableau

Needleman et Wunsch – Java

Needleman et Wunsch – Python

Interface graphique en Java

Interface graphique avec SWING

TogoWS

Eclipse

UniProt

11. Index

Index des figures

Index lexical