Sarra Ghabri
Antoine Haas
Jason Golda

Rapport du projet de Bioinformatique

Rappel concis du sujet

Le projet consiste à la recherche et au téléchargement des gènes et génomes des eucaryotes, procaryotes (séparés en bacteria et archaea), virus plasmids et organelles, selon la base de données de GenBank (https://www.ncbi.nlm.nih.gov/genome/browse).

Ensuite, en copiant l'arborescence du site, recréer l'ensemble des répertoires et effectuer des analyses sur l'ensemble des données :

- Les nombres et fréquences des trinucléotides dans les trois phases des gènes.
- Les nombres de trinucléotides en phase préférentielle dans les trois phases des gènes.
- Les nombres et fréquences des dinucléotides dans les deux phases des gènes.
- Les nombres de dinucléotides en phase préférentielle dans les deux phases des gènes.

Des contraintes ont également été apportées sur le formatage des données et les fonctionnalités devant être présentent sur le logiciel. De plus, des conseils sur le calcul des informations statistiques ont été fournis.

La structure du projet

Bibliothèques utilisées

- Afin de générer les fichiers Excel, nous avons utilisé la bibliothèque POI (version 3.15).

- Pour les appels http vers le serveur distant, nous avons opté pour Google HTTP Client Library for Java (version 1.21.0).

- Dans le but de stocker dans des fichiers json les informations qui ont déjà été téléchargées, nous utilisons json-simple (version 1.1.1).

- Pour utiliser l'injection de dépendance dans le projet, nous utilisons Google Guice (version 3.0).

- Nous avons également utilisé la collection de bibliothèques fournies par Guava (version 19.0) afin de faciliter diverses implémentations en utilisant les classes fournies par cet outil.

Arborescence Locale

En ouvrant le projet fourni, une arborescence des documents sera visible.

- Le répertoire « src » contient les fichiers sources du projet.

- Le « default package » contient le fichier Main.java contenant lui-même la méthode « main ». C'est donc ce fichier qu'il faut exécuter avec Eclipse pour lancer le logiciel.
- Le répertoire « controller » possède le fichier MainController.java, seul contrôleur de l'application, gérant la vue et l'ensemble la gestion de la redirection vers les fonctionnalités.
- Le répertoire « model » contient toutes les classes utilisées pour représenter les objets sur lesquels on effectue les statistiques.
- Les dossiers contenus dans « services » gèrent les services utilisés dans toute l'application (interfaces et implémentations), comme par exemple les requêtes http, le parser, la gestion du statut de l'api, etc…
- Le répertoire « utils » comporte l'utilitaire de gestion de l'archivage des données.
- Finalement, « views » contient la gestion de l'interface graphique.

- Le répertoire « resources » contient la configuration des noms des répertoires à générer selon les options choisies par l'utilisateur de l'application (on reviendra plus tard sur ce point).

- Les répertoires « Referenced Libraries », « JRE System Library » et « lib » contiennent les bibliothèques dont nous nous sommes servis dans ce projet, ainsi que le détail des modules utilisés pour chacune d'elles.

- Le répertoire « Results » comporte l'arborescence contenant tous les fichiers Excel de statistiques, générés à partir des téléchargements.

- Le répertoire « Gene » contient les gènes valides au format .txt

- Le répertoire « Genome » contient les séquences entières des gènes au format .txt

- Le répertoire « updates » possède pour chaque Kingdom le fichier json contenant chaque téléchargement effectué, ainsi que la date. Ceci nous permet de comparer avec la date de dernière modification pour pouvoir choisir s'il est nécessaire ou non de télécharger un gène.

Le logiciel

La moitié de gauche de l'écran comporte l'arborescence des fichiers. Elle se met à jour au fur et à mesure du téléchargement (ou est déjà remplis si des téléchargements ont précédemment eu lieux).

Sur la moitié droite

- La partie Option se décompose en deux parties :
    - La section « Kingdoms » permet de choisir le ou les Kingdom(s) à télécharger, parmi les eucaryotes (Eukaryota), virus (Viruses), organelles (Organelles) et procaryotes (Prokaryotes).

      Le téléchargement des procaryotes générera deux arborescences « Archaea » et « Bacteria » à la place d'une seule qui serait nommée Prokaryota, afin de bien pouvoir les distinguer.

      Les Viroïds n'étant finalement pas nécessaires, n'ont pas été rajoutés aux téléchargements possibles.

    - La section « Zip » permet d'archiver au format .zip les génomes ou/et les gènes valides.

- La partie Progression permet de visualiser le fichier au cours de téléchargement « Downloading » et le dernier fichier téléchargé « Downloaded ».

- La dernière partie contient le status du téléchargement :
    - « API Online » s'il n'y a rien à signaler lors du téléchargement
    - « There is an issue with the API » si un problème est survenu pendant le téléchargement (si le serveur ne répond plus ou s'il renvoie un message d'erreur).
    - « API Online – Could not get all data (server issues) » arrive lorsque le téléchargement reprend après une erreur. Ce message signifie que le reste des données sera téléchargé, mais que certaines seront manquantes suite à une erreur du service distant.

La partie basse de l'écran
- Contient une barre de progression permettant de voir combien de fichiers sont encore à télécharger, ainsi qu'une estimation (vague) du temps restant. Cette barre de progression peut parfois mettre quelque temps avant de s'actualiser.

- Un bouton Execute, pour pouvoir lancer le téléchargement avec les options préalablement sélectionnées.

- Un bouton Interrupt, afin de stopper proprement le téléchargement. Il sera nécessaire d'attendre plusieurs secondes avec que tous les calculs s'arrêtent.

Travail réalisé et problèmes rencontrés

Nous avions malencontreusement oublié d'effectuer la 4^{e} étude statistique : « Les nombres de dinucléotides en phase préférentielle dans les deux phases des gènes. » au début du projet et n'avons finalement pas eu le temps de la rajouter sur la fin.

Pour la répartition du travail,
Antoine s'est principalement chargé de l'étude statistique des fichiers.
Jason a géré le téléchargement des fichiers et le parsing.
Sarra a mis en place l'interface graphique et sa gestion, puis a aidé alternativement Antoine et Jason sur les blocages ponctuels.

Pour la gestion de projet, nous avons utilisé un dépôt git, ce qui nous a permis de gérer plus efficacement les merges des différentes parties.