université PARIS-SACLAY

M2 AMI2B Biologie Computationnelle

Analyse, Modélisation et Ingénierie de l'Information Biologique et Médicale
Projet Programmation WEB

Bryan Brancotte – Olivier Lespinet

Conception d'une application web permettant de réaliser l'annotation et l'analyse fonctionnelle de génomes bactériens

---

Données disponibles

Vous disposez sur eCampus dans d'une archive (data_web.zip) contenant 3 génomes d'Escherichia coli déjà annotés (k12, CFT073 et o157h7) et un nouveau génome à annoter (new_coli). Chaque génome est présenté sous la forme de trois fichiers : le génome complet, la liste des CDS (fichier _ CDS) et la liste des peptides (_pep).

Initialisation : Chargement des données

Les données seront à charger dans l'application. Les séquences sont uniques et on stocke une seule annotation fonctionnelle par séquence. Une annotation est un texte qui décrit la fonction (cf. fichiers fasta des génomes déjà annotés).

Les fonctionnalités principales attendues sont les suivantes.

Gestion des utilisateurs et des projets

L'accès à l'outil ne peut se faire que par des utilisateurs enregistrés dans le système. Chaque utilisateur a un email qui l'identifie ainsi qu'un mot de passe sécurisé. Un utilisateur a un prénom et un nom, un numéro de téléphone. Un utilisateur peut avoir trois rôles : lecteur, annotateur et validateur. Un lecteur peut lire les informations de la base (poser ses requêtes grâce à un formulaire). Un annotateur peut faire tout ce que fait un utilisateur et en plus annoter des séquences. Un validateur peut faire tout ce que fait un annotateur et en plus valider des annotations des annotateurs.

L'administrateur du système a accès à la liste des utilisateurs, il est le seul à pouvoir créer ou supprimer un utilisateur et leur affecter des rôles. Il peut voir la date (et l'heure) de dernière connexion d'un utilisateur au système.

Un administrateur a tous les droits.

Chaque nouvel utilisateur devra remplir un formulaire avec ses informations (email, nom, prénom…), il choisira le rôle qu'il souhaite avoir dans le système : utilisateur, annotateur ou validateur. Lorsqu'il a rempli tous les champs il clique sur un bouton (par exemple "créer l'utilisateur") ce qui aura pour effet de créer l'utilisateur.

- Pour simplifier l'outil : ces demandes sont par défaut valides.
- En option (fonctionnalité non prioritaire) : Lorsqu'une demande de création de nouvel utilisateur est faite, l'administrateur valide ou non cette création d'utilisateur.

Une fois connecté tout utilisateur doit pouvoir accéder aux fonctionnalités suivantes.

Sélection des informations relatives à un génome

L'utilisateur du système peut accéder aux données du système en sélectionnant les informations par le biais d'un formulaire. On considère ici un unique formulaire qui contiendra en attributs tous les champs d'intérêt présents dans l'ensemble des types de fichiers Fasta (séquence nucléotidiques, peptidiques, nom de gènes, de transcrits, description…). Si l'utilisateur remplit plusieurs champs, il conviendra de les combiner par un « ET ».
Par exemple, l'utilisateur peut indiquer « %ATGC% » pour l'attribut relatif à la séquence et « eColi » pour l'attribut relatif à l'espèce pour récupérer la liste des génomes ayant le motif ATGC dans leur séquences et dont l'espèce contient le mot eColi. On fixera une taille minimale de la sous-séquence nucléotidique ou peptidique à 3 résidus et une taille maximale à la taille moyenne d'un gène.

Le formulaire précisera aussi le type de sortie attendue (un seul type possible)

- génome
- gene/proteine

Dans le cas où plusieurs résultats sont possibles, la liste des résultats (d'un type donné) est renvoyée. Cette liste est cliquable.

- Lorsqu'un génome est sélectionné on accède à une page de visualisation du génome.
- Lorsque gène/protéine est sélectionné on accède à une page avec l'ensemble des informations relatives à ce gène et à sa protéine associée

Note importante sur l'évolution du système : le système ne traitera que des génomes procaryotes.

***En option** : de l'auto-complétion pourra être proposée pour les valeurs d'espèce, de noms de protéine et de nom de gènes en fonction des valeurs existantes dans la base. L'auto-completion est la fonctionnalité que vous avez sur les sites web qui complète pour vous un mot à partir du début de ce mot (par exemple vous tapez le début d'un nom de protéine et on vous propose automatiquement les noms qui commencent par les lettres saisies). C'est une fonctionnalité qui est complexe à mettre en œuvre : à ne faire que si **vous êtes déjà très à l'aise** en programmation Web (AJAX).*

Visualisation des informations génomiques

Lorsque l'utilisateur clique sur des résultats de recherches de génomes, il doit pouvoir visualiser le génome et notamment visualiser la position des gènes sur le génome et la séquence associée à chaque gène. **La visualisation est libre**, différentes stratégies pourront être explorées : utiliser des couleurs qui changent d'un gène à l'autre ou mettre certaines séquences en gras, en italique… Une fonctionnalité souhaitable est de permettre de rendre le génome cliquable : on voudrait pouvoir cliquer sur un gène donné pour accéder à toutes ses informations. On souhaiterait aussi dans le génome pouvoir voir le nom des gènes du génome facilement.

En option : Visualisation des domaines d'une séquence protéique donnée

Lorsqu'une séquence protéique est sélectionnée par l'utilisateur, il doit pouvoir visualiser les domaines de sa séquence. Ces informations peuvent être obtenues dans la base internationale PFAM.

Extraction des données déjà annotées

L'ensemble des données présentes dans la base peuvent être extraites sous la forme de fichiers plats.

Une fois la sélection (filtre) des données effectuée, l'utilisateur peut choisir les champs à extraire, le séparateur sera le « ; ». Un fichier .txt sera ainsi extrait.

Comparaison et alignement de séquences

L'accès aux outils d'alignements de séquences est important. L'utilisateur doit pouvoir facilement lancer une recherche d'alignement de séquences à partir d'une séquence sélectionnée. Cet accès à un Blast se fait uniquement à partir des séquences de la base.

Note importante : Il ne s'agit pas ici de recoder Blast mais plutôt d'accéder à une API distante (Blast NCBI, …).

Accès aux informations complémentaires des banques de données

D'autres banques de données disponibles peuvent être interrogées. L'utilisateur doit pouvoir accéder à ces informations complémentaires sur demande : il accède à une liste de banques de données disponibles dans une liste déroulante, il en choisit une. La sélection de la banque entraîne l'ouverture d'une nouvelle page Web sur la banque choisie interrogée grâce à l'attribut qui permet de faire le lien avec la banque (identifiant, séquence, mots clés, nom du gène, nom du transcrit…).

Il convient de déterminer les bases que vous êtes prêt à pouvoir mettre en relation avec les données lors de votre livrable 1.

Annotation des nouveaux génomes

On distinguera deux types de génomes : les génomes déjà annotés et validés qu'aucun utilisateur ne peut (ré)annoter et les génomes non annotés qui peuvent être annotés par les annotateurs de notre système.

Lorsqu'un génome non annoté est ajouté à la base, le validateur affecte les séquences à annoter à des annotateurs. Chaque séquence ne peut être affectée qu'à un annotateur.

Lorsque les annotateurs se loguent au système ils peuvent accéder à une interface dans laquelle ils accèdent aux séquences qui leur ont été affectées. Une fois effectuée, chaque annotation doit être validée par le validateur avant d'être considérée comme validée dans la base.

En option : *un mail est envoyé aux annotateurs pour les informer qu'une liste de séquences à annoter leur a été attribuée.*

En option : Forum des annotateurs

Les annotateurs peuvent discuter entre eux des annotations qu'ils effectuent : ils peuvent choisir un ou plusieurs interlocuteurs dans une liste d'annotateurs et de validateurs et s'envoyer des messages sur un forum.

Validation des annotations

Les utilisateurs de profil validateur ont accès à une liste de toutes les annotations en attente de validation, ils peuvent approuver ou rejeter chacune d'entre elles avec un commentaire.

En option : *un mail est envoyé aux annotateurs ayant contribué pour les informer de la décision.*

Visualisation des annotations en cours

Tous les profils utilisateurs peuvent voir les annotations en cours de validation, mais leur forme est clairement différenciée quand on consulte les pages d'information.

Gestion du code source

Utilisez git pour versionner vos source, github ou gitlab. Vous pouvez mettre votre projet en publique ou en privé. Dans les deux cas ajouter bryan-brancotte au projet afin qu'il puisse voir vos sources et l'évolution de votre projet.

Les issues et/ou les projects pourront vous aider dans la gestion de votre projet, à vous de savoir si vous voulez vous en servir.

Sécurité dans votre application

Une attention particulière sera portée sur la sécurité de votre application : est-ce que l'accès aux vues est bien contrôlé, est-ce que les entrées utilisateur sont bien filtrées pour ne pas avoir d'injection de code, est-ce qu'aucun vrai mots de passe n'est écrit en dure dans les sources, …

2- Déroulement Projet, jalons et évaluation

Les séances de suivi sont effectuées par Olivier Lespinet (OL) ou Bryan Brancotte (BB) selon les jours.

04 Décembre : Démarrage du projet . Formez des groupes de 3 à 4 ou 4 à 5.

11 Décembre (BB) : Suivi des projets et questions par groupes

08 janvier(BB) : Suivi des projets et questions par groupes

15 Janvier (OL) : Suivi des projets et questions par groupes

17 janvier: rendu Livrable 1 (bryan.brancotte@pasteur.fr)

[Projet-Web Jalon 1] noms) vous enverrez un zip avec les sous-repertoires suivants

- Fichier groupe.md avec
  - Lien vers le projet git
  - noms/prénoms/mails de votre équipe projet
- BD : Un document comportant un diagramme des classes UML
- Web : Le site contient une ébauche du thème graphique, les principales vues sont partiellement en place même si elles ne contiennent pas d'information, les boutons ne sont pas forcement tous fonctionnels mais permettent de se rendre compte de l'utilisation futur. Lorsque les vues sont manquantes, des maquettes donnent une idée de ce que vous envisagez.

- Fonctionnement : Une description du fonctionnement global de l'outil du point de vue de l'utilisateur (on débute sur http://localhost:8000/ puis si on clique sur le bouton « recherche » la page… s'ouvre). Cette description peut prendre la forme de votre choix (document rédigé, slides explicites, schémas commentés…)
- Projet : un fichier simple avec la liste des fonctionnalités que vous comptez faire (parmi les fonctionnalités optionnelles) et la liste des bases extérieures auxquelles vous proposerez l'accès.

22 janvier (BB) : Retour sur le rendu 1, suivi des projets.

29 janvier (OL) : Suivi des projets et questions par groupes.

5 février (BB) : Suivi des projets et questions par groupes.

8 février : rendu Livrable (bryan.brancotte@pasteur.fr et olivier.lespinet@i2bc.paris-saclay.fr)

[Projet-Web Jalons 2] noms) vous enverrez un .zip avec les sous-repertoires et fichiers suivants

- Fichier groupe.md avec
  - Lien vers le projet git
  - noms/prénoms/mails de votre équipe projet
- Répertoire source: ensemble des sources du projet. En se plaçant dans le répertoire contenant le fichier manage.py, on doit pouvoir faire le python manage.py migrate, puis python manage.py import-my-data et ensuite python manage.py runserver afin d'utiliser l'application.
- Répertoire Doc : (1) documentation orientée utilisateur de l'outil (équivalent 3-5 pages), (2) documentation technique décrivant sur quelles installations l'outil repose (3 pages) et (3) retour d'expérience sur l'organisation du projet (1/2 page)

12 février: Soutenance du projet

Critères d'évaluation

1) Documentation : 30%

2) Présentation orale (10 min + 5 min questions) : 30%

3) Qualité des sources : 40% - Attention : avant de considérer des fonctionnalités optionnelles vous devez vous assurer que les fonctionnalités attendues sont opérationnelles et de bonne qualité.

Horaires et groupes constitués

9h00 – 9h30 Open Desk

9h30 – 10h00 Groupe 1

10h00 – 10h30 Groupe 2

10h30 – 11h00 Groupe 3

11h00 – 11h30 Groupe 4

11h30 – 12h30 Open Desk

Groupe 1 : LOTH, BEIGEAUD, MAILLIE, LE ROY

Groupe 2 : HANNA, DEMEULLE, BOZIER, GUALDONI

Groupe 3 : CHAMBE, MEYDAN,PARIZOT, CARRE

Groupe 4 : JOURDAIN, FAYE, Jaffar HUSSEIN, MELIGA, AMIENS