

Algorithmique et bioinformatique

Enoncé du projet

Université de Mons - Année académique 2014-2015

Enseignants : O. Delgrange & D. Maslowski

1 Le sujet

A partir d'une collection de fragments (chacun d'une longueur variant entre 500 et 700 pb), il vous est demandé de concevoir un programme d'assemblage de ceux-ci, qui utilise l'approximation de type Greedy, et qui fournit une séquence cible attendue.

2 Consignes

- Le projet se fera par groupe de 2 étudiants. Veillez à répartir correctement le travail entre chaque étudiant ;
- Les notes théoriques sur l'assemblage de fragments ainsi que trois collections de fragments sont disponibles sur la plate-forme e-learning¹.
- Les complications prises en compte lors du séquençage des séquences sont : les erreurs de séquençage et l'orientation inconnue ;
- Les fragments sont donnés dans le format fasta : la séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre qui doit commencer par le caractère '>'. Plusieurs séquences peuvent être mises dans un même fichier ;
- Les séquences cibles fournies par votre programme seront données dans le format fasta. La ligne de titre sera formatée comme suit :

```
> Groupe-num_groupe Collection num_collection Longueur longueur_sequence_cible
```

- L'implémentation se fera en *Java*.

3 Outil

Afin de tester la qualité de vos résultats, l'utilisation de l'outil *dotmatcher* (outil de dotplot de deux séquences) est conseillée. Cet outil fonctionne en ligne de commande et est disponible :

- sous linux : télécharger et installer le package **emboss** (aussi disponible dans les dépôts) ;
- sous Windows : téléchargeable à cette adresse :
www.interactive-biosoftware.com/embosswin/embosswin.html.

L'utilisation en ligne de commande est la suivante :

```
dotmatcher SEQUENCE_1 SEQUENCE_2 -threshold 50
```

où :

- **SEQUENCE_1** est le nom du fichier contenant votre séquence résultat ;
- **SEQUENCE_2** est le nom du fichier contenant la séquence attendue (que vous trouverez sur l'e-learning) ;
- **-threshold 50** est un paramètre nécessaire pour obtenir une bonne visualisation du résultat.

Une démonstration de *dotmatcher* sera réalisée lors de la première séance de travaux pratiques.

1. <https://applications.umons.ac.be/moodle/course/view.php?id=178>

4 Marchine disponible

Vous aurez à votre disposition une machine de l'université afin de tester votre programme sur des entrées volumineuses.

5 Dates

- **Vendredi 6 mars 2015 : composition des groupes**

Pour cette date, vous aurez communiqué la composition de votre groupe à Dany Maslowski (dany.maslowski@umons.ac.be).

- **Vendredi 22 mai 2015 à 16h : remise des fichiers et du rapport**

- *Projet*

Vous déposerez via la plate-forme e-learning une archive (zip, jar, tgz, ...) contenant :

1. les fichiers .java ;
2. les fichiers .fasta reprenant les séquences cibles demandées.

- *Rapport*

Votre rapport contiendra :

1. la répartition des tâches au sein du groupe ;
2. une brève explication de chaque étape de votre démarche ;
3. les points forts, les points faibles et les erreurs connues de votre programme ;
4. une interprétation des résultats obtenus ;
5. une conclusion comprenant une réflexion sur le projet (apports, difficultés rencontrées, ...).

Vous déposerez via la plate-forme e-learning une version électronique de votre rapport (au format PDF).

- **Evaluation orale**

Une évaluation orale du travail sera réalisée pour chaque groupe. Un horaire vous sera communiqués par la suite.

6 Remarque

1. Votre présence sera **obligatoire** lors de la séance du lundi 27 avril 2015 à 15h15. Un contrôle de l'avancement du travail y sera effectué ;
2. Si vous avez des questions, pendant toute la durée du projet, vous prendrez contact par e-mail avec D. Maslowski (dany.maslowski@umons.ac.be) ;
3. La date de dépôt est stricte. Le site n'acceptera plus de dépôt après 16h.