

Reconstruire l'ADN mitochondrial de la 1ère femme

Objectif de ce travail

3 types de mutations

A chaque fois que l'ADN est recopié, il peut y avoir des erreurs de copies appelées « mutations ». Ces mutations peuvent être de 3 types différents : la plus fréquente est la substitution d'une base (T,G,A,C) par une autre. Il peut aussi y avoir une élimination (ou délétion) ou une insertion ; ces deux dernières étant moins fréquentes car pour insérer des bases, il faut en avoir à insérer quelque part et quand on en élimine, que fait-on du déchet produit ? Dans ces deux derniers cas, une élimination ou une insertion peut concerner un nombre plus ou moins important de base (T,G,A,C).

Transmission de l'ADN muté

Chaque enfant reçoit une partie de l'ADN de ces deux parents. Cet ADN reçu contient les mutations reçues par les parents à leur naissance, et les mutations qui ont eu lieu au cours de leur vie de parents jusqu'à la fécondation de l'ovule (pour l'ADN mitochondrial) qui donnera naissance à l'enfant.

ADN mitochondrial (ADNmt)

Le sigle ADNmt représente la partie mitochondriale de l'ADN humain. Cette partie de l'ADN ne se transmet que de la mère aux enfants et donc se retrouve dans les lignées maternelles (le même travail pourrait être fait sur le chromosome de la masculinité (le chromosome 22) pour une transmission selon la lignée paternelle). Comme l'ADNmt et les mutations se transmettent de mère en fille, l'ADNmt de chaque femme vivant aujourd'hui est celui de la première femme sur lequel sont venues agir les mutations. On supposera dans ce travail qu'une mutation est aléatoire et qu'il n'y a pas de zones privilégiées où les mutations vont avoir lieu (en d'autres mots, cela revient à dire que la probabilité de mutation est uniforme sur l'ensemble de l'ADNmt).

En 2003 a eu lieu le premier séquençage complet d'un ADN humain. Aujourd'hui, nous disposons des ADNmt de plusieurs milliers de femmes, c'est à dire que nous disposons de l'ADNmt de la première femme avec des mutations aléatoires par dessus. **Votre travail est de retrouver l'ADNmt de la première femme.**

Le compte-rendu de votre projet

Vous devrez rendre une archive informatique (au format zip) sous le nom groupe01.zip (le numéro 01 changera éventuellement) qui contiendra l'arborescence suivante :

```
groupe01/
├── css
│   ├── prism.css
│   ├── prism.js
│   └── style.css
├── groupe.txt
├── historique.txt
├── index.html
└── py
    ├── ADMmt_eve_mitochondriale.py
    ├── analyse_donnees.py
    ├── comptage_mutation.py
    └── tri_des_donnees.py
```

Toutes les réponses aux questions posées sont à mettre dans le fichier `index.html` que vous pouvez modifier

comme vous le souhaitez.

■ Travail de groupe

Vous allez travailler en duo et vous allez avoir chacun plusieurs rôles.

■ Liste des rôles

Le rédacteur : c'est celui qui centralise les réponses (et les sauvegarde en cas de perte). Il veille à ce que les réponses soient accessibles au groupe même en cas de maladie.

Le vérificateur : c'est celui qui vérifie que ce qui est demandé dans l'énoncé est bien fait, que les programmes font ce qu'ils doivent faire et qu'ils génèrent tous les fichiers qu'ils doivent générer. C'est aussi celui qui vérifie qu'il ne manque pas de réponses aux différentes questions. Il garde un oeil sur ce qui est demandé sur la grille de correction.

L'intégrateur : c'est celui qui intègre les participations des autres membres du groupe dans les programmes à rendre.

Le planificateur : c'est celui qui planifie le travail du groupe en fonction des avancées de chaque membre du groupe. En début de cours, il fait le bilan de ce qui a été fait les dernières fois, il sait où en est le travail de groupe, qui fait quoi en début de séance.

■ Répartition des rôles

Répartissez-vous les rôles. Vous remplirez un fichier `historique.txt` qui expliquera séance par séance qui a fait quoi et quand. Les rôles ne sont pas figés.

■ A la maison

Préparez ce que vous avez à faire en fonction de vos rôles à la maison pour être efficace en cours.

■ Données fournies

■ Fichier de données

Le fichier `mtDNA.fasta` contient 62534 séquences d'ADN mitochondrial (1 Go) téléchargées depuis le site <https://www.mitomap.org/MITOMAP> (base de donnée sur le génome mitochondrial humain).

Certaines de ces séquences sont **ambiguës** (il y a des lettres autres que T, G, C ou A à cause d'une indétermination) ou **partielles** (il en manque une partie).

■ Le format fasta

Le format fasta est un des formats utilisés pour stocker un ADN. Vous pouvez l'ouvrir avec n'importe quel éditeur de texte, les informations étant au format texte. En voici les 3 premières lignes d'un exemple :

```
>OR166265.1 Homo sapiens haplogroup V15 mitochondrion, complete genome
GATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GCATGCACGCGATAGCATTGCGAGACGCTGGAGCCGAGCACCCCTATGTCGAGTATCTGTCTTTGATTC
```

Attention, ouvrir un fichier de 1 Go dans un éditeur texte mobilise les ressources de votre machine. En ligne de commande, la commande `head -100 mtDNA.fasta` affiche les 100 premières lignes du fichier fasta.

■ Lire le format fasta avec le module biopython

Pour lire le format fasta, nous allons utiliser le module biopython (`pip3 install biopython`) :

```
from Bio import SeqIO
for record in SeqIO.parse('mtDNA.fasta', 'fasta') :
    print(record.id, len(record), record.description)
```

Dans l'interpréteur python, `help(record)` vous donnera des informations (rechercher par exemple `id`, `seq`, `description` ou `name` ...).

■ Ecrire une séquence ADN créée dans un fichier fasta

```
from Bio import SeqIO
from Bio.Seq import Seq
from Bio.SeqRecord import SeqRecord

# pour écrire une séquence ADN qu'on vient de déterminer
chaine_de_caractere_cree = '....'
s = SeqRecord(Seq(chaine_de_caractere))
s.id = "Nom Prénom ou Membres du groupe"
s.description = "ADNmt de la première femme"
SeqIO.write(s, "eve.fasta", 'fasta')
```

■ Copier des séquences dans un nouveau fichier fasta

```
from Bio import SeqIO

# pour écrire une série de séquences ADN
liste = []
for record in SeqIO.parse('mtDNA.fasta', 'fasta') :
    if condition :
        liste.append(record)
SeqIO.write(liste, "mon_fichier.fasta", 'fasta')
```

■ La grille de notation

La grille de notation est fournie en annexe de ce document.

Elle précise les points les plus importants du projet en terme de note. Jetez-y un oeil.

■ Analyse des données

Toutes les réponses aux questions seront écrites dans le fichier `index.html` et les différents fichiers demandés devront être au bon endroit dans l'arborescence de l'archive `groupe01.zip` qui sera rendue.

(1) On ne travaillera dans ce projet que sur des séquences que l'on peut garantir **complètes** et **non ambiguës** d'ADNmt. Si besoin, relire le début de l'énoncé pour savoir ce qu'est une séquence ambiguë.

Trouver un moyen de sélectionner ces données via, entre autre, `record.description`.

Vous devez trouver 46181 séquences garanties complètes non ambiguës et les écrire dans un fichier `sequences.fasta`.

Mettre le fichier `mtDNA.fasta` décompressé dans le répertoire `groupe01/py/`.

Compléter le fichier `groupe01/py/tri_des_donnees.py`.

L'exécution de ce programme doit créer le fichier `sequences.fasta`.

Pour suivre l'exécution de votre programme, vous pouvez vous inspirer du code ci-dessous :

```
COMPTEUR, MAXI = 0, 62534
for record in SeqIO.parse('mtDNA.fasta', 'fasta') :
    COMPTEUR += 1
    if COMPTEUR % 1000 == 0 :
        print(COMPTEUR, '/', MAXI)
    print(record.id, len(record), record.description)
```

(2) Ecrire un programme `groupe01/py/analyse_donnees.py` qui lit toutes les séquences sélectionnées à la question précédente et écrit un fichier `analyse.txt` qui contiendra des lignes du genre « longueur de séquence : nombre de séquence ayant cette longueur ».

Les lignes du fichier seront classées par longueur de séquence croissante.

Dans votre compte-rendu, vous expliquerez quelle structure de donnée vous utilisez et pourquoi.

Vérifier que vous trouvez bien les mêmes valeurs que celles-ci dessous :

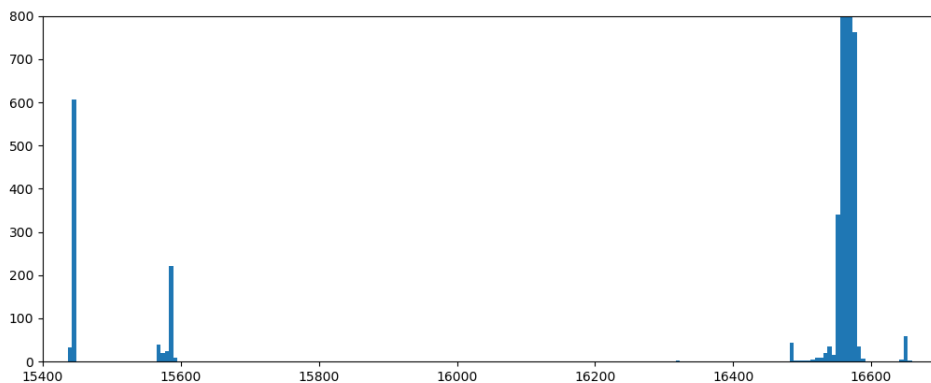
15437 : 24
15445 : 12
15446 : 578

- (3) Parmi les séquences trouvées,
- quelle est la longueur de la plus courte ?
 - de la plus longue ?
 - de la plus fréquente ?
 - de la seconde plus fréquente ?

(4) Créer un tableau `longueurs` qui contient les longueurs de chaque séquence complète sélectionnée. Si plusieurs séquences ont la même longueur, cette longueur doit apparaître autant de fois qu'il y a de séquence.

(5) Représenter l'historgramme où il y a la longueur de la séquence ADNmt en abscisse et le nombre de séquence d'ADNmt en ordonnée. On enregistrera la figure sous le nom `histogramme.png`. Ci-dessous un code pour vous inspirer et le résultat à obtenir. Vous pouvez mettre différents zooms dans votre compte-rendu s'ils vous semblent pertinents.

```
import matplotlib.pyplot as plt
plt.figure(figsize=(12.4,4.8))
plt.xlim([15400, 16700])
# plt.ylim([0,800])
plt.hist(longueurs, bins=200) # le tableau longueurs de la question 4
plt.savefig('histogramme.png')
plt.show()
```



(6) On prendra comme référence la longueur de séquence la plus fréquente et on supposera que c'est la longueur de la séquence de la première femme.

Vous avez déjà déterminer cette valeur qui va servir de référence. Quelle est-elle ?

Ecrire un fichier (dont le nom sera `xxx.fasta` où `xxx` correspond à la valeur de référence citée ci-dessus) contenant toutes les séquences de cette longueur parmi les séquences garanties complètes et non ambiguës.

(7) Vérifier que l'exécution de votre fichier `groupe01/py/analyse_donnees.py` génère sans erreur tous les fichiers attendus :

- `analyse.txt` (question 2)

- `histogramme.png` et éventuellement d'autres fichiers images (question 5)
- `xxx.fasta` (question 6)

(8) Il y a 3 sortes de mutations.

Quelle est la conséquence de chaque mutation sur la longueur d'une chaîne d'ADN ?

A quel type de mutation êtes-vous confrontés en ne sélectionnant que des séquences de même longueur ?

■ En route vers l'ADNmt de la première femme

(9) Quel est l'ADNmt de la première femme ?

Vous écrirez un fichier `eve.fasta` au format `fasta` à l'aide du programme

`groupe01/py/ADMmt_eve_mitochondriale.py` pour présenter votre réponse.

L'exécution de votre fichier `ADMmt_eve_mitochondriale.py` doit générer le fichier `eve.fasta`.

(10) Votre compte-rendu de projet se fera sous forme d'une page web intitulée `index.html` qui devra être appelée une feuille de style `style.css`.

Vous y expliquerez votre démarche pour trouver l'ADNmt de la première femme, il devra y avoir des images, la réponse aux questions, vos démarches, ...

Votre fichier sera analysé par le programme « tidy » (`tidy index.html` en ligne de commande) qui vous fournira des avertissements et des erreurs.

(11) Créer un fichier `groupe.txt` contenant le nom de chaque personne dans le groupe à mettre dans le répertoire `groupe01`.

■ Compter les mutations

(12) Pour chaque séquence considérée de `xxx.fasta`, on souhaite déterminer le nombre de mutation entre la séquence de l'ADNmt de la première femme et la séquence considérée.

C'est un vrai travail de réflexion, de programmation, qui n'est pas immédiat.

A vous d'essayer de comparer basiquement une première séquence et celle de la première femme, puis d'améliorer au fur et à mesure des difficultés rencontrées.

Pour vous aider à commencer, lire le fichier `groupe01/py/comptage_mutation.py`.

Modifier ce qui doit être modifié.

Que doit-on rechercher dans le fichier produit pour trouver toutes les fois où les séquences d'ADN sont différentes ?

Exécuter le programme.

Décrire dans votre compte-rendu ce que vous observez pour un `numero_sequence` 1 et 3.

(13) On veut trouver le nombre moyen de mutation entre une séquence actuelle (contenue dans `xxx.fasta`) et la séquence de la première femme. On supposera que si on connaît le nombre de mutation de 400 séquences actuelles, la moyenne sera représentative de la réalité.

Il faut donc être capable de compter les mutations pour 400 séquences minimum.

Il y a plusieurs stratégies possibles. A vous de réfléchir et de proposer votre stratégie.

Votre stratégie sera expliquée dans votre compte-rendu.

Vous la testerez ensuite en programmant.

Si vous donnez le nombre de mutation d'une séquence actuelle, il faudra qu'il y ait un fichier `mutations.txt` qui contiendra des lignes du type

`record.id : nombre de mutation`

Vous écrirez dans le fichier réponse une conclusion : entre ce que vous aviez prévu et ce que vous avez réellement fait, que s'est-il passé ?

(14) L'étude de la transmission des mutations génération par génération propose des taux de mutation par génération :

- un taux de 1 mutation toutes les 33 générations a été mesuré dans cet article : Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, Holland MM. A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet. 1997 Apr;15(4) :363-8. doi : 10.1038/ng0497-363. PMID : 9090380.
- un taux mesuré de 1 mutation pour 25 à 40 générations et un taux issu des modèles évolutionnistes standards (arbres phylogénétiques) de 1 mutation pour 600 générations, soit 1 mutation pour 12000 ans dans l'article : Gibbons, Ann. "Calibrating the Mitochondrial Clock." Science 279 (1998) : 28 - 29.

Donner une estimation de l'âge de la première femme à l'aide des taux mesurés et à l'aide des taux évolutionnistes standard de mutation. Discuter vos résultats.

(15) Au lieu de considérer la longueur de séquence d'ADNmt la plus fréquente, on peut considérer la deuxième longueur de séquence la plus fréquente.

Construire le fichier `eve2.fasta`. Comparer les deux séquences `eve.fasta` et `eve2.fasta`.

S'il vous reste du temps, à vous de voir s'il est préférable de travailler le fichier de réponse `index.html` ou de compter le nombre moyen de mutation entre cette nouvelle séquence de la 1ère femme et les séquences actuelles. Retrouve-t-on les mêmes résultats ?

(16) Vérifier que votre dossier ne contient plus aucun fichier `fasta` et fichiers accessoires (l'exécution de vos programmes doit générer tous les fichiers), puis créer l'archive `groupe01.zip` et la rendre au professeur. Modifier éventuellement le numéro de l'archive en accord avec le professeur.

Votre archive, une fois créée, ne doit pas faire plus que quelques ko.

Grille d'évaluation du projet Eve mitochondriale

9 Fichiers historique.txt et groupe.txt présents ? Membres / 1

Analyse des données

1 Sélection des données complètes et non ambiguës / 2

2 Fichier analyse.txt / 3

Occurrences des longueurs de séquences + longueur de séquence croissante

3 Longueur de la séquence la plus courte / 3

Longueur de la séquence la plus longue

Longueur de la séquence la plus fréquente

5 Fichier histogramme.png des longueurs des séquences / 3

8 Sur quel type de mutations travaillons-nous ? / 1

L'ADN mitochondrial de la première femme

9 Fichier eve.fasta : validité de la séquence proposée ? / 3

Comment construire l'ADN mitochondrial de la première femme ? / 15

10 Fichier index.html (tidy) / 15

Le html est bien formaté ?

Le style n'est défini que dans le fichier style.css ?

La présentation est cohérente et est plaisante ?

La répartition du travail et le fonctionnement du groupe ?

Compter les mutations

13 Combien de séquences ont été comparées avec celle de la 1ère femme ? / 15

Référence des séquences permettant de vérifier le travail fait ?

Comment a-t-on compté les mutations ?

Explications des difficultés rencontrées ?

Nombre moyen de mutation ?

14 Estimation de l'âge de la 1ère femme / 2

15 Résultats similaires pour la deuxième séquence la plus fréquente ? Cette question est facultative