

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 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 mère en fille (le même travail pourrait être fait sur le chromosome de la masculinité (le chromosome 22) pour une transmission père-fils).

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 changera) qui contiendra l'arborescence suivante :

```
groupe01/
├── analyse.txt
├── css
│   ├── prism.css
│   ├── prism.js
│   └── style.css
├── eve.fasta
├── groupe.txt
├── img
│   └── histogramme.png
├── index.html
└── py
    ├── ADMmt_eve_mitochondriale.py
    └── analyse_donnees.py
```

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

Données fournies

Fichier de données

Un 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 étant ambiguës ou partielles, il vous est fourni un fichier `mtDNA_GL.fasta` qui contient 46181 séquences complètes d'ADNmt.

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
GATCACAGGTCTATCACCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGG
GCATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTCGCAGTATCTGTCTTTGATTC
```

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, `help(record)` vous donnera des informations (rechercher par exemple `id`, `seq`, `description` ou `name` ...).

Analyse des données

(1) On ne travaillera dans ce projet que sur les séquences complètes d'ADNmt. Trouver un moyen de sélectionner ces données via `record.description`. Combien de séquence avez-vous trouvé ?

(2) Ecrire un programme `analyse_donnees.py` qui lit toutes les séquences sélectionnées à la question précédente et écrit un fichier `analyse.txt` qui contient des lignes du genre « longueur de séquence : nombre de séquence ayant cette longueur ». Vérifier que vous trouvez bien les mêmes valeurs.

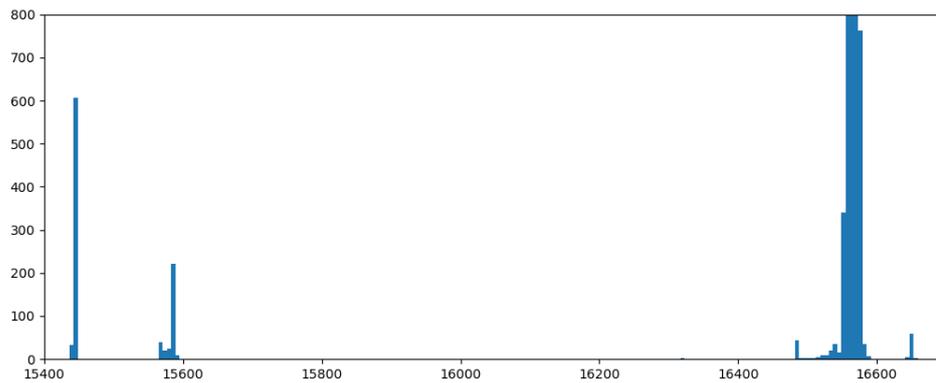
```
15437 : 24
15445 : 12
15446 : 578
```

(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'histogramme où il y a la longueur de la séquence ADNmt en ordonnée 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)
```



(6) On prendra la longueur de séquence la plus fréquente comme référence et on supposera que c'est la longueur de la séquence de la première femme. Il y a 3 sortes de mutations. A quel type de mutation êtes-vous confrontés ?

En route vers l'ADNmt de la première femme

(7) Quel est l'ADNmt de la première femme ? Vous écrirez un fichier `eve.fasta` au format `fasta` à l'aide du programme `ADMmt_eve_mitochondriale.py` pour présenter votre réponse. Vous pouvez vous inspirer du code ci-dessous.

```
from Bio import SeqIO
from Bio.Seq import Seq
from Bio.SeqRecord import SeqRecord
chaine_de_caractere = 'ATCTGATGTCAGTAC'*100
s = SeqRecord(Seq(chaine_de_caractere))
s.id = "Nom-Prénom"
s.description = "ADNmt de la première femme"
SeqIO.write(s, "ADMmt_eve_mitochondriale.fasta", 'fasta')
```

(8) Votre compte-rendu de projet se fera sous forme d'une page web intitulée `index.html` qui devra 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, la répartition du travail dans le groupe, ... Cf l'archive informatique reçue contenant ce travail.

(9) Créer un fichier `groupe.txt` contenant le nom de chaque personne dans le groupe.

Compter les mutations

(10) Pour chaque séquence considérée, 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 chaque séquence considérée, il faudra être capable de dire combien de mutation est-ce que vous avez compté, à quel endroit et il faudra que la personne qui vous relit soit capable de retrouver la séquence correspondante dans vos données.

(11) 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.

(12) Au lieu de considérer la longueur de séquence d'ADNmt la plus fréquente, on peut considérer la deuxième longueur de séquence la plus fréquente. Retrouve-t-on des résultats similaires ?

Grille d'évaluation du projet Eve mitochondriale

9 Le fichier groupe.txt est présent ? Membres / 1

Analyse des données

1 Sélection des données complètes / 2

2 Fichier analyse.txt / 3

Occurrences des longueurs de séquences

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

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

L'ADN mitochondrial de la première femme

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

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

8 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

10 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 ?

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

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