Hadoop et Mapreduce TD/TP

Exercice1

Rappelons que l'ADN (acide désoxyribonucléique) est une molécule dont la structure (primaire) peut être vue comme une chaîne de caractères. Les caractères sont de quatre sorte : A, T, C et G et correspondent aux quatre nucléotides (adénine, thymine, cytosine, guanine) de la molécule.
Construire une job Mapreduce permettant à partir de la bactérie Escherichia coli de calculer le nombre de A, de T, de G et de C dans la chaîne.

Par exemple, pour la chaine suivante :

ACACACAGT

Le résultat sera :

```
A    4
C    3
G    1
T    1
```

Exercice2

Construire une job Mapreduce permettant à partir de la bactérie Escherichia coli de calculer la position et le nombre de chaque codon (Triplet de nucléotides) dans la chaîne.

Par exemple, pour la chaine suivante :

ACACACAGT

Le résultat sera :

```
ACA    1    3
CAC    2    2
ACA    3    3
CAC    4    2
ACA    5    3
CAG    6    1
AGT    7    1
```

Exercice 3

Supposons que la bactérie est atteinte par un virus qui affecte la machinerie de la réplication aléatoirement en changeant la manière dont chaque nucléotide est recopié : chaque A peut être répliqué comme AAA, chaque C peut être répliqué comme CCCC, chaque G peut être répliqué comme GGGG, et chaque T peut être répliqué comme TTT.

Ecrire une job Mapreduce qui prend en entrée la bactérie Escherichia coli et qui renvoie le nombre et la position des réplications effectués.

Par exemple, pour la chaine suivante :

ACACAAACAGGGGTCGTTT

Le résultat sera :

```
AAA     1    5
GGGG    1    10
TTT     1    17
```