

Informatique – Biologie L2 – TP2 : Programmation Python

Jean-Baptiste LAMY (jiba@tuxfamily.org) d'après Aurélien Mazurie (parties 1 & 2)

16 février 2007

Nous allons travailler sur les séquences d'ADN suivantes (brin codant, que vous pouvez recopier en Python ou télécharger sur le campus virtuel, adresse <http://www-limbio.smbh.univ-paris13.fr/campusvirtuel/>) :

```
adn1 = "atgagtgaacgtctgagcattaccccgctggggccgtatatcggcgca"
adn2 = "atgagtgaacgtctgagcattaccccgctggggccgtatatcggcgcacataacgg"
adn3 = "atgagtgaacgt"
adn4 = "atgagt"
adn5 = "atgagcgagaggctgagcattaccccgctggggccgtatatcggcgcacataacccaagggt"
```

1 Introduction

Quel est la longueur de l'ADN 1 ?

Afficher successivement les différentes paires de base de l'ADN 1.

Afficher successivement les différentes paires de base de l'ADN 1 et leur position.

Résultat attendu :

```
position 1 : a
position 2 : t
...
```

1.1 Calculer le pourcentage de G+C de l'ADN 1

Pour cela, il faut compter le nombre de G et de C dans l'ADN 1, puis le divisant par la taille de la séquence.

2 Traduction de séquences d'ADN

2.1 Transcrire l'ADN 1 en ARNm (on l'appellera ARN 1)

2.2 Traduire la séquence d'ARN 1

Aidez-vous de la table de traduction suivante (que vous pouvez recopier en Python ou télécharger sur le campus virtuel) :

```
code_genetique = {
    'uuu' : 'F', 'ucu' : 'S', 'uau' : 'Y', 'ugu' : 'C',
    'uuc' : 'F', 'ucc' : 'S', 'uac' : 'Y', 'ugc' : 'C',
    'uua' : 'L', 'uca' : 'S', 'uaa' : '*', 'uga' : '*',
    'uug' : 'L', 'ucg' : 'S', 'uag' : '*', 'ugg' : 'W',
    'cuu' : 'L', 'ccu' : 'P', 'cau' : 'H', 'cgu' : 'R',
    'cuc' : 'L', 'ccc' : 'P', 'cac' : 'H', 'cgc' : 'R',
    'cua' : 'L', 'cca' : 'P', 'caa' : 'Q', 'cga' : 'R',
    'cug' : 'L', 'ccg' : 'P', 'cag' : 'Q', 'cgg' : 'R',
    'auu' : 'I', 'acu' : 'T', 'aau' : 'N', 'agu' : 'S',
    'auc' : 'I', 'acc' : 'T', 'aac' : 'N', 'agc' : 'S',
    'aua' : 'I', 'aca' : 'T', 'aaa' : 'K', 'aga' : 'R',
    'aug' : 'M', 'acg' : 'T', 'aag' : 'K', 'agg' : 'R',
    'guu' : 'V', 'gcu' : 'A', 'gau' : 'D', 'ggu' : 'G',
    'guc' : 'V', 'gcc' : 'A', 'gac' : 'D', 'ggc' : 'G',
    'gua' : 'V', 'gca' : 'A', 'gaa' : 'E', 'gga' : 'G',
    'gug' : 'V', 'gcg' : 'A', 'gag' : 'E', 'ggg' : 'G',
}
```

À quel type de donnée correspond cette table ?

- 2.3 Écrire des fonctions pour les 3 tâches qui ont été réalisées précédemment (calcul du pourcentage de GC, transcription de l'ADN, traduction de l'ARN)
- 2.4 Utiliser ces fonctions pour calculer le pourcentage de GC, transcrire et traduire l'ADN 2
- 2.5 Modifier la fonction de traduction pour qu'elle s'arrête au codon stop
- 2.6 L'ADN 2 et l'ADN 5 codent-ils pour la même protéine ?

3 Calcul du point de fusion (Tm) de l'ADN

Plusieurs formules existent pour calculer le point de fusion (Tm) d'une chaîne d'ADN.

3.1 Règle de Wallace

Cette règle fonctionne sur des oligonucléotides très courts (≤ 15 bases). La formule est la suivante :

$$Tm = 8 + 2 \times (nbAT) + 4 \times (nbGC)$$

où nbAT est le nombre de AT, et nbGC le nombre de GC.

Écrire une fonction pour calculer le point de fusion d'un ADN avec la règle de Wallace. Écrire une fonction pour calculer le point de fusion d'un ADN avec la formule de Howley. Utiliser votre fonction pour calculer le point de fusion de l'ADN 3.

3.2 Calcul par le contenu en GC (Howley *et al.*)

Cette formule est valable pour les longs oligonucléotides (> 10 bases). La formule est la suivante :

$$Tm = 67,5 + (0,34 \times \%GC) - (395/nbbases)$$

où %GC est le pourcentage de GC, et nbbases le nombre de base dans la séquence.

Écrire une fonction pour calculer le point de fusion d'un ADN avec la formule de Howley. Utiliser votre fonction pour calculer le point de fusion de l'ADN 3.

3.3 Méthode du plus proche voisin

Cette méthode est valable pour les séquences entre 20 et 60 nucléotides. Elle s'appuie sur les propriétés thermodynamiques des dinucléotides (paires de nucléotides voisins).

$$Tm = \frac{1000 \times dH}{A + dS + R \times \ln(\frac{C}{4})} - 273,15 + 16,6 \times \log(cK^+)$$

où :

- dH est la somme des enthalpies de chaque dinucléotide
- dS est la somme des entropies de chaque dinucléotide
- A est l'entropie de la formation de l'hélice de l'ADN (constante -10,8 cal)
- R est la constante des gaz parfaits (1,984 cal/grad x mol)
- C est la concentration en oligonucléotide (nous prendrons 250 pmol/l)
- cK⁺ est Concentration en ions potassium dans la solution (nous prendrons 50 mmol/l)

Les enthalpies et entropies des dinucléotides sont données par les dictionnaires suivants :

```
enthalpies_des_dinucleotides = {
  "aa" = -9.1,
  "ag" = -7.8,
  "ac" = -6.5,
  "at" = -8.6,
  "ga" = -5.6,
  "gg" = -11.0,
  "gc" = -11.1,
  "gt" = -6.5,
  "ca" = -5.8,
  "cg" = -11.9,
  "cc" = -11.0,
  "ct" = -7.8,
  "ta" = -6.0,
  "tg" = -5.8,
  "tc" = -5.6,
  "tt" = -9.1,
```

```

}
entropies_des_dinucleotides = {
  "aa" = -24.0,
  "ag" = -20.8,
  "ac" = -17.3,
  "at" = -23.9,
  "ga" = -13.5,
  "gg" = -26.6,
  "gc" = -26.7,
  "gt" = -17.3,
  "ca" = -12.9,
  "cg" = -27.8,
  "cc" = -26.6,
  "ct" = -20.8,
  "ta" = -16.9,
  "tg" = -12.9,
  "tc" = -13.5,
  "tt" = -24.0,
}

```

Comment faire pour lister les dinucléotides d'une séquence d'ADN ?

Écrire une fonction pour calculer le point de fusion d'un ADN avec la méthode du plus proche voisin. Pour calculer le log et le ln, il faut :

- placer "import math" au début de votre fichier
- ln s'obtient alors avec "math.log"
- log s'obtient alors avec "math.log10"

Utiliser votre fonction pour calculer le point de fusion de l'ADN 1.

3.4 Fonction générique

Écrire une fonction qui calcule le point de fusion d'un ADN, en utilisant automatiquement la formule la plus appropriée. En utilisera en priorité la méthode du plus proche voisin, puis la formule de Howley, puis celle de Wallace. Utiliser cette fonction pour calculer le point de fusion des ADN 1, 2, 3 et 4.

3.5 Fonction générique (2)

Modifier la fonction précédente pour faire la moyenne des formules de Howley et de Wallace, dans le cas où les deux peuvent être utilisées. Utiliser cette fonction pour calculer le point de fusion des ADN 1, 2, 3 et 4.