3I019 - Examen écrit – 1er section 28 mai 2018 (Corrigé)

H. Richard et J. Bernardes

durée: 2 heures

Les documents ne sont pas autorisés. Vous pouvez utiliser la calculatrice si nécessaire.
Les téléphones portables doivent être éteints.
Le barême est donné à titre indicatif.

Exercice 1 - Biologie Moléculaire

(Question 1) Si une molécule d'ADN contient 8 % d'adénine (A) et 42 % de guanine (G), elle contient également

a. 8 % d'uracile (U) et 42 % de cytosine (C).
b. 42 % d'uracile (U) et 8 % de cytosine (C).
c. 8 % de thymine (T) et 42 % de cytosine (C). (X)
d. 42 % de thymine (T) et 8 % de cytosine (C).

(Question 2) Un codon correspond à un triplet

a. de nucléotides d'acides ribonucléiques de transfert (d'ARNt)
b. de nucléotides d'acide désoxyribonucléique (ADN)
c. de nucléotides d'acide ribonucléique messager (ARNm) (X)
d. d'acides aminés

(Question 3) La traduction conduit à la formation...

a. d'une séquence d'ARNm identique au brin transcrit où les U remplacent les T
b. d'une séquence d'ARNm complémentaire du brin transcrit de l'ADN
c. d'une séquence d'acides aminés associés par des liaisons peptidiques **(X)**
d. d'une séquence d'ADN complémentaire du brin transcrit
e. d'une séquence d'ADN complémentaire du brin non transcrit

(Question 4) Expliquer les différences entre un génome, un gène et une protéine.

(Question 5) ADN et protéines sont des polymères présents chez toutes les êtres vivants qui peuvent être représentées sous la forme d'une chaine de caractère. Pour chacune de ces molécules indiquez à quoi correspondent les lettres et combien de lettre sont utilisées.

Exercice 2 - Détection/Prediction de gènes, modèle background et modèle Bernoulli

(Question 6) Parmi les informations suivantes, lesquelles sont vraies et aident à prédire les gènes dans le génome d'un organisme **eucaryote** ?

a) Un gène est un segment d'ADN continu qui commence par un codon start et finit par un codon stop
b) De longs "Open Reading Frames" dans un génome eucaryote ont une forte probabilité de correspondre à un véritable gène (X)
c) Pour bien prédire des gènes dans une espèce A nous pouvons utiliser un génome de référence d'une espèce B similaire ou proche de A (X)
d) c'est plus simple prédire la localisation d'un gène dans les organismes eucaryotes que dans les organismes procaryotes.

(Question 7) Considérons la figure ci dessous, où les régions qui ont les mêmes propriétés sont représentées de la même couleur:

Sélectionnez la (ou les) réponse(s) correcte(s).

a) Il s'agit du gène d'un organisme procaryote
b) Il s'agit du gène d'un organisme eucaryote (X)
c) La lettre A correspond au terme promoteur, B à intron et C à exon.
d) La lettre A correspond au terme exon, B à intron et C à promoteur.
e) La lettre A correspond au terme intron, B à exon et C à promoteur.
f) La lettre A correspond au terme promoteur, B à exon et C à intron. (X)

(Question 8) Soit le segment d'ADN 5' -ACATGTATCGTGATATAA- 3' et le tableau du code génétique en annexe :

a) donner la séquence d'ADN répliquée.
ACATGTATCGTGATATAA

b) donner la séquence transcrite en ARN.
ACAUGUAUCGUGAUUAA

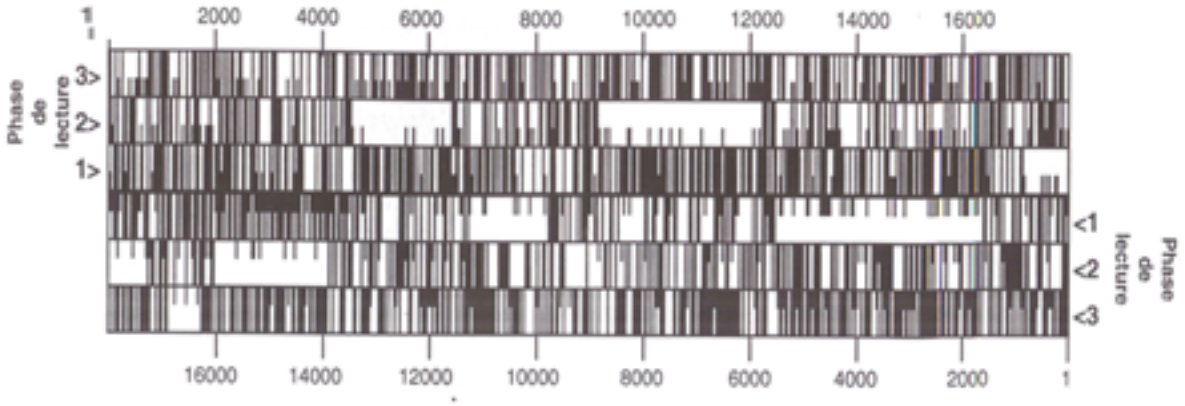
c) donner les séquences protéique dans les 6 cadres ouverts de lecture (Open Read Frame (ORF))".

5'3' Frame 1 T C I V I
5'3' Frame 2 H V S Stop L
5'3' Frame 3 Met Y R D Stop
3'5' Frame 1 L I T I H
3'5' Frame 2 Stop S R Y Met
3'5' Frame 3 N H D T C

d) donner la séquence d'acide aminé de la protéine codée, sur quel brin et dans quel cadre la trouvez vous ?

5'3' Frame 3 Met Y R D Stop

(Question 9) Qu'est-ce qu'une séquence codante ? La figure ci-dessous est issue d'un algorithme permettant de détecter ce type de séquence. Expliquer le principe de cet algorithme et la représentation obtenue. Combien de séquences codantes sont détectable sur la figure. Les annoter sur la figure à l'aide de flèches.



(Question 10) Un étudiant en bioinformatique a utilisé deux méthodes, notées A et B, pour prédire la localisation d'un gène sur une séquence. Pour simplifier nous attribuons à chaque nucléotide d'un génome la valeur 0 ou 1, 0 si le nucléotide ne fait pas partie d'un gène et 1 si il en fait partie. Voici un exemple

```
genome = 'ACTAGTGCATCGTACGT'
gene   = '00011111111110011'
```

les prédictions des deux méthodes sont les suivantes:
Méthode A : `00001111111000111`
Méthode B : `00111111110011111`,

Pour chaque méthode calculer:
a) La sensibilité =Sen = (le taux de vrais positifs) / (le taux de vrais positifs + le taux de faux négatifs)

```
gene = 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1
metA = 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1
```

```
gene = 0 0 0 1 1 1 1 1 1 1 1 1 0 0 1 1
metB = 0 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1
```

Sen A = 11/(11 + 10) = 9 / (9 + 3) = 9/12
Sen B= 11/(11 + 10) = 10 / (10 + 2) = 10/12

b) La spécificité = Spe = (le taux de vrais négatifs) / (le taux de vrais négatifs + le taux de faux positifs)

Spe A = 00 /(00 + 01) = 4 / (4 + 1) = 4/5
Spe B = 00 /(00 + 01) = 2 / (2 + 3) = 2/5

c) La valeur prédictive = VP = VP = (le taux de vrais positifs) / (le taux de vrais positifs + le taux de faux positifs)

VP A = 11 /(11 + 01) = 9 / (9 + 1) = 9/10
VP B = 11 /(11 + 01) = 10 / (10 + 3) = 10/13

d) Quelle méthode est la plus performante ? Pour répondre à cette question vous pouvez calculer le F-score donné par:

F-score = 2 * (Sen * VP) / (Sen + VP)

F-score A = 2 * (9/12 * 9/10)/ (9/12 + 9/10) = 0.82
F-score B = 2 * (10/12 * 10/13)/ (10/12 + 10/13) = 0.80

(Question 11) Expliquez le modèle background et donnez le code python ou pseudo code d'une fonction qui étant donnée une séquence S en forme de chaine de caractère renvoie le modèle de background sous la forme d'un dictionnaire.

```
def background(S):
    dic = {'A': 0, 'C' : 0, 'G' : 0, 'T' : 0}

    for c in dic:
       dic[c] = S.count(c)
    return dic
```

(Question 12) Etant donné un dictionnaire contenant la probabilité d'apparition de chaque codon, par exemple : `modeleCodons` = {'TTT' : 0.002, 'TTC' :0.004, …}, une séquence de codons sous forme de liste S=['TTT', 'TCA', TGA', …], et le modèle de background en forme de dictionnaire, background = ['A' : 0.25, 'C' : 0.24, 'G' : 0.23, 'T' : 0.27]. Donnez le code python ou pseudo code de la fonction `rapportVraisemblance(S, modeleCodons, background)` qui renvoie le rapport de vraisemblance pour une séquence S;

Rappel : le rapport de vraisemblance pour une séquence S est:
P(S | modeleCodons) / P(S | background)

```
def rapportVraisemblance(S, modeleCodons, background):
    pmc = 1 #probabilité modele codon
    pmb = 1 #probabilité modele background
    for codon in S:
        pmc = pmc * modeleCodons[codon]
        for n in codon:
            pmb = pmb * background[n]
    return pmc/pmb
```

(Question 13) Le génome des mammifères est caractérisé par sa forte hétérogénéité dans sa composition. Par exemple, la moyenne en GC d'un fragment de 100 kb du génome humain peut être aussi faible que 35% ou aussi élevée que 60%, une gamme qui est deux fois plus large que celle généralement observée par exemple chez les poissons. Donner le code python ou le pseudocode pour la fonction `GCcontent(genome, k)` qui renvoie la moyenne de pourcentage GC observé dans les fragment de taille k.

```
def GCcontent(genome, k):
    sumGC = 0
    for i in range(len(S) - k + 1):
        seq = S[i:k+i]
        countC = seq.count('C')
        countG = seq.count('G')
        percCG = (countG + countC)/k
        sumGC = sumGC + percCG

    return sumGC/(len(S) - k + 1)
```

Exercice 3 - Alignement par paires, dot plot et matrices de substitutions

(Question 14) Parmi ces propositions, laquelle **ne s'applique pas** à l'alignement par paire ?

a. Les résidus (nucléotides, acides aminés) sont superposés de façon à maximiser la similarité entre les séquences.
b. Il existe deux sortes de mutations : Substitutions (mismatch) et insertions/délétions (indels ou gaps)
c. L'alignement renvoyé est sensible au système de score utilisé.

d. L'alignement par paire utilise une heuristique pour aligner plus rapidement les séquences. (X)

(Question 15) Quel alignement est utilisé pour prédire si deux séquences sont homologues (si elles ont la même fonction)? Quel alignement est utilisé pour prédire si deux séquences ont le même motif (une région conservée)?

a) local, global
b) local, local
c) global, global
d) global, local (X)

(Question 16) Parmi ces propositions, laquelle **s'applique** à des matrices dot-plot?

a. Le dot-plot permet seulement la visualisation de similarités local.
b. Le dot-plot permet la visualisation de répétitions directes ou inversées. (X)
c. Le dot-plot produit un alignement global.
d. Le dot-plot est une méthode simple et rapide, de complexité quadratique.

(Question 17) Donner la matrice dot-plot de deux séquences A=ACCAGT et B=TGCAG sans considérer une fenêtre et en considérant une fenêtre de taille 3, que remarquez-vous ?

	A	C	C	A	G	T
T						
G						
C			*			
A				*		
G					*	

	A	C	C	A	G	T
T						*
G					*	
C		*	*			
A	*			*		
G					*	

(Question 18) Donner le score des alignements ci-dessous, les coûts d'événement sont match=2, mismatch=1, ouverture de gap = -2, et extension de gap = -0.5

alignement 1	alignement 2	alignement 3
G T T A C G A C - T T C C G - -	G T T A C G A C - - - A C C G -	G T T A C G A C G T T A - - - -

match = 4*2=8 mismatch=1 OpenGap=(-2)*2 GapExt= -0.5 score = 4.5	match = 2*2=4 mismatch=2*1 OpenGap=(-2)*2 GapExt=2*(-0.5) score = 1	match = 4*2=8 mismatch=0 OpenGap=-2 GapExt=(-0.5)*3 score = 4.5

(Question 19) On considère les séquences Seq1= AACGT et Seq2= AATCG, avec le système de score suivant, match=2, mismatch=1, gap=-1

a. Donnez l'alignement **global** optimal et montrez tous les calculs nécessaires (matrice d'alignement)

```
          A    A    C    G    T
     0   -2   -4   -6   -8  -10
A   -2    1   -1   -3   -5   -7
A   -4   -1    2    0   -2   -4
T   -6   -3    0    1   -1   -1
C   -8   -5   -2    1    0   -2
G  -10   -7   -4   -1    2    0

le path est marque en bold et sousligné.

AATCG-
** **
AA-CGT
```

b. Quel est le score de l'alignement et le pourcentage d'identité après alignement ?

score est 0
pourcentage d'identité est 4/6

c. Il y a d'autre alignement optimal? Justifiez votre réponse.

non, cet alignement est le meilleur chemin trouvé dans la matrice.

(Question 20) En utilisant la matrice de substitution ci-dessous et un coût de gap à -5, quel alignement a le meilleur score?

D	6				
E	2	5			
F	-3	-3	6		
G	-1	-2	-3	6	
W	-4	-3	1	-2	11
	D	E	F	G	W

Alignement 1	Alignement 2
DFDW-FE DF-WDGF	DEDW-FE FE-WDWE

Alignement 1: 6 + 6 -5 + 11 -5 -3 -3 = 23 - 16 = 7

Alignement 2 : 2 + 5 -5 +11 -5 +1 +5 = 14

l'alignement 2 a le score plus élevé.

(Question 21) Ecrivez la fonction `calculeScore(seq1, seq2, systemeScore)` (en python ou pseudo-code) qui, étant donner deux séquences alignées seq1 et seq2, et un système de score sur le format d'une liste ([2, 1, -1], premier élément cout d'un match, deuxième mismatch et troisième gap) renvoie le score du alignement.

```
def calculeScore(seq1, seq2, systemeScore):
    score = 0
    for i in range(0, len(seq1)):
        if seq1[i] == seq2[i]:
            score = score + systemeScore[0]
        elsif seq1[i] != seq2[i]:
            score = score + systemeScore[1]
        else:
            score = score + systemeScore[2]
    return score
```

Exercice 4 – Détection de motifs : Algorithmes table de hashage, force brute, Matrices poids position

(Question 22) Soit la séquence suivante de taille 40 ci-dessous, nous savons qu'un motif de taille 5 a été implanté 4 fois dans cette séquence et qu'il peut y avoir une substitution sur ce motif, notre but est de le trouver.

ACGCTGGATCGCTGAACCCCAAAGCTGGCAGCTGGAATGT

a. Quel est le motif implanté? Donnez la table de hachage pour la séquence et une taille de mot de 5. Quel est le mot apparaissant le plus souvent ?
GCTGG:3,CGCTG:2,CTGGA:2,AGCTG:2,ACGCT:1,TGGAT:1,GGATC:1,GATCG:1,ATCGC:1,TCGCT:1,GCTGA:1,CTGAA:1,TGAAC:1,GAACC:1,AACCC:1,ACCCC:1,CCCCA:1,CCCAA:1,CCAAA:1,CAAAG:1,AAAGC:1,AAGCT:1,CTGGC:1,TGGCA:1,GGCAG:1,GCAGC:1,CAGCT:1,TGGAA:1,GGAAT:1,GAATG:1,AATGT:1,ATGTC:1,TGTCA:1,GTCAG:1,TCAGG:1

Le motif est GCTGG (3 fois) avec une version GCTGA qui apparaît 1 fois.

b. Ecrire la fonction python pour construire une table de hashage de tous les mots d'une taille donnée

c. Est ce que le motif apparaît autant de fois qu'attendu ? En considérant le fait qu'il peut y avoir une substitution, proposez une autre version du motif.
Il doivent trouver GCTGA qui apparaît une fois et est le seul mot à une distance de Hamming de 1.

d. (bonus) Que remarquez vous pour les mots ayant le plus grand nombre d'occurrences ?
ici ils pourraient commenter sur le fait que le vrai motif créé des motifs artificiels qui le chevauchent juste par hasard.

e. Ecrire le code python qui, utilisant l'algorithme force brute, trouve le motif consensus sachant qu'il y a un nombre connu, k, d'apparitions du motif dans la séquence. Donnez, en la justifiant, la complexité de l'algorithme brute force dans ce cas.
Il s'agit adaptation du code vu en cours, pas trop dur.

(Question 23) Calculer le modèle de "background" (composition en nucléotides) pour la séquence d'ADN suivante :

a) ACTGCAGTTCTCTAAT
nA= 4, nC = 4, nG=2, nT = 6 donc p = (0.25, 0.25, 0.125, 0.375)

b) Quel est le modèle de "background" si on rajoute des pseudo comptages à 1 ?
nA=5, nC=5, nG=3, nT=7 donc p=(0.07, 0.36, 0.21 ,0.36)

(Question 24) Le facteur de transcription PAX5 chez l'homme est décrit par la liste de séquences suivantes (une sélection est donnée) :

```
TTGGGCAGCCAAGCATGAC
AGAGGCAACCAAGCGTGAC
GGAGTCACCCAAGCGTGAC
ATGCCCAGTCAAGCATGAC
GAGTTCAGCCAAGCGTAGC
CTGGGCAGCAGAGCATGAC
GAAAGCAACCAACCATGAC
GAAATCAGTGATGCATGAC
CAGTTCAATCAAGCATGAC
AGGTTCAGTGAACCGTGAC
```

a. Donnez le motif consensus et la matrice de fréquence et poids position

Le modèle est soit la matrice des fréquences soit la matrice poids position si on connait le background. On a comme matrice de comptage:

pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	3	4	4	2	0	0	10	3	0	1	9	9	0	0	6	0	1	9	0
C	2	0	0	1	1	10	0	1	6	7	0	0	2	10	0	0	0	0	10
G	4	3	6	4	4	0	0	6	0	2	1	0	8	0	4	0	9	1	0
T	1	3	0	3	5	0	0	0	4	0	0	1	0	0	0	10	0	0	0

b. Donnez le modèle background ou modèle nul.

c. Calculez le score pour la séquence suivante: `GAGTTCAGCCAAGCGTAGC`

J'ai pris un des motifs de la liste au dessus ;)

d. Est ce que le motif vous semble très prédictif ? Calculez l'entropie pour les positions 1 et 2 du motif, voir la formule d'entropie en annexe.

Annexes

Code génétique

Le code génétique au niveau de l'ARNm.

1e nucléotide (en 5')	2e nucléotide U	2e nucléotide C	2e nucléotide A	2e nucléotide G	3e nucléotide (en 3')
U	Phe:F	Ser:S	Tyr:Y	Cys:C	U
	Phe:F	Ser:S	Tyr:Y	Cys:C	C
	Leu:L	Ser:S	STOP	STOP	A
	Leu:L	Ser:S	STOP	Trp:W	G
C	Leu:L	Pro:P	His:H	Arg:R	U
	Leu:L	Pro:P	His:H	Arg:R	C
	Leu:L	Pro:P	Gln:Q	Arg:R	A
	Leu:L	Pro:P	Gln:Q	Arg:R	G
A	Ile:I	Thr:T	Asn:N	Ser:S	U
	Ile:I	Thr:T	Asn:N	Ser:S	C
	Ile:I	Thr:T	Lys:K	Arg:R	A
	Met:M	Thr:T	Lys:K	Arg:R	G
G	Val:V	Ala:A	Asp:D	Gly:G	U
	Val:V	Ala:A	Asp:D	Gly:G	C
	Val:V	Ala:A	Glu:E	Gly:G	A
	Val:V	Ala:A	Glu:E	Gly:G	G

formule entropie

$$E = -\sum_{i=1}^{K} p_i \log(p_i)$$