Université Batna 2 **Année universitaire : 2020-2021**

Faculté des Sciences de la nature et de la vie

Département d'Ecologie et environnement

M1 Biotechnologie végétale

Corrigé type de l'Examen de Bioinformatique

Exercice 1 : (14 pts)

I- Cochez la ou les bonnes réponses : **(5 pts = chaque réponse sur un point)**

1- La banque de données GenBank (1 pts)

a- est une banque de données généraliste de séquences protéiques

b- contient des données la plus exhaustive possible √

c- diffusé par EBI

d- est une banque de données spécialisée de séquences nucléiques

e- contient des données homogènes et répartie par thématique

2- Dans les matrices de substitutions (nucléiques et protéiques) (2 pts : 2* 1pts)

a- La transition est beaucoup plus favorable que la transversion √

b- pour les matrices BLOSUMS (ex : BLOSUM62) le numéro indique le pourcentage d'identité√

c- Pour la matrice PAM250 le chiffre indique que 50 mutations sont acceptées dans chaque 100 résidus

d- Les matrices BLOSUMS et PAMs sont des matrices asymétriques

3- Parmi les banques de données suivant, lequel(s) est/sont regroupé(s) dans les banques généralistes de séquences nucléiques (2 pts : 2* 1pts)

a- DDBJ √ b- UniProt c- TrEMBL d- PDB e-EMBL √

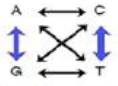
II- Voici l'alignement obtenu pour deux séquences :

```
ACTGGTCT--GACCTG--
CTTG--CTGACTT--AGG
```

1- Calculez le score de cet alignement dans les cas suivants : **(3,5 pts) répartie comme suit**

Match= 4, Mis-match : (transition = 1 ; transversion = -1), Gap= -2

Score Total = Σ Score élémentaires - Σ Score pénalités…….. **(0,5 pts)**



P(transition) > *P*(transversion) → **1 pts = 0,5*2**

A	C	T	G	G	T	C	T	-	-	G	A	C	C	T	G	-	-
C	T	T	G	-	-	C	T	G	A	C	T	T	-	-	A	G	G
-1	**1**	**4**	**4**	**-2**	**-2**	**4**	**4**	**-2**	**-2**	**-1**	**-1**	**1**	**-2**	**-2**	**1**	**-2**	**-2**

Après les calculs : on trouve un **score = 0** → **1 pts sur le tableau, 1 pts sur le calcul**

2- Donnez le score de l'alignement ci dessous en utilisant la matrice de substitution **BLOSUM 62** (**Gap = -5**).

```
RDISLV---KNAGI

RNI-LVSDAKNVGI
```

Le score de l'alignement est la somme des scores élémentaires. En regardant la matrice on trouve :

Score = 5 +1 + 4 + (-5) + 4 + 4 + (-5) + (-5) + (-5) + 5 + 6 +0 + 6 +4 → (0,125* 14 = 1,75 pts)

= 39 – 20 → score = 19 → (0,25 pt)

3- Pour obtenir la fiche 1, vous avez réalisé une requête sur le site serveur de SwissProt dont le résultat est le suivant :

```
>sp|P05231|IL6_HUMAN Interleukin-6 precursor (IL-6) - Homo sapiens (Human).
MNSFSTSAFGPVAFSLGLLLVLPAAFPAPVPPGEDSKDVAAPHRQPLTSSERIDKQIRYI
LDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKDGCFQSGFNEETCLVKIITGLL
EFEVYLEYLQNRFESSEEQARAVQMSTKVLIQFLQKKAKNLDAITTPDPTTNASLLTKLQ
AQNQWLQDMTTHLILRSFKEFLQSSLRALRQM
```

Fiche1

a) Quelle est le format de ce fichier ? Justifier → **(2 pts = 2*1)**

FASTA (1 pts) : toute en haut c'est le titre précédé par le symbole > SP qui se poursuit en sautant la ligne par la séquence brute écrite succinctement sans espace, ni chiffre. → (1 pts = 0,25*4)

b) Quelle est la nature de cette séquence ? **Protéique = succession d'acides aminés → 1 pts**

c) Que signifie le caractère « > SP » ?

« > SP » signifie que : Plusieurs séquences peuvent être mises dans un même fichier. → 0,5 pts

Exercice 2: (6 pts)

1- Expliquer la différence entre les banques de données **EMBL** et **trEMBL** ? → **(2 pts = 0,5*4)**

EMBL est la banque de données européenne généraliste de séquences d'acides nucléiques. TrEMBL est elle aussi une banque de données généraliste mais elle contient des séquences protéiques (Tr pour Traduced). Elle contient des séquences protéiques traduite automatiquement à partir des séquences codantes contenues dans EMBL.

2- Les banques de données GenBank, EMBL, DDBJ sont interconnectés, expliquer ? → (2 pts = 0,5*4)

Parce que les 3 banques échangent systématiquement leur informations ce qui fait, il suffit de consulter une de ces 3 banques pour accéder au contenu de ces trois banques en même temps.

3- La Bioinformatique est l'approche ***in silico*** de l'analyse de l'information biologique, **expliquez que signifie *in silico*** ?

***In silico* se réfère à l'outil informatique, c'est-à-dire l'utilisation des processeurs, logéciels,… etc pour analyser, traiter l'information biologique contenue essentiellement dans les séquences nucléiques (séquences d'ADN, ARN) et protéiques. → (2 pts = 0,5*4)**