

Corrigé type de l'Examen de Bioinformatique

Exercice 1 : (10 pts)

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

1- Les bases de données spécialisées contiennent des données **(2 pts : 2* 1pts)**

a- homogènes b-hétérogènes c- répartie par thématique

2- Dans les matrices Nucléiques et Protéiques **(2 pts : 2* 1pts)**

a- La transversion est beaucoup plus favorable que la transition

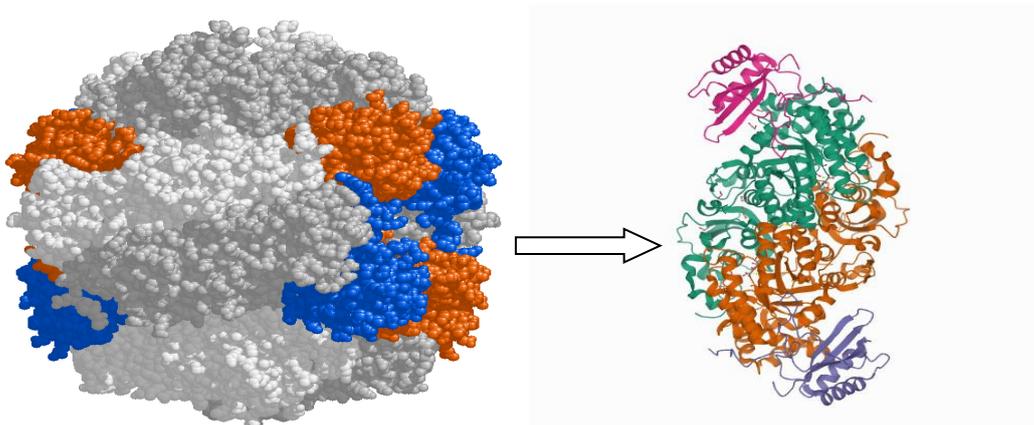
b- basées sur l'identité absolue (n'acceptent aucune possibilité de mutation)

c- pour les matrices PAMs (ex : PAM250) le numéro indique le pourcentage d'identité

d- basées sur la notion de similarité

e- appelés aussi matrice de score

3- La molécule de Rubisco ci contre est obtenue par la banque de donnée **(1 pts)**



a- GenBank b- UniProt c- TrEMBL d- PDB e-SwissProt

II- Voici l'alignement obtenu pour deux séquences :

CGTTAACG---ACTGTCT

CG-TATCGGCCACTATCT

1- Calculez le score de cet alignement dans les cas suivants : Match= 3, Mismatch= -1, Gap= -2,5 **(3 pts)**

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

C	G	T	T	A	A	C	G	-	-	-	A	C	T	G	T	C	T
C	G	-	T	A	T	C	G	G	C	C	A	C	T	A	T	C	T
3	3	-2,5	3	3	-1	3	3	-2,5	-2,5	-2,5	3	3	3	-1	3	3	3

$$\text{Score} = \Sigma \text{ Match} - \Sigma \text{ Mismatch} + \text{Gap}$$

$$= 12*3 + (-1 -1 -2,5 (4))$$

$$\text{Score} = 24$$

1 pts sur le tableau, 1 pts sur le calcul, 0,5 sur le résultat

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

	R	A	S	V	E	T	D	P	-	-	L	T
	T	T	S	L	Q	T	T	K	N	L	A	T
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓

Score = -1 + 0 + 4 + 1 + 2 + 5 + (-1) + (-1) + (-5) + (-5) + (-1) + 5 → (0,125* 14 = 1,75 pts)

= 17 - 14 = +3 → (0,25 pt)

Le score de l'alignement est la somme des scores élémentaires.

Exercice 2: (10 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- Classez les banques et les bases de données suivantes : Pfam, DDBJ, UniProt, Flybase, GenBank, KEEG Pathway

Banques de données Généralistes	Bases de données spécialisées
DDBJ, GenBank	Pfam, Flybase
UniProt → (1,5 pts = 0,5*3)	KEEG Pathway → (1,5 pts = 0,5*3)

3- Citez une différence entre la matrice Dot plot et les matrices de Substitutions ? → (2 pts = 0,5*4)

Dot plot : basée sur l'identité absolue (n'accepte aucune possibilité de mutation), Identité = même résidu.
Les matrices de Substitutions : basées sur la notion de similarité c'est-à-dire certains mutations/changement vont être acceptables.

4- Que signifie les numéros venant juste après PAM et BLOSUM (ex : PAM 1 et PAM 250 ou BLOSUM62 et BLOSUM30) ? (3 pts)

PAM : le numéro indique le nombre de mutations acceptés dans chaque 100 résidu par exemple : PAM250, 250 veut dire 250 mutations sont acceptés dans chaque 100 acides aminés. (1,5 = 0,5*3)

BLOSUM : le numéro indique le pourcentage d'identité par exemple BLOSUM62, 62 pour dire on garde uniquement les BLOks ayant un % d'identité > ou égale à 62% (1,5 = 0,5*3)