Réponses aux questions de cours

1. Définir les Banques de données primaires et secondaires ?

Banques primaires ~ archives
Banques secondaires ~ données vérifiées (corrigées et annotées)

2. Dans le format EMBL, GenBank / DDBJ. Que signifie le code DT et DE.

DT est spécifique au format EMBL et contient la date de création de l'entrée (première ligne) ainsi que la date de dernière modification (deuxième ligne).

DE (DEFINITION) contient ce que l'on appelle la définition de la séquence. Il s'agit de quelques lignes, fournies par les auteurs, décrivant sommairement le contenu de l'entrée (noms des gènes, fonction des protéines pour lesquelles ils codent, etc.)

3. Identifiez ce format de banque de données biologique et définir ces paramètres :

; Dro5s-T.Seq Length: 120 April 6, 1989 21:22 Check: 9487 ...

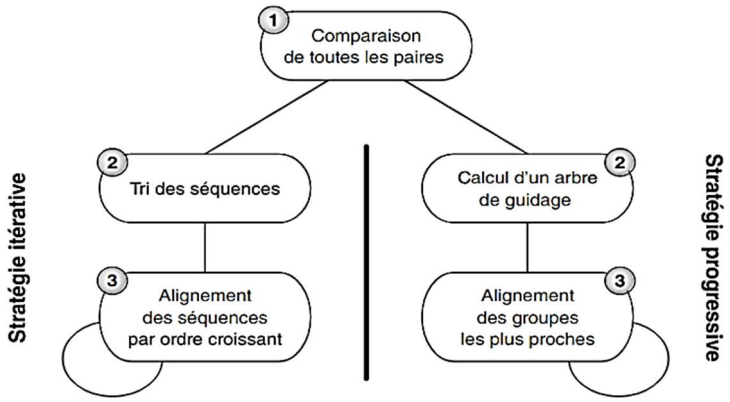
pir:ccho (1-104) , 104 bases, 7DA79498 checksum.

GDV EKG KKI FVQ KCA QCH TVE KGG KHK TGP NLH GLF GRK TGQ (etc ...) TWK EET LME YLE NPK KYI PGT KMI FAG IKK KTE RED LIA YLK KAT NE

Format Fitch : La 1ère ligne contient le nom de la séquence. Les lignes suivantes contiennent la séquence, découpée en 20 blocs (par ligne) de 3 caractères, séparés par un espace.

4. Pourquoi remède-t-on a un Alignement multiple itératif ou progressif au lieu de la programmation Dynamique. Expliquer le principe de fonctionnement de ces derniers et leur principale différence.

Dans l'alignement multiple de n séquences, la programmation dynamique n'est pas le choix le plus judicieux car il s'agira de trouver un chemin dans une matrice de dimension n, chose très contréniente en pratique. Les méthodes d'alignement multiple peuvent être classées en deux catégories selon l'approche utilisée, itérative ou progressive.



Alignement multiple itératif

La stratégie la plus simple pour calculer un alignement multiple, est celle dite itérative, elle utilise trois étapes :

1. Dans une première phase : Calculer un score de similarité entre toutes les paires de séquences par comparaisons des séquences deux à deux ; on obtient un ensemble de scores d'alignement qui sont regroupés dans une matrice dite de similarité ;

2. Dans la deuxième phase, cette matrice est utilisée pour trier les séquences, généralement des plus proches ou similaires aux plus éloignées ;

3. En troisième et dernière phase, cette liste est parcourue itérativement pour construire l'alignement multiple final, c'est-à-dire que les deux plus proches séquences sont alignées (itération 1). À partir de cet alignement, on calcule un « profil », qui est en quelque sorte une séquence consensus, puis on aligne la troisième séquence avec ce profil (itération 2). Un nouveau profil est calculé avec ces trois séquences, et la quatrième séquence est alignée avec ce profil (itération 3), etc.

4. L'algorithme prend fin quand toutes les séquences ont été alignées.

Alignement multiple progressif

Les algorithmes progressifs consistent à réaliser les alignements multiples en alignant progressivement les séquences. Il s'agit en quelques sortes d'algorithmes de type divide-and-conquer. D'une façon générale, tous les algorithmes progressifs reposent sur le même principe : commencer par aligner des sous-groupes de séquences puis essayer de les combiner entre eux pour former des alignements contenant de plus en plus de séquences. L'algorithme s'arrête lorsque toutes les séquences ont été regroupées pour former l'alignement multiple complet.

5. Le modèle de bases de données biologiques est de type relationnel. Ce modèle repose sur les 12 règles de Codd. Définir les 4 règles suivantes :

Règle 8

Indépendance physique : Les modifications au niveau physique (comment les données sont stockées, si dans les rangées ou les listes sont liées etc…) ne nécessitent pas un changement d'une application basée sur les structures.

Règle 9

Indépendance logique : Les changements au niveau logique (tables, colonnes, rangées, etc) ne doivent pas exiger un changement dans l'application basée sur les structures. L'indépendance de données logiques est plus difficile à atteindre que l'indépendance de donnée physique.

Règle 10

Indépendance d'intégrité : Des contraintes d'intégrité doivent être indiquées séparément des programmes d'application et être stockées dans le catalogue. Il doit être possible de changer de telles contraintes au fur et à mesure sans affecter inutilement les applications existantes.

Règle 11

Indépendance de distribution : La distribution des parties de la base de données à de diverses localisations doit être invisible aux utilisateurs de la base de données. Les applications existantes doivent continuer à fonctionner avec succès : quand une version distribuée du système de gestion de bases de données est d'abord présentée ; et quand des données existantes sont redistribuées dans le système.

Exercice :

Soient les deux séquences nucléotides suivantes : U = VTEERDEF et V = ITSHEAL
Utiliser la matrice PAM 250 (au verso du sujet) pour effectuer un alignement de ces deux séquences protéiques par programmation dynamique (algorithme de NEEDLEMAN & WUNCH).

	V	T	E	E	R	D	A	F
I	4	0	-2	-2	-2	-2	-2	1
T	0	3	0	0	-1	0	0	-2
S	-1	0	0	0	0	0	0	3
H	-2	-1	1	1	2	1	1	-2
E	-2	0	4	4	-1	3	4	-5
A	0	1	0	0	-2	0	0	-4
L	2	-2	-3	-3	-3	-4	-3	2

	V	T	E	E	R	D	A	F
I	4	4	4	4	4	4	2	1
T	4	4	4	4	4	0	0	-2
S	4	4	4	4	4	4	2	3
H	4	4	4	1	2	4	1	-2
E	1	0	4	4	-1	3	4	-5
A	0	1	0	0	-2	2	2	-4
L	2	-2	-3	-3	-3	-4	-3	2

V	T	___	E	E	R	D	A	F
I	T	S	H	E	___	___	A	L