université de BORDEAUX	**ANNEE UNIVERSITAIRE 2017-2018** **SESSION 1 D'AUTOMNE DECEMBRE 2017** **MENTION** : BIOINFORMATIQUE **Code UE :** 4TBI703U **Intitulé de l'épreuve : Omiques et Bioinformatiques** **Date :** 20/12/2017 **Heure : 11H30 Durée :** 1h30 Documents : ~~autorisés~~ / non autorisés Epreuve de M/Mme : R. Uricaru, P. Thébault et JC Taveau	**Collège Sciences et technologies** **Masters**

Exercice 1 :

Grâce à des nouvelles technologies dites de séquençage aujourd'hui on peut *lire* le génome d'un individu rapidement et à moindre coût. Le séquençage produit un grand nombre de petits segments du génome appelés lectures (*reads* en anglais). Un *read* est donc une chaîne de bases nucléiques (soit de caractères parmi les quatre : 'A', 'T', 'G', 'C'). *Par exemple un read peut être défini par la séquence « AGGGTCCCGAGAT ».* Le but de ces exercices est de manipuler un ensemble de N reads modélisé par une liste de N chaînes de caractères.

Note 1: Pensez à réutiliser vos fonctions.
Note 2: Seules les fonctions et méthodes Python suivantes sont autorisés: List.append(), len(), print(), range()
Note 3: Pour ajouter un élément à la fin d'une liste, on utilise la méthode append().
Exemple: my_liste.append('ACGT') ajoute à la fin de la liste my_liste, l'élément 'ACGT'.

1.1. Écrire une fonction **average(reads)** qui calcule la longueur moyenne des reads.
*Exemple : **average(['AGGCT','GGAT','GGCAAA'])** renverra le résultat **5***

1.2. Écrire une fonction **threshold(reads)** qui prend en argument un ensemble de N **reads** passé en paramètre de la fonction et qui retourne ceux qui ont une longueur supérieure ou égale à la longueur moyenne des N **reads**.
*Exemple : **threshold(['AGGCT','GGAT','GGCAAA'])** renverra ['AGGCT','GGCAAA']*

1.3. Écrire une fonction **countNucl(seq,symbol)** qui compte le nombre de nucléotides `**symbol**` dans la sequence `**seq**`.
*Exemple **: countNuc('AGGCT','G')** renverra 2*

1.4. Écrire une fonction **ratioGC(reads)** qui, pour un ensemble de N reads passés en paramètre de la fonction, calcule le taux moyen de GC (proportion de G et C dans la chaîne de caractères) des N **reads**.
*Exemple **: ratioGC('AGGCT', 'GGAT','GGCAAA'**) renverra 0.53333 équivalent à ((3/5 +2/4 +3/6)/3)*

1.5. Écrire une fonction **match(seq1,seq2)** prenant en arguments deux séquences de même longueur et retournant un score de matchs (1 point si identiques et 0 si différents).
*Exemple : **match('AGGCA','GGCAA')** renverra 2*

1.6. Écrire une fonction **removeEnds(reads, adaptor)** qui, pour un ensemble de N reads et un **adaptator** (adaptateur : courte chaîne de bases nucléiques) passés en paramètre de la fonction, filtre (enlève) les extrémités des **reads** correspondants à l'adaptateur. Ici on s'intéresse uniquement aux matchs parfaits entre les extrémités des reads et la séquence de l'adaptateur. La fonction devra retourner le nouvel ensemble de reads
*Exemple : **removeEnds(['TTTCAGGC', 'GGATTTTC','GGCAAATTTC'], 'TTTC')***
renverra ['AGGC', 'GGAT','GGCAAA']

Exercice 2

Une étude a été menée par Zeng *et al.* en 2016 (*doi:10.1038/srep33031)* sur la diversité et présence de certains gènes bactériens, impliqués dans la dégradation du sulfure, dans les fjords des régions de l'Arctique. Les échantillons recueillis ont été ensuite séquencé en utilisant des *primers* spécifiques des gènes Ddd (DddL, DddA, DddP …). A noter ces séquences n'étaient pas, au début de l'étude, présentes dans les bases de données du NCBI. **On s'intéressera dans les extraits suivants exclusivement aux gènes DddP.**

2.1 Expliquez en vous appuyant sur l'extrait suivant de la méthode analytique mise en œuvre pour construire l'arbre les étapes clés et quels outils ont été utilisés pour l'obtenir?

Data analysis

As the known DddP genes with a similarity >81% from different bacteria belong to the same genus, sequences showing more than 81% amino acid identity with each other were grouped in the same cluster. … using a BLASTX search against NCBI databases (http://www.ncbi.nlm.nih.gov)…. Sequence alignment and phylogenetic tree building were completed using respectively MUSCLE and Neighbour-joining method. Bootstrap tests of phylogeny were run with 1000 replicates.

2.2- A partir de l'extrait suivant, et en vous appuyant sur la figure de l'arbre, vous devrez décrire l'arbre obtenu et expliquer sur quelles observations s'appuient l'hypothèse d'un transfert horizontal (les clusters représentent un groupe de gènes identifiés dans les zones de prélèvements (3 stations : K1, K2 et K3). Les autres gènes sont issus des bases de données du NCBI) :

The dddP gene is one of the most frequently detected ddd genes in marine bacteria and is mainly found in the Roseobacter and SAR116 clades of Alphaproteobacteria18,20,43. However, evidence for horizontal gene transfer of dddP to some Gammaproteobacteria and fungal species has been reported. In the present study, the DddP sequence of Pseudomonas sp. DMSP-1…), was found to be closely related to Cluster P2 and Roseicitreum antarcticum ZS2-28 (Fig. 4), suggesting a possible inter-class horizontal gene transfer of dddP between Alpha and Gammaproteobacteria. In addition, comparing with the absence of the genus Roseovarius, which dominated dddP genotypes in this study, the genera Sulfitobacter and Loktanella were the dominant members of the Alphaproteobacteria in bacterioplankton community in Kongsfjorden31, suggesting a possible inter-genus horizontal gene transfer of dddP in Alphaproteobacteria…..

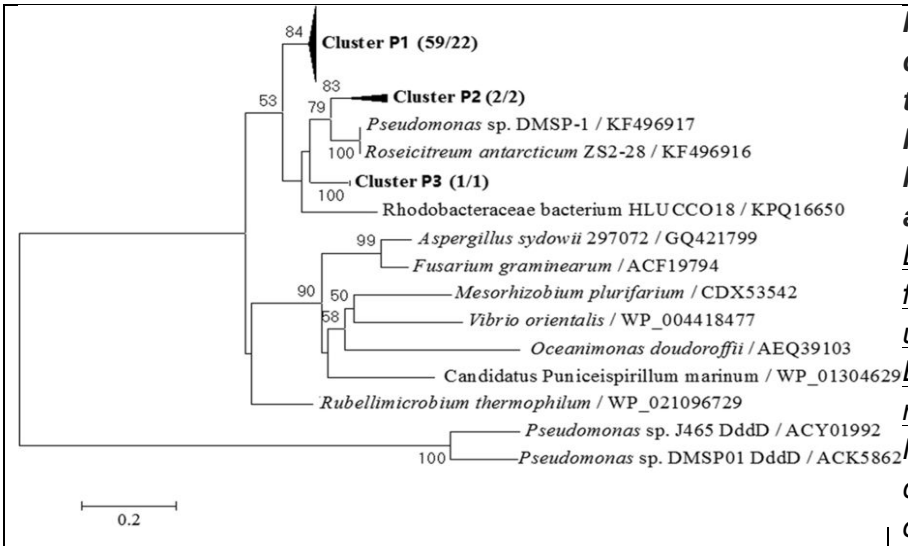

Figure 4 : Phylogenetic tree of deduced DddP sequences from two clone libraries of seawater in Kongsfjorden plus those from known bacterial species available in NCBI.
DddD sequences from Pseudomonas species were used as an outgroup.
Bootstrap values of <50 have been removed for clarity.
Numbers in parentheses following cluster names indicate the number of sequences found in stations K1 and K3, respectively. The scale bar indicates evolutionary distance.

2.3- Expliquez à quoi correspondent dans la légende *outgroup* et *bootstrap*.