université de BORDEAUX	**ANNEE UNIVERSITAIRE 2016 / 2017** **SESSION D'AUTOMNE DECEMBRE 2016** **MENTION** : MASTER BIOINFORMATIQUE **Code UE :** 4TBI703U **Intitulé de l'épreuve : Omiques et Bioinformatique** **Date :** 13/12/2016 **Heure : 10h00** **Durée :** 2H Documents : autorisés Epreuve de : Jean-Christophe TAVEAU /Patricia THEBAULT	**Collège Sciences et Technologies** **Masters**

Exercice 1 (5 points)

Ecrivez en Python une fonction **`containsORF(txt)`** qui retourne le booléen *True* si la séquence de type *String* contient un ORF c'est à dire un codon START suivi d'un codon STOP. On rappelle que le codon START est ATG et les codons STOP sont TAA, TAG ou TGA.
Cette fonction doit être implantée avec une **seule boucle** et comme on se limite à un seul cadre de lecture, à chaque itération, on avancera au codon suivant.
Voici quelques exemples de séquence et d'ORF trouvé (ou non):

```
0             1             2             3             4
012 345 678 901 234 567 890 123 456 789 012 345 678 910 123 456 78
CTG ATG TTC CAT TAC CAG TAC AAC AAA CTA TGA TTC CAT TAC CAG TAC A   # ORF: 3-30
CTG ATG TTC CAT TAC CAG TAC AAC AAA CTT ATG ATT CCA TAA CCA GTA CA  # ORF: 3-39
CTG TAA TTC CAT TAC CAG TAC AAC ATG CTA TGA TTC CAT TAA CAG TAC A   # ORF: 24-30
CTG TAA TTC CAT TAC CAG TAC AAC ATC CTA TAG TTC CAT TAA CAG ATG A   # ORF: None
CTG ATG TTC CAT TAC CAG TAC AAC AAA CTA AGA TTC CAT TAC CAG TAC A   # ORF: None
```

Exemple d'utilisation dans un script Python :

```
seq='CTGATGTTCCATTACCAGTACAACAAACTATGATTCCATTACCAGTACA'
print(containsORF(seq)) # returns True
```

Suggestion d'algorithme: Pour chaque codon de la séquence, on recherche **d'abord** le 1er codon START dans la séquence. Une fois trouvé, on passe en mode de recherche de codon STOP. Une fois trouvé le 1er codon STOP juste après le codon START, on retourne *True*, dans tous les autres cas, on retourne *False* (start non trouvé, stop avant start, stop non trouvé).

> **Note** : Pour simplifier la correction, on matérialisera les indentations par une flèche.

Exercice 2 (5 points) – Analyse de Cas

Vous disposez des 4 séquences suivantes.

(1) GTACGAG (2) GTAAAG (3) GTAGAG (4) GTAAG

1. Recherchez à la main le meilleur alignement multiple entre les 4 séquences suivantes. Vous utiliserez le système de score d'alignement global suivant : match +1 / mismatch : -1 / indel : -1
2. Déduisez de cette alignement multiple la matrice de distance.
3. Vous utiliserez la méthode **UPGMA** pour construire un arbre phylogénétique à partir de cette matrice.
4. Quelle stratégie proposez-vous pour évaluer la qualité de votre arbre ?

Exercice 3 (10 points) – Analyse de Cas

Voici un extrait de l'article "*Application and Accuracy of Molecular Phylogenies*" (1994) de David M. Hillis, John P. Huelsenbeck et Clifford W. Cunningham (Science, **264**, pp 671).

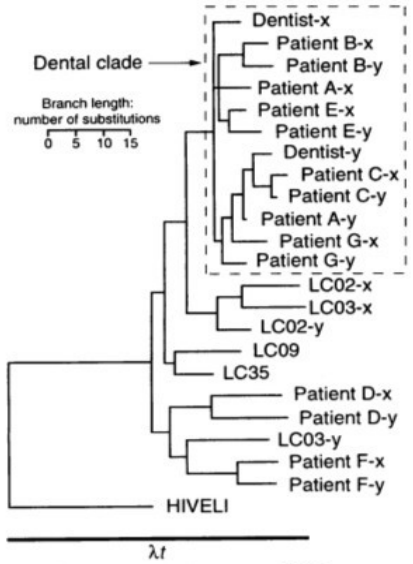

Fig. 3. Estimated phylogeny of HIV sequences from a Florida dentist, seven of his HIV-seropositive patients, and four individuals from the local population (LC) whose HIV sequences were most similar to those of the dentist (*47*). The outgroup (HIVELI) is an African HIV-1 sequence. Two divergent HIV sequences (labeled x and y) were examined from most individuals. The dental clade consists of patients whose HIV sequences are closer to those of the dentist than to those of any of the local controls. Branch lengths are proportional to the number of inferred evolutionary changes averaged across all possible character reconstructions (from *MacClade*) (*20*). The bar labeled λt is the distance from the root to the most divergent tip; it also indicates the divergence scale for the simulations in Fig. 4.

Table 1. Number of nucleotide substitutions across the tree shown in Fig. 3, as estimated from HIV sequence data (*4*). Values were derived from the averages across all equally parsimonious character-state reconstructions by use of the program MacClade (*20*); minimum and maximum number of substitutions across reconstructions are shown in parentheses.

From	To			
	A	G	C	T
A	—	80.00 (66, 94)	40.62 (29, 53)	17.67 (13, 22)
G	41.90 (28, 56)	—	4.46 (3, 6)	3.00 (0, 6)
C	23.08 (12, 35)	1.93 (1, 3)	—	21.83 (15, 29)
T	10.34 (5, 15)	12.67 (9, 16)	23.50 (16, 31)	—

patient was identified, the dentist wrote an open letter to his other patients in which he encouraged them to be tested for HIV infection. To date, 10 seropositive patients have been identified (*23*). However, some of these patients have other risk factors for HIV, so the question arises as to which, if any, of the patients were infected by the dentist rather than from another source. Sequences of the *gp120* gene of HIV encoding the C2-V3 domains were obtained from DNA amplified from peripheral mononuclear cells from the dentist, from seropositive patients, and from control individuals from the local population. Phylogenetic

Modeling Complex Phylogenies: HIV and the Florida Dental Case

Consider the recently reported case of the Florida dentist suspected of transmitting human immunodeficiency virus (HIV) to some of his patients (*4*). After a probable case of HIV transmission from dentist to

A partir du document ci-joint:

1) Faire un résumé succinct du contexte de l'article?
2) Expliquez l'analyse et la démarche scientifique associée qui a été conduite par Hillis et *al.(1994).*
3) Décrire l'arbre et ses principaux clades. A quelle population se réfère chacun des principaux clades ?
4) En quoi la région de l'arbre dite "Dental Clade" prouve t-elle la contamination par le dentiste ? Justifiez votre réponse. Quelles sont les précautions prises par les auteurs dans leur démarche pour que leurs résultats soient inattaquables?
5) A l'époque, les 1ers résultats phylogénétiques ont été critiqués en reprochant que les logiciels utilisés ne tenait pas compte des mutations particulières du HIV (Table 1). En vous appuyant sur vos connaissances en biologie, en quoi cette table de mutations est-elle inhabituelle?

Définition du terme **clade** en Biologie : Embranchement, groupe d'êtres vivants rassemblant à la fois un ancêtre commun et tous ses descendants.