

<div> <div>université</div> <div>de BORDEAUX</div> </div>	<div> ANNEE UNIVERSITAIRE 2024-2025 SESSION 1 D'AUTOMNE DECEMBRE 2023 MENTION : BIOINFORMATIQUE Code UE : 4TBI703U Intitulé de l'épreuve : Omiques et bioinformatique Date : 17/12/2024 Heure : 11h30 Durée : 1h30 Documents : autorisés / non autorisés Épreuve de M/Mme : Slim Karkar, Jean Christophe Taveau, Raluca Uricaru </div>	<div> Collège Sciences et technologies Masters </div>
-----------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------

Exercice 1 (10 points) : Phylogénie

Figure extraite de : Ochman et al. Examining bacterial species under the specter of gene transfer and exchange. PNAS. 2005.

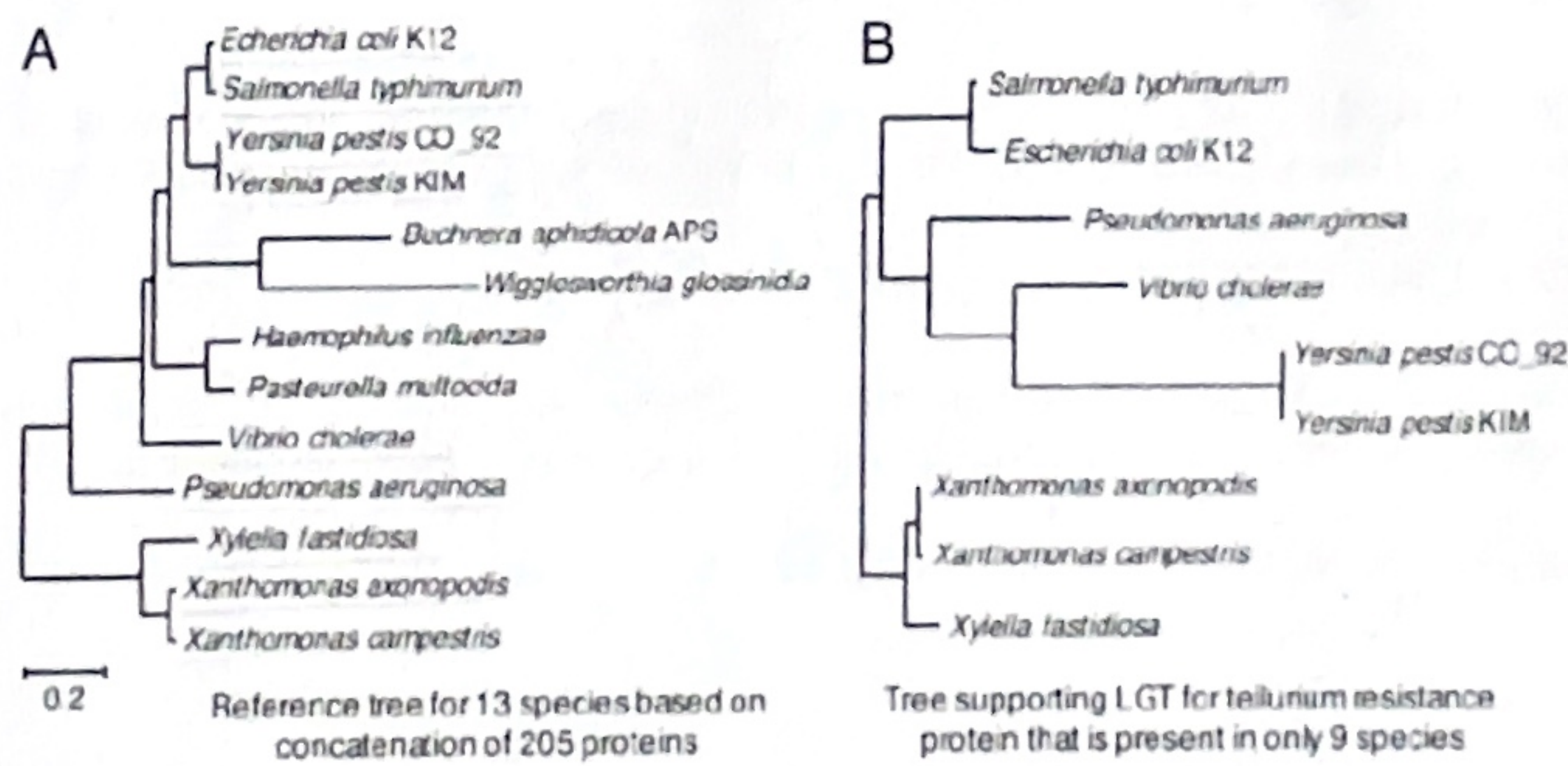


Figure 1 : (A) Arbre obtenu par Neighbor-joining à partir de la concaténation de 205 gènes à copie unique communs aux 13 espèces de Gammaprotéobactéries. Notez que 203 des 205 gènes soutiennent individuellement la même topologie. (B) Parmi la petite proportion de protéines montrant un soutien statistique pour une topologie alternative se trouve la protéine de résistance au tellure. Des homologues du gène codant pour cette protéine ont été détectés dans seulement 9 des 13 espèces. La résistance au tellure chez les bactéries fait référence à leur capacité à tolérer ou neutraliser la toxicité du tellure, un métalloïde, souvent par le biais de mécanismes génétiques ou biochimiques qui transforment les composés toxiques du tellure en formes moins nocives.

1. Afin d'identifier les différences de topologie entre ces deux arbres, proposez une version de l'arbre basé sur les 205 gènes concaténés (Figure A) restreint aux organismes qui présentent le gène de résistance au tellure (Figure B). Expliquez ce que chacune de ces différences impliquent en termes d'évolution et de relations phylogénétiques.
2. À partir de l'arbre B, proposez une hypothèse sur le scénario évolutif du gène de résistance au tellure. Justifiez votre raisonnement en fonction des différences entre les arbres A et B.
3. Des homologues du gène de résistance au tellure ont été détectés dans 9 des 13 espèces. Quels outils et base de données bioinformatiques permettent de détecter ces homologues ? Décrivez la démarche pour les détecter.
4. Quelles séquences génétiques sont classiquement utilisées pour construire l'arbre de la vie ?

5. Est-ce que l'arbre de la (Figure A) est un arbre de la vie ? Justifiez. Si vous pensez qu'il ne s'agit pas d'un arbre de la vie, quel est le type de cet arbre ? Même question pour l'arbre de la (Figure B).

Exercice 2 (10 points) : Des séquences et du Python

1. Écrire une fonction **find_motif(sequence, motif)** qui retourne une liste contenant toutes les positions (indices) où le motif apparaît dans la séquence.
Ici vous n'utiliserez aucune fonction avancée de Python : uniquement les boucles, les listes et les tests d'égalité sur les caractères et les chaînes de caractères sont autorisés.
Exemple : find_motif("ATGCGATATGCT", "ATG") renverra [0, 7].
2. Écrire une fonction **period(sequence, p)** qui renvoie **True** si la séquence est de période **p** et **False** sinon. La séquence a **p** comme période si elle est composée de répétitions successives et non chevauchantes de **p**. Justifiez l'utilisation (ou non) de **find_motif**.
Exemple : period("ATGCATGC", "ATGC") renverra True, alors que period("TATTATTTATT", "TATT") renverra False.
3. Écrire une fonction **gc_content(sequence)** qui calcule le pourcentage de bases G et C dans une séquence d'ADN.
Exemple : gc_content("ATGCATGC") renverra 50%.
4. Écrire une fonction **gc_pairs(sequence)** qui retourne une liste contenant toutes les positions où la paire **GC** apparaît dans la séquence en étant précédée par un **A** ou un **T** et suivie également par **A** ou **T**. Justifiez l'utilisation (ou non) de **gc_content**.
Exemple : gc_pairs("ATGCATGCCTGGCAGCT") renverra [2, 14].
5. Écrire une fonction **nucleotide_diversity(sequence1, sequence2)** qui calcule la diversité nucléotidique entre deux séquences de même taille. La diversité nucléotidique se définit ici par la proportion de positions où les séquences diffèrent.
Exemple : nucleotide_diversity("ATGCTTCAGA", "ATCCATCCGC") renverra 0.4 (4 différences sur 10 positions).