**NOM PRENOM**

**Master 1 Bioinfo / Master 1 Biotech**

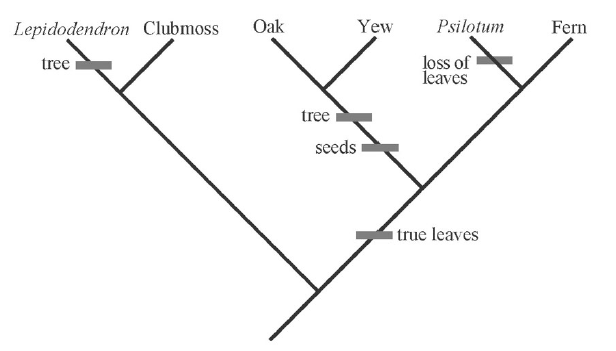
**Contrôle terminal d’Evolution Moléculaire (EMBIAE-EMBTAE) – Juin 2020 (Durée 2h30)**

**A la fin de l’épreuve, envoyez vos copies aux deux adresses suivantes :** [**Gwennaele.Fichant@ibcg.biotoul.fr**](mailto:Gwennaele.Fichant@ibcg.biotoul.fr)**,** [**bonhomme@lrsv.ups-tlse.fr**](mailto:bonhomme@lrsv.ups-tlse.fr)

**Analyse de cladogrammes : *(2 points)* Répondre en bleu en dessous de chacun des cladogrammes (2 points)**

**1.** Parmi les quatre arbres suivants, quel est celui pour lequel les relations de parentés sont différentes ? Justifiez votre réponse.

|  |  |
| --- | --- |
| question1a_TD | question1b_TD |

**2.** Supposons que dans l’arbre ci-contre, l’ancêtre était une herbe (pas un arbre) sans feuille (leaves) ni graine (seeds). En considérant que tous les changements évolutifs des états de ces caractères sont indiqués sur l’arbre, quelle feuille de l’arbre correspondrait à une espèce avec des feuilles mais pas de graine et qui ne serait pas un arbre ?

**Problème** ***(11 points)*** **écrivez vos réponses en bleu après chaque question**

La capacité de bioluminescence des Copepods, qui sont les taxa les plus nombreux de la faune zoo-planctonique trouvée dans les océans, a été étudiée. Un fort élément de preuve de la présence de bioluminescence dans un organisme particulier est l'identification moléculaire et l'analyse fonctionnelle de la luciférase. Des échantillons de plancton ont donc été collectés et les Copepods vivants ont été sélectionnés. Pour déterminer les espèces de Copepods vivantes présentes dans les échantillons, les séquences d'ARNr 18S ont été obtenue par PCR et comparées à celles présentes dans les bases de données. Ces séquences auxquelles ont été ajoutées des séquences d'organismes n'appartenant pas aux taxa des Copepods ont permis d'établir un arbre phylogénétique des espèces (Figure 1). Il a été réalisé à l'aide de la méthode PhyML.

De même, les séquences cDNA des luciférases présentes dans les ARN totaux de ces organismes ont été amplifiées en utilisant des primers déterminés à partir des régions conservées identifiées à partir des séquences de ce gène présentes dans les banques de données. Pour les analyses évolutives, les séquences en acides aminés de la luciférase déduites de celles des cDNA obtenus ont été utilisées (Figure 2A : arbre obtenu avec la méthode PhyML, Figure 2B : arbre obtenu avec la méthode Neighbor Joining (NJ)). Parmi les espèces de Copepods identifiées dans les échantillons, les gènes codant pour la luciférase ont été identifiés par cette étude uniquement dans : *Metridia pacifica*, *Metridia longa*, *Metridia okhotensis*, *Pleuromamma abdominalis*, *Lucicutia ovaliformis*, *Heterostylites major*, *Heterorhabdus tanneri* et *Gaussia princeps*. Des études antérieures avaient montré la présence de ce gène dans les espèces de Copepods suivantes : *Pleuromamma scutullata*, *Pleuromamma xiphias*, *Metridia asymmetrica*, *Metridia curticauda* et *Metridia lucens* GU594642.

**1.** Analyse de l'arbre phylogénétique des espèces (Figure 1) :

1. Pourquoi avoir utilisé pour la construction de cet arbre des séquences d'espèces de Crustacés n'appartenant pas aux taxa des Copepods ?
2. Pourquoi avoir utilisé les séquences de l'ARNr 18S (équivalent eucaryote de l'ARNr 16S procaryote) et non une approche de type super matrice permettant une reconstruction phylogénétique utilisant un ensemble de séquences de gènes orthologues ?
3. Sur cet arbre, les espèces de Copepods possédant des gènes de luciférase forment-elles un groupe monophylétique, paraphylétique ou polyphylétique ? Argumentez votre réponse.

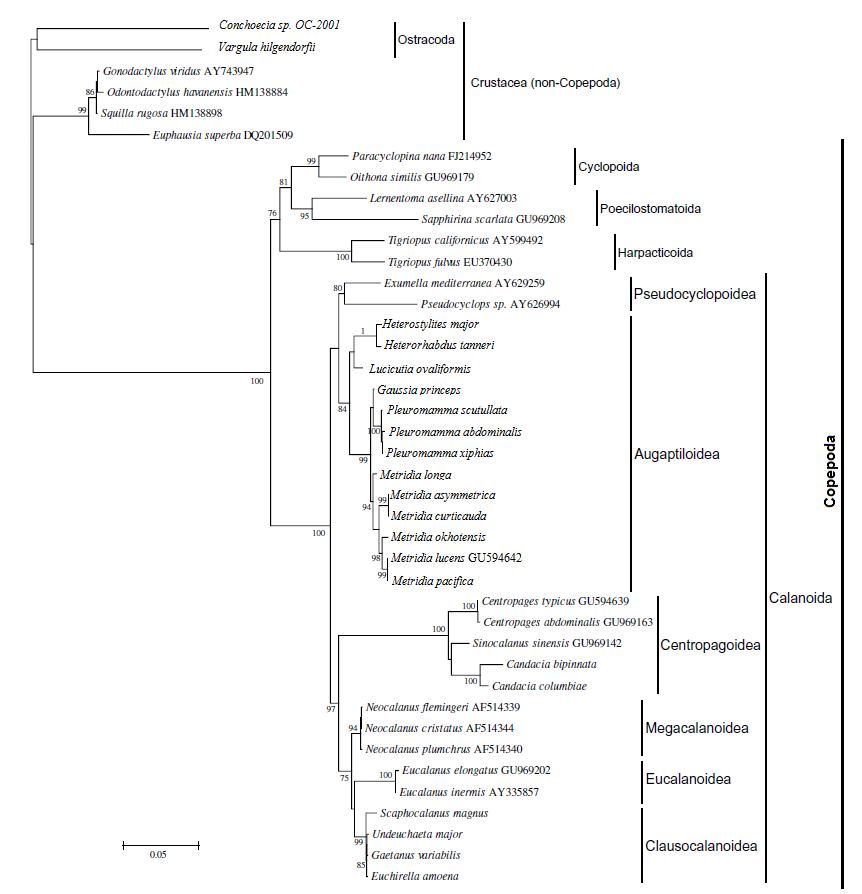
**2.** Analyse des arbres construits à partir des séquences protéiques de luciférase obtenues dans cette étude (Figure 2). L'arbre de la Figure 2A a été obtenu avec la méthode PhyML et celui de la figure 2B avec la méthode Neighbor Joining (NJ)

1. Le modèle évolutif le plus adapté aux données s'est avéré être le modèle WAG accompagné d'une correction par la loi Gamma. Pourquoi utilise-t-on une correction par la loi Gamma ? Pourquoi avoir utilisé pour les analyses évolutives les séquences en acides aminés de la luciférase et non les séquences des cDNA ?
2. Les deux arbres obtenus sont-ils congruents ? Argumentez votre réponse.
3. Sur l’arbre de la Figure 2A, quelles sont les branches que vous considérez peu fiables ? Répondez en les identifiant par le nombre présent au-dessus. Argumentez votre réponse.
4. Les gènes codant pour les séquences MoLuc1 et MoLuc2 sont-ils homologues, orthologues, paralogues? Argumentez votre réponse.
5. En vous référant à l’arbre de la figure 1, au niveau de quel nœud ancêtre supposeriez-vous que le gène de la luciférase a été acquis ? Pourquoi ?
6. On se concentrera sur l'arbre obtenu avec PhyML (Figure 2A). Quelle évolution le gène de la luciférase a-t-il subit dans les espèces analysées ? Pour répondre à cette question, vous vous attacherez à décrire les évènements de duplication, perte de gènes et transferts horizontaux qui ont pu se produire. Vous préciserez sur quelles branches de l'arbre ces évènements ont eu la plus grande chance de s'être produits. Est-ce que des évènements de même type (duplication ou perte de gènes) se seraient produits de façon indépendante ? Si oui, lesquels.

|  |  |
| --- | --- |
| **arbre_PhyML_lucferase** | **arbre_NJ_luciferase** |

**Figure 2 : A) arbre obtenu avec la méthode PhyML, B) arbre obtenu avec la méthode NJ**

Les noms des séquences correspondent à l'abréviation du nom de l'espèce suivi de Luc (pour luciférase) suivi d'un chiffre si plusieurs gènes codant pour cette protéine ont été identifiés. Les abréviations des noms d'espèces sont : M pour *Metridia longa*, Mo pour *Metridia okhotensis*, Mp pour *Metridia pacifica*, Pa pour *Pleuromamma abdominalis*, G pour *Gaussia princeps*, Lo pour *Lucicutia ovaliformis*, Ht pour *Heterorhabdus tanneri* et Hm pour *Heterostylites major*. Les différents taxa sont reportés (traits verticaux).



**Figure 1 : arbre phylogénétique des espèces obtenu avec la méthode PhyML**

Les différents taxa sont reportés (traits verticaux).

**Quizz : Surlignez en bleu la bonne réponse VRAI ou FAUX**(0.25 point par question, attention : Quizz à points négatifs !)

1. La sélection positive agit sur une mutation qui est favorable à l’individu et qui va tendre à se fixer à terme dans la population : **vrai faux**
2. La sélection négative sur un gène génère un excès d’allèles rares aux sites polymorphes de ce gène : **vrai faux**
3. La sélection négative opère sur un domaine protéique important dont la conservation est requise pour la fonction de la protéine : **vrai faux**
4. La sélection balancée maintient le polymorphisme dans un gène par un mécanisme d’avantage aux hétérozygotes : **vrai faux**
5. La néofonctionalisation d’un gène dupliqué est associée à un processus de sélection positive : **vrai faux**
6. Dans un alignement d’une séquence codante, le ratio dN/dS (proportion de substitutions non synonymes sur la proportion de substitutions synonymes) est égal à 1 en cas d’évolution neutre, < 1 en cas de sélection négative (purifiante), et > 1 en cas de sélection positive (adaptative) : **vrai faux**

**Problème 1 *(5,5 points)* écrivez vos réponses en bleu après chaque question**

L’ARNm du gène *P* vient d’être séquencé chez 50 lignées de la plante modèle *Arabidopsis thaliana*, suite à son identification comme élément majeur dans la dégradation par les plantes d’une protéine fongique toxique. Suite à l’alignement de ces séquences codantes (500 paires de bases), 30 sites polymorphes ont pu être identifiés.

Le ratio de substitutions non-synonymes sur synonymes (ω=dN/dS) sur ces séquences est égal à 1, et la statistique de Tajima (D) est égale à 0.

1. Décrivez et expliquez ce qu’indiquent les statistiques ω et D ? Que déduisez-vous des valeurs prises ici par ces statistiques ?

Une analyse plus fine de la séquence montre que :

- ω = 0 et D = -3 sur les 200 premières paires de bases

- ω = 2 et D = +3 sur les 300 dernières paires de bases.

1. Quelle explication pouvez-vous donner ?

Une analyse bioinformatique de la protéine *P* prédite montre l’existence de deux domaines protéiques :



L’annotation fonctionnelle des domaines de la protéine *P* indique que l’un des deux domaines est impliqué dans des interactions avec d’autres protéines, l’autre domaine est impliqué dans l’adressage de la protéine *P* et de son interactant vers un organite cellulaire impliqué dans la dégradation protéique.

1. Au regard de l’ensemble de ces résultats, assignez à chaque domaine (X et Y) les annotations fonctionnelles décrites ci-dessus. Explicitez votre raisonnement.