

TD 2 Evolution Moléculaire

Exercice 1

Le tableau suivant montre la distribution de 4 caractères dans 4 organismes différents.

- Quelle méthode de reconstruction d'arbre utiliseriez vous pour établir les relations phylogénétiques existant entre ces espèces?
- Reconstituez cet arbre en expliquant les principes de la construction et en détaillant chaque étape.

		Caractères			
		Xylème et phloème (1)	Bois (2)	Graines (3)	Fleurs (4)
Espèces	Mousse	0	0	0	0
	Fougère	1	0	0	0
	Pin	1	1	1	0
	Chêne	1	1	1	1

0 signifie que le caractère n'est pas observé et 1 que le caractère est observé.

Problème (basé sur les travaux publiés dans Desmond and Gribaldo, 2009, *Genome Biol. Evol.*, 364-81, doi:10.1093/gbe/evp036)

La voie de biosynthèse des stérols a été étudiée de façon approfondie chez les vertébrés, les plantes terrestres et les champignons. Elle fait intervenir un grand nombre d'enzymes (une quinzaine) dont les séquences sont disponibles chez des représentants des vertébrés, plantes terrestres et champignons.

Pour connaître la distribution taxonomique et l'évolution des enzymes impliquées dans cette voie de biosynthèse, nous nous proposons d'analyser un large échantillon de lignées eucaryotes pour lesquelles les génomes complets sont disponibles (38 lignées). Curieusement, quelques bactéries ont la capacité de synthétiser des stérols, parmi elles *Methylococcus capsulatus* (γ -Proteobacteria), *Gemmata obscuriglobus* (Plantomycetales), *Stigmatella aurantiaca* et *Plesiocystis pacifica* (δ -Proteobacteria). Leurs génomes ont été ajoutés à l'analyse.

1) Décrivez la démarche bioinformatique que vous adopteriez pour réaliser cette étude (de la construction des jeux de données à la construction des arbres).

Le résultat de votre démarche vous a permis d'obtenir l'arbre sur les protéines ERG5 (Figure 1) et l'arbre sur les protéines ERG7 (Figure 2). Pour analyser vos résultats, vous allez utiliser l'arbre consensus (Figure 3) représentant la phylogénie des espèces eucaryotes étudiées.

2) Qu'est ce qu'un arbre consensus ?

3) A quoi correspond les nombre figurant sur les branches des arbres de ERG7 et ERG5 ? Quel est l'intérêt de calculer ces valeurs ?

4) Pourquoi avons-nous travaillé à partir des séquences protéiques des enzymes et non pas à partir des séquences de leurs gènes ?

5) Lors de la construction de l'arbre sur les protéines ERG7, des séquences homologues bactériennes (SHC) n'étant pas impliquées dans la synthèse de stérols ont été ajoutées à l'analyse. Quel est l'intérêt d'avoir rajouté ces séquences ? Est-ce que les quatre génomes procaryotes étudiés possèdent un exemplaire de cette séquence SHC (ne pas répondre par oui ou non) ?

6) ERG7 est une enzyme indispensable dans la voie de biosynthèse des stérols où elle intervient dans la deuxième étape. Nous nous concentrerons sur la partie de l'arbre correspondant aux séquences de ERG7. En comparant l'arbre obtenu avec celui de la phylogénie des eucaryotes, indiquez sur quelle(s) branche(s) de l'arbre des eucaryotes la capacité à synthétiser les stérols a été perdue.

7) Toujours sur la partie de l'arbre correspondant aux séquences de ERG7 peut-on identifier des duplications ou des transferts horizontaux? Argumenter votre réponse.

8) Les gènes eucaryotes codant pour ERG7 sont-ils homologues, orthologues, paralogues? Argumenter votre réponse.

9) En analysant la distribution taxonomique des protéines ERG5 (Figure 1), à quel nœud de l'arbre des eucaryotes situeriez-vous la présence du gène codant pour cet enzyme (quelle est le plus ancien ancêtre possédant ce gène). Quelle évolution ce gène a-t-il subit par la suite (perte, duplication).

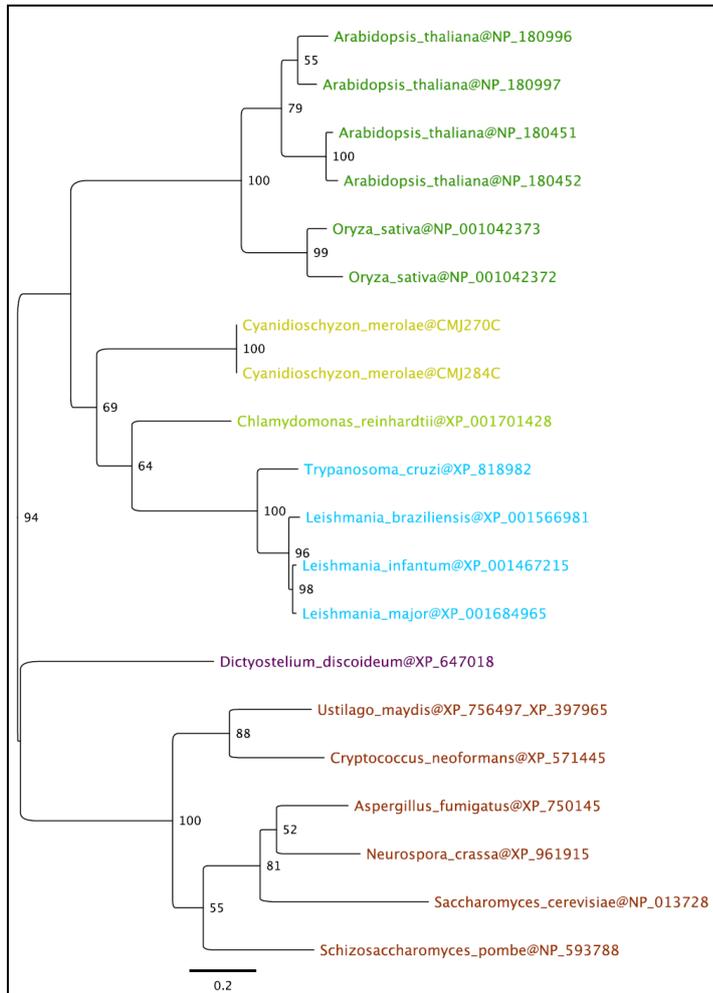


Figure 1 : arbre obtenu sur les protéines ERG5 en utilisant la méthode PhyML, le modèle WAG et une correction Gamma 4 catégories. La proportion de sites invariants a été estimée à partir des données. Les noms des feuilles correspondent au nom de l'espèce suivi du numéro d'accès de la séquence protéique dans la banque RefSeq. Les couleurs des groupes taxonomiques sont identiques entre les deux arbres ERG5 et ERG7 mais différentes de celles de l'arbre des espèces.

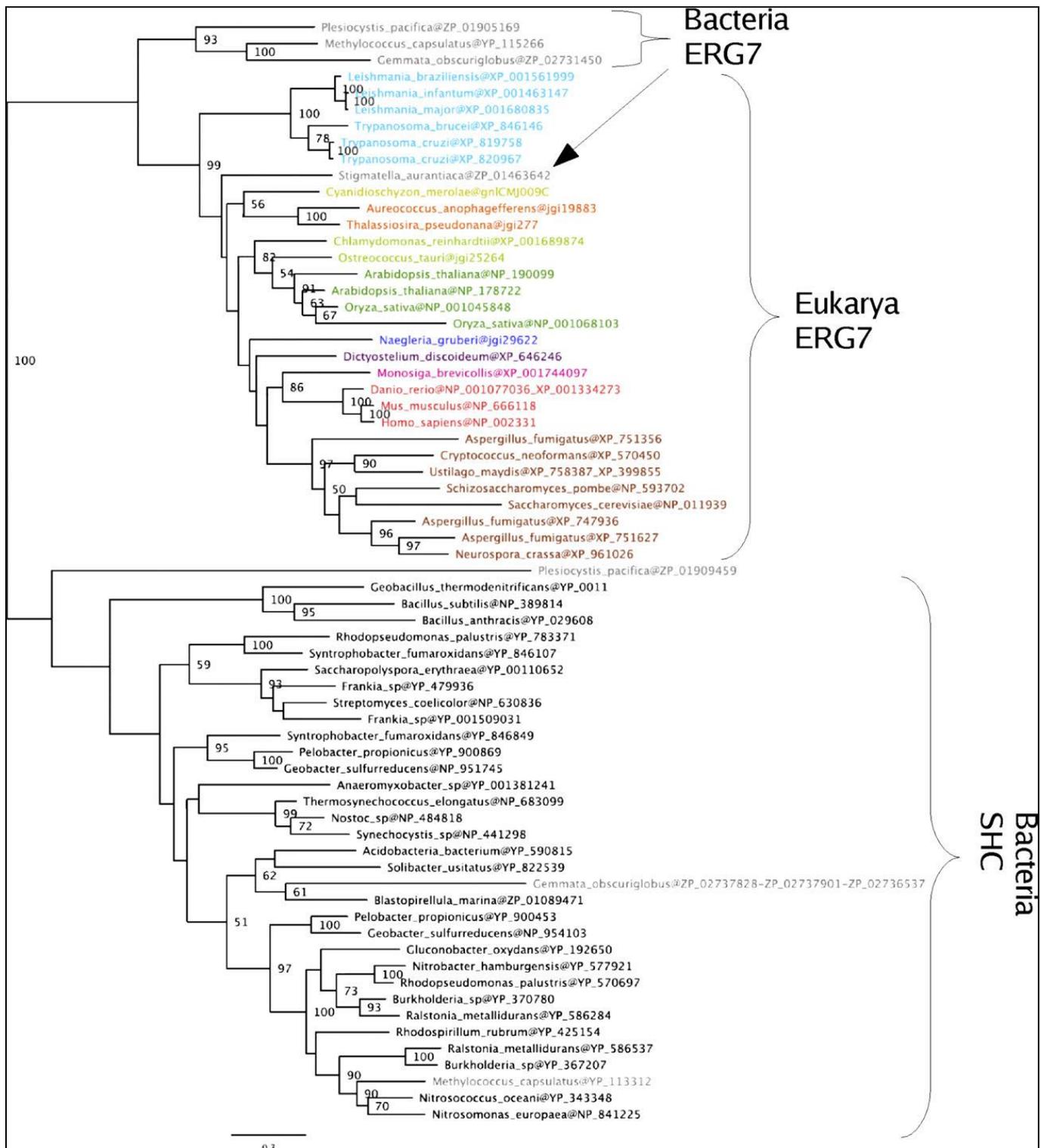


Figure 2 : arbre obtenu sur les protéines ERG7 et SHC, protéines homologues bactériennes non impliquées dans la synthèse de stérols, en utilisant la méthode PhyML, le modèle WAG et une correction Gamma 4 catégories. La proportion de sites invariants a été estimée à partir des données. Les noms des feuilles correspondent au nom de l'espèce suivi du numéro d'accès de la séquence protéique dans la banque RefSeq. Les couleurs des groupes taxonomiques sont identiques entre les deux arbres ERG5 et ERG7 mais différentes de celles de l'arbre des espèces.

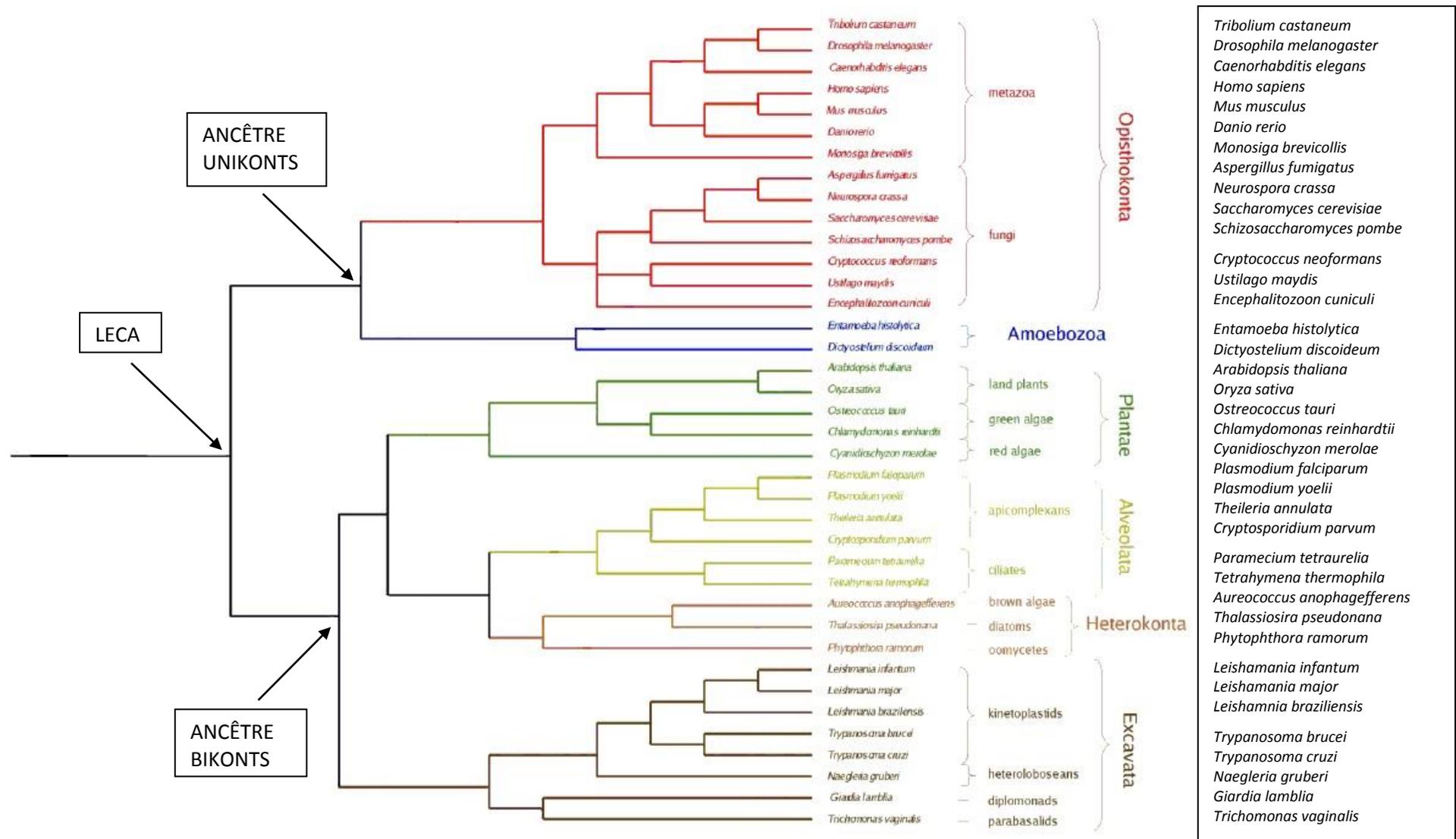


Figure 3 : Arbre consensus représentant la phylogénie des espèces eucaryotes étudiées et enraciné entre les unikonts et les bikonts (deux divisions des eucaryotes). LECA (Last Eukaryote Common Ancestor) correspond au nœud ancêtre le plus ancien de l'arbre.