# Evolution et biodiversité des microorganismes : contrôle des connaissances 1

## Projet personnel : annotation d’un gène hypothétique chez *Thermococcus kodakarensis*

**Consignes** : Votre mission est d’analyser la séquence du gène hypothétique qui vous a été attribué grâce à différents outils et bases de données vus précédemment en TD. Vous allez devoir *documenter* et *commenter* les différentes étapes de votre recherche puis, à la fin, *conclure* sur une fonction éventuelle du gène. Rapportez-vous au fichier excel déposé sur E-campus « CC1-DEVOIR ANNOTATION DE GENE » pour découvrir le nom du gène que vous devez analyser. Utilisez la **couleur rouge** pour donner vos réponses. Une fois votre travail terminé, vous devez déposer votre fichier sur l’espace devoir intitulé « CC1-RENDU\_DEVOIR ». Le nom de votre fichier doit contenir le nom du gène que vous avez étudié, votre nom et prénom (**TKXXXX\_NOM\_PRENOM**). La note du travail personnel est établie sur 20 points et compte pour 40% de la note totale. Cette note sera reportée pour la 2-ème session. Vous devez rendre votre travail **au plus tard le lundi 30 décembre, 12h**.

1. Pour débuter copiez-collez ci-dessous la fiche GenBank de votre gène. (1 point)

2. ORFfinder : faites une analyse commentée et concluez sur la présence éventuelle de codons START alternatifs. (2 points)

3. BLASTp : faites une recherche uniquement chez les Thermococcales, puis dans l’ensemble de la base de données, commentez les résultats. Si BLAST trouve dans votre séquence des domaines conservés, notez-le et décrivez leur fonction. A partir des résultats chez les Thermococcales choisissez trois séquences représentatives (justifiez votre choix) en plus de la vôtre et faites un alignement (4 points).

4. PFAM : faites une recherche avec votre séquence ou avec un mot clé. Documentez et commentez les résultats pour les onglets « domain organisation » et HMM logo. Comparez le HMM logo à votre alignement (Q3) et commentez ces résultats. (2 points)

5. HMMER : faites une recherche avec votre séquence contre la base de données UniProt. Documentez et commentez le nombre de résultats obtenu et la répartition taxonomique des résultats (2 points)

6. AlphaFold2 : utilisez le ColabFold pour modéliser la structure 3D de votre séquence. Si votre séquence est > 800 AA cela ne sera peut-être pas possible (mais essayez tout de même), passez à la question suivante. Alternativement, vous pouvez consulter AlphaFold Protein Structure Database pour trouver la prédiction de structure pour la protéine que vous étudiez (3 points).

7. Foldseek : faites une recherche avec la structure modélisée par AlphaFold2, documentez et commentez les résultats. (3 points)

8. STRING : faites la recherche avec votre séquence, donnez le réseau et commentez-le. Présentez ensuite les données qui vous semblent les plus informatives. (4 points)

9. Conclusion : pouvez-vous prédire une fonction pour votre gène ? Justifiez votre réponse. (4 points)

10. Question bonus : si vous avez la possibilité de tester votre prédiction, quelles expériences feriez-vous ? (3 points)