# Evolution et biodiversité des microorganismes : travaux dirigés

## Recherche de fonction(s) pour des gènes de fonction inconnue chez *Thermococcus kodakarensis*

**Objectif** : Savoir se servir des outils d’analyse de séquences afin de prédire la fonction d’un gène hypothétique.

**Consignes** : Donnez vos réponses en couleur pour mieux les repérer, utilisez la police Courier New (taille de lettre constante) pour écrire les séquences. Pour les informations concernant les logiciels utilisés rapportez-vous aux **TUTORIELS**.

Lors de ce TD nous vous proposons d’annoter les séquences génomiques issues de *Thermococcus kodakarensis* KOD1. Il s’agit d’une Euryarchée hyperthermophile (température optimale de croissance 85°C) et anaérobie stricte. Cette archée est l’un des modèles d’étude le plus utilisé par des chercheurs qui s’intéressent aux mécanismes moléculaires et leur évolution chez les Archées. *T. kodakarensis* possède un petit génome d’environ 2 Mpb et code environ 2300 protéines. Dans les conditions du laboratoire *T. kodakarensis* KOD1 se divise toutes les 30 min, aussi vite qu’*E. coli*. Actuellement, il existe plusieurs outils génétiques permettant de modifier son génome.

Au cours de ce TD vous allez utiliser plusieurs outils d’analyse de séquences et de structures protéiques afin de prédire la fonction d’un gène de fonction inconnue. Ces prédictions pourront par la suite être testées au laboratoire grâce à des approches génétiques, biochimiques, ou encore des approches utilisant le séquençage à haut débit (RNA-seq, ChIP-seq etc.). Dans le cadre de ce TD vous allez tester l’utilisation et la qualité de ces outils sur l’exemple d’un gène codant la protéine Pcc1 qui fait partie d’un complexe conservé chez les eucaryotes et les archées.

1. **Utilisez le navigateur Chrome**. Trouvez dans la base de données NCBI la séquence du génome de *T. kodakarensis* KOD1 dont le numéro d’accession est NC\_006624.1 Ouvrez la fiche GenBank, faites apparaître toutes les caractéristiques (« customize view » 🡪 »customize »🡪 »all features ») et cherchez dedans (ctrl + F) le mot « CDS » : combien de résultats avez-vous obtenus ? Faites la même recherche pour le mot « hypothetical » et notez le nombre de résultats. Que vous apprennent ces données ?

2. Cherchez (toujours ctrl + F) maintenant dans la fiche GenBank le gène *TK1253*. Il s’agit du gène codant la protéine Pcc1. Quelles sont ses coordonnées génomiques ? Quelle est la taille de la protéine prédite en acides aminés ?

3. En utilisant maintenant l’option « change region shown » (voir TUTORIELS) récupérez la séquence contenant *TK1253* ainsi que 200 pb en amont et en aval. Copiez-collez la séquence sous format FASTA ci-dessous.

4. Nous allons maintenant utiliser l’outil ORFfinder du NCBI (voir TUTORIELS) pour traduire cette séquence en 6 cadres de lecture. Il s’agit ici de vérifier si la partie codante a été correctement annotée ou si des codons start alternatifs peuvent être trouvés. Choisissez l’option "ATG and alternative initiation codons" et le code génétique 11 (Bacteria, Archaea, plant plastids). Copiez-collez ci-dessous le résultat de votre recherche (capture d’écran).

5. Quelle ORF correspond à la protéine codée par *TK1253*? Justifiez. Sur quel brin d’ADN est codé *TK1253* et quel est le cadre de lecture utilisé ?

6. A votre avis est-ce que l’hypothèse d’un codon start alternatif est plausible ? Pouvez-vous en conclure que *TK1253* est correctement annoté ? Justifiez votre réponse.

7. Récupérez la séquence protéique de la protéine TK1253 (Pcc1) sous format FASTA et copiez-collez la ci-dessous.

8. Nous allons maintenant utiliser BLASTp pour chercher les orthologues de Pcc1 chez les Thermococcales. Au préalable, trouvez dans NCBI Taxonomy les noms des trois genres appartenant à l’ordre des Thermococcales et notez les ci-dessous. Vous pouvez ignorer les catégories « unclassified » ou « environmental samples »

9. Faites maintenant votre recherche BLASTp contre les Thermococcales. Trouvez-vous des orthologues chez tous les genres de cet ordre ? Que signifie cela en termes d’histoire évolutive de Pcc1 au sein de ce groupe d’organismes ? Gardez la page des résultats BLAST pour plus tard.

10. Question préliminaire : comment peut-on définir un domaine protéique ?

11. Est-ce que BLAST trouve des domaines conservés dans votre séquence ? Comment peut on expliquer cela ?

12. D’autres méthodes (autre que les recherches par BLAST) peuvent être utilisés pour détecter les protéines contenant des domaines conservés. C’est le cas de PFAM qui utilise HMM (hidden markov model) pour détecter des résidus signature au sein d’un groupe de séquences homologues. PFAM (https://www.ebi.ac.uk/interpro) répertorie toutes les superfamilles, familles et domaines protéiques connues ainsi que des alignements de séquences pour chaque famille ou domaine. Faites une recherche avec le mot-clé Pcc1 dans la fenêtre prévue à cet effet (cliquez sur « Search » ensuite « by text »). Quel est le numéro d’accession de la base de données PFAM pour la famille Pcc1 ?

13. Regardez maintenant les informations contenues dans l’onglet « domain architectures » (à gauche). Ici le PFAM répertorie toutes les protéines contenant la séquence étudiée en tant que domaine protéique isolé ou fusionné à d’autres domaines. Sous quelle forme trouve-t-on Pcc1 majoritairement ?

14. Il peut être informatif de s’intéresser à des protéines « fusion » ou la protéine de fonction inconnue fait partie d’un polypeptide plus grand contenant un ou plusieurs domaines de fonction connue. Une telle fusion peut indiquer que la protéine d’intérêt est associée avec les domaines de fonction connue dans une même voie métabolique. Il y a-t-il des protéines fusion contenant le domaine Pcc1 qui vous semblent intéressantes à mentionner ? Cliquez sur les domaines qui vous intéressent pour apprendre leur fonction. Que pouvez-vous déduire de ces données ?

15. Regardez maintenant les informations contenues dans l’onglet « Profile HMM ». Copiez-collez l’image ci-dessous. A partir de l’alignement d’un grand nombre de séquences PFAM va construire un profil de séquence typique pour la famille Pcc1 en utilisant le HMM. Le graphique montre la fréquence d’acides aminés pour chaque position. Répertoriez ci-dessous les résidus conservés. Ces résidus sont potentiellement importants pour la fonction du Pcc1.

16. Vous allez maintenant utiliser le serveur HMMER (<https://www.ebi.ac.uk/Tools/hmmer/>) pour trouver les orthologues de la protéine TK1253 et étudier leur distribution au sein du vivant. Ce serveur utilise HMM (hidden markov model) pour trouver les orthologues éloignés. Si vous souhaitez connaitre le principe de fonctionnement de ce site vous pouvez consulter la page Wikipedia dédiée (<https://en.wikipedia.org/wiki/HMMER>). Entrez la séquence protéique de TK1253 (Pcc1) dans la fenêtre prévue à cet effet et cochez la base de séquences « UniProtKB ». Combien de résultats avez-vous obtenu ? Comment sont distribuées ces séquences au sein du vivant ? Pour répondre à cette question utilisez l’onglet « Taxonomy » en haut de la page. Copiez-collez l’image montrant la distribution de la TK1253 au sein du vivant.

17. Vous avez sans doute entendu parler du logiciel AlphaFold2. Il s’agit d’un algorithme de « machine learning » (intelligence artificielle) qui utilise les réseaux neuronaux pour prédire grâce à un échantillon de structures résolus la structure tridimensionnelle des protéines uniquement à partir d’une séquence protéique. Dernièrement, une nouvelle version de ce programme a été publié (DOI: 10.1038/s41586-021-03819-2) avec des résultats de qualité de prédiction époustouflants à tel point qu’on parle d’une révolution dans le domaine de la biologie structurale ! Si vous possédez un compte gmail vous pouvez utiliser l’interface Colab qui permet d’utiliser facilement AlphaFold2 : **https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=kOblAo-xetgx**

Reportez-vous au TUTORIELS pour quelques consignes d’utilisation et pour savoir comment interpréter les résultats. Faites une capture d’écran de principaux résultats (le modèle 3D, le graphique montrant l’alignement des séquences et le graphique montrant la distribution de l’indicateur IDDT au sein de la séquence). Télécharger le fichier .zip contenant tous les résultats. Vous allez en avoir besoin pour l’exercice suivant.

18. Récupérez le fichier pdb qui correspond au modèle le plus fiable « unrelaxed\_rank\_1\_model». Vous allez utiliser ce fichier pour effectuer une recherche avec l’outil Foldseek (<https://search.foldseek.com/search>). Ce serveur vous permettra de chercher parmi des centaines de millions de structures prédites par AlphaFold2 et parmi les structures déposées dans la PDB (donc des structures réelles déterminées expérimentalement. Faites une capture d’écran avec la liste des meilleurs résultats trouvées dans toutes les bases de données (« ALL DATABASES »). Quelles informations vous donne cette liste ? Affichez la superposition de structure pour le premier résultat (en cliquant sur l’icône « Alignment »), et commentez-la notamment en terme de sa robustesse.

19. Affichez maintenant la liste qui correspond aux résultats trouvés dans la PDB (onglet « PDB100 »). Existe-t-il des structures de Pcc1 déterminées expérimentalement ?

Pour terminer nous allons interroger la base de données STRING (<https://string-db.org/>) qui répertorie les réseaux d’interaction physique et/ou fonctionnelle pour de nombreuses protéines. Pour créer les réseaux protéiques STRING s’appuie sur les prédictions de contextes génomiques, sur les données expérimentales (comme par exemple les expériences de co-purification à grande échelle ou les données de transcriptome) et sur les recherches dans la littérature scientifique.

20. Question préliminaire : à quoi correspond la notion de contexte génomique ou de synthénie ? Que peut suggérer la conservation d’une synthénie particulière chez de nombreuses espèces et comment cela peut être intéressant pour l’annotation des gènes hypothétiques ?

21. Sur la page d’accueil de STRING (https://string-db.org) cliquez sur « search » puis à gauche sur « protein by sequence » puis faites la recherche avec la séquence protéique de TK1253. Copiez-collez ci-dessous le résultat graphique montrant le réseau.

22. La couleur des lignes reliant les différentes protéines a une signification. Le code couleur est indiqué sous l’image. Quelle est la nature des interactions entre le TK1253 et les autres protéines selon ce code couleur ?

23. Cliquez sur « Viewers » pour afficher les différentes catégories d’informations, copiez-collez ci-dessous la catégorie qui vous semble la plus informative. Justifiez votre choix et commentez les données.

24. En conclusion, qu’indique ce réseau ? Dans quel processus moléculaire pourrait participer TK1253 ? Comment tester cette hypothèse ?