# Evolution et biodiversité des microorganismes : travaux dirigés

## L’évolution des protéines universelles Sua5/TsaC : quand la simplicité est l’ultime sophistication ?

**Objectifs** : Savoir chercher les séquences à partir de la base de données NCBI. Savoir utiliser les logiciels d’alignement de séquences. Savoir construire les arbres phylogénétiques et les interpréter.

**Consignes** : Donnez vos réponses en couleur pour mieux les repérer, utilisez la police Courier New (taille de lettres constante) pour écrire les séquences. Nous vous recommençons fortement d’utiliser Firefox, nous avons constaté que ce browser est le plus compatible avec les outils en ligne que nous utilisons.

Les protéines de la famille Sua5/TsaC participent à la voie de biosynthèse d’une modification universelle d’ARNt nommé thréonyl-carbamoyl-adénosine (t6A). Cette modification est trouvée uniquement à la position 37 (à côté de l’anticodon) au sein de tous les ARNt décodant les codons du type ANN (ou N = A, C, T, G) (Fig. 1A). t6A contribue à la bonne lecture du code génétique, entre autre, en stabilisant l’interaction entre l’anticodon et le codon. La perte de t6A est létale pour les organismes modèle procaryotes ; chez les eucaryotes elle provoque des phénotypes sévères y compris chez l’homme.

Fig. 1. Structure et voie de biosynthèse de t6A.

A) Positionnement au sein d’ARNt et structure chimique de t6A. B) Voie de biosynthèse de t6A. La première étape de la réaction est catalysée par les enzymes de la famille Sua5/TsaC, cette réaction conduit à la formation d’un intermédiaire réactionnel instable le thréonine-carbamoyle-AMP. C) La famille Sua5/TsaC est composé de deux variantes protéiques, l’une, TsaC est une protéine constituée d’un seul domaine, l’autre, Sua5, est composé d’un domaine TsaC-like relié par une boucle « linker » à un deuxième domaine nommé SUA5 de fonction inconnue.

Les protéines de la famille Sua5/TsaC catalysent la première étape de la réaction conduisant à la formation de t6A (Fig. 1B). Alors que les protéines TsaC sont composées d’un seul domaine, les protéines Sua5 sont composés d’un domaine TsaC-like et d’un deuxième domaine nommé SUA5 de fonction inconnue (Fig 1C). Ces deux domaines sont reliés par une boucle « linker ». Les travaux d’analyse structurale et biochimique ont démontré que le linker et le domaine SUA5 étaient indispensables à l’activité de la protéine Sua5 issue de l’archée *Pyrococcus abyssi* (Pichard-Kostuch et al., 2018, RNA). La mutation des motifs conservés P227-G228-M229 et H234-Y235 en alanine au sein du linker provoque la perte d’activité *in vitro* et *in vivo*. Le motif K56-X-R58/S143-X-N145 (toutes les positions des acides aminés correspondent à la séquence de Sua5 de *P. abyssi*) au sein de TsaC et TsaC-like est essentiel pour la liaison à l’ATP et est conservé au sein de l’ensemble des protéines TsaC/Sua5. Ce motif est également essentiel pour l’activité enzymatique.

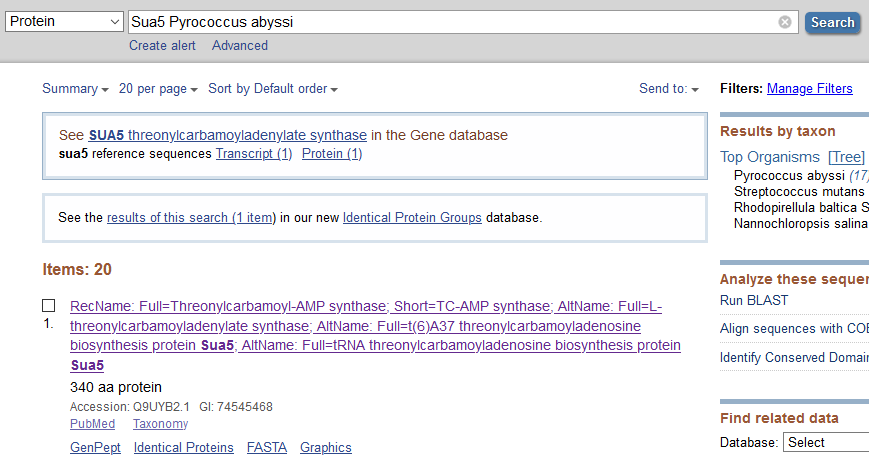
L’étude de distribution de Sua5 et TsaC au sein du vivant a démontré : (i) qu’au sein des trois domaines on trouve à la fois Sua5 et TsaC ; (ii) que la distribution de ces protéines ne suit pas la phylogénie des organismes ; (iii) que, sauf rares exceptions, chaque organisme code pour un des deux variants ; parmi ces exceptions on trouve les champignons du genre *Aspergillus*.

Au cours de ces travaux dirigés vous allez tenter de répondre à la question suivante : pourquoi existe-t-il deux versions d’enzymes, Sua5 et TsaC, pour la synthèse de la t6A ? Pour essayer d’y répondre nous allons tenter dans la partie I d’établir l’histoire évolutive de cette famille d’enzymes grâce à la construction d’arbres phylogénétiques. Dans la partie II, vous allez essayer de mettre en évidence un évènement de transfert de gène horizontal au sein de la famille Sua5/TsaC. Dans la partie III nous allons utiliser l’alignement de séquences pour identifier les résidus caractéristiques d’un variant ou de l’autre, vous vous appuierez sur ces résidus pour effectuer une classification cladistique.

Si vous voulez savoir comment nous avons mené l’étude pour répondre à cette question et avec quel résultat, vous pouvez lire notre article : Pichard-Kostuch et al. “The universal Sua5/TsaC family evolved different mechanisms for the synthesis of a key tRNA modification.” *Frontiers in microbiology* vol. 14 1204045. 21 Jun. 2023, doi:10.3389/fmicb.2023.1204045

PARTIE I

1. Trouvez dans le NCBI la séquence protéique de Sua5 issue de *Pyrococcus abyssi*, une euryarchée hyperthermophile. Copiez-collez la séquence sous format FASTA :



Cliquer sur lien FASTA pour avoir accès à la séquence

>WP\_010868714.1 threonylcarbamoyl-AMP synthase [Pyrococcus abyssi]

MTIIINVRERIEEWKIRIAAGFIREGKLVAFPTETVYGLGANALDENAVKRIFEAKGRPADNPLIIHIAS

FEQLEVLAKEIPEEAEMLAKRFWPGPLTLVLPKSEVVPRVITGGLDTVAVRMPAHEIALKLIELSERPIA

APSANISGKPSPTSAHHVAEDFYGKIECIIDGGETRIGVESTVIDLTEWPPVLLRPGGLPLEEIEKVIGE

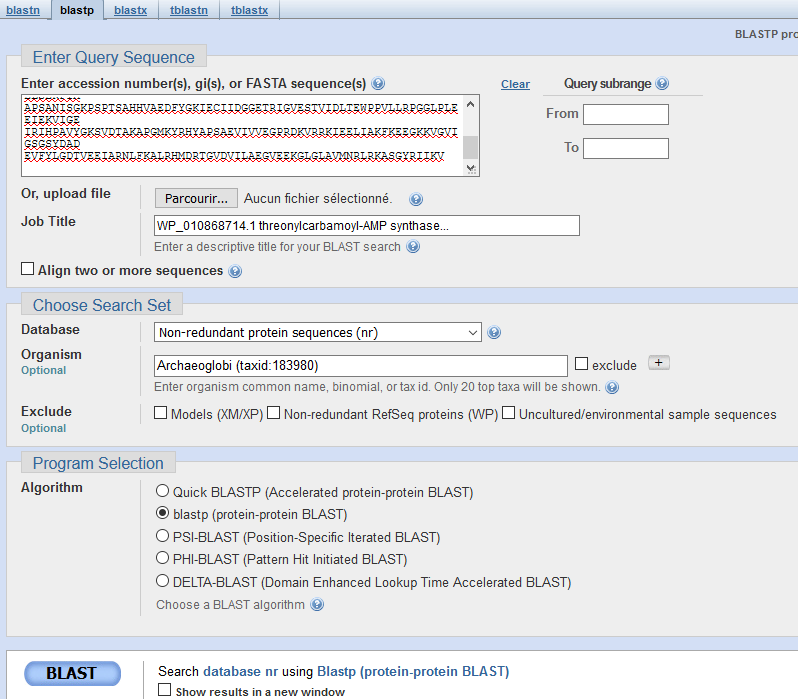
IRIHPAVYGKSVDTAKAPGMKYRHYAPSAEVIVVEGPRDKVRRKIEELIAKFKEEGKKVGVIGSGSYDAD

EVFYLGDTVEEIARNLFKALRHMDRTGVDVILAEGVEEKGLGLAVMNRLRKASGYRIIKV

2. Quel est le numéro d’accession pour cette protéine ?

Q9UYB2.1

3. Trouvez maintenant les orthologues de Sua5 chez les Archées de l’ordre Archaeoglobi. Effectuez pour cela un BLAST protéique contre la base de données « non-redundant protein sequences » en restreignant la recherche uniquement au sein d’Archaeoglobi. Trouvez l’alignement obtenu pour *Archaeoglobus veneficus* et copiez-collez cet alignement ci-dessous. (vous pouvez utiliser ctrl + F pour chercher le mot « veneficus » dans la page des résultats).



En effectuant cette recherche vous devriez obtenir environ 70 séquences.



4. Qu’est-ce qu’un alignement de séquences ?

C’est l’écriture de deux (ou plus) séquences l’une sur l’autre de façon à faire apparaitre des identités (ou similitudes).

5. Qu’est que c’est un indel ? Est-ce que cet alignement en contient ? Justifiez.

C’est une délétion ou insertion d’au moins un acide aminé dans la séquence qui s’est produite au cours de l’évolution. Oui, dans l’alignement on peut identifier des indels indiqués par le signe - .

6. Au vue de l’alignement, la protéine de *A. veneficus* est-elle homologue à celle de *P. abyssi* ? Justifiez.

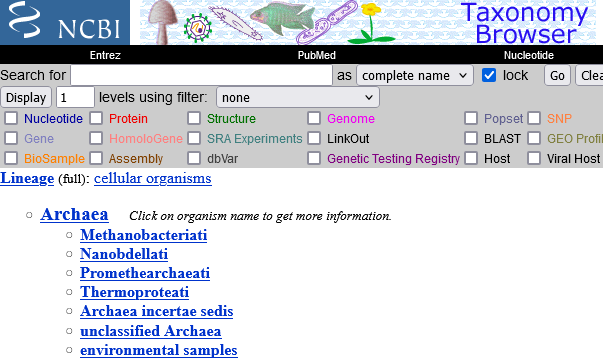
Oui, l’identité de séquence est élevée : 50 % et la E-value est de 4x10-98, c’est-à-dire que la probabilité de trouver une séquence non-homologue produisant un alignement de même qualité est très basse (1 / 2x10101).

7. S’agit-il d’un orthologue ou d’un paralogue? Justifiez.

Il s’agit d’un orthologue, car les deux protéines se trouvent au sein de deux espèces différentes.

8. Les Archées sont actuellement classées dans quatre superphylum. Citez les. Vous trouverez cette information dans la base de données NCBI taxonomy. Utilisez le niveau taxonomique le plus bas (display = 1) pour afficher uniquement les superphyla. Vous pouvez ignorer les groupes « candidats » ou les groupes avec un positionnement taxonomique incertain (incertae sedis ou unclassified ou environmental samples)

En cherchant avec mot clé « Archaea » dans cette base de données vous trouverez une liste des différents groupes dont les quatre groupes majeurs (superphylum) : Promethearchaeati (anciennement Asgard), Thermoproteati (anciennement TACK), Nanobdellati (anciennement DPANN), Methanobacteriati (anciennement Euryarchaeota)



9. Votre prochain objectif sera de construire un arbre phylogénétique pour les séquence représentatives de la famille TsaC/Sua5 chez les Archées. Pour ce faire, vous avez choisi, en plus de la séquence de Sua5 de *P. abyssi*, les séquences ci-dessous :

Asgard :

>Asgard\_RLI59347.1 threonylcarbamoyl-AMP synthase [Candidatus Thorarchaeota archaeon]

MVRIFRWPLRESEREMLSAILMNGGVVVYPTDTIYGIGGIIYDRRVVDKIYKIKERLRDKGLPILFGSIE

KIEEIAILNPLAKKLAQRFWPGMLTLVVPLKRRELRFLTGLEDKIAVRIPSDEIVLEVIKLAGGAIIGTS

ANLSGKLPCTNIKCVISQLGERMDAIVDGGERGTGFPSTIVEIQDDKKIKIIREGSIEVEKIKEAIGLEF

S

DPANN

>DPANN\_RLG13225.1 threonylcarbamoyl-AMP synthase [Candidatus Pacearchaeota archaeon]

MEVEIIKVKEKLADNELIQAIRYLEQEKTVVFPTDTVYGIGADLFKEKAVKNLFEIKKRPLDKPINALVS

SMKQAEMIVENIPDSAKKLIKEFWPGGLTIILKKKDIVPDIVTAGRETIGVRKPNNEIILQILEKFGKPL

AATSANISGLTSPITAKEVLQQLGEKSYLLIDGGKTIEKQPSTIIDLSSKKPLILRKGSVSVEKIKEILP

SIIE

TACK

>TACK\_AIF13504.1 Sua5/YciO/YrdC/YwlC family protein (rimN, SUA5) [uncultured marine thaumarchaeote KM3\_62\_H02]

MKISCNDVDIQIATKAINDGAIVVFPTDTVYGLGCNPYNHDAVLSLYEIKKRKKTKPFPVLGYSKKELEK

IAEFNSLEEKIAEKFWPGPITLILKVKDKEIQKSLDLEGKIAVRVPNNQCILALLKECKLLVGTSANISG

TAPFNDPKECGENLSGYDLLIDGGVISSQGESTIVEIENNDVKILRKGSVSEEMIKELT

Euryarchaeota:

>Euryarcheeota\_WP\_042702101.1 threonylcarbamoyl-AMP synthase [Methanobrevibacter arboriphilus]

MKIVQMDQNNPDLDLIDEAIEVLTSGGVVLYPTDTVYGLGANVFNEKAVEKVYNIKNRDYFKPLSVCVSS

IDEILLIADVGNNKTHQILKNNLPGPFTFIFYKKEPIHNYATKNHKVGIRIPENIISRKLTQSFPITTTS

ANLSGKKTLNNPDEIINQLNEGIDFVIDVGRLKQSEPSTIVDLTKKEPKILRKGSGILKSF

>WP\_010868714.1 threonylcarbamoyl-AMP synthase [Pyrococcus abyssi]

MTIIINVRERIEEWKIRIAAGFIREGKLVAFPTETVYGLGANALDENAVKRIFEAKGRPADNPLIIHIAS

FEQLEVLAKEIPEEAEMLAKRFWPGPLTLVLPKSEVVPRVITGGLDTVAVRMPAHEIALKLIELSERPIA

APSANISGKPSPTSAHHVAEDFYGKIECIIDGGETRIGVESTVIDLTEWPPVLLRPGGLPLEEIEKVIGE

IRIHPAVYGKSVDTAKAPGMKYRHYAPSAEVIVVEGPRDKVRRKIEELIAKFKEEGKKVGVIGSGSYDAD

EVFYLGDTVEEIARNLFKALRHMDRTGVDVILAEGVEEKGLGLAVMNRLRKASGYRIIKV

Expliquez quel est le raisonnement derrière ce choix de séquences par rapport à votre objectif, décrire l’histoire évolutive de la famille TsaC/Sua5 chez les Archées ?

Pour parvenir à cet objectif vous devez inclure, dans l’idéal, les séquences issues de l’ensemble de la diversité phylogénétique des Archées. Dans la pratique vous ne pouvez pas utiliser toutes les séquences disponibles dans les bases de données car votre jeu de données risque d’être biaisé (par exemple les Halobactéries sont fortement surreprésentées parmi les génomes séquencés d’Archées) et trop important en nr. de séquences pour pouvoir être traité par les logiciels. Par conséquent, vous devez sélectionner un nombre restreint de séquences pour chaque phylum d’Archées. Pour simplifier l’exercice ici, nous avons choisi seulement une séquence par superphylum d’archées, pour une vraie étude il aurait fallu inclure plusieurs dizaines de séquences par superphylum.

10. Utilisez maintenant Clustal Omega (**https://www.ebi.ac.uk/jdispatcher/msa/clustalo**) pour aligner les 5 séquences que vous avez choisies. Copiez-collez le résultat ci-dessous. Ensuite, téléchargez le fichier d’alignement sous forme clustal (cliquer sur le bouton « download » à gauche), nommez le « clustalo-PaSua5\_Archaea » et sauvegardez-le sur votre bureau.

WP\_042702101.1 --MKIVQMDQNNPDLDLIDEAIEVLTSGGVVLYPTDTVYGLGANVFNEKAVEKVYNIKNR 58

RLI59347.1 -MVRIFRW--PLRES-EREMLSAILMNGGVVVYPTDTIYGIGGIIYDRRVVDKIYKIKER 56

AIF13504.1 -------MKISCNDV-DIQIATKAINDGAIVVFPTDTVYGLGCNPYNHDAVLSLYEIKKR 52

WP\_010868714.1 -MTIIINVRERIEEW-KIRIAAGFIREGKLVAFPTETVYGLGANALDENAVKRIFEAKGR 58

RLG13225.1 MEVEIIKVKEKLADN-ELIQAIRYLEQEKTVVFPTDTVYGIGADLFKEKAVKNLFEIKKR 59

: : . \* :\*\*:\*:\*\*:\* .. .\* ::: \* \*

WP\_042702101.1 DYFKPLSVCVSSIDEILLIADVGNNKTHQILKNNLPGPFTFIFYKKEPIHNYAT-KNHKV 117

RLI59347.1 LRDKGLPILFGSIEKIEEIAI-LNPLAKKLAQRFWPGMLTLVVPLKRRELRFLTGLEDKI 115

AIF13504.1 KKTKPFPVLGYSKKELEKIAE-FNSLEEKIAEKFWPGPITLILKVKDKEIQKSLDLEGKI 111

WP\_010868714.1 PADNPLIIHIASFEQLEVLAKEIPEEAEMLAKRFWPGPLTLVLPKSEVVPRVITGGLDTV 118

RLG13225.1 PLDKPINALVSSMKQAEMIVENIPDSAKKLIKEFWPGGLTIILKKKDIVPDIVTAGRETI 119

: : \* .: :. . : :. \*\* :\*::. . .:

WP\_042702101.1 GIRIPENIISRKLTQS--FPITTTSANLSGKKTLNNPDEIINQLNEGIDFVIDVGRLKQS 175

RLI59347.1 AVRIPSDEIVLEVIKLAGGAIIGTSANLSGKLPCTNIKCVISQLGERMDAIVDGGERGTG 175

AIF13504.1 AVRVPNNQCILALLKEC-KLLVGTSANISGTAPFNDPKECGENLSG-YDLLIDGGVISSQ 169

WP\_010868714.1 AVRMPAHEIALKLIELSERPIAAPSANISGKPSPTSAHHVAEDFYGKIECIIDGGETRIG 178

RLG13225.1 GVRKPNNEIILQILEKFGKPLAATSANISGLTSPITAKEVLQQLGEKSYLLIDGGKTIEK 179

.:\* \* . : : : \*\*\*:\*\* . .:: ::\* \*

WP\_042702101.1 EPSTIVDLTK-KEPKILRKGSGILKSF--------------------------------- 201

RLI59347.1 FPSTIVEIQDDKKIKIIREGSIEVEKIKEAIGLEFS------------------------ 211

AIF13504.1 GESTIVEIEN-NDVKILRKGSVSEEMIKELT----------------------------- 199

WP\_010868714.1 VESTVIDLTE-WPPVLLRPGGLPLEEIEKVIGEIRIHPAVYGKSVDTAKAPGMKYRHYAP 237

RLG13225.1 QPSTIIDLSS-KKPLILRKGSVSVEKIKEILPSIIE------------------------ 214

\*\*:::: . ::\* \*. : :

WP\_042702101.1 ------------------------------------------------------------ 201

RLI59347.1 ------------------------------------------------------------ 211

AIF13504.1 ------------------------------------------------------------ 199

WP\_010868714.1 SAEVIVVEGPRDKVRRKIEELIAKFKEEGKKVGVIGSGSYDADEVFYLGDTVEEIARNLF 297

RLG13225.1 ------------------------------------------------------------ 214

WP\_042702101.1 ------------------------------------------- 201

RLI59347.1 ------------------------------------------- 211

AIF13504.1 ------------------------------------------- 199

WP\_010868714.1 KALRHMDRTGVDVILAEGVEEKGLGLAVMNRLRKASGYRIIKV 340

RLG13225.1 ------------------------------------------- 214

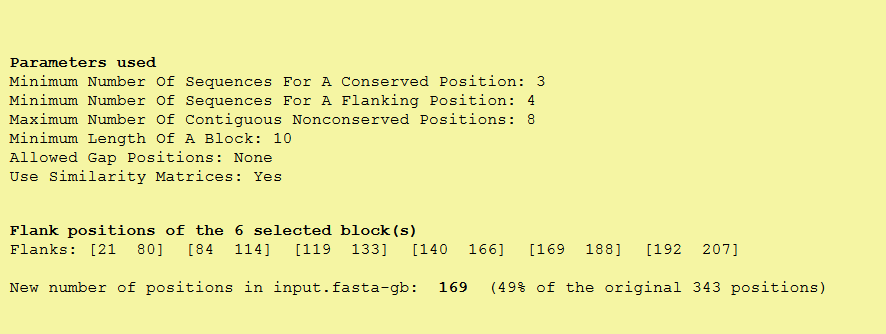
11. Identifiez les motifs conservés (K56-X-R58/S143-X-N145) dans les séquences que vous avez choisies et soulignez-les en jaune. Pourquoi est-il important que les séquences que vous avez choisies comportent ces résidus conservés ? Si besoin, rapportez-vous au début du TD pour vous rappeler la fonction de ces résidus.

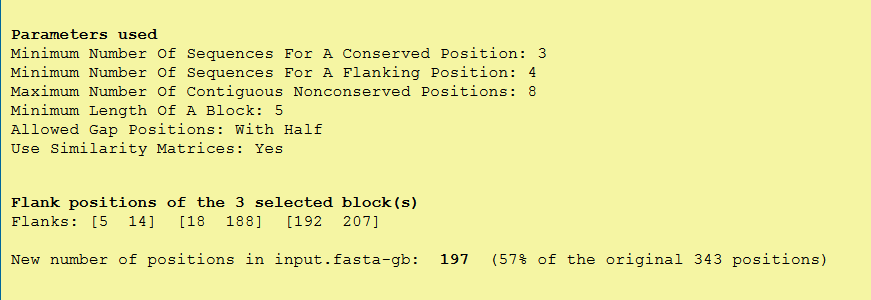
Le motif K56-X-R58/S143-X-N145 est important pour la fixation de l’ATP et donc pour l’activité catalytique. La présence de ces résidus conservés indique qu’il s’agit d’une protéine fonctionnelle et non pas d’une protéine codée par un pseudogène. Les pseudogènes sont généralement soumis à des taux élevés d’évolution de séquences (comparé à des gènes correspondants fonctionnels) et pour cette raison doivent être éliminés de l’analyse phylogénétique.

12. Avant de procéder à la construction d’un arbre phylogénétique vous allez utiliser un programme nommé GBlocks pour choisir les résidus qui seront utilisés. Expliquez pourquoi.

Parce que pour construire un arbre phylogénétique il est nécessaire d’utiliser les sites qui correspondent uniquement aux résidus homologues, or les alignements comportent toujours des régions ou l’alignement (détermination de l’homologie) est incertain (par exemple les régions comportant des indels ou des régions peu conservés) il est donc souhaitable de les éliminer. Cependant, il faut trouver un juste équilibre entre la quantité des sites utilisés (la quantité du signal phylogénétique) et la qualité des sites (la qualité du signal phylogénétique).

13. Allez sur Gblocks (**http://www.phylogeny.fr/one\_task.cgi?task\_type=gblocks**) et lancez-le pour votre alignement (« clustalo-PaSua5\_Archaea ») avec les paramètres par défaut (ne cocher aucune case). Copiez-collez ci-dessous le résumé de l’analyse (en bas de la fenêtre jaune). Dans un nouvel onglet faites l’analyse Gblocks de nouveau en utilisant les paramètres les moins stricts pour la sélection des blocks de séquences (cocher les 3 options dans le menu « less stringent »). Comparez les deux résultats. Si vous deviez choisir l’une de deux possibilités laquelle choisiriez-vous et pourquoi ? Sauvegardez (clique droit 🡪 enregistrez la cible de lien sous…) les deux fichiers **sous format FASTA** sur votre bureau. Nommez les GBlock\_Sua5\_Archaea\_default.fasta et GBlock\_Sua5\_Archaea\_moins\_strict.fasta





Pour condition 1 (défaut) : on voit que la sélection de blocks est plus stricte (pas d’indels admis) cela donne 6 blocks avec 169 positions sur 343 (49%)

Pour condition 2 (sélection moins stricte) : 3 blocks sélectionnés, comportant des indels et aussi des régions peu conservées (block 5-14), en tout 197 positions sélectionnés 57%

Je choisirai la condition 1 car le rapport qualité d’information/quantité d’information me semble meilleur (en inspectant les séquences des blocks).

14. Vous allez maintenant construire un arbre phylogénétique en utilisant le logiciel IQTree disponible sur **https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html** Ce logiciel utilise le maximum de vraisemblance pour la construction d’arbres phylogénétiques. En quoi consiste cette approche ?

Elle consiste à probabiliser entièrement le processus évolutif (définir les probabilités de tous les évènements évolutifs possibles depuis n’importe quelle séquence ancestrale jusqu’aux feuilles), et à trouver par une approche mathématique itérative quel est le scénario évolutif (topologie de l’arbre et longueur de ses branches), qui a la plus forte probabilité d’avoir donné naissance aux séquences analysées.

15. Revenons maintenant sur votre analyse phylogénétique : utilisez IQtree pour obtenir l’arbre phylogénétique de vos séquences pour les deux fichiers fasta que vous avez enregistrés (GBlock\_Sua5\_Archaea\_default.fasta et GBlock\_Sua5\_Archaea\_moins strict.fasta). Dans les options, substitution model, choisissez « Find best and apply », cela signifie que l’algorithme cherchera le meilleur modèle évolutif pour votre alignement et l’appliquera pour la construction de l’arbre. Choisissez ensuite 1000 réplicas de ultrafast bootstrap. Lancez les deux analyses en parallèle dans deux onglets différents. Vous recevrez un e-mail lorsque les résultats seront disponibles.

Pendant que le programme s’exécute répondez aux questions suivantes :

16. Qu’est-ce qu’un bootstrap et à quoi sert -il?

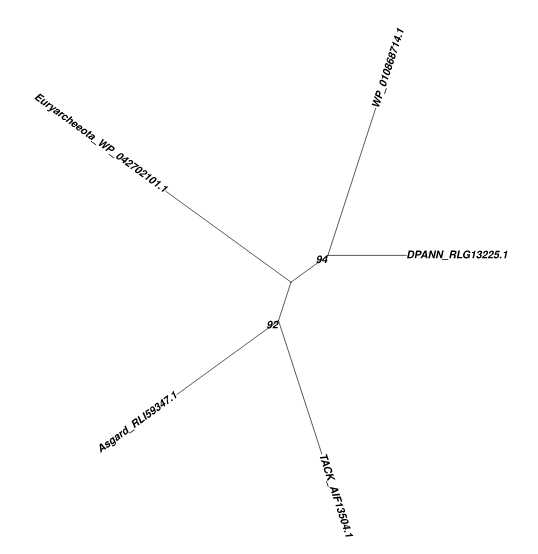
L’analyse de bootstrap est une méthode non-paramétrique d’évaluation de robustesse d’un arbre phylogénétique. Pour ce faire l’algorithme va recréer un grand nombre de jeux de données (alignement de séquences, habituellement 100) par ré-échantillonnage numérique à partir du jeu de données réel (l’alignement de séquences obtenu par un programme d’alignement). Pour chaque jeu de données l’algorithme va construire un arbre phylogénétique, en utilisant l’une des méthodes de reconstruction phylogénétique (Maximum Vraisemblance/Bayes); puis il va compter la proportion d’arbres qui montrent un même regroupement phylogénétique (une même branche dans l’arbre) et l’exprimer en pourcentage.

17. Quelle est la définition d’un clade selon la phylogénie moléculaire ?

Dans un arbre phylogénétique, un clade ou un groupe monophylétique est défini comme un groupe d’organismes (ou séquences) comprenant tous les descendants d’un ancêtre commun

18. Copiez-collez les arbres obtenus sous format graphique et sous format Newick pour vos deux analyses IQ-Tree (en les identifiant) et répondez aux questions suivantes :

GBlocks défaut :

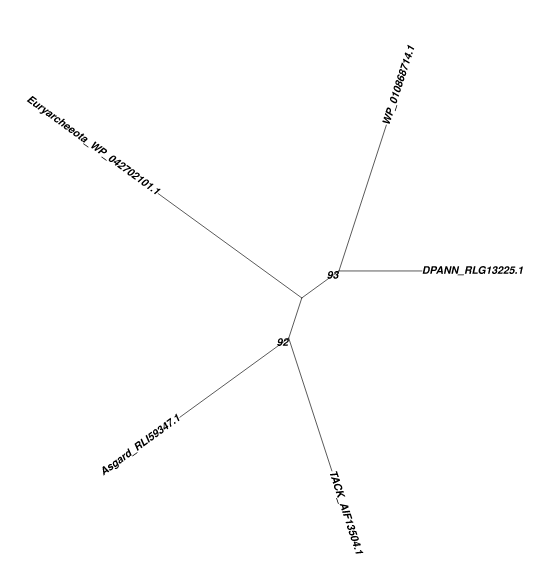


Modèle évolutif utilisé : BLOSUM62+I

Newick format :

(Euryarcheeota\_WP\_042702101.1:0.5907841102,(Asgard\_RLI59347.1:0.4788284980,TACK\_AIF13504.1:0.5343790424)92:0.1543566347,(DPANN\_RLG13225.1:0.2990881085,WP\_010868714.1:0.5908956618)94:0.1738811184);

GBlocks moins stricte



Modèle évolutif utilisé : BLOSUM62+I

Newick format : (Euryarcheeota\_WP\_042702101.1:0.7193776787,(Asgard\_RLI59347.1:0.5451121903,TACK\_AIF13504.1:0.5655785752)92:0.1712604314,(DPANN\_RLG13225.1:0.3358285745,WP\_010868714.1:0.6214251740)93:0.1860205374);

19. Les deux arbres sont-ils comparables ? Justifiez. Comment expliquez-vous cela ?

Oui, les deux arbres sont comparables : la topologie consensus est identique et la robustesse d’arbre (les valeurs de bootstrap) est comparable. Cela suggère également que les paramètres par défaut de GBlocks permettent d’éliminer les positions non-informatives sans perte d’information phylogénétique, en tout cas pour les séquences que nous avons choisies. Les algorithmes du type GBlocks sont surtout utiles pour des alignements où plusieurs séquences ont été concaténées. Dans ce cas, le nombre de positions est suffisamment élevé (plusieurs milliers) pour que l’analyse ne soit pas trop impactée par la perte d’une partie des résidus. En effet, dans notre cas nous disposons seulement d’une centaine des résidus, en règle générale cela n’est pas suffisant (lorsqu’on étudie les séquences assez divergentes) pour établir avec la robustesse la topologie de l’arbre. En effet, vous remarquerez que seulement deux nœuds sur trois sont soutenus par les valeurs de BS > 90%.

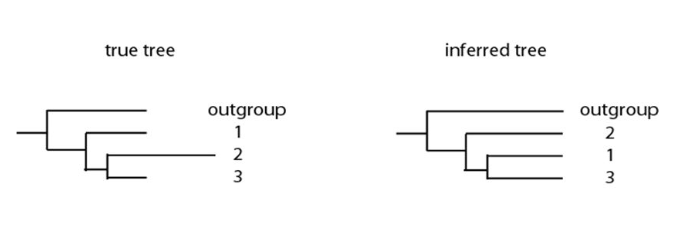
20. Pouvez-vous orienter dans le temps les arbres que vous avez obtenus (c’est-à-dire identifier le nœud le plus ancien) ? Justifiez.

Non, car ces arbres ne sont pas enracinés.

21. Pour enraciner cet arbre que faudrait-il faire ?

Inclure dans l’analyse une séquence qui n’appartient pas au groupe étudié, ici les Archées.

22. L’attraction des longues branches est l’un des artéfacts le plus connus de la reconstruction phylogénétique. A quoi correspond cet artéfact ? Aidez-vous du schéma ci-dessous pour répondre à la question.



Sur le schéma on voit que les séquences 2 et 3 forment un clade, cependant le taux d’évolution de la séquence 2 est supérieur à celui de la séquence 3 (longue branche). La séquence qui a été choisi pour l’outgroup a elle aussi une longue branche (ce qui est normal car elle est divergente par rapport aux séquences du groupe étudié). La séquence 2 étant également divergente elle va être « attirée » vers la séquence outgroup pour donner une topologie incorrecte de l’arbre.

23. Pour enraciner votre arbre en limitant l’artefact d’attraction de longues branches, choisiriez-vous une Sua5 eucaryote ou bactérienne ? Détaillez votre raisonnement en s’appuyant sur l’arbre universel classique à trois domaines de Carl Woese.

Selon l’arbre de C. Woese, les eucaryotes sont phylogénétiquement plus proches des Archées (car forment un groupe sœur avec les archées) que des Bactéries, par conséquence les séquences de Sua5 eucaryotes devraient, en théorie, être moins divergentes comparées à celles des bactéries. Il est donc plus judicieux d’utiliser une Sua5 Eucaryote. En réalité c’est plus compliqué que le raisonnement ci-dessus, en effet il est possible que le taux d’évolution des protéines TsaC/Sua5 eucaryotes est supérieur (pour une raison inconnue) au celui des orthologues bactériens. Dans quel cas il faut tester à la fois les séquences eucaryotes et celles issues des bactéries comme groupe externe.

24. Pour enraciner votre arbre vous avez refait une nouvelle recherche BLAST en utilisant la séquence de Sua5 de *P. abyssi* en cherchant uniquement chez les Eucaryotes. Voici la séquence que vous avez choisi :

>EZG56502.1 Sua5/YciO/YrdC/YwlC family protein [Gregarina niphandrodes]

MERFRVEMEDDAPAVYEEAGRRIREGKLVAFPTETVYGLGGNGLDVEAVKSIYEMKRRPADDPVILHVID

FQDAVELWDLALPLRRVIAGLQEAFWPGPLTIVAKANSQVPSEVTAGTGKVAVRCPAHPVARRLIAAARV

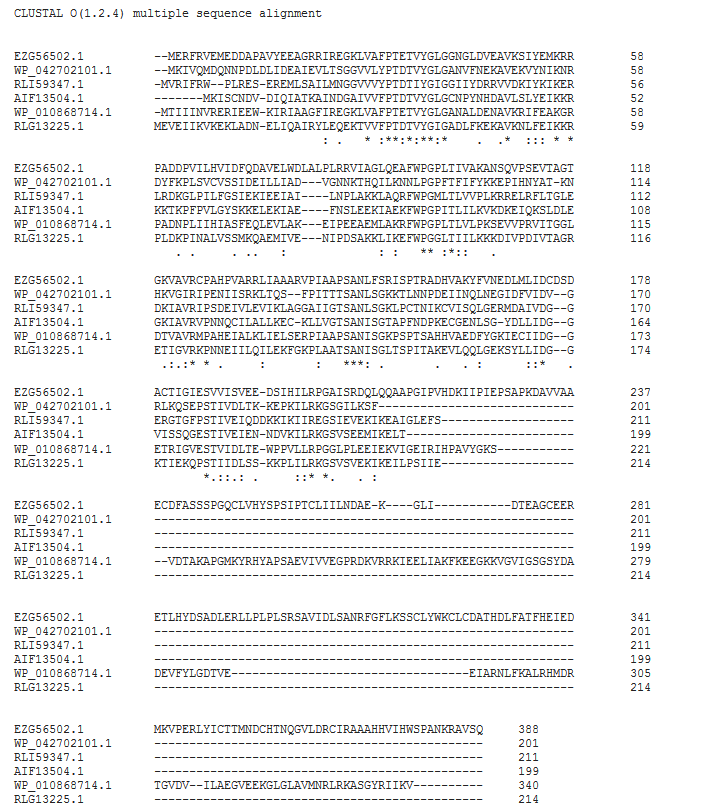
PIAAPSANLFSRISPTRADHVAKYFVNEDLMLIDCDSDACTIGIESVVISVEEDSIHILRPGAISRDQLQ

QAAPGIPVHDKIIPIEPSAPKDAVVAAECDFASSSPGQCLVHYSPSIPTCLIILNDAEKGLIDTEAGCEE

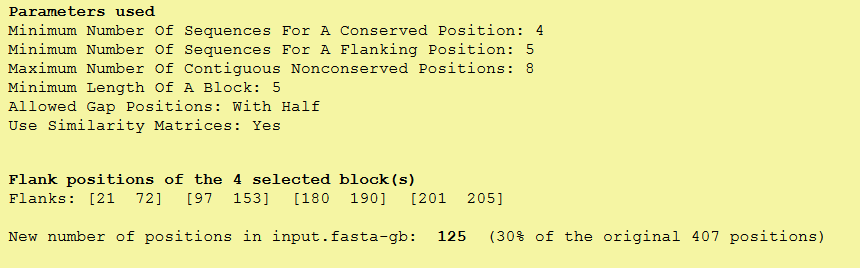
RETLHYDSADLERLLPLPLSRSAVIDLSANRFGFLKSSCLYWKCLCDATHDLFATFHEIEDMKVPERLYI

CTTMNDCHTNQGVLDRCIRAAAHHVIHWSPANKRAVSQ

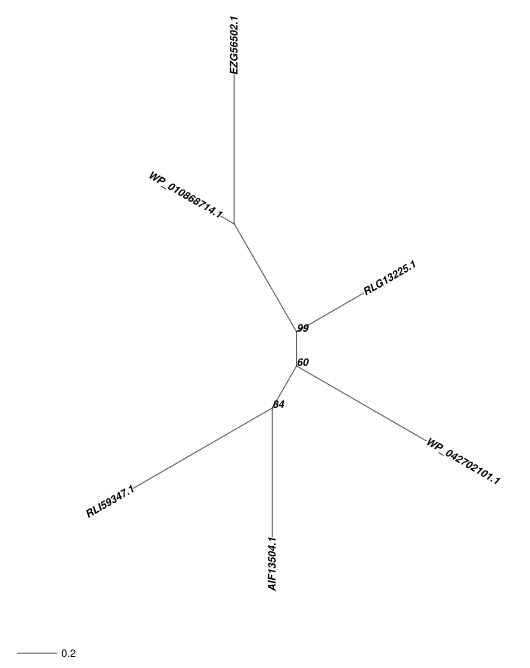
25. Effectuez de nouveau un alignement de vos séquences d’Archées y compris celle de *P. abyssi* en incluant la séquence eucaryote. Vous trouverez les 6 séquences dans l’annexe de ce document. Faites ensuite un GBlock avec les paramètres les moins stricts (cocher les trois cases) et utilisez IQTree pour construire l’arbre. Comparer cet arbre avec les deux arbres précédents. En parallèle, faites un calcul IQ-Tree en utilisant l’alignement ClustalOmega sans analyse GBlocks. Documentez et commentez toutes les étapes de votre analyse. N’oubliez pas de **conserver vos arbres sous format Newick**, vous en aurez besoin pour la question suivante.



On remarque que les motifs importants sont conservés dans la séquence eucaryote que j’ai choisie.



Ici on voit que le nombre de résidus retenus est plus faible que précédemment 125 au lieu de 197, la séquence eucaryote s’aligne moins bien avec les séquences d’archées que les séquences d’archées entre elles – ce qui est logique.

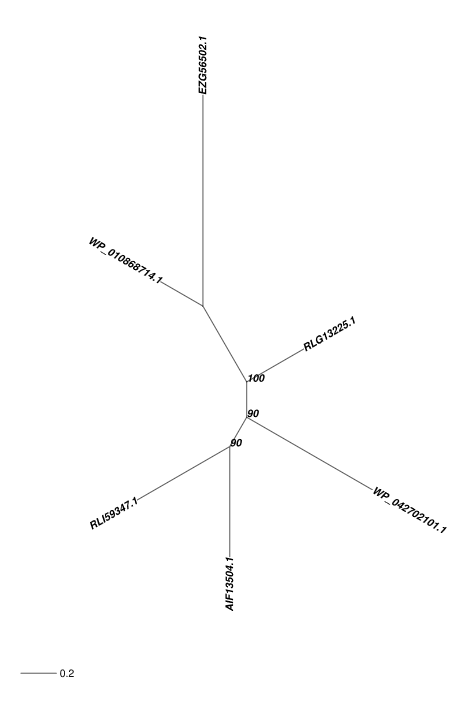


Modèle d’évolution de sequences utilize : cpREV+G4

(EZG56502.1:0.7425949968,((WP\_042702101.1:0.7486135311,(RLI59347.1:0.7995056069,AIF13504.1:0.6434639600)84:0.2425738990)60:0.1706197764,RLG13225.1:0.3882477012)99:0.6228776512,WP\_010868714.1:0.0767037302);

Ici on voit qu’on peut identifier un seul nœud soutenu de manière robuste (99%) mais à l’intérieur du clade émergeant de ce nœud les relations de parenté entre les séquences ne sont pas résolues. On peut supposer que l’analyse faite par GBlock a éliminé une partie de résidus contenant le signal phylogénétique valide.

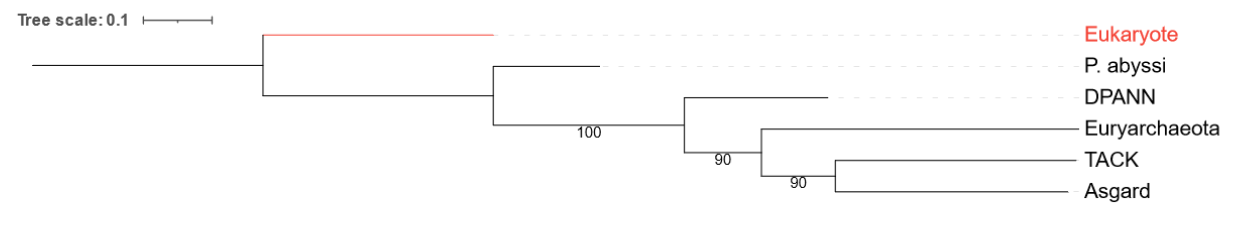
Voici le résultat obtenu avec l’alignement clustal Omega sans l’analyse GBlocks :



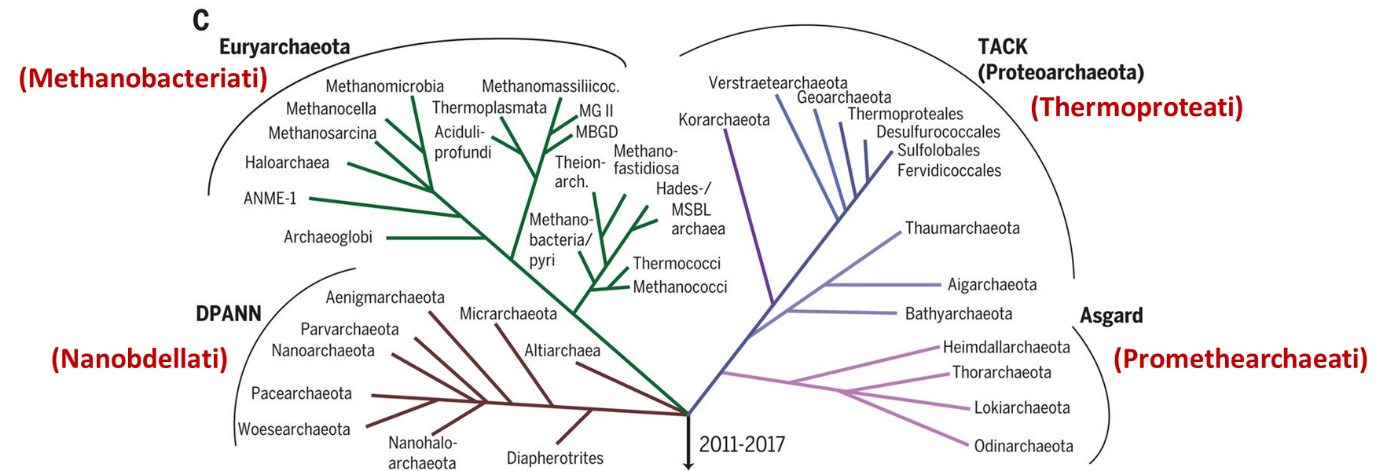
(EZG56502.1:1.1975938300,((WP\_042702101.1:0.8257192245,(RLI59347.1:0.6059517405,AIF13504.1:0.6261481430)90:0.1921924856)90:0.2003884697,RLG13225.1:0.3733046536)100:0.4968503087,WP\_010868714.1:0.2763189135);

On remarque maintenant que l’arbre a la même topologie que précédemment mais que les branches sont mieux soutenues, cela suggère que l’analyse GBlocks a en effet enlevé des acides aminés informatifs.

26. Vous allez maintenant éditer l’arbre le plus robuste (contenant votre séquence outgroup) que vous avez obtenu grâce à l’interface iTOL (<https://itol.embl.de/>). Pour cela chargez (bouton Upload) le forma Newick de l’arbre. Vous devez placer la racine de l’arbre, afficher les valeurs de bootstrap sur les branches, changer les noms des séquences pour indiquer leur groupe taxonomique et l’exporter sous forme de fichier svg que vous nommerez « Arbre ML PaSua5\_Archaea\_outgroup.svg ». Insérez l’arbre ci-dessous.



27. Commentez l’arbre que vous avez obtenu. Est-il congruent avec l’arbre ci-dessous montrant les relations de parenté entre les 4 groupes majeurs d’Archées ? Pouvons-nous nous servir des protéines Sua5/TsaC comme marqueurs pour établir la phylogénie des Archées ? La phylogénie universelle ? Pour information, les DPANN (Nanobdellati) correspond à des nanoarchées (petites cellules) dont les génomes sont de taille réduite (< 1Mpb), ces organismes se positionnent à la base de l’arbre des archées mais leur positionnement est actuellement débattu car leurs séquences subissent un taux d’évolution élevé et présentent donc des longues branches dans les reconstructions phylogénétiques.



Voici comment on peut interpréter votre arbre :

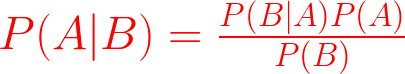
Il présente un clade solidement soutenu (BS = 100 %) comportant les Sua5/TsaC des 4 groupes d’archées, au sein de ce clade on trouve les DPANN à la base puis un groupe avec les Eury les TACK et les Asgard ou les deux derniers forment un groupe sœur. Cela correspond tout à fait à la phylogénie actuelle de principaux groupes d’Archées. Cependant, la Sua5 de *P. abyssi,* qui est pourtant une Euryarchée, se trouve exclue de ce groupe, cela est surprenant mais il y a peut-être une explication : parmi les séquences que nous avons étudié seulement la Sua5 de *P. abyssi* et la séquence outgroup eucaryote correspondent aux protéines longues Sua5. Les autres séquences sont des variantes courtes TsaC. Cela suggère que les séquences longues Sua5 et courtes TsaC se ségrégent en deux groupes distincts. Si cela est vrai, les protéines Sua5 et TsaC ne peuvent pas être utilisés comme marqueurs pour inférer la phylogénie universelle des organismes (malgré le fait qu’elles sont universelles). Cela suggère aussi que les deux versions de protéines comportent des résidus spécifiques et conservés pour l’une ou l’autre version. Dans la deuxième partie de ces travaux dirigés nous allons nous intéresser à ces résidus spécifiques de chaque variante.

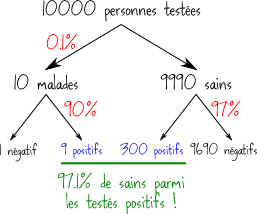
# Question bonus pour les curieux

En cours nous avons vu une autre méthode probabiliste très utilisée pour la construction d’arbres phylogénétiques : l’approche Bayésienne. La formule de Bayes (ci-dessous) permet de calculer les *probabilités à posteriori* d’une hypothèse sachant les données. Cette formule est également utilisée par exemple en médecine pour calculer la probabilité d’être malade d’après les résultats d’un test de dépistage. Voici les données :

* Votre test pour le dépistage d’une forme de cancer est positif
* Ce type de cancer ne touche que 0.1% de la population
* *« Si vous avez un cancer, le test sera positif dans 90% des cas »*
* *«  Si vous n’en avez pas, il sera négatif dans 97% des cas »*

Calculez la probabilité d’avoir un cancer si vous êtes testés positif, P(C | +).





On écrit la formule de Bayes ainsi :

P (C | +) = P (+ | C) x P (C) / P (+)

Probabilité *à priori* d’avoir cancer, P (C) = 0.1%

Probabilité (vraisemblance) d’être positif si vous avez cancer, P(+ | C) = 90%

Probabilité d’être testé positif 309/10000, P (+) = 3.09%

**Probabilité *à posteriori* d’avoir cancer sachant que je suis testé positif P (C | +) = 90% x 0.1 % / 3.09% = 2.9%**

ANNEXE

>WP\_010868714.1 threonylcarbamoyl-AMP synthase [Pyrococcus abyssi]

MTIIINVRERIEEWKIRIAAGFIREGKLVAFPTETVYGLGANALDENAVKRIFEAKGRPADNPLIIHIAS

FEQLEVLAKEIPEEAEMLAKRFWPGPLTLVLPKSEVVPRVITGGLDTVAVRMPAHEIALKLIELSERPIA

APSANISGKPSPTSAHHVAEDFYGKIECIIDGGETRIGVESTVIDLTEWPPVLLRPGGLPLEEIEKVIGE

IRIHPAVYGKSVDTAKAPGMKYRHYAPSAEVIVVEGPRDKVRRKIEELIAKFKEEGKKVGVIGSGSYDAD

EVFYLGDTVEEIARNLFKALRHMDRTGVDVILAEGVEEKGLGLAVMNRLRKASGYRIIKV

>Asgard\_RLI59347.1 threonylcarbamoyl-AMP synthase [Candidatus Thorarchaeota archaeon]

MVRIFRWPLRESEREMLSAILMNGGVVVYPTDTIYGIGGIIYDRRVVDKIYKIKERLRDKGLPILFGSIE

KIEEIAILNPLAKKLAQRFWPGMLTLVVPLKRRELRFLTGLEDKIAVRIPSDEIVLEVIKLAGGAIIGTS

ANLSGKLPCTNIKCVISQLGERMDAIVDGGERGTGFPSTIVEIQDDKKIKIIREGSIEVEKIKEAIGLEF

S

>DPANN\_RLG13225.1 threonylcarbamoyl-AMP synthase [Candidatus Pacearchaeota archaeon]

MEVEIIKVKEKLADNELIQAIRYLEQEKTVVFPTDTVYGIGADLFKEKAVKNLFEIKKRPLDKPINALVS

SMKQAEMIVENIPDSAKKLIKEFWPGGLTIILKKKDIVPDIVTAGRETIGVRKPNNEIILQILEKFGKPL

AATSANISGLTSPITAKEVLQQLGEKSYLLIDGGKTIEKQPSTIIDLSSKKPLILRKGSVSVEKIKEILP

SIIE

>TACK\_AIF13504.1 Sua5/YciO/YrdC/YwlC family protein (rimN, SUA5) [uncultured marine thaumarchaeote KM3\_62\_H02]

MKISCNDVDIQIATKAINDGAIVVFPTDTVYGLGCNPYNHDAVLSLYEIKKRKKTKPFPVLGYSKKELEK

IAEFNSLEEKIAEKFWPGPITLILKVKDKEIQKSLDLEGKIAVRVPNNQCILALLKECKLLVGTSANISG

TAPFNDPKECGENLSGYDLLIDGGVISSQGESTIVEIENNDVKILRKGSVSEEMIKELT

>Euryarcheeota\_WP\_042702101.1 threonylcarbamoyl-AMP synthase [Methanobrevibacter arboriphilus]

MKIVQMDQNNPDLDLIDEAIEVLTSGGVVLYPTDTVYGLGANVFNEKAVEKVYNIKNRDYFKPLSVCVSS

IDEILLIADVGNNKTHQILKNNLPGPFTFIFYKKEPIHNYATKNHKVGIRIPENIISRKLTQSFPITTTS

ANLSGKKTLNNPDEIINQLNEGIDFVIDVGRLKQSEPSTIVDLTKKEPKILRKGSGILKSF

>EZG56502.1 Sua5/YciO/YrdC/YwlC family protein [Gregarina niphandrodes]

MERFRVEMEDDAPAVYEEAGRRIREGKLVAFPTETVYGLGGNGLDVEAVKSIYEMKRRPADDPVILHVID

FQDAVELWDLALPLRRVIAGLQEAFWPGPLTIVAKANSQVPSEVTAGTGKVAVRCPAHPVARRLIAAARV

PIAAPSANLFSRISPTRADHVAKYFVNEDLMLIDCDSDACTIGIESVVISVEEDSIHILRPGAISRDQLQ

QAAPGIPVHDKIIPIEPSAPKDAVVAAECDFASSSPGQCLVHYSPSIPTCLIILNDAEKGLIDTEAGCEE

RETLHYDSADLERLLPLPLSRSAVIDLSANRFGFLKSSCLYWKCLCDATHDLFATFHEIEDMKVPERLYI

CTTMNDCHTNQGVLDRCIRAAAHHVIHWSPANKRAVSQ