LIÈGE université Library

https://lib.uliege.be

https://matheo.uliege.be

Mémoire

Auteur : Roomans, Célyne
Promoteur(s) : Baurain, Denis
Faculté : Faculté des Sciences
Diplôme : Master en bioinformatique et modélisation, à finalité approfondie
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/16278

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Recherche de protéines de *S-layer* archéennes par homologie de séquence et analyses phylogénétiques

Université de Liège

Faculté des sciences

Département des Sciences de la Vie

Unité de Phylogénomique des Eucaryotes

Promoteur **:** Pr. Denis Baurain

Mémoire présenté par

CÉLYNE ROOMANS

En vue de l'obtention du grade de

Master en Bioinformatique et Modélisation, à finalité

Année académique 2021-2022

Remerciements

Je tiens tout particulièrement à remercier mon promoteur, le Prof. Denis Baurain, pour son encadrement, ses conseils et son soutien sans faille tout au long de la réalisation de ce travail. De tout cœur, merci.

Mes remerciements vont également à toutes les autres personnes sans qui la réalisation de ce mémoire n'aurait pas été possible.

Enfin, je remercie mes collègues de l'Unité de Phylogénomique des Eucaryotes.

Résumé

La paroi cellulaire de la presque totalité des archées décrites jusqu'à présent se compose d'une couche protéique superficielle, appelée *S-layer*. Cette structure correspond à un réseau cristallin bidimensionnel recouvrant l'entièreté de la cellule et contenant un ou deux types de protéines. La séquence en acides aminés des protéines composant la *S-layer* (SLPs) confère à ces dernières la capacité intrinsèque de d'auto-assembler en réseau monomoléculaire. Cependant, la diversité des séquences des protéines de *S-layer* est énorme, si bien que peu – voire pas – de similitudes sont observables entre organismes étroitement apparentés. Ce constat amène à s'interroger sur l'origine évolutive de cette structure.

Un travail de littérature a d'abord été réalisé afin de relever les protéines de *S-layer* archéennes ayant été décrites jusqu'à présent. Les séquences protéiques correspondantes ont été récoltées et regroupées en 49 *clusters* selon leur similarité de séquences. Une recherche par homologie de séquences (`BLASTp`) contre une base de données d'archées représentatives a ensuite été réalisée dans le but d'enrichir les *clusters* et découvrir si des SLPs étaient représentées dans différents groupes taxonomiques. Les résultats obtenus montrent que les séquences de SLPs divergent fortement et que, par conséquent, aucune protéine n'est présente dans plusieurs groupes taxonomiques d'archées.

Des profils HMM pour les 33 *clusters* pertinents ont été construits puis utilisés afin d'identifier de nouvelles séquences et d'étudier de façon plus sensible la distribution phylogénétique de chaque SLP au sein de leur groupe taxonomique, voire au-delà. Cette analyse a mis en évidence des cas de recouvrement entre *clusters* qui ont, par conséquent, été fusionnés, étendant *de facto* la distribution de plusieurs SLPs. Des inférences phylogénétiques des 7 *clusters* obtenus après fusion ont finalement été effectuées pour étudier l'évolution de ces différentes SLPs. Les arbres résultants montrent que 1) seules les SLPs provenant d'organismes étroitement apparentés aux séquences de référence se sont vues attribuées une annotation automatique et 2) des duplications de gènes plus ou moins récentes ont eu lieu au cours de l'évolution.

En conclusion, nos analyses suggèrent que les SLPs sont effectivement très diverses, mais que leur distribution taxonomique limitée est peut-être en partie liée à une divergence importante de leur séquence primaire, plutôt qu'à une hétérogénéité réelle. Cette hypothèse devrait être testée au travers de l'identification de SLPs additionnelles à intégrer dans nos arbres.

Table des matières

Tables des illustrations

Liste des abréviations

SLP	protéine de *S-layer*
SSU rRNA	petite sous-unité de l'ARN ribosomique
UPE	Unité de Phylogénomique des Eucaryotes

1 Introduction

1.1 Le domaine des Archaea

A première vue, les archées et les bactéries se ressemblent fortement. Au point que les archées étaient autrefois considérées comme des bactéries atypiques capables de se développer dans des environnements extrêmes et/ou de produire du méthane. Certaines caractéristiques propres aux archées, telles que la présence d'étherlipides dans la membrane plasmique ou l'absence de peptidoglycane dans la paroi cellulaire, étaient d'ailleurs perçues comme des curiosités du domaine bactérien [1].

La découverte des archées en tant que troisième domaine du vivant a été faite par Carl Woese à la fin des années 1960. A l'époque, ses travaux cherchaient à comprendre comment la vie cellulaire avait évolué jusqu'aux formes existant actuellement sur Terre. Dans un premier temps, Woese s'aperçut que l'ARN ribosomique – et plus particulièrement la petite sous-unité de l'ARN ribosomique (SSU rRNA) – constituait un marqueur d'évolution idéal. Effectivement, il présente un taux de mutation lent, exerce la même fonction chez tous les organismes et, puisqu'il interagit avec une multitude de protéines, il est peu probable que les gènes qui l'encodent soient transférés entre individus d'espèces différentes. Suite à cette première observation, Woese tenta de construire un arbre phylogénétique universel intégrant toute la vie cellulaire [2]. Pour y parvenir, il commença par séquencer l'ARN ribosomique 16S, c'est-à-dire le SSU rRNA des archées et des bactéries, de tous les microorganismes dont il disposait. C'est ainsi qu'en étudiant une bactérie productrice de méthane, Woese découvrit les Archaea, un domaine du vivant différent des Bacteria et des Eucarya connus jusqu'alors [3]. En 1977, il formula l'existence de ces trois lignées ancestrales évolutivement équidistantes les unes des autres (**Figure 1**) [4]. L'introduction initiale de ce concept des trois domaines, bien que largement accepté désormais, suscita de vives critiques de la part de la communauté scientifique. Le scepticisme auquel Woese fut confronté se dissipa progressivement durant les années 1980, période durant laquelle des recherches approfondies sur l'enveloppe cellulaire des archées aboutirent à la mise en évidence de différences majeures entre les bactéries et les archées [1].

Finalement, les archées furent officiellement présentées comme le troisième domaine du vivant en 1990 ; soit près de 110 ans après que le premier rapport d'un isolat archéen dans la littérature scientifique semble été fait lorsqu'une espèce de *Halococcus*, appelée *Sarcinamorrhuae*, a été décrite ; et plusieurs milliards d'années après avoir évolué en tant que l'une des trois lignées primitives [2].

1.2 La paroi des archées

La paroi cellulaire, qu'elle soit bactérienne ou archéenne, entoure la cellule à l'extérieur de la membrane cytoplasmique. Elle maintient la forme et l'intégrité structurale de la cellule, et assure les interactions

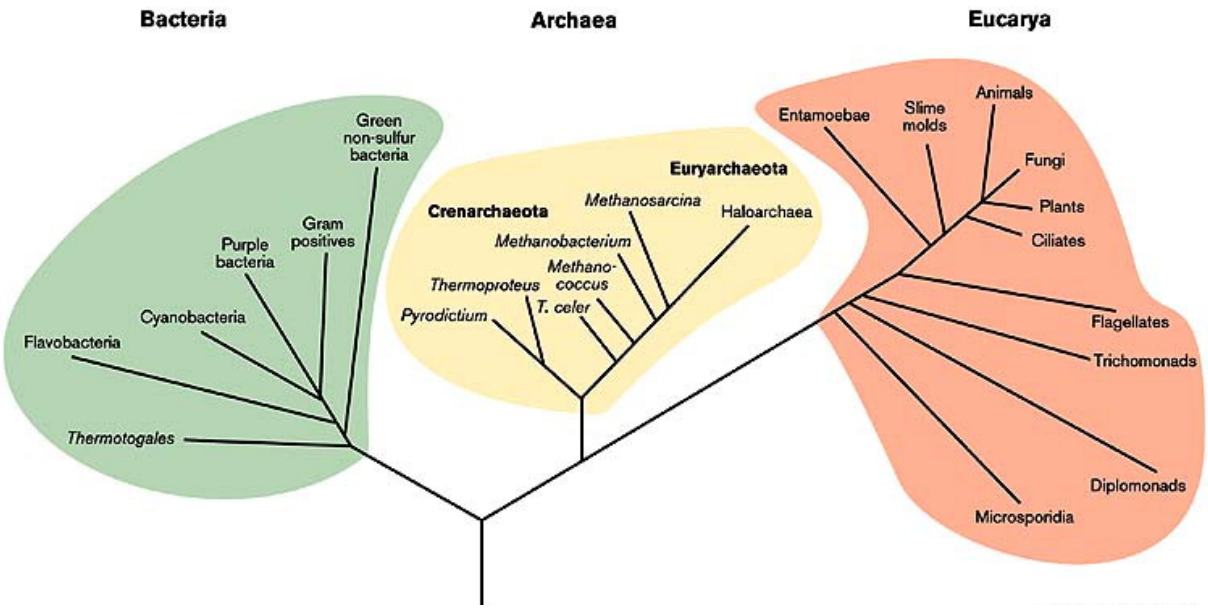


***Figure 1 : Arbre phylogénétique universel**, dont la topologie est basée sur des comparaisons de séquences d'ARN ribosomique [5].*

avec l'environnement [5,6].

Toutefois, la composition et l'organisation de la paroi cellulaire des archées est fondamentalement différente de celle des bactéries. Chez les bactéries, la paroi est principalement constituée de peptidoglycane, un exopolymère constitué de polysaccharides reliés entre eux par des pentapeptides. Cette couche de peptidoglycane est éventuellement agrémentée de polymères secondaires tels que des acides téichoïques, d'une membrane externe contenant des lipopolysaccharides (chez les bactéries dites Gram négatives), ou encore d'une capsule polysaccharidique (ex. *Klebsiella pneumoniae*) ou protéique (ex. *Bacillus anthracis*). A l'inverse, le domaine des archées se distingue par une grande diversité de types de paroi cellulaire et par l'absence de peptidoglycane.

La paroi archéenne peut être schématisée comme un assemblage, plus ou moins complexe, de diverses couches : *S-layer*, pseudomuréine, membrane externe, gaine protéique… (**Figure 2**) [1]. Cette diversité de structures permet aux archées, outre de maintenir la forme et l'intégrité de leurs cellules, de se développer dans des environnements situés aux frontières les plus extrêmes de la vie sur Terre en termes de température, de pH, de salinité et d'anaérobiose. Quoiqu'abondantes dans les milieux hostiles, les archées colonisent également des habitats plus modérés, comme les eaux douces, les sols, ou encore le tractus gastro-intestinal des animaux [7,8]. Toutes les archées connues présentent une paroi, à l'exception de *Ferroplasma acidiphilum* et *Thermoplasma* spp. Bien qu'extrêmophiles, ces organismes se sont adaptés à leurs habitats respectifs en l'absence de paroi cellulaire [6].

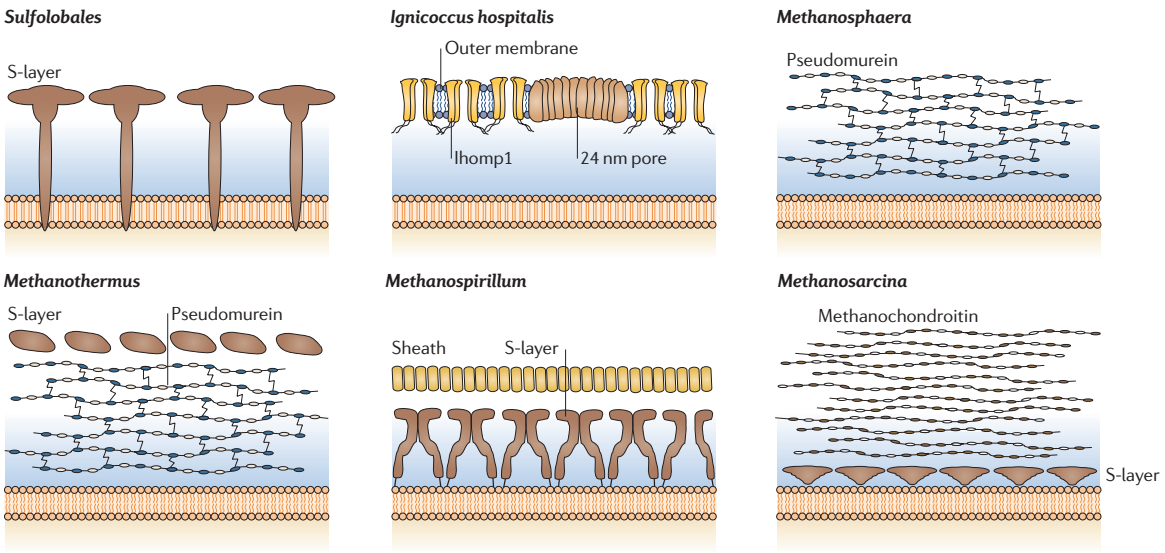


Figure 2 : Représentation schématique des types de parois cellulaires archéennes les plus courants*, y compris les genres les plus pertinents [1].*

1.3 Les S-layers

De toutes les structures composant la paroi archéenne répertoriées jusqu'à présent, la *S-layer* est la plus courante [6]. Effectivement, elle est presque omniprésente chez les archées, où elle peut constituer la seule structure de la paroi cellulaire en dehors de la membrane plasmique, et se retrouve également chez quelques bactéries (**Figure 3**) [9]. La *S-layer*, pour *Surface layer*, est la couche la plus externe de beaucoup de cellules procaryotes ; elle est donc en contact direct avec l'environnement de ces dernières [10]. Albers et Meyer (2011) ont suggéré que les *S-layers,* de par leur large distribution et leur composition simple, pourraient être les plus anciennes structures de la paroi cellulaire à avoir évolués.

La *S-layer* est un réseau cristallin bidimensionnel contenant un ou deux types de protéines ou glycoprotéines, appelées protéines de *S-layer* (SLPs). La plupart des *S-layers* ont une épaisseur de 5 à 25 nm, mais elle peut atteindre 70 nm chez *Staphylothermus marinus*, et présentent une surface extérieure plutôt lisse [1].

Selon l'organisme, les mailles du réseau cristallin se composent d'une (p1), deux (p2), trois (p3), quatre (p4) ou six (p6) SLPs identiques liées de manière non covalente, et présentent un espacement centre-à-centre allant de ~4 à ~35 nm. Le réseau de la *S-layer* présente une symétrie oblique (p1 ou p2), une symétrie carrée (p4) ou une symétrie hexagonale (p3 ou p6) (**Figure 4** et **Figure 5**) [1,9,11,12] Cette organisation régulière fait que la *S-layer* est abondamment ponctuée de pores uniformément distribués, de taille et de morphologie identiques, couvrant jusqu'à 70 % de la surface totale de la cellule. Dans de nombreuses *S-layers*, plusieurs catégories distinctes de pores, d'un diamètre compris entre 2 et 8 nm, ont été observées [10].

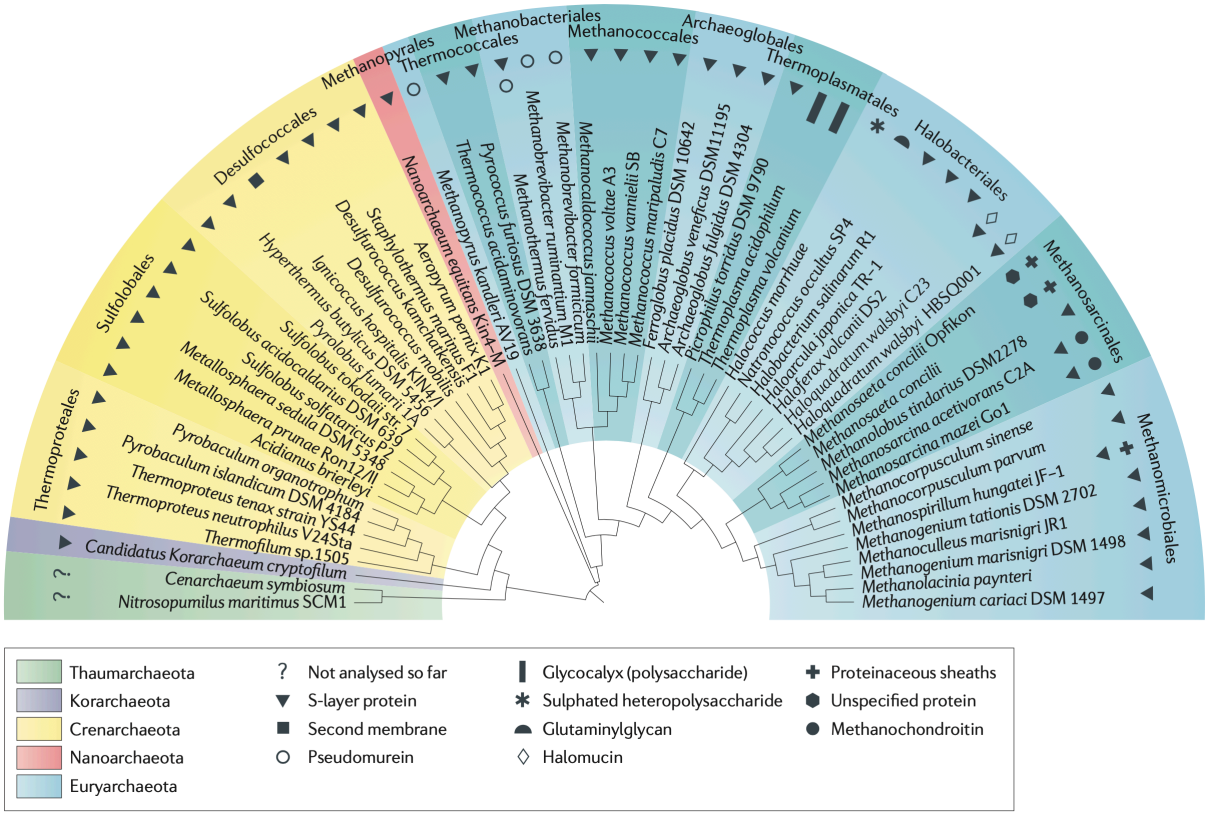


Figure 3 : Diversité des types de paroi cellulaire dans le domaine des Archaea. *L'arbre phylogénétique est basé sur l'alignement de la séquence complète de l'ARN 16S* [1]. *Notons que la biodiversité des Archaea a été fort étendue depuis 2011, notamment grâce aux études métagénomiques, mais les organismes correspondants (par exemple les Asgard supposées proches des Eucarya) ne sont pas encore bien connus [13,14].*

La majorité des Euryarchaeota, qui est le phylum comptant le plus d'espèces représentatives cultivées en laboratoire, possèdent une *S-layer* à symétrie hexagonale (p6). En revanche, les Crenarchaeota ne présentent pas de type de symétrie prédominante et arborent des réseaux hexagonaux (par exemple, p3 pour *Sulfolobus islandicus* ; p6 pour *Pyrobaculum aerophilum*) et carrés (p4 pour *Staphylothermus marinus*). Les autres phyla d'archées, y compris les Nanoarchaeota, les Thaumarchaeota, et les Asgard, restent peu caractérisés pour le moment [9,10].

La fonction de la *S-layer* était initialement incertaine, cependant il est désormais établi qu'elle sert de couche protectrice, de tamis moléculaire, de piège à molécules et à ions, et qu'elle intervient dans la reconnaissance des surfaces et le maintien de la forme des cellules [9].

La biogenèse de la *S-layer* reste encore assez méconnue. Néanmoins, elle nécessite que les SLPs soient transférées du cytoplasme, où elles sont synthétisées, vers la surface de la cellule. Chez les procaryotes, les domaines hydrophobes de la chaîne polypeptidique naissante sont reconnus pendant la traduction, ce qui a pour effet de diriger le complexe ribonucléique protéique vers la membrane plasmique. Les

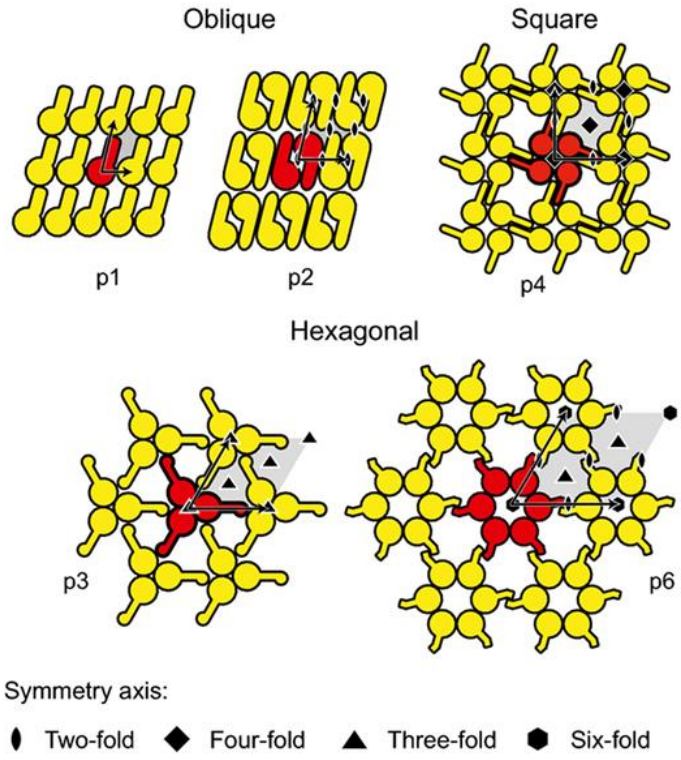


Figure 4 : Représentation des trois types de symétrie de S-layer possibles. *Les protéines de S-layer constitutive d'une maille sont mise en évidence en rouge [12].*

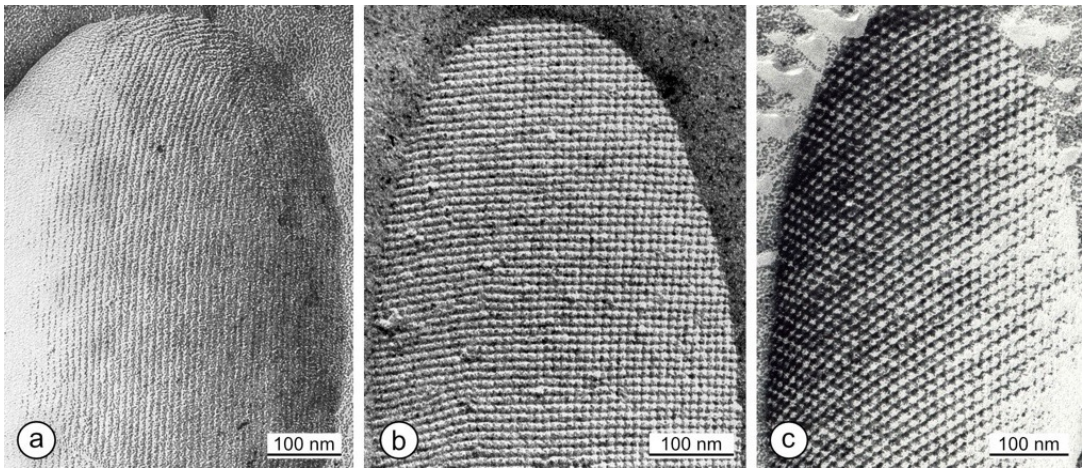


Figure 5 : Images de microscopie électronique à transmission *montrant une symétrie oblique (a), carrée (b) ou hexagonale (c) [11].*

SLPs sont ensuite transportées à travers la membrane, puis fixées à la surface cellulaire via des domaines transmembranaires ou des liaisons covalentes à des ancres lipidiques. Une fois à la surface de la cellule archéenne, les SLPs subissent des modifications post-traductionnelles – telles que l'élimination du peptide signal, la N- et la O-glycosylation – et polymérisent pour former la *S-layer* (**Figure 6**) [13].

Dans la mesure où les SLPs ont une grande proportion d'acides aminés non polaires, il est très probable que des interactions hydrophobes interviennent dans le processus d'assemblage [12]. Des expériences de microscopie optique en temps réel combinée à des marquages fluorescents ont montré que, chez *Haloferax volcanii,* l'insertion des SLPs a principalement lieu au milieu de la cellule [14]. L'assemblage de la *S-layer* chez les procaryotes serait donc étroitement lié au cycle cellulaire, en particulier aux mécanismes de division et d'élongation cellulaires [10].

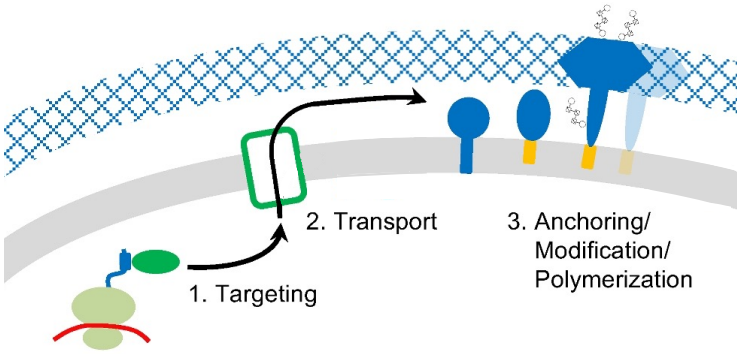


Figure 6 : Aperçu schématique des étapes de la biogenèse de la S-layer archéenne. *Les SLPs sont représentées en bleu, les protéines impliquées dans le processus de biogenèse de la S-layer en vert, et les ancrages membranaires en orange. Adapté de [13].*

1.4 Les protéines de *S-layer*

Toutes les SLPs, qu'elles soient bactériennes ou archéennes, se caractérisent par leur remarquable capacité intrinsèque à s'auto-assembler en réseau bidimensionnel, faisant ainsi d'elles des candidates prometteuses pour des applications nanotechnologiques [1].

Il a été démontré que la formation du réseau constitutif de la *S-layer* est déterminée uniquement par la structure tertiaire des SLPs et, dès lors, par la séquence des chaînes polypeptidiques [15]. Ces protéines ont une longueur de 400 à 2500 acides aminés de long et sont fortement enrichies en résidus hydrophobes et acides. Cependant, les SLPs ne présentent quasiment aucune similarité au niveau de leur séquence. Si des similitudes ont tout de même été détectées dans des groupes taxonomiques spécifiques, il arrive très fréquemment que même les SLPs d'organismes étroitement apparentés divergent fortement au niveau de leur séquence en acides aminés [10].

En plus de former un réseau, les SLPs doivent interagir avec, selon les organismes, la membrane ou la paroi procaryote pour rester accrochées à la cellule et assurer leur rôle de couche de surface. Ainsi, en général, les SLPs archéennes possèdent deux domaines : un domaine d'ancrage à la surface cellulaire en C-terminal et un domaine d'auto-assemblage en N-terminal [10,16]. Des structures atomiques de la

région d'assemblage des SLPs, résolues par cristallographie aux rayons X et cryomicroscopie électronique, ont montré que ce domaine est riche en feuillets β et forme des mailles serrées en initiant des contacts multiples le long du réseau. Faute de données structurales relatives au domaine d'ancrage des SLPs, les mécanismes de fixation de la *S-layer* à la surface cellulaire restent encore mal compris d'un point de vue atomique [10].

Chez la plupart des archées, la *S-layer* est la seule structure composant la paroi cellulaire ; celle-ci est donc directement attachée à la membrane plasmique. Les SLPs crenarchéennes, comme celles retrouvées dans l'ordre des Sulfolobales, sont généralement ancrées à la membrane plasmique par un segment transmembranairehydrophobe de type pilier situé du côté C-terminal de la protéine (**Figure 2** et **Figure 7a**). Les SLPs euryarchéennes (par exemple, *Methanosarcina* spp., *Methanospirillum* spp.), quant à elles, semblent plutôt arrimées par attachement covalent d'une partie lipidique en C-terminal avec la membrane cellulaire (**Figure 2** et **Figure 7b**). En revanche, quelques espèces d'archées (par exemple, *Methanothermus fervidus*) possèdent une couche de pseudomuréine ou de méthanochondroïtine entre la membrane plasmique et la *S-layer*. Dans ces cas précis, la *S-layer* s'ancre à cette couche intermédiaire (**Figure 2** et **Figure 7c**) [10,12].

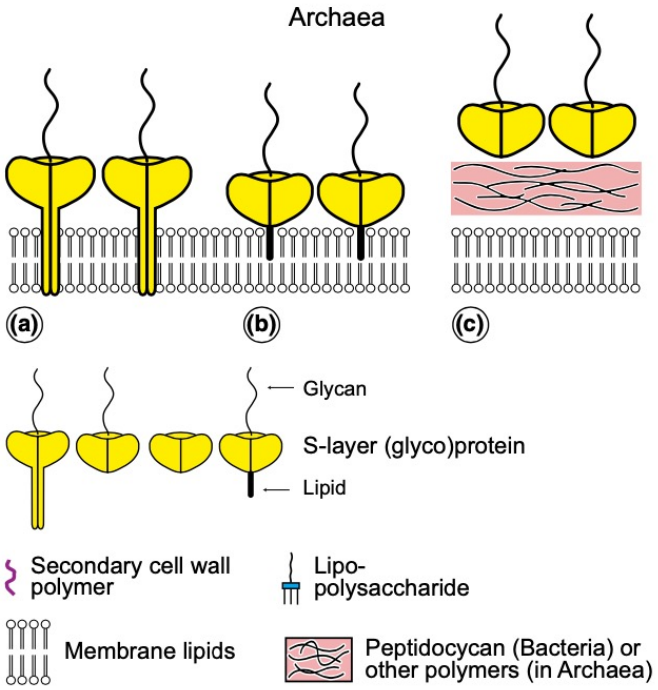


Figure 7 : Représentation des différents modes d'ancrage des SLPs à la surface cellulaire des archées. *Chez les archées dont la paroi n'est composée que d'une S-layer, les SLPs s'ancrent à la membrane plasmique, soit par l'intermédiaire de domaines transmembranaires hydrophobes de type pilier (a), soit via des lipides modifiés (b). Chez les quelques archées possédant une couche intermédiaire entre la membrane plasmique et la S-layer, les SLPs s'ancrent à cette couche supplémentaire (c) [12].*

Dans certains cas, la *S-layer* se compose de deux SLPs différentes. Dans le phylum des Euryarchaeota, *Haloarcula hispanica* est la seule espèce d'*Haloarcula* à posséder deux SLPs : Slg1 et Slg2 [17]. L'hyperthermophile *Thermococcus stetteri* présente une double *S-layer*, chaque couche étant constituée d'une SLP différente [18]. Du côté des Crenarchaeota, la *S-layer* des Sulfolobales affiche une symétrie hexagonale (p3) et ressemble à une canopée dont la hauteur dépend de l'organisme. Elle se compose de deux SLPs, SlaA et SlaB, hautement glycosylées et constituant respectivement le feuillage et les troncs de cette canopée. Les SlaB s'assemblent en trimère, chacun d'eux surmonté d'un dimère de SlaA (**Figure 8**). De plus, SlaB avoir été conservé au cours de l'évolution, mais pas SlaA [19]. Autre curiosité, la *S-layer* de *Staphylothermus marinus* forme une canopée culminant à 70 nm de la membrane cellulaire et se composant de deux SLPs assemblées en tétrabrachion [20].

En plus d'être extrêmement diverses en termes de séquences et de structures, les SLPs présentent un niveau supplémentaire de diversité par le biais de modifications post-traductionnelles. Premièrement, presque toutes les SLPs d'archées sont synthétisées avec un peptide signal et/ou un signal de tri. Ces segments sont généralement clivés après la translocation des SLPs à travers la membrane plasmique. Parfois, chez certains organismes comme *Staphylothermus marinus,* la *S-layer* comporte deux SLPs obtenues par un traitement protéolytique d'une seule protéine précurseur. Deuxièmement, la lipidation des protéines est un type modification post-traductionnelle dans lequel des fragments lipidiques sont liés de manière covalente aux protéines. Ainsi, celle-ci est essentielle à l'attachement des SLPs d'*Haloferax volcanii* à la membrane plasmique. Troisièmement, un grand nombre de SLPs archéennes sont N- et O-glycosylées. Le rôle de ces glycosylations demeure obscur, mais elles semblent liées à la survie des organismes dans les environnements extrêmes [10].

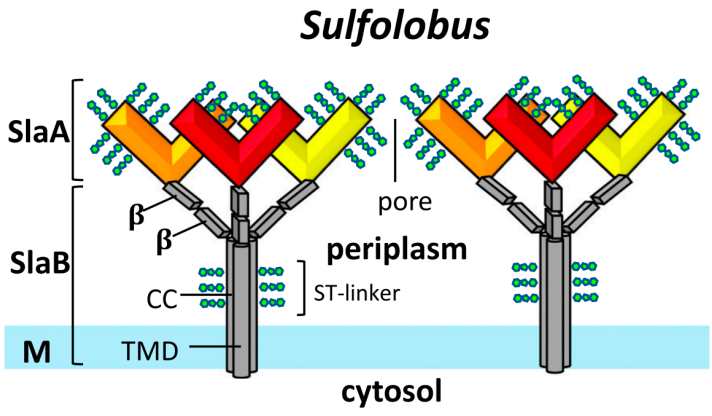


Figure 8 : Représentation de la S-layer de Sulfolobus*, composée de 2 sous-unités protéiques : des dimères de SlaA (rouge, orange, jaune) et des trimères de SlaB. Les protéines SlaA et SlaB sont fortement glycosylées (vert) [19].*

En conclusion, malgré leur composition en acides aminés similaire, les SLPs révèlent une immense diversité de séquences, de structures et de modifications post-traductionnelles. Le fait qu'elles n'aient quasiment aucune similarité au niveau de leur séquence suggère qu'elles ont des origines évolutives multiples et indépendantes [10].

2 Objectifs

L'objectif principal de ce mémoire est d'investiguer la diversité de séquence des SLPs d'archées et de jeter les bases pour étudier leur origine et leur évolution.

Pour y parvenir, des SLPs archéennes déjà décrites dans la littérature vont être recherchées. D'autres SLPs vont ensuite être recherchées par des méthodes basées sur l'homologie de séquence. Des analyses phylogénétiques vont finalement être réalisée sur les SLPs ainsi découvertes.

3 Matériel et méthodes

3.1 Environnement

3.1.1 Hardware

Outre mon ordinateur portable personnel, toutes les manipulations réalisées durant ce mémoire ont été effectuées sur le cluster de calcul *durandal*. Il s'agit d'un système Flex d'IBM/Lenovo composé d'un nœud de calcul x440, onze nœuds x240 et possédant 228 cœurs physiques (456 cœurs logiques) ainsi que 2.9 TB de RAM et 162 TB d'espace de stockage. *durandal* utilise un système d'exploitation Linux CentOS 6.6.

3.1.2 UNIX

Les manipulations ont été effectuées dans un interpréteur de commandes (shell UNIX, `bash`), dont le rôle est d'interpréter les commandes lui étant fournies par l'utilisateur via une interface en ligne de commandes.

3.1.3 Perl

Tous les scripts utilisés dans ce travail ont été rédigés en Perl, qui est un langage de programmation permettant de traiter aisément de l'information textuelle. Les scripts de ma composition sont repris en annexe et les programmes non référencés ont été développés au sein de l'Unité de Phylogénomique des Eucaryotes (UPE).

3.2 Traitement des séquences de SLPs initiales

3.2.1 Recherche et récupération de SLPs de référence

Un inventaire des SLPs ayant déjà été décrites dans la littérature a tout d'abord été fait. De cette manière, 90 identifiants de protéines ont été trouvés : 25 provenant de la banque de données Swiss-Prot (UniProtKB, https://www.uniprot.org), 5 de GenBank (https://www.ncbi.nlm.nih.gov/genbank/) et 60 de NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/). Les identifiants ont été enregistrés dans un fichier `TXT` et les séquences protéiques correspondantes récupérées au format `FASTA` grâce à la commande `efetch` (NCBI E-utilities, version 10.4) [21].

```
for i in $(cat SLPs_ids.txt)
do efetch -id $i -db protein -format fasta
done > SLPs.fasta
```

3.2.2 Déréplication

La déréplication réduit un jeu de données en se basant sur les fortes similarités qui existent entre certaines séquences. Son intérêt est d'éviter la redondance, mais également de faciliter l'exploitation des données. Les séquences de SLPs initiales ont été dérépliquées avec le programme `CD-HIT` (version 4.6) [22]. Brièvement, ce dernier regroupe les protéines dont l'identité de séquence est supérieure à un seuil défini par l'utilisateur. Les séquences qui lui sont données en entrée sont d'abord triées par ordre décroissant de longueur. La plus longue devient le représentant du premier *cluster*. Ensuite, chaque séquence restante est comparée aux représentants des *clusters* existants. Si la similarité avec un représentant est supérieure à un seuil donné, la séquence est ajoutée dans ce groupe. Sinon, un nouveau *cluster* est défini avec cette séquence comme représentant.

Lorsqu'il est utilisé en ligne de commande, `CD-HIT` renvoie deux fichiers. Le premier reprend les séquences de protéines des représentants de chaque *cluster* au format `FASTA`, tandis que le second est un rapport contenant le nombre de *clusters* et la composition de chacun d'entre eux (en termes d'identifiants). Dans le cadre de mon mémoire, j'ai écrit un script Perl, `cdhit-clustering.pl`, qui lance le programme `CD-HIT` et renvoie les séquences protéiques composant chaque *cluster* dans des fichiers `FASTA` distincts (**Annexe 1**). L'argument `--identity` correspond au seuil d'identité de séquence globale requis par `CD-HIT`, et `--in` au fichier `FASTA` d'entrée. Ainsi, les 90 séquences protéiques de référence ont été dérépliquées à 65% d'identité de séquences.

```
./cdhit-clustering.pl --identity 0.65 --in SLPs.fasta
```

La déréplication a ainsi réparti les 90 séquences de SLPs initiales dans 49 *clusters* différents, chacun correspondant en principe à un type de SLP distinct du point de vue de la séquence protéique.

3.3 Recherche de nouvelles SLPs dans le domaine des archées

Afin de couvrir toute la diversité des archées et de déterminer si certaines SLPs étaient représentées dans différents groupes taxonomiques, des `BLASTp` des représentants de chaque cluster ont été réalisés contre une banque de données assemblée par l'UPE à partir de protéomes d'archées représentatives provenant de NCBI RefSeq. En outre, cette manipulation permet d'enrichir les 49 *clusters* en séquences homologues aux SLPs de référence.

```
for SEQUENCE in SLPs-cdhit65/cluster*
do
     # Extract the representative sequence (first sequence of each file)
     head -2 $SEQUENCE > seq_rep

     # Format the sequence for BLASTp
```

```
    ali2fasta.pl seq_rep

    # Format name of the outfile
    NAME=$(basename $SEQUENCE .fasta)

    # Do BLASTp
    blastp \
    -query seq_rep.fasta \
    -db /media/vol2/home/vlupo/thesis/archaea/edmee/archee_proteomes/databases_
    archaea/db_all_archaea_proteomes/all_archaea_proteomes \
    -outfmt 5 \
    -out $NAME.fmt5

    # Remove useless files
    rm -f seq_rep
    rm -f seq_rep.fasta
done
```

Les résultats des `BLASTp` sont ensuite visualisés grâce au programme `ompa-pa.pl`. Celui-ci fournit une interface graphique représentant les *hits* (c'est-à-dire les séquences retrouvées par `BLASTp`) en fonction de leur longueur et du logarithme négatif de leur *e-value* vis-à-vis de la séquence *query*. Il prend en argument la base de données utilisée par `BLASTp` et un fichier *color* (`CLS`) attribuant une couleur à chaque classe d'archées. De plus, les zones du graphique contenant des séquences d'intérêt peuvent être sélectionnées manuellement via cette interface.

```
ompa-pa.pl cluster*.fmt5 \
--database=/media/vol2/home/vlupo/thesis/archaea/edmee/archee_proteomes/database
s_archaea/db_all_archaea_proteomes/all_archaea_proteomes \
--report-type=blastxml \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
--colorize=/media/vol2/home/vlupo/thesis/archaea_364.cls \
--min-cov=0.5 \
--max-copy=20 \
--print-plots
```

Pour chacun des 49 *clusters*, les *hits* ayant une longueur relativement proche de celle de la séquence *query* et une faible (= très significative) *e-value* ($< 10^{-50}$) ont été sélectionnés et leurs identifiants sauvegardés dans des fichiers `IDL`. Les 49 sélections ont ensuite été examinées plus en détail, toujours via l'interface graphique de `ompa-pa.pl`, pour déterminer les groupes de protéines à étudier par la suite. Parmi les 49 *clusters*, les 33 *clusters* comportant un nombre suffisant de *hits* ont été sélectionnées pour être enrichis et nouvelle fois et être soumis à des analyses phylogénétiques. Les noms de ces 33 *clusters* ont été enregistrés dans le fichier nommé `selected_clusters.txt`.

3.4 Assemblage de bases de données de protéomes complets

Seuls quatre groupes taxonomiques (Halobacteria, Methanococci, Methanomicrobia et Sulfolobales) sont représentés dans les 33 *clusters* sélectionnés au point précédent. En vue d'enrichir une nouvelle fois les *clusters*, des bases de données contenant chacune des séquences protéiques d'Halobacteria, de Methanococci, de Methanomicrobia et de Sulfolobales ont été construites. Une banque de données représentant toute la diversité des archées, y compris les Asgard, a également été assemblée dans le but de chercher des SLPs potentiellement distribuées dans l'ensemble du domaine des Archaea.

3.4.1 ToRQuEMaDA

Pour générer des banques de données reprenant toute la diversité des Archaea, des Halobacteria, des Methanococci, des Methanomicrobia et des Sulfolobales, des génomes représentatifs de ces groupes doivent être sélectionnés. L'outil `ToRQuEMaDA` (*Tool for Retrieving Queried Eubacteria, Metadata and Dereplicating Assemblies*) permet de télécharger, stocker et produire des listes de génomes dérépliqués représentatifs de la diversité procaryote [23]. Ce programme se compose de plusieurs scripts Perl à lancer l'un après l'autre.

La première étape du pipeline de `ToRQuEMaDA` a consisté à télécharger les génomes et les protéomes archéens de GenBank n'étant pas encore stockés dans la base de données MySQL 5 du programme et à mettre à jour la taxonomie à l'aide du script `tqmd_download.pl`.

Ensuite, les différentes métriques nécessaires à `ToRQuEMaDA` pour dérépliquer et sélectionner les meilleurs génomes représentatifs ont été calculées grâce au second script (`tqmd_update.pl`) : la qualité des génomes a été mesurée avec `QUAST` [24] ; les ARN ribosomiques 16S ont été prédits et dérépliqués selon leur identité de séquences avec `RNAmmer` [25] et `CD-HIT`, respectivement ; le niveau de contamination des génomes a été estimé avec `Forty-Two` [26]. La richesse de l'annotation des génomes a également été évaluée à l'aide d'un script interne de l'UPE.

Finalement, le troisième script, qui est le script principal de `ToRQuEMaDA`, a été utilisé. C'est lui qui assure la déréplication et la sélection des génomes à proprement parler. La déréplication peut être modulée via les diverses options acceptées par le script. Parmi les arguments utilisés, `--negative-gca-list` permet de ne pas sélectionner de génomes ayant été retirés de GenBank (après leur mise à disposition initiale) ni de métagénomes ; `--distance-threshold` définit le seuil d'identité de séquence au-delà duquel deux génomes sont regroupés ; `--requires-SSU-rRNA` ordonne au programme de ne prendre en compte que les génomes pour lesquels au moins une séquence d'ARN ribosomique 16S a été prédite par le sous-script `RNAmmer` de `tqmd_update.pl` (**Annexe 2**).

A la sortie du pipeline, 5 listes de génomes représentatifs et dérépliqués ont été obtenus. Chacune de ces

listes a ensuite été utilisée pour générer une base de données de protéomes complets.

3.4.2 Récupération des protéomes

Les listes de génomes représentatifs des Archaea, des Halobacteria, des Methanococci, des Methanomicrobia et des Sulfolobales générées par `ToRQuEMaDA` sont récupérées. Ces dernières comportent respectivement 150, 185, 37, 107 et 45 génomes.

```
cp -s ~/thesis_Master2/TQMD_SLPs_genbank/TQMD/results_TQMD/Archaea_genbank_mash1
8_kmer14_SSU_20220727/*result tqmd_archaea.gca

cp -s ~/thesis_Master2/TQMD_SLPs_genbank/TQMD/results_TQMD/Halobacteria_genbank_
mash07_kmer14_20220721/*result tqmd_halobacteria.gca

cp -s ~/thesis_Master2/TQMD_SLPs_genbank/TQMD/results_TQMD/Methanococci_genbank_
mash01_kmer14_20220721/*result tqmd_methanococci.gca

cp -s ~/thesis_Master2/TQMD_SLPs_genbank/TQMD/results_TQMD/Methanomicrobia_genba
nk_mash01_kmer14_20220721/*result tqmd_methanomicrobia.gca

cp -s ~/thesis_Master2/TQMD_SLPs_genbank/TQMD/results_TQMD/Sulfolobales_genbank_
mash01_kmer14_20220721/*result tqmd_sulfolobales.gca
```

Ensuite, la taxonomie relative aux génomes de chaque liste a été récupérée avec le script `fetch-tax.pl`.

```
for GCA in tqmd_archaea.gca tqmd_halobacteria.gca tqmd_methanococci.gca \
tqmd_methanomicrobia.gca tqmd_sulfolobales.gca
do
     # Télécharge la taxonomie relative à ces génomes
     fetch-tax.pl $GCA \
     --item-type=taxid \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/
done
```

Enfin, les protéomes complets des organismes de chaque liste ont été récupérés à l'aide d'une série de `rsync` et concaténés dans des fichiers `FAA`.

3.4.3 Création des bases de données BLAST

Des banques de données ont finalement été assemblées à partir des fichiers `FAA` à l'aide de la commande `makeblastdb` (version 2.2.28+).

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
     makeblastdb -in db_${CLASS}_proteomes.faa \
     -dbtype prot -out db_${CLASS}_proteomes \
     -parse_seqids
done
```

De cette manière, 5 bases de données ont été créées en vue des recherches par homologie et des analyses phylogénétiques (**Tableau 1**).

Groupe taxonomique	Nom	Contenu
Archaea	`db_Archaea_proteomes`	287 058 protéines issues de 150 génomes
Halobacteria	`db_Halobacteria_proteomes`	664 219 protéines issues de 185 génomes
Methanococci	`db_Methanococci_proteomes`	63 577 protéines issues de 37 génomes
Methanomicrobia	`db_Methanomicrobia_proteomes`	299 025 protéines issues de 107 génomes
Sulfolobales	`db_Sulfolobales_proteomes`	111 640 protéines issues de 45 génomes

***Tableau 1 : Résumé des bases de données BLAST** créées avec makeblastdb.*

3.5 Recherche de nouvelles SLPs dans certains groupes taxonomiques

3.5.1 Récupération et alignement des séquences

Pour chacun des 33 *clusters* retenus, les séquences protéiques sélectionnées par l'intermédiaire de l'interface de `ompa-pa.pl` et conservées par les filtres inhérents au programme ont été récupérées au format `FASTA` grâce à la commande `blastdbcmd`.

```
cat selected_clusters.txt | while read F
do
     blastdbcmd \
     -db /media/vol2/home/vlupo/thesis/archaea/edmee/archee_proteomes/databases_
     archaea/db_all_archaea_proteomes/all_archaea_proteomes
     -entry_batch $F-1.idl \
     -out $F.fasta
done
```

L'identifiant complet de chaque séquence protéique a ensuite été restauré, notamment avec `change-ids-ali.pl`.

```
perl -i -nle 's/>lcl\|(.*?) unnamed protein product/>\1/; print' *.fasta
```

```
cat *.fasta | grep \> | cut -c2- | cut -f1 -d'|'|sort -u > 33clusters.gcf

fetch-tax.pl 33clusters.gcf \
--item-type=taxid \
--org-mapper \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/

change-ids-ali.pl *.fasta \
--org-mapper=33clusters.org-idm \
--mode=abbr2long \
--out=_longid
```

Les séquences de chaque *cluster* ont été finalement alignées grâce à l'algorithme `LINSI` du package `MAFFT` (v7.453) [27].

```
for CLUSTER in cluster*_longid.fasta
do
     # Format name of the outfile
     NAME=$(basename $CLUSTER _longid.fasta)
     OUT="${NAME}_aligned.fasta"

     # Align the sequences
     linsi --thread 40 $CLUSTER > $OUT
done
```

3.5.2 Enrichissement des 33 *clusters*

De nouvelles SLPs ont été recherchées dans l'intégralité du domaine des archées et dans certains groupes taxonomiques (mentionnés précédemment). Le but de cette manipulation était, encore une fois, de déterminer si certaines SLPs étaient représentées dans différents groupes taxonomiques, et d'enrichir les 33 *clusters* sélectionnés au point 3.3. A partir des alignements obtenus à l'opération précédente, des profils HMM (*Hidden Markov Model*) ont été construits grâce à la commande `hmmbuild` du package `HMMER` [28].

```
for IN in cluster*_aligned.fasta
do
     # Format name of the outfile
     NAME=$(basename $IN .fasta)
     OUT="profiles-HMM/${NAME}.hmm"

     # Construct profile
```

```
      hmmbuild $OUT $IN
done
```

Via la commande `hmmsearch` [28], ces profils ont été exploités pour rechercher des séquences de possibles SLPs dans les 5 bases de données assemblées au point 3.4.3. Les 33 profils HMM ont été soumis à un `hmmsearch` contre la banque de données d'Archaea `db_Archaea_proteomes`. Des `hmmsearch` des profils correspondants à des Halobacteria ont été réalisés contre la banque de données des Halobacteria (`db_Halobacteria_proteomes`). Il en a été fait de même pour les profils de Methanococci, de Methanomicrobia et de Sulfolobales.

```
# Archaea
cat selected_clusters.txt | while read LINE
do
      # Format name of the outfile
      IN="profiles-HMM/${LINE}_aligned.hmm"
      OUT="hmmsearch/Archaea/${LINE}_aligned.hmms"

      # Search in the DB
      hmmsearch --domtblout $OUT $IN \
      db_Archaea_proteomes/db_Archaea_proteomes.faa
done

# Halobacteria – Methanococci – Methanomicrobia – Sulfolobales
for CLASS in Halobacteria Methanococci Methanomicrobia Sulfolobales
do
      cat ${CLASS}_clusters.txt | while read LINE
      do
            # Format name of the outfile
            IN="profiles-HMM/${LINE}_aligned.hmm"
            OUT="hmmsearch/${CLASS}/${LINE}_aligned.hmms"

            # Search in the DB
            hmmsearch --domtblout $OUT $IN \
             db_${CLASS}_proteomes/db_${CLASS}_proteomes.faa
      done
done
```

Les résultats des `hmmsearch` ont été visualisés et traités par `ompa-pa.pl` comme cela avait déjà été fait après les recherches par `BLASTp` (voir 3.3). Certaines protéines récupérées se retrouvent dans plusieurs *clusters* car elles partagent des similarités de séquence (**Annexe 3**). Les *clusters* concernés par ces chevauchements ont donc été fusionnés en vue des analyses phylogénétiques (**Tableau 2**), par concaténation des fichiers `IDL` résultant de `ompa-pa.pl`.

```
## cluster40-46.idl
cat cluster40-B1GT62.idl cluster46-A4YHQ9.idl | sort | uniq > cluster40-46.idl

## cluster15-17.idl
cat cluster15-WP_230375448.idl cluster16-WP_048108772.idl \
cluster17-WP_135388159.idl | sort | uniq > cluster15-17.idl

## cluster43-44.idl
cat cluster43-WP_004040159.idl cluster44-WP_066957445.idl | sort | uniq \
> cluster43-44.idl

## cluster45-47.idl
cat cluster45-WP_011845030.idl cluster47-WP_211530663.idl | sort | uniq \
> cluster45-47.idl

## cluster32-39.idl
cat cluster32-Q50833.idl cluster35-A6URZ5.idl cluster36-Q8X235.idl \
cluster37-Q58232.idl cluster39-WP_013799875.idl | sort | uniq > cluster32-39.idl

## cluster12-19.idl
cat cluster12-G0HV86.idl cluster19-A0A0K1IRS6.idl | sort | uniq \
> cluster12-19.idl

## cluster09-23.idl
cat cluster09-WP_011570500.idl cluster10-WP_129452348.idl cluster14-Q5V7F4.idl \
cluster18-Q9C4B4.idl cluster20-B0R8E4.idl cluster21-WP_007188624.idl \
cluster22-P25062.idl cluster23-WP_014555115.idl | sort | uniq > cluster09-23.idl
```

***Clusters* se chevauchant**	→	***Cluster* fusionné**
Clusters 40 et 46	→	`cluster40-46`
Clusters 15, 16 et 17	→	`cluster15-17`
Clusters 43 et 44	→	`cluster43-44`
Clusters 45 et 47	→	`cluster45-47`
Clusters 32, 35, 36, 37 et 39	→	`cluster32-39`
Clusters 12 et 19	→	`cluster12-19`
Clusters 9, 10, 14, 18, 20, 21, 22, 23	→	`cluster09-23`

Tableau 2 : Résumé des clusters fusionnés.

3.5.3 Récupération des séquences protéiques

Pour chaque *cluster* fusionné, les séquences protéiques correspondant aux identifiants enregistrés dans

les fichiers `IDL` sont extraites des bases de données utilisées au point 3.5.2 lors des `hmmsearch` grâce à la commande `blastdbcmd` (version 2.2.28+). Les séquences en acides aminés récupérées ont été utilisées pour construire des arbres phylogénétiques (voir 3.6.2).

```
# Halobacteria
for CLUSTER in cluster09-23 cluster12-19
do
     blastdbcmd \
     -db db_Halobacteria_proteomes \
     -entry_batch ${CLUSTER}.idl \
     -out ${CLUSTER}.fasta
done

# Methanococci
blastdbcmd \
-db db_Methanococci_proteomes \
-entry_batch cluster32-39.idl \
-out cluster32-39.fasta

# Methanomicrobia
for CLUSTER in cluster15-17 cluster43-44 cluster45-47
do
     blastdbcmd \
     -db db_ Methanomicrobia_proteomes \
     -entry_batch ${CLUSTER}.idl \
     -out ${CLUSTER}.fasta
done

# Sulfolobales
blastdbcmd \
-db db_Sulfolobales_proteomes \
-entry_batch cluster40-46.idl \
-out cluster40-46.fasta
```

3.6 Inférence d'arbres phylogénétiques

3.6.1 Des génomes représentatifs sélectionnés par ToRQuEMaDA

3.6.1.1 Extraction des protéines ribosomiques

Pour chaque base de données assemblée au point 3.4.3, un arbre phylogénétique a été inféré à partir des protéines ribosomiques des organismes représentés. L'utilisation de `Forty-Two` en mode métagénomique a permis d'extraire les protéines ribosomiques des organismes. Pour y parvenir, `Forty-`

Two cherche des orthologues à 90 protéines ribosomiques de référence (procaryotes et eucaryotes) dans les séquences protéiques des banques de données. J'ai écrit un script Bash (`forty-two.sh`) qui crée les fichiers de configuration requis par `Forty-Two` et lance le programme de manière automatisée (**Annexe 4**). Les protéines ribosomiques extraites par `Forty-Two` sont enregistrées dans 90 fichiers `FASTA` (un fichier par protéine ribosomique de référence). Ces fichiers sont ensuite alignés avec le programme `MAFFT` (v7.453) afin d'être utilisés plus tard dans le processus [29].

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
     for FASTA in $CLASS/ribo_prots_Fasta/*.fasta
     do
            FILE=$(echo $FASTA |cut -d'/' -f10)
            echo "In: $FILE"

            BASE=$(echo $FILE |cut -d'.' -f1)
            OUTPUT="$BASE"_align.fasta

            echo "Mafft in process..."
            mafft --anysymbol --auto --reorder $FASTA > $OUTPUT

            echo "Out: $OUTPUT"
     done
done
```

Après cela, les 90 fichiers `FASTA` sortant de `Forty-Two` (non alignés) sont traités avec le programme `SCaFoS` (*Selection, Concatenation and Fusion of Sequences for phylogenomics*) [30]. Cet outil suit un protocole en trois étapes.

Lors de la première étape, les organismes et leur fréquence d'apparition dans les fichiers `FASTA` sortant de `Forty-Two` ont été inventoriés afin de sélectionner les génomes à garder pour la suite des analyses. La seconde étape a permis de retirer les séquences protéiques provenant des organismes n'ayant pas été sélectionnés à la première étape. L'intérêt de la manipulation présentée ici étant de réaliser un arbre ribosomique des génomes sélectionnés par `ToRQuEMaDA`, tous les organismes ont été conservés. Ces deux manipulations successives ont également été automatisées à l'aide d'un script Bash (`scafos_1-2.sh`, **Annexe 5**).

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
     # Scafos_1
     cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${CLASS}
     scafos.pl in=ribo_prots_Fasta/ out=scafos1/
```

```
    grep -v "#" scafos1/scafos1-freq.otu | sed 's/ (.*//' > ${CLASS}.otu

    # Scafos_2
    scafos.pl in=ribo_prots_Fasta/ out=scafos2/ otu=${CLASS}.otu
done
```

La troisième et dernière étape de SCaFoS permet, pour chaque organisme, de concaténer les séquences en acides aminés des protéines ribosomiques extraites par Forty-Two. Les fichiers à concaténer doivent avoir été préalablement alignés, puis filtrés avec ali2phylip.pl. Un script Bash reprenant les différentes lignes de commandes nécessaires a été créé (scafos_3.pl, **Annexe 6**).

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
    # Ali2phylip
    cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${CLASS}/mafft/
    for FILE in * _align.fasta
    do
        ali2phylip.pl --bmge-mask=loose --ali --min=0.3 --max=0.5 $FILE
    done

    # Scafos 3
    cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${CLASS}/
    scafos.pl in=ali2phylip/ out=scafos3/ otu=${CLASS}.otu format=fp g=25
done
```

3.6.1.2 Construction d'arbres ribosomiques

Le fichier ALI (aligné) résultant de la troisième étape de SCaFoS est converti en FASTA par ali2fasta.pl puis utilisé pour construire un arbre phylogénétique basé sur les protéines ribosomiques grâce au programme iqtree-mpi (v1.6.12) [31].

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
    # Pour construire l'arbre un fichier FASTA est requis
    ali2fasta.pl $CLASS/scafos3/scafos3.ali

    name=$(echo "$CLASS" | tr A-Z a-z)

    # Construction de l'arbre
    iqtree-mpi-1612 -s $CLASS/scafos3/scafos3.fasta -v -nt 20 -pre ${name}-lg4x
    -bb 1000 -wbtl -m LG4X -seed 12345
```

```
done
```

3.6.1.3 Formatage des arbres

Les cinq arbres résultants sont formatés grâce au programme `format-tree.pl` afin d'être importés et visualisés dans `iTOL` (https://itol.embl.de).

```
for CLASS in Archaea Halobacteria Methanococci Methanomicrobia Sulfolobales
do
     format-tree.pl ${name}-lg4x.treefile \
     --annotate \
     --map-ids \
     --colorize=/media/vol2/home/croomans/thesis_Master2/colors_cls/${CLASS}
     _colors.cls \
     --itol \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
     --collapse=genus
done
```

3.6.2 Des *clusters* fusionnés

Des arbres phylogénétiques des possibles SLPs appartenant aux sept *clusters* issus de la fusion d'autres *clusters* ont été construits.

3.6.2.1 Alignement des séquences

Les séquences protéiques des sept *clusters* d'intérêt récupérées au point 3.5.3 ont été alignées à l'aide de l'algorithme `LINSI` (v7.453) [27].

```
for CLUSTER in cluster09-23 cluster12-19 cluster15-17 cluster32-39 \
cluster40-46 cluster43-44 cluster45-47
do
     linsi --thread 20 ${CLUSTER}.fasta > ${CLUSTER}_aligned.fasta
done
```

3.6.2.2 Construction des arbres préliminaires

Les alignements des séquences des sept *clusters* fusionnés ont ensuite été filtrés via le programme `ali2phylip.pl`. L'option `--max` désigne la proportion maximale de *non-gap* qu'un site peut présenter (par rapport au nombre de séquences) pour être éliminé, tandis que `--min` représente la proportion minimale de *non-gap* qu'une séquence doit avoir (par rapport à la séquence la plus longue) pour être

conservée. De plus, l'option `--map-ids` commute l'identifiant des séquences en « seqN » et crée le fichier `IDM` correspondant ; et `--ali` détermine le format du fichier de sortie.

```
ali2phylip.pl *.fasta --max=0.3 --min=0.3 --ali --map-ids
```

Les fichiers obtenus ont été convertis au format `FASTA` par `ali2fasta.pl` puis soumis à `iqtree-mpi` (v1.6.12) pour construire des arbres phylogénétiques [31].

```
for CLUSTER in cluster09-23 cluster12-19 cluster15-17 cluster32-39 \
cluster40-46 cluster43-44 cluster45-47
do
     ali2fasta.pl ${CLUSTER}.ali

     iqtree-mpi-1612 -s ${CLUSTER}.fasta -v -nt 35 -pre ${CLUSTER}_lg4x -bb 1000
     -wbtl -m LG4X -seed 12345
done
```

3.6.2.3 Formatage des arbres préliminaires

Les sept fichiers `TREEFILE` produits sont formatés avec `format-tree.pl` avant d'être importés et visualisés dans `iTOL` (https://itol.embl.de). Les fichiers *color* (`CLS`) utilisés varient selon la classe taxonomique du *cluster*.

```
# Halobacteria
for TREE in cluster09-23 cluster12-19
do
     format-tree.pl ${TREE}_lg4x.treefile \
     --map-ids \
     --annotate \
     --colorize=halobacteria_colors.cls \
     --itol \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
     --collapse=label
done

# Methanococci
format-tree.pl cluster32-39_lg4x.treefile \
--map-ids \
--annotate \
--colorize=methanococci_colors.cls \
--itol \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
```

```
--collapse=label

# Methanomicrobia
for TREE in cluster15-17 cluster43-44 cluster45-47
do
     format-tree.pl ${TREE}_lg4x.treefile \
     --map-ids \
     --annotate \
     --colorize= methanomicrobia_colors.cls \
     --itol \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
     --collapse=label
done

# Sulfolobales
format-tree.pl cluster40-46_lg4x.treefile \
--map-ids \
--annotate \
--colorize= sulfolobales _colors.cls \
--itol \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
--collapse=label
```

3.6.2.4 Construction des arbres finaux

A l'issue de cette analyse, seul un arbre relatif au `cluster40-46` a dû être recalculé. En effet, il présentait une branche (`Sulfodiicoccus_acidiphilus_GCA_003967175.1@BBD73200.1`) plus longue que toutes les autres et provenant d'une souche présente deux fois dans l'arbre. Cette séquence a donc été supprimée et l'inférence d'un second arbre phylogénétique réalisée.

```
cp cluster40-46.fasta cluster40-46_v2.fasta

# Élimination manuelle de la séquence

linsi --thread 20 cluster40-46_v2.fasta > cluster40-46_v2_aligned.fasta

ali2phylip.pl cluster40-46_v2_aligned.fasta --max=0.3 --min=0.3 --ali --map-ids

ali2fasta.pl cluster40-46_v2_aligned-a2p.ali

iqtree-mpi-1612 -s cluster40-46_v2_aligned-a2p.fasta -v -nt 35 \
-pre cluster40-46_v2_lg4x -bb 1000 -wbtl -m LG4X -seed 12345
```

3.6.2.5 Annotation des séquences

Une annotation a été attribuée aux protéines grâce au programme `annotate.pl`. Pour annoter les séquences protéiques avec ce script, des protéines de référence sont nécessaires. Les représentantes des 49 *clusters* initiaux ont donc été utilisées comme référence.

```
for F in cluster09-23 cluster12-19 cluster15-17 cluster32-39 cluster40-46_v2
cluster43-44 cluster45-47
do
     ./annotate.pl ${F}_aligned_longid.fasta \
     --ref-file ref_prots.fasta \
     --ref-regex '^\w+\|\w+.*\|(\w+)\s' \
     --write-ann-file \
     --identity=50 \
     > ${F}_annotation.log
done
```

3.6.2.6 Formatage final des arbres

Les arbres sont finalement formatés avec `format-tree.pl` afin de faire apparaitre les annotations attribuées aux organismes à l'étape précédente. Les fichiers résultants sont importés et visualisés dans `iTOL`. Comme précédemment, les fichiers *color* (`CLS`) utilisés varient selon la classe taxonomique du *cluster*.

```
# Halobacteria
for TREE in cluster09-23 cluster12-19
do
     format-tree.pl annotated_tree/${TREE}_lg4x.treefile \
     --map-ids \
     --annotate \
     --colorize=halobacteria_colors.cls \
     --itol \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
     --collapse=label \
     --ladderize=desc
done

# Methanococci
format-tree.pl annotated_tree/cluster32-39_lg4x.treefile \
--map-ids \
--annotate \
--colorize=methanococci_colors.cls \
--itol \
```

```
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
--collapse=label \
--ladderize=desc

# Methanomicrobia
for TREE in cluster15-17 cluster43-44 cluster45-47
do
     format-tree.pl annotated_tree/${TREE}_lg4x.treefile \
     --map-ids \
     --annotate \
     --colorize= methanomicrobia _colors.cls \
     --itol \
     --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
     --collapse=label \
     --ladderize=desc
done

# Sulfolobales
format-tree.pl annotated_tree/cluster40-46_v2_lg4x.treefile \
--map-ids \
--annotate \
--colorize=sulfolobales_colors.cls \
--itol \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
--collapse=label \
--ladderize=desc
```

3.7 Krona

En parallèles des manipulations déjà décrites, des diagrammes circulaires des abondances relatives des différents groupes taxonomiques d'archées présentes dans la base de données de `ToRQuEMaDA` et sélectionnés par ce dernier ont été visualisées à l'aide de `Krona` [32].

4 Résultats

4.1 Diversité taxonomique

4.1.1 Des génomes d'archées complets

A la date du 15 mai 2022, le nombre de génomes d'archées enregistrés dans la base de données de ToRQuEMaDA s'élevait à 9849 (**Figure 9**). Ces génomes se répartissent dans cinq groupes principaux : les *Euryarchaeota* (34%), le *TACK group* (22%), le *DPANN group* (20%), les *Candidatus Thermoplasmatota* (16%), le *Asgard group* (4%). Les génomes restants (4%) sont des Archaea ne possédant pas encore de taxonomie, des *Candidatus Hydrothermarchaeota*, ou des métagénomes (provenant d'échantillons environnementaux).

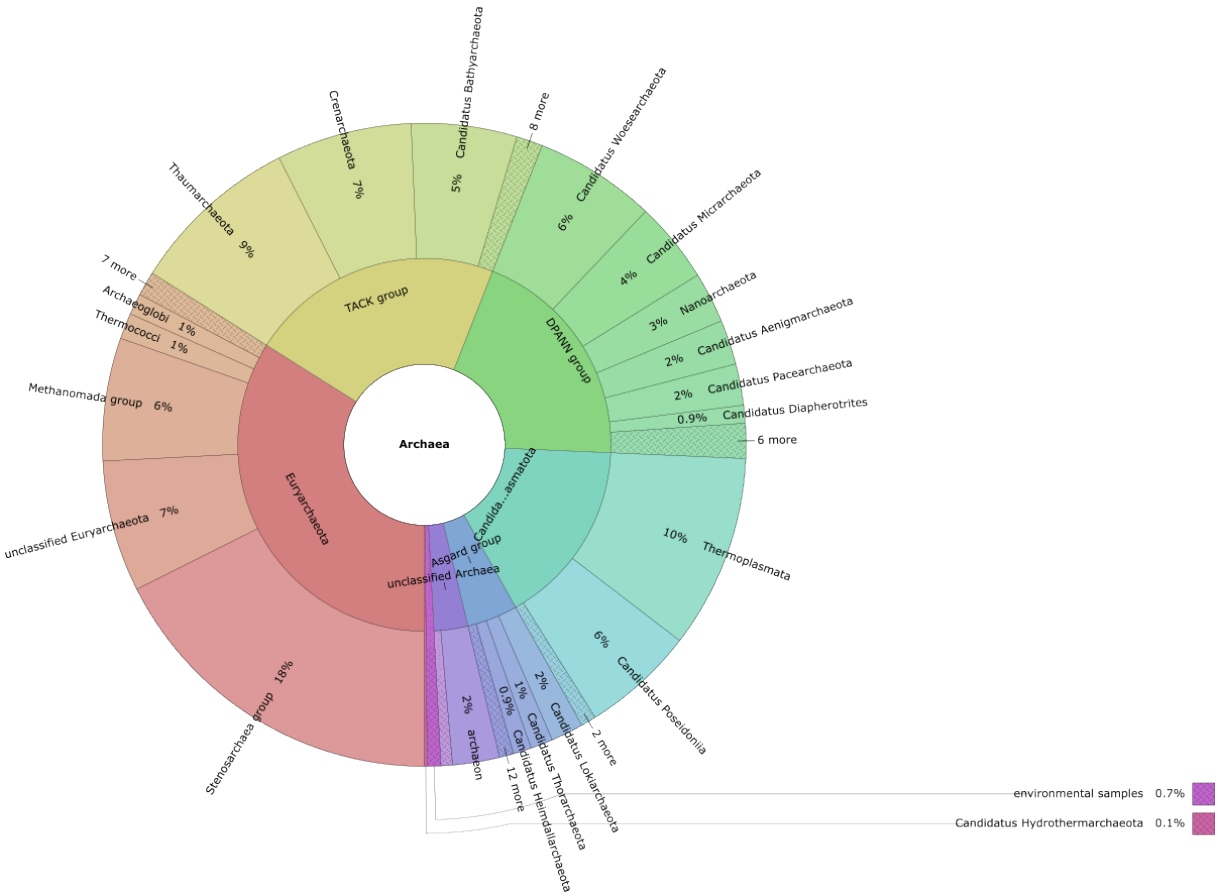


Figure 9 : Diagramme Krona illustrant la diversité taxonomique des 9849 génomes archéens *de la base de données de ToRQuEMaDA en date du 15 mai 2022.*

4.1.2 Des génomes sélectionnés par ToRQuEMaDA

Au cours des recherches de nouvelles SLPs par homologie de séquence, quatre groupes archéens ont été étudiés plus en détail : les Halobacteria, les Methanococci, les Methanomicrobia et les Sulfolobales. Des

génomes représentatifs et dérépliqués de ces groupes et du domaine des Archaea ont été sélectionnés par ToRQuEMaDA afin de construire des banques de données de séquences protéiques où rechercher des SLPs.

Un total de 150 génomes archéens ont ainsi été sélectionnés par ToRQuEMaDA. Parmi ceux-ci, 42% sont des *Euryarchaeota*, 36% des *TACK group*, 9% des *Candidatus Thermoplasmatota*, 4% des *DPANN group*. De plus, seul un génome d'*Asgard group* (du genre *Candidatus Prometheoarchaeum*) a été choisi (**Figure 10**).

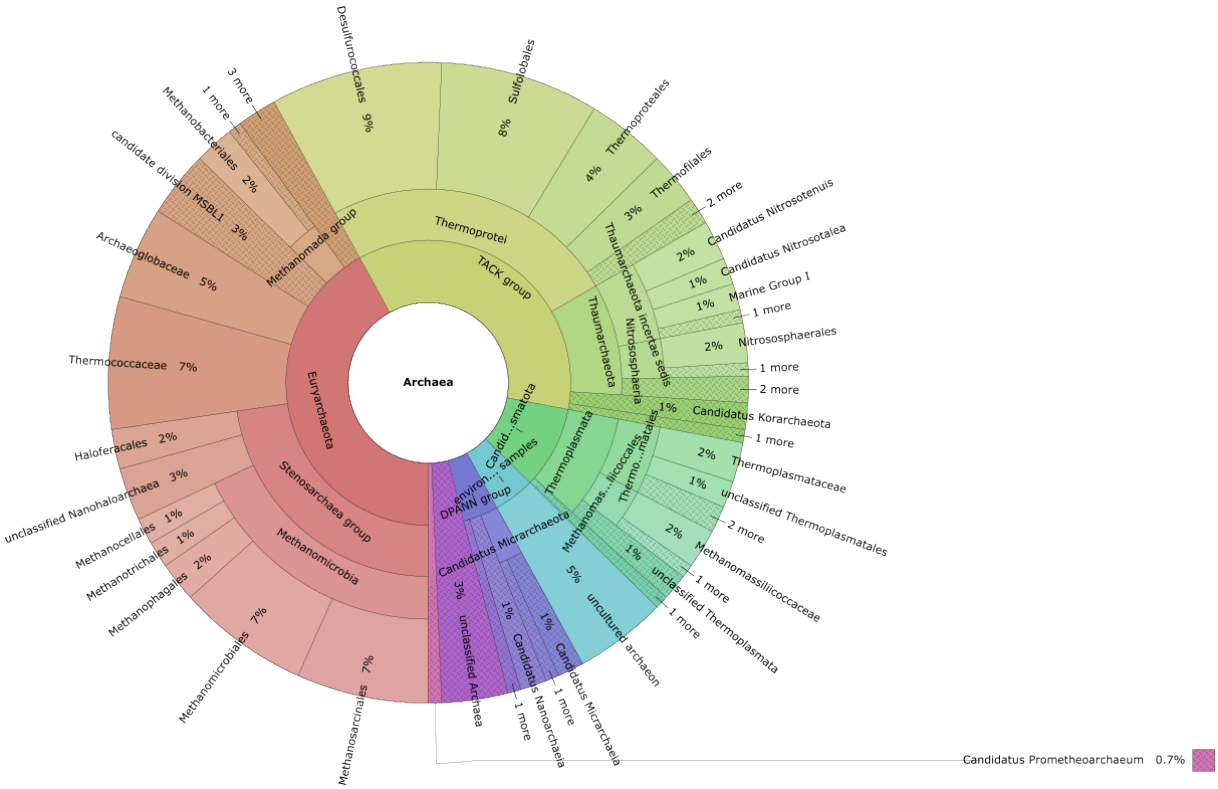


Figure 10 : Diagramme Krona illustrant la diversité taxonomique des 150 génomes archéens *sélectionnés par ToRQuEMaDA.*

En ce qui concerne les *sets* de génomes d'Halobacteria, de Methanococci, de Methanomicrobia et de Sulfolobales sélectionnés par ToRQuEMaDA, ceux-ci comportent respectivement 185, 37, 107 et 45 génomes (**Annexe 7**). Les génomes d'Halobacteria de sélectionnés par ToRQuEMaDA compte 38% d'Halobacteriales, 38% d' Haloferacales, et 24% de Natrialbaceae. Ceux de Methanococci se répartissent dans trois groupes : les Methanococcaceae (65%), les Methanocaldococcaceae (30%) et les Methanofervidicoccus (5%). Les génomes de Methanomicrobia se subdivisent en Methanosarcinales (57%), Methanomicrobiales (31%), Methanophagales (7%), Methanocella (3%) et Methanothrix (3%). Enfin, 97% des génomes de Sulfolobales sont des Sulfolobaceae. Ces derniers se répartissent entre plusieurs genres : *Sulfolobus* (31%), *Acidianus* (21%), *Metallosphaera* (17%), *Saccharolobus* (14%),

Sulfuracidifex (5%), *Sulfurisphaera* (5%), *Stygiolobus* (5%) et *Sulfodiicoccus* (2%).

4.2 Recherche de SLPs archéennes

4.2.1 Par BLAST

La première tâche de ce projet était de dresser un inventaire des SLPs décrites dans la littérature. De cette façon, 90 protéines ont été trouvées. Après une déréplication à 65% d'identité, ces 90 protéines initiales ont été réparties dans 49 *clusters*, chacun correspondant à un type de SLP distinct au niveau de leur séquence en acides aminés. La majorité de ces 49 *clusters* ne contient qu'une ou deux séquences.

Des séquences homologues aux représentants de ces 49 *clusters* ont été recherchées parmi les archées (`BLASTp`). Cette manipulation a, d'une part, permis d'enrichir les *clusters* et, d'autre part, d'étudier la distribution taxonomique des différents types de SLPs. L'interface graphique de `ompa-pa.pl` montre que peu de séquences homologues ont été trouvées pour certaines SLPs. Par exemple, seules deux séquences homologues à la SLP de *Picrophilus torridus* (NCBI RefSeq : WP_011177134) et une homologue à la SLP de *Thermococcus stetteri* (NCBI RefSeq : WP_209476963) ont été trouvées (**Figure 11**). De plus, aucun type de SLPs n'est représenté dans plusieurs groupes taxonomiques. Par exemple, les séquences homologues à la SLP de *Methanosarcina* acetivorans (UniProt : Q8TSG7) restent cantonnées au groupe des Methanomicrobia (**Figure 12**).

Parmi les 49 *clusters*, 16 ont été laissés de côté car ils ne contenaient que peu de *hits*, et donc peu de séquences protéiques à étudier. Dans les 33 *clusters* restants, 10 ne comportaient que des Halobacteria, 6 des Methanococci, 11 des Methanomicrobia et 6 des Sulfolobales. Les SLPs de ces quatre groupes d'archées ont donc été étudiées plus en profondeur.

4.2.2 Par profils HMM

Des profils HMM des 33 *clusters* sélectionnés ont été construits. Des séquences homologues à ces derniers ont été recherchées dans la base de données d'archées (`db_Archaea_proteomes`) et dans les bases de données propres au groupe taxonomique de chaque profil (`db_Halobacteria_proteomes`, `db_Methanococci_proteomes`, `db_Methanomicrobia_proteomes`, `db_Sulfolobales_proteomes`). Le passage par `ompa-pa.pl` montre, encore une fois, que la distribution taxonomique de chaque SLPs est restreinte à un seul groupe taxonomique. A titre d'exemple, les séquences homologues à la SLP d'*Haloarcula hispanica* (UniProt : G0HV86) décrite dans la littérature appartiennent à la classe des Halobacteria (**Figure 13**). En outre, les graphiques `ompa-pa.pl` de certains *clusters* ont la même allure, ce qui témoigne de redondance entre eux. Par exemple, les graphiques des SLPs d'*Halobacterium salinarum* (UniProt : B0R8E4) et *Haloarcula californiae* (NCBI RefSeq : WP_007188624) se ressemblent fortement (**Figure 14**). Les *clusters* concernés par ces chevauchements ont été fusionnés.

Ainsi, sept nouveaux *clusters* ont été obtenus : `cluster09-23` (49 protéines), `cluster12-19` (222 protéines), `cluster15-17` (252 protéines), `cluster32-29` (71 protéines), `cluster40-46` (30 protéines), `cluster43-44` (64 protéines) et `cluster45-47` (33 protéines).

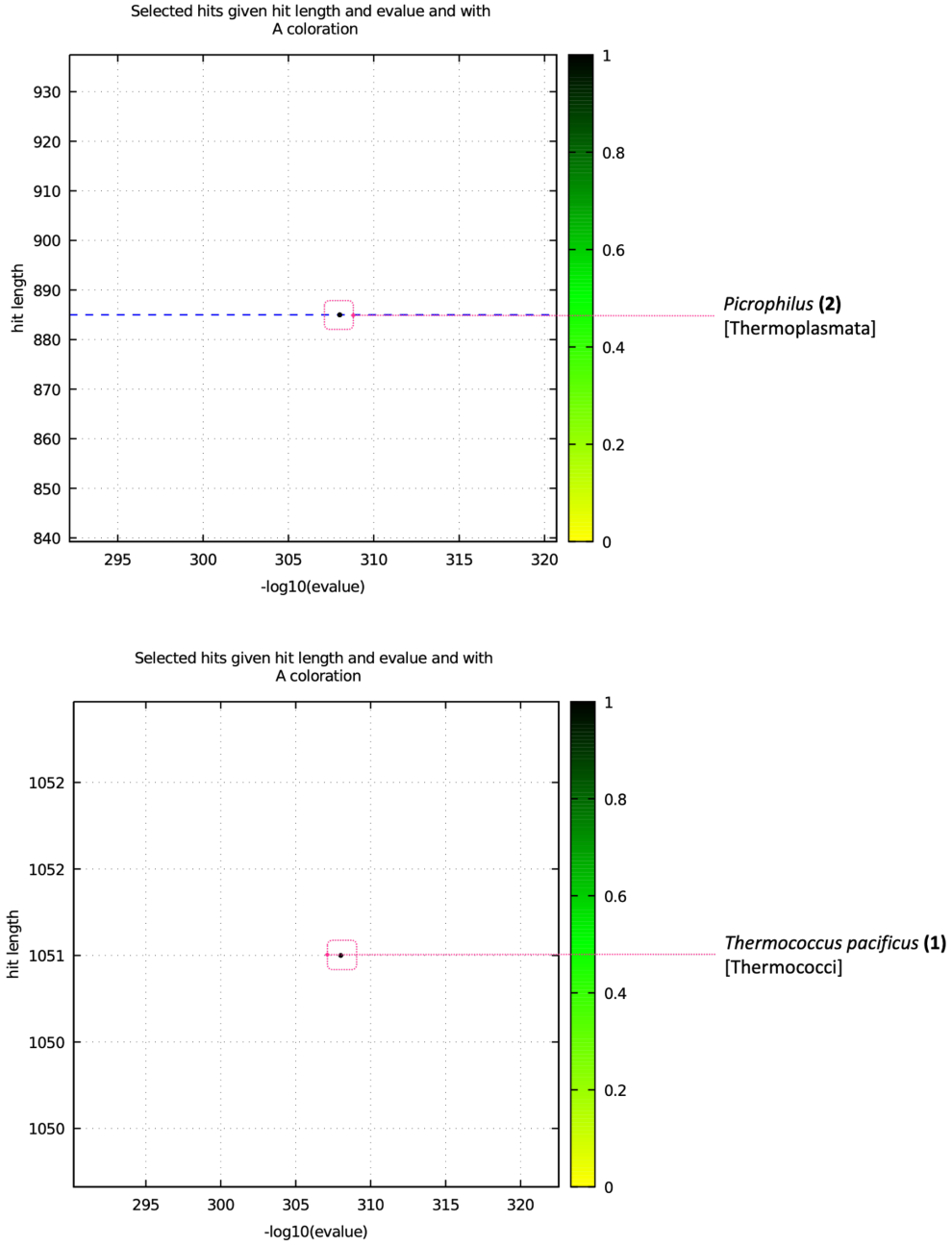


Figure 11 : Graphiques ompa-pa.pl *représentant les séquences homologues (représentées par des points) aux SLPs de Picrophilus torridus (en haut) et de Thermococcus stetteri (en bas) trouvées à l'aide de BLASTp.*

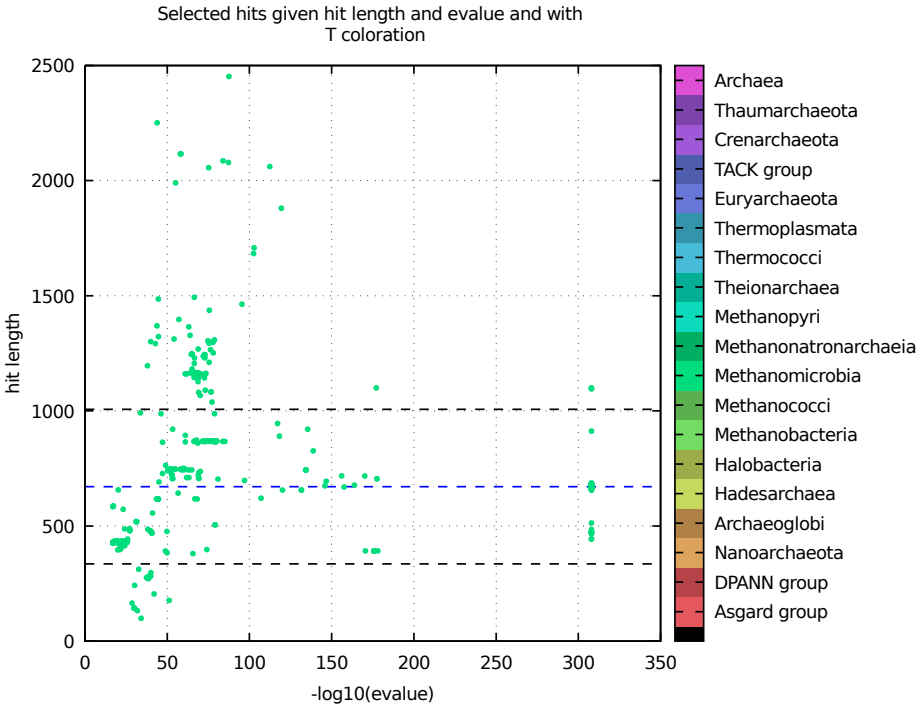


Figure 12 : Graphiques ompa-pa.pl *représentant les séquences homologues (représentées par des points) à la SLP de Methanosarcina acetivorans (UniProt : Q8TSG7) trouvées à l'aide de BLASTp.*

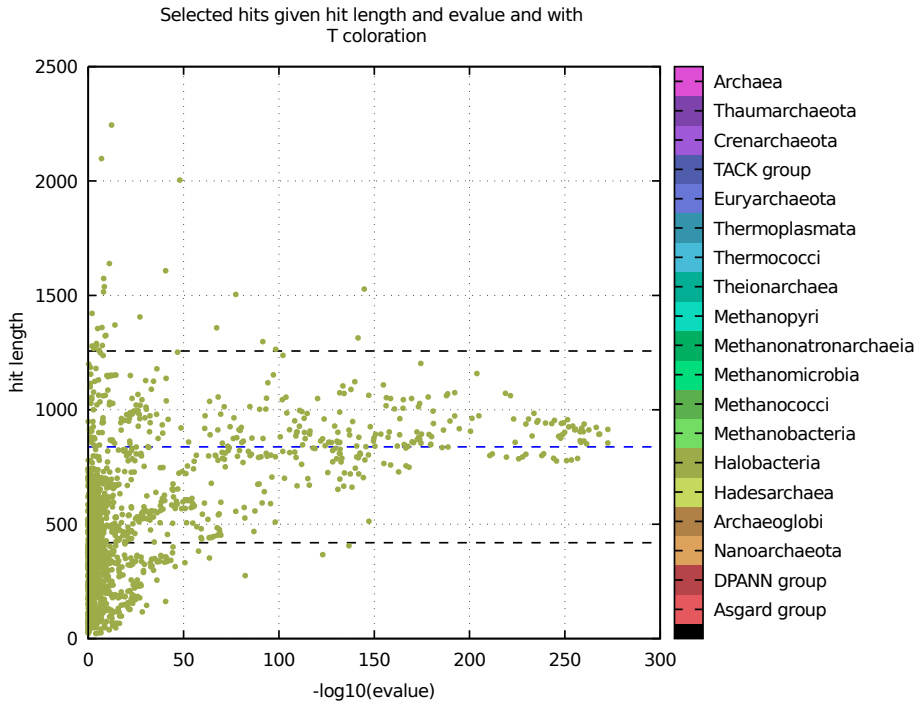


Figure 13 : Graphiques ompa-pa.pl *représentant les séquences homologues (représentées par des points) à la SLP d'Haloarcula hispanica (UniProt : G0HV86) trouvées à l'aide de hmmsearch.*

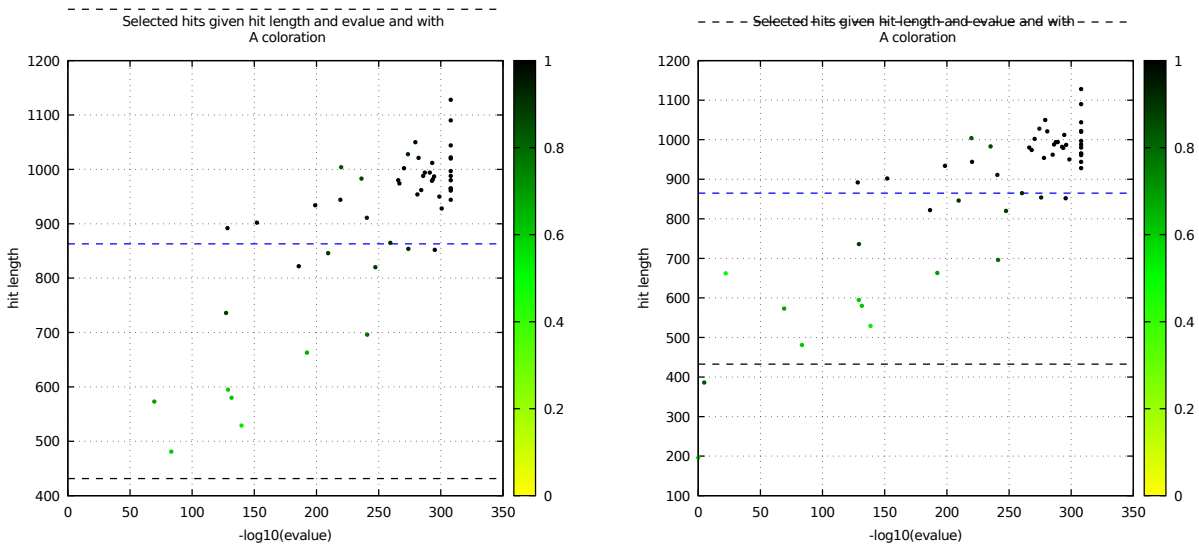


Figure 14 : Graphiques ompa-pa.pl *représentant les séquences homologues (représentées par des points) aux SLPs de Halobacterium salinarum (à gauche) et de Haloarcula californiae (à droite) trouvées à l'aide de hmmsearch.*

4.3 Annotations de séquences protéiques

Les protéines représentantes des 49 *clusters* ont été utilisées comme protéines de référence pour `annotate.pl` afin d'annoter les séquences protéiques des sept nouveaux *clusters*. Les séquences protéiques possédant un pourcentage d'identité d'au moins 50% et une *e-value* inférieure à 10^{-10} par rapport à une séquence de référence se sont vue attribuées un *tag* correspondant à l'annotation de cette dernière.

L'annotation des séquences a mené à l'ajout de six *tags* différents sur 16 des 49 séquences protéiques du `cluster09-23` ; cinq *tags* sur 19 des 222 séquences du `cluster12-19` ; quatre *tags* sur 117 des 252 séquences du `cluster15-17`, sept *tags* sur 36 des 71 séquences du `cluster32-29` ; trois *tags* sur 12 des 30 protéines du `cluster40-46` ; deux *tags* sur 22 des 64 séquences du `cluster43-44` ; et deux *tags* sur 12 des 33 séquences du `cluster45-47`.

4.4 Arbres phylogénétiques

Dans un premier temps, des arbres phylogénétiques ont été construits à partir des séquences protéiques annotées des sept *clusters* issus de fusions.

Dans l'arbre relatif au `cluster40-46`, les *Acidianus* se retrouvent à deux endroits différents. Presque toutes les séquences du premier groupe ont été annotées (ce groupe a d'ailleurs été utilisé pour enraciner l'arbre) alors que celles du second groupe n'ont reçu aucun *tag*. Toutes les séquences du genre *Metallosphaera* se trouve dans un groupe monophylétique et ont été annotées avec le même *tag*. Enfin,

seule une séquence de *Sulfolobus* a été annotée. Il est également intéressant de constater qu'aucun génome n'est représenté plusieurs fois (**Figure 15**).

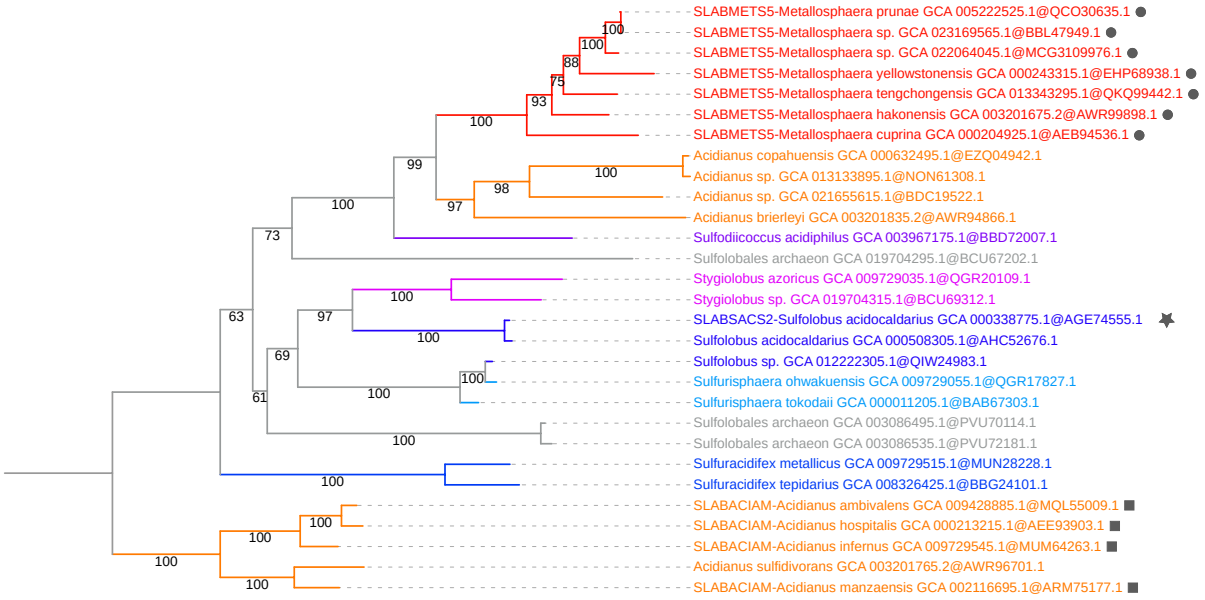


Figure 15 : Arbre phylogénétique final des protéines annotées du cluster40-46 *(29 seqs x 472 sites ; ML LG4X). Chaque couleur représente un genre différent et chaque symbole correspond à une annotation distincte. L'enracinement a été fait sur le groupe d'Acidianus annotés.*

En ce qui concerne l'arbre du `cluster40-46`, les *Haloplanus* se retrouvent à divers endroits. Le groupe monophylétique de 12 protéines issues de génomes d'*Haloplanus* a été utilisé pour enraciner l'arbre. Les protéines issues de génomes présents en double et en triple correspondent à des duplications récentes ; tandis que les protéines provenant d'organismes représentés à quatre ou cinq reprises dans l'arbre sont répartis dans deux groupes monophylétiques (**Figure 16**).

Le `cluster15-17`, le `cluster43-44` et le `cluster45-47` ne contiennent que des Methanomicrobia. Cependant, on observe que le `cluster15-17` ne comporte que des Methanosarcinales et des Methanotrichales, alors que seuls des Methanomicrobiales sont présents dans le `cluster43-44` et le `cluster45-47` (**Annexe 8**). L'arbre du `cluster32-29` n'a pas été discuté, mais il visible dans l'**annexe 8**.

Les arbres phylogénétiques basés sur les protéines ribosomiques des *sets* de génomes d'Archaea, d'Halobacteria, de Methanococci, de Methanomicrobia et de Sulfolobales sélectionnés par `ToRQuEMaDA` sont représentés en **annexe 9**. Il sont utiles pour comparer la phylogénie des SLPs.

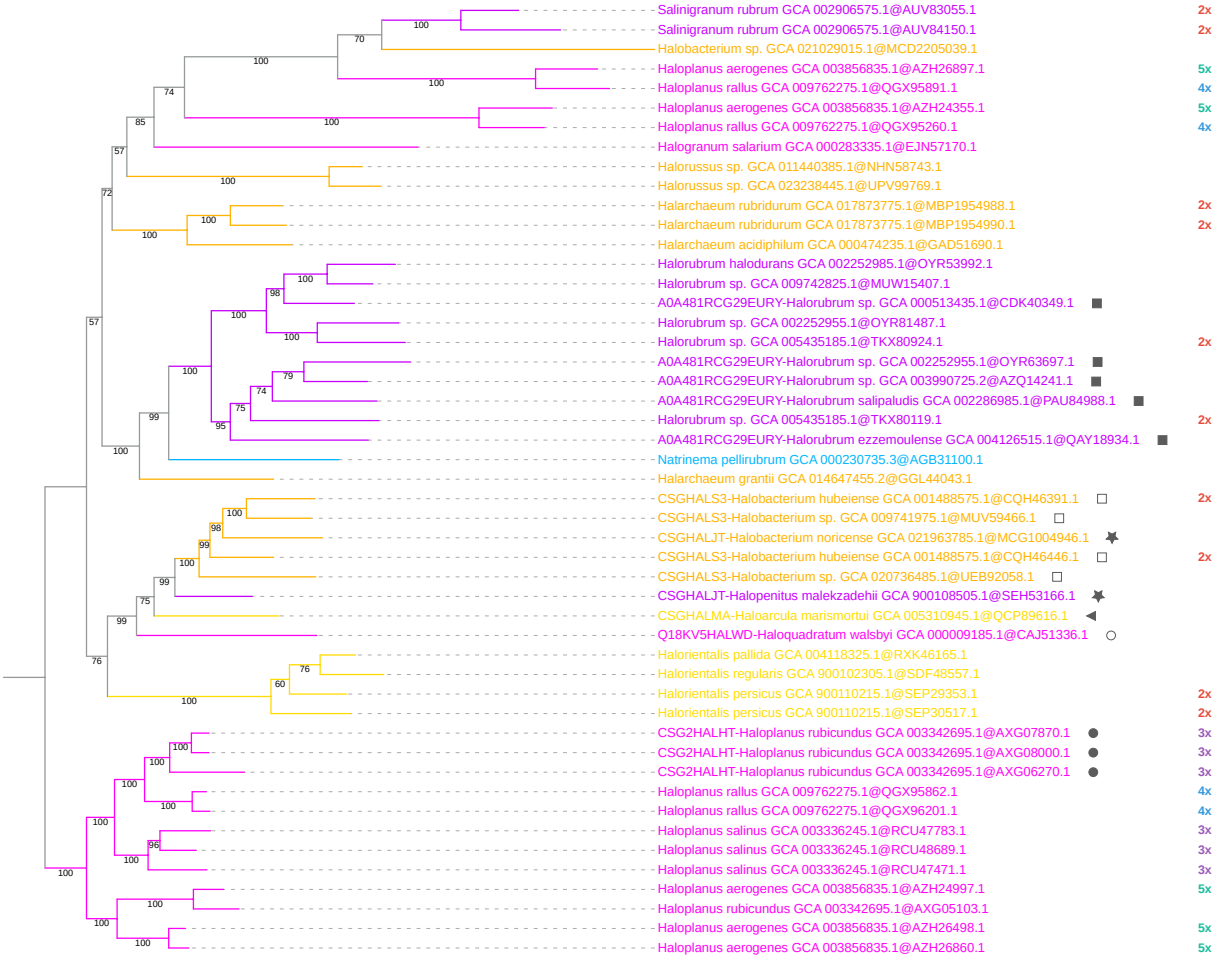


Figure 16 : Arbre phylogénétique final des protéines annotées du cluster09-23 *(49 seqs x 1019 sites ; ML LG4X). Chaque couleur représente une famille différente, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur le groupe des Haloplanus contenant 12 feuilles.*

5 Discussion

5.1 Abondances relatives et diversité des organismes

Les proportions relatives des différents groupes taxonomiques représentés dans la banque de données de `ToRQuEMaDA` se rapportent aux nombre de génomes archéens ayant été étudiés et non à l'abondance de ces organismes dans l'environnement. Par conséquent, il y a un biais en faveur des groupes d'archées fortement étudiés.

L'utilisation de banques de données dérépliquées et représentatives de la diversité des Archaea, des Halobacteria, des Methanococci, des Methanomicrobia et des Sulfolobales, ce biais a été réduit. En effet, les génomes trop similaires ont été rassemblés au sein du même groupe, et seuls les génomes représentants de chaque groupe ont été sélectionnés par `ToRQuEMaDA`. Par conséquent, le fonctionnement même de par `ToRQuEMaDA` explique les différences d'abondances relatives entre les organismes archéens présents dans la banque de données du programme (**Figure 9**) et ceux qu'il a sélectionné (**Figure 10**).

Finalement, on constate que la diversité des Archaea et des quatre groupes sélectionnés est conservée dans les génomes choisis par `ToRQuEMaDA`. En effet, des représentants de toutes les familles et de tous les ordres sont retrouvés. Ainsi, l'utilisation ne `ToRQuEMaDA` permet de dérépliquer et de sélectionner des séquences génomiques, sans porter atteinte à la diversité des organismes.

5.2 Diversité et homologie de séquence des SLPs

La déréplication à 65% d'identité des 90 séquences de SLPs initiales avec `cdhit-clustering.pl` a généré 49 *clusters* distincts, chacun correspondant potentiellement à un type de SLP différent. Parmi ces *clusters*, 30 ne comportent qu'une seule séquence. Cette première analyse suggère que les séquences protéiques des SLPs sont très diverses. Malgré cette forte diversité de séquences, aucune SLPs représentée dans plusieurs groupes taxonomiques à la fois n'a été trouvée. Elles restent toutes cantonnées à un groupe (ex. Halobacteria, Methanomicrobia). En outre, le fait que certains *clusters* se chevauchent amène à s'interroger sur cette diversité.

Un constat commun aux 49 *clusters* enrichis par des `BLASTp` et aux 33 recherches par profils HMM réalisées ensuite peut être fait : certains clusters restent pauvres en séquences. Ce qui signifie que ces SLPs ne présentent qu'une faible homologie de séquence. En résumé, le fait que peu de séquences homologues aient été trouvées pour certains types de SLPs témoignent de la spécificité de ces dernières.

5.3 Arbres phylogénétiques

Les arbres phylogénétiques réalisés à partir des séquences protéiques des sept *clusters* issus de fusions

montrent des duplications récentes et plus anciennes des gènes encodant les SLPs. Egalement, les arbres ribosomiques et les arbres de SLPs ne sont pas superposables, ce qui signifie que les protéines ribosomiques et les SLPs n'ont pas évolués en parallèle, ni de la même manière.

Enfin, les arbres de SLPs mettent en évidence les distributions taxonomiques limitées de certaines protéines. Par exemple, les arbres relatifs aux `cluster15–17`, `cluster43–44` et `cluster45–47` ne contiennent que des SLPs issue de Methanomicrobia, mais ne représentent que certaines familles. Cette dernière observation confirme bien la distribution taxonomique restreinte de SLPs..

6 Conclusion

Au terme de ce mémoire, il a été démontré que la plupart des SLPs ne présentent que peu, voire pas, de similarité de séquence, même si elles sont exprimées par des organismes étroitement apparentés. Les SLPs archéennes sont donc très diverses au niveau de leur séquence en acides aminés. Cependant, leur distribution taxonomique limitée est peut-être en partie liée à une divergence importante de leur séquence primaire, plutôt qu'à une réelle hétérogénéité.

Cette dernière hypothèses pourrait être testée au travers de l'identification de SLPs additionnelles à intégrer aux arbres phylogénétiques déjà construits.

Références bibliographiques

1 Albers, S.-V. and Meyer, B.H. (2011) The archaeal cell envelope. *Nat Rev Microbiol* 9, 414–426

2 Cavicchioli, R. (2010) Archaea — timeline of the third domain. *Nature Publishing Group* 9,

3 Noller, H. (2013) Carl Woese (1928–2012). *Nature* 493, 610–610

4 Woese, C.R. (1996) Phylogenetic trees: Whither microbiology? *Current Biology* 6, 1060–1063

5 Dörr, T. *et al.* (2019) Editorial: Bacterial Cell Wall Structure and Dynamics. *Front Microbiol* 10, 2051

6 Klingl, A. *et al.* (2019) Archaeal Cell Walls. pp. 471–493, Springer, Cham

7 Chaban, B. *et al.* (2006) Archaeal habitats--from the extreme to the ordinary. *Can J Microbiol* 52, 73–116

8 Suchodolski, J.S. (2013) Gastrointestinal Microbiota. *Canine and Feline Gastroenterology* DOI: 10.1016/B978-1-4160-3661-6.00002-X

9 Rodrigues-Oliveira, T. *et al.* (2017) Archaeal S-Layers: Overview and Current State of the Art. *Front Microbiol* 8, 2597

10 Bharat, T.A.M. *et al.* (2021) Molecular Logic of Prokaryotic Surface Layer Structures. *Trends Microbiol* 29, 405–415

11 Pum, D. *et al.* (2021) Patterns in Nature—S-Layer Lattices of Bacterial and Archaeal Cells. *Crystals (Basel)* 11, 869

12 Sleytr, U.B. *et al.* (2014) S-layers: principles and applications. *FEMS Microbiol Rev* 38, 823–864

13 Pohlschroder, M. *et al.* (2018) Archaeal cell surface biogenesis. *FEMS Microbiol Rev* 027, 694–717

14 Abdul-Halim, M.F. *et al.* (2020) Lipid Anchoring of Archaeosortase Substrates and Midcell Growth in Haloarchaea. *mBio* 11,

15 Pum, D. *et al.* (2013) S-layer protein self-assembly. *Int J Mol Sci* 14, 2484–501

16 Pum, D. and Sleytr, U.B. (2014) Reassembly of S-layer proteins. *Nanotechnology* 25, 312001

17 Lu, H. *et al.* (2015) Identification of the S-layer glycoproteins and their covalently linked glycans in the halophilic archaeon *Haloarcula hispanica*. *Glycobiology* 25, 1150–1162

18 Gongadze, G.M. *et al.* (1993) *Regular Proteinaceous Layers of Thermococcus stetteri Cell Envelope*, 27

19 Gambelli, L. *et al.* (2019) Architecture and modular assembly of *Sulfolobus* S-layers revealed by electron cryotomography. *Proceedings of the National Academy of Sciences* 116, 25278–25286

20 Peters, J. *et al.* (1996) Hyperthermostable surface layer protein tetrabrachion from the archaebacterium Staphylothermus marinus: evidence for the presence of a right-handed coiled coil derived from the primary structure. *J Mol Biol* 257, 1031–41

21 Kans, J. *Entrez Direct: E-utilities on the Unix Command Line*,

22 Li, W. *et al.* (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18, 77–82

23 Léonard, R.R. *et al.* (2021) ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* 9, e11348

24 Gurevich, A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–5

25 Lagesen, K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35, 3100–8

26 van Vlierberghe, M. *et al.* (2021) Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project. *BMC Res Notes* 14, 306

27 Katoh, K. and Toh, H. (2007) PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* 23, 372–4

28 Finn, R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39, W29-37

29 Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30, 772–80

30 Roure, B. *et al.* (2007) SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evol Biol* 7, S2

31 Nguyen, L.-T. *et al.* (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32, 268–74

32 Ondov, B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC*

Bioinformatics 12, 385

Annexe 1 : Script Perl du programme cdhit-clustering.pl

```
1   #!/usr/bin/env perl
2   # PODNAME: cdhit-clustering.pl
3
4   use Modern::Perl '2011';
5   use autodie;
6   use Getopt::Euclid qw(:vars); # qw(:vars) pour notation $ARGV_argument
7   use Smart::Comments;
8   use File::Basename;
9   use Path::Class 'file';
10  use Bio::MUST::Core;
11  use Bio::MUST::Drivers;
12  use aliased 'Bio::MUST::Drivers::CdHit';
13  use aliased 'Bio::MUST::Core::SeqId';
14
15  ### Running cd-hit
16  my $report = CdHit->new( seqs => $ARGV_in , cdhit_args => {-c => $ARGV_identity} );
17
18  ### Setup outfiles
19  my ($basename, $dir, $suffix) = fileparse($ARGV_in, qr{\.[^.]*}xms);
20  my $identity = $ARGV_identity * 100;
21  $dir = $dir . $basename . '-cdhit' . $identity . '/';
22  mkdir $dir;
23
24  ### Parsing report
25  my $i = 0;
26
27  # For each cluster…
28  for my $repr ( $report->all_cluster_names ){
29
30  # Create an outfile for this cluster and open it
31  my $cluster_id = repr_short_id($repr);
32  my $outfile = file($dir, 'cluster' . sprintf ("%02d", $i) . '-' . $cluster_id .
    $suffix);
33
34  open my $out, '>', $outfile;
35
36  # Retrieve the sequence of the representative
37  my $seq1 = $report->get_seq_with_id($repr)->seq;
38
39  # Write the representative's ID and its sequence in the outfile
40  say {$out} '>' . $repr;
41  say {$out} join q{}, $seq1;
42
43  for my $member ( map { $_->full_id } @{ $report->seq_ids_for($repr) } ) {
44
45  # Retrieve the sequence of each members of the cluster
46  my $seq2 = $report->get_seq_with_id($member)->seq;
```

```
47
48      # Write the members ID and their sequences in the outifle
49      say {$out} '>' . $member;
50      say {$out} join q{}, $seq2;
51      }
52
53      close $out;
54      $i++;
55      }
56
57      sub repr_short_id {
58      my $repr = shift;
59      my $id = $repr;
60
61      # ID from Swissprot
62      ($id) = $repr =~ m/\Asp\|([A-Z0-9]{6,})\.[0-9]{1}\|.*/xms
63      if $repr =~ m/\Asp\|[A-Z0-9]{6,}\.[0-9]{1}\|.*/xms;
64
65      # ID from RefSeq
66      ($id) = $repr =~ m/\A([A-Z]{2}_[0-9]{9})\.[0-9]{1}.*/xms
67      if $repr =~ m/\A[A-Z]{2}_[0-9]{9}\.[0-9]{1}.*/xms  ;
68
69      # ID from GenBank
70      ($id) = $repr =~ m/\A([A-Z0-9]{8})\.[0-9]{1}.*/xms
71      if $repr =~ m/\A[A-Z0-9]{8}\.[0-9]{1}.*/xms;
72
73      return $id;
74      }
75
76      __END__
77      =pod
78
79      =head1 NAME
80      cdhit-clustering.pl -
81
82      =head1 VERSION
83      version
84
85      =head1 USAGE
86      cdhit-clustering.pl --in=<infile> --identity=<identity>
87
88      =head1 REQUIRED ARGUMENTS
89
90      =over
91
92      =item --in [=] <infile>
93      Path to input FASTA file.
94      =for Euclid:
95        infile.type: readable
96
```

```
=item --identity [=] <identity>
Sequence identity threshold.
=for Euclid:
  identity.type: number, identity > 0 && identity <= 1

=back

=head1 OPTIONAL ARGUMENTS

=over

=item --version

=item --usage

=item --help

=item --man
Print the usual program information

=back

=cut
```

Annexe 2 : Sripts Perl du pipeline de ToRQuEMaDA

- **Premier script : `tqmd_download.pl`**

```
tqmd_download.pl \
  --tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
  --coll-name=20220512_genbank \
  --config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.
ini \
  --setup-taxdir \
  --filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Archaea.idl \
  --source=genbank
```

- **Deuxième script : `tqmd_update.pl`**

```
# Qualité des génomes (QUAST)
tqmd_update.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Prep_quast_archaea_genbank_20220512 \
--calc=quast \
2> prep_quast_archaea_genbank_20220512

# Richesse de l'annotation
tqmd_update.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Prep_annotation_archaea_genbank_20220512 \
--calc=annotation \
2> prep_annotation_archaea_genbank_20220512

# Niveau de contamination des génomes (Forty-Two)
perl /media/vol2/scratch/tqmd/tqmd_files/test_versions/bug_fixe_FT/bin/tqmd_upda
te.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--config-42=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config_FT.ini \
--temp-dir=Prep_FT_archaea_genbank_20220512 \
--calc=fortytwo \
--pack-size=10000 \
--max-array=10 \
--threads=4 \
2> prep_FT_archaea_genbank_20220512

# Prédiction des ARNr 16S (RNAmmer)
tqmd_update.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Prep_rnammer_archaea_genbank_20220512 \
```

```
--calc=rnammer \
2> prep_rnammer_archaea_genbank_20220512

# Déréplication des ARNr 16S (CD-HIT)
tqmd_update.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Prep_cdhit_archaea_genbank_20220512 \
--calc=cdhit \
--cdhit-threshold=0.975
2> prep_cdhit_archaea_genbank_20220512
```

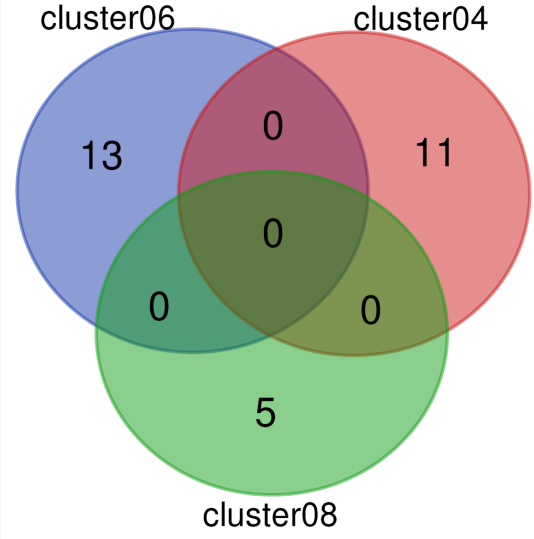
- **Troisième script : `tqmd_cluster.pl`**

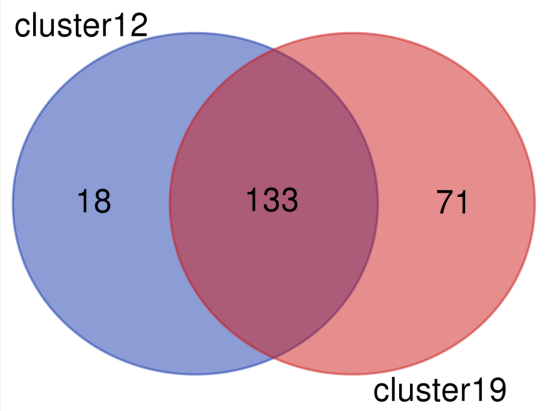
```
# Sélection de génomes représentatifs d'archées
tqmd_cluster.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Archaea_genbank_mash18_kmer14_SSU_20220727 \
--filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Archaea.idl \
--source=genbank \
--kmer-engine=mash \
--dist-metric=JI \
--kmer-size=14 \
--dist-threshold=0.18 \
--negative-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_Archaea_genbank/excluded_GCA_archaea_genbank_20220512.txt \
--priority-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_SLPs_genbank/GCA_list_Asgard.txt \
--kmer-canonical \
--requires-SSU-rRNA \
--ranking-formula='-quast.N.per.100.kbp, +quast.largest.contig.ratio, -42.contam.perc, +42.added.ali' \
--dividing-scheme=taxonomic \
--clustering-mode=strict \
--min-round=1 \
--max-round=10 \
--max-array=100 \
2> archaea_genbank_mash18_kmer14_SSU_20220727

# Sélection de génomes représentatifs d'Halobacteria
tqmd_cluster.pl \
--tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
--config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
--temp-dir=Halobacteria_genbank_mash07_kmer14_20220721 \
--filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Halobacteria.idl \
--source=genbank \
--kmer-engine=mash \
--dist-metric=JI \
```

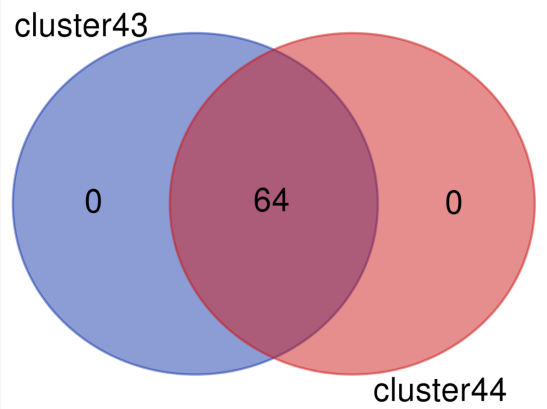
```
    --kmer-size=14 \
    --dist-threshold=0.07 \
    --negative-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_Archaea_genbank/excluded_GCA_archaea_genbank_20220512.txt \
    --kmer-canonical \
    --requires-SSU-rRNA \
    --ranking-formula='-quast.N.per.100.kbp, +quast.largest.contig.ratio, -42.contam.perc, +42.added.ali' \
    --dividing-scheme=taxonomic \
    --clustering-mode=strict \
    --min-round=1 \
    --max-round=10 \
    --max-array=100 \
    2> halobacteria_genbank_mash07_kmer14_20220721

    # Sélection de génomes représentatifs de Methanococci
    tqmd_cluster.pl \
    --tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
    --config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
    --temp-dir=Methanococci_genbank_mash01_kmer14_20220721 \
    --filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Methanococci.idl \
    --source=genbank \
    --kmer-engine=mash \
    --dist-metric=JI \
    --kmer-size=14 \
    --dist-threshold=0.01 \
    --negative-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_Archaea_genbank/excluded_GCA_archaea_genbank_20220512.txt \
    --kmer-canonical \
    --requires-SSU-rRNA \
    --ranking-formula='-quast.N.per.100.kbp, +quast.largest.contig.ratio, -42.contam.perc, +42.added.ali' \
    --dividing-scheme=taxonomic \
    --clustering-mode=strict \
    --min-round=1 \
    --max-round=10 \
    --max-array=100 \
    2> methanococci_genbank_mash01_kmer14_20220721

    # Sélection de génomes représentatifs de Methanomicrobia
    tqmd_cluster.pl \
    --tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
    --config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
    --temp-dir=Methanomicrobia_genbank_mash01_kmer14_20220721 \
    --filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Methanomicrobia.idl \
    --source=genbank \
    --kmer-engine=mash \
    --dist-metric=JI \
    --kmer-size=14 \
```

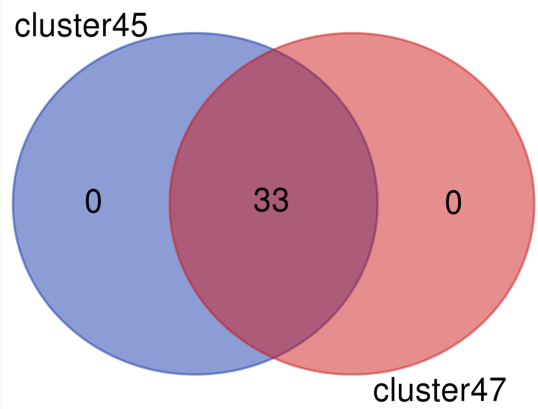
```
78  --dist-threshold=0.01 \
79  --negative-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_Archaea_genbank/exc
    luded_GCA_archaea_genbank_20220512.txt \
80  --kmer-canonical \
81  --requires-SSU-rRNA \
82  --ranking-formula='-quast.N.per.100.kbp, +quast.largest.contig.ratio,
    -42.contam.perc, +42.added.ali' \
83  --dividing-scheme=taxonomic \
84  --clustering-mode=strict \
85  --min-round=1 \
86  --max-round=10 \
87  --max-array=100 \
88  2> methanomicrobia_genbank_mash01_kmer14_20220721
89
90  # Sélection de génomes représentatifs de Sulfolobales
91  tqmd_cluster.pl \
92  --tqmd-dir=/media/vol2/scratch/tqmd/tqmd_prokaryotes/ \
93  --config=/media/vol2/scratch/tqmd/tqmd_files/configuration_files/tqmd_config.ini \
94  --temp-dir=Sulfolobales_genbank_mash01_kmer14_20220721 \
95  --filter=/media/vol2/scratch/tqmd/tqmd_files/filter_files/Sulfolobales.idl \
96  --source=genbank \
97  --kmer-engine=mash \
98  --dist-metric=JI \
99  --kmer-size=14 \
100 --dist-threshold=0.01 \
101 --negative-gca-list=/media/vol2/home/croomans/thesis_Master2/TQMD_Archaea_genbank/exc
    luded_GCA_archaea_genbank_20220512.txt \
102 --kmer-canonical \
103 --requires-SSU-rRNA \
104 --ranking-formula='-quast.N.per.100.kbp, +quast.largest.contig.ratio,
    -42.contam.perc, +42.added.ali' \
105 --dividing-scheme=taxonomic \
106 --clustering-mode=strict \
107 --min-round=1 \
108 --max-round=10 \
109 --max-array=100 \
110 2> sulfolobales_genbank_mash01_kmer14_20220721
```

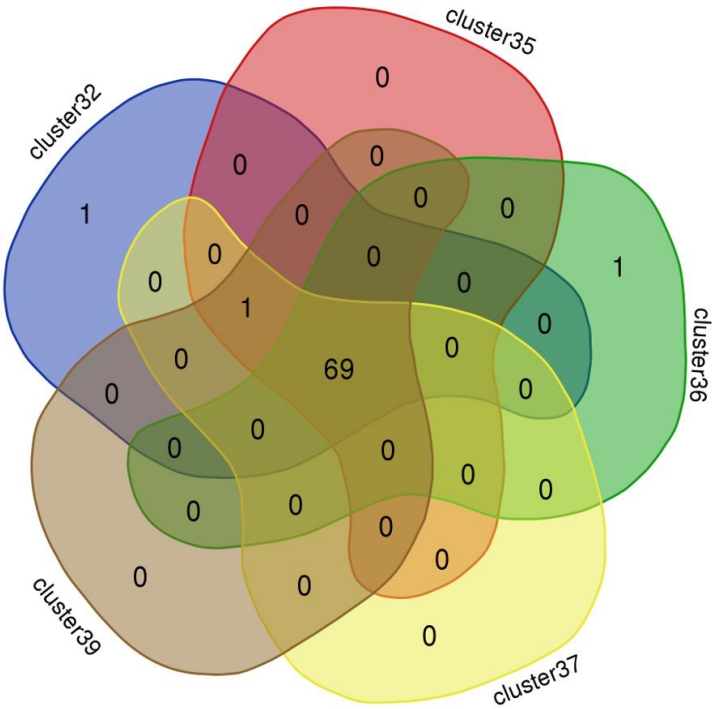
Annexe 3 : Diagrammes de Venn représentant les chevauchements entre les différents *clusters*

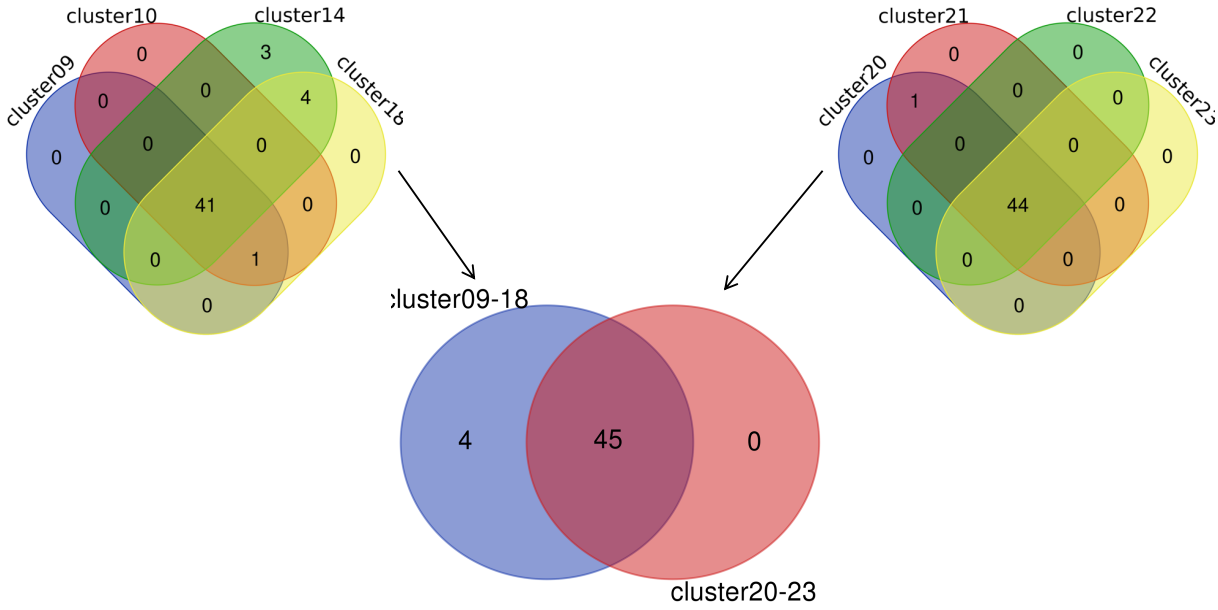









cluster35
cluster32
cluster36
cluster39
cluster37
69


cluster10
cluster14
cluster09
41
cluster21
cluster22
cluster20
44
cluster09-18
4
45
0
cluster20-23

Annexe 4 : Script Bash du programme forty-two.sh

```
#!/bin/bash

name=$(echo "$1" | tr A-Z a-z)
NAME=$(echo ${name^})

# Lien vers les protéomes
cp -s ~/thesis_Master2/TQMD_SLPs_genbank/proteomes/${name}_proteomes/proteomes/*_protein.faa .

# Bank-mapper
## Création fichier IDM
ls *.faa | perl -nle '($gcf) = m/(GC[AF]\_\d{9}\.\d{1}_.*)\.faa/ ; print "$gcf"' > liste-faa.idl
cut -f1-2 -d"_" liste-faa.idl > liste-gca.idl
fetch-tax.pl liste-gca.idl --taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ --item-type=taxid --missing=MISSING
cut -f2 liste-gca.tax > liste-gca-name.idl
paste liste-gca-name.idl liste-faa.idl > bank-mapper.idm

## Création une DB BLAST à partir des protéomes récupérés
for REFORG in *.faa
do
  makeblastdb -in $REFORG -dbtype prot -out `basename $REFORG .faa` -parse_seqids
done

# Génération des fichiers de configuration
mkdir ../forty-two && cd $_

yaml-generator-42.pl \
--run_mode=metagenomic \
--out_suffix=42_$NAME \
--queries /media/vol2/home/mvanvlierberghe/databases/ribo_prots/prokaryotes/queries.idl \
--evalue=1e-05 \
--homologues_seg=yes \
--max_target_seqs=10000 \
--templates_seg=no \
--bank_dir /media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}/proteomes \
--bank_suffix=.psq \
--bank_mapper /media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}/proteomes/bank-mapper.idm \
--code=1 \
--ref_brh=on \
--ref_bank_dir /media/vol2/home/mvanvlierberghe/databases/ref_banks/prokaryotes \
--ref_bank_suffix=.psq \
--ref_bank_mapper
```

```
/media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_Archaea/proka_ref_bank_mapper.idm \
--ref_org_mul=0.3 \
--ref_score_mul=0.99 \
--trim_homologues=off \
--merge_orthologues=off \
--aligner_mode=off \
--ali_keep_old_new_tags=off \
--ali_keep_lengthened_seqs=keep \
--tax_reports=on \
--taxdir=/media/vol2/home/croomans/thesis_Master2/taxdump-20220517/ \
--megan_like \
--tol_check=off

# Lancement de 42
## Création des jobs
for ALI in /media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_Archaea/ribo_prots_Prokaryotes/*.ali ; do echo "forty-two.pl $ALI --config=config-42_${NAME}.yaml --verbosity=6"; done > ${name}_42_species-tree.cmds

for FTLIST in ${name}_42_species-tree.cmds ; do tpage --define T=`wc -l $FTLIST | cut -f1 -d" "` --define TC=100 --define list=$FTLIST /media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_Archaea/jobarray-42-envvars.tt > $FTLIST-array.sh; done

## Lancement des jobs
qsub ${name}_42_species-tree.cmds-array.sh
```

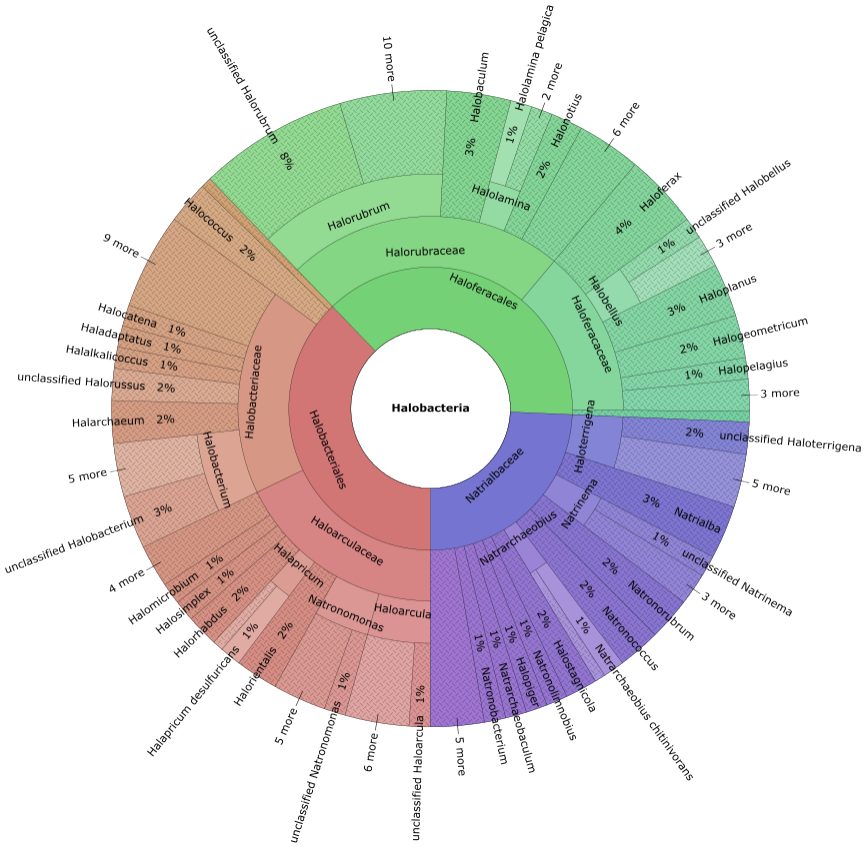
Annexe 5 : Script Bash du programme scafos_1-2.sh

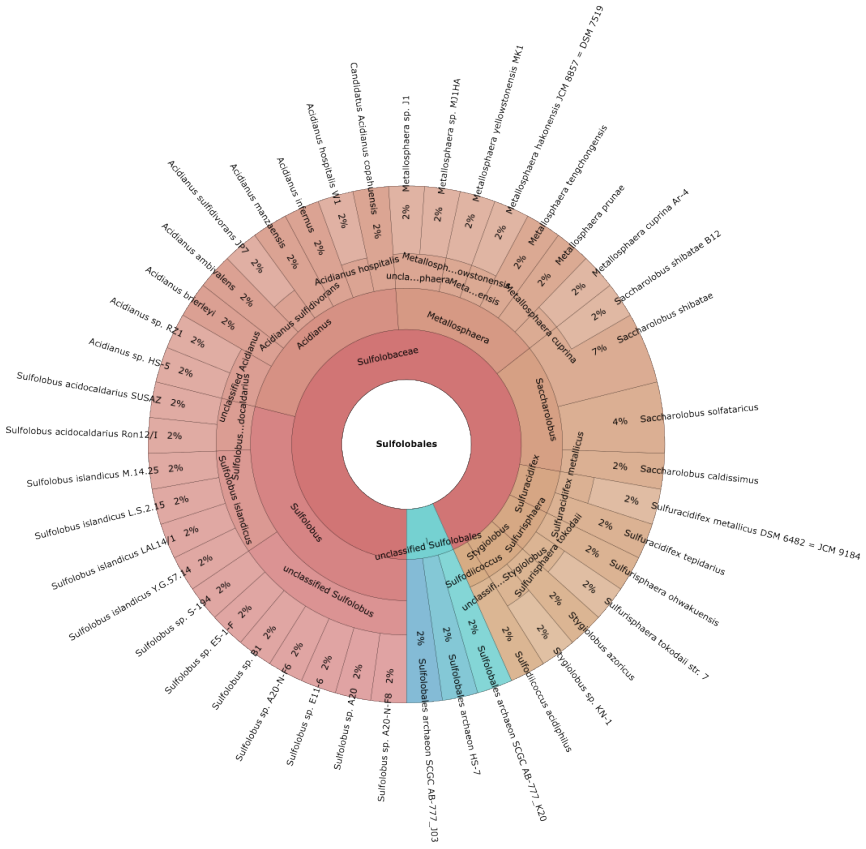
```
#!/bin/bash

name=$(echo "$1" | tr A-Z a-z)
NAME=$(echo ${name^})

# Formatage des fichiers sortants de 42
cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}/forty-two/

mkdir log
rm -f ${name}_42_species-tree.cmds-array.sh.o*
mv ${name}_42_species-tree.cmds-array.sh.* log/

mkdir tax-report && cd $_
mv
/media/vol2/home/croomans/thesis_Master2/phylo_ribosomes_Archaea/ribo_prots_Prokaryotes/
*_${NAME}.tax-report .

mkdir ../../ribo_prots_Ali

for FILE  in *_${NAME}.tax-report ; do perl -F"\t" -anle 'next if /^#/; print q{>} .
$F[0] . qq{\n} . $F[12]' $FILE > $FILE.ali; done

mv *.ali ../../ribo_prots_Ali && cd $_

## Retrait de la double extension : .tax-report.ali -> .ali
perl ~/thesis_Master2/phylo_ribosomes_Archaea/rename_tax-report_ali.pl *

## Enlève les lignes vides des fichiers .ali
perl -i -nle 's/^>$// ; print' *.ali

ali2fasta.pl *.ali

mkdir ../ribo_prots_Fasta
mv *.fasta ../ribo_prots_Fasta

# Scafos_1
cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}
scafos.pl in=ribo_prots_Fasta/ out=scafos1/

grep -v "#" scafos1/scafos1-freq.otu | sed 's/ (.*//' > ${NAME}.otu

# Scafos_2
scafos.pl in=ribo_prots_Fasta/ out=scafos2/ otu=${NAME}.otu
```

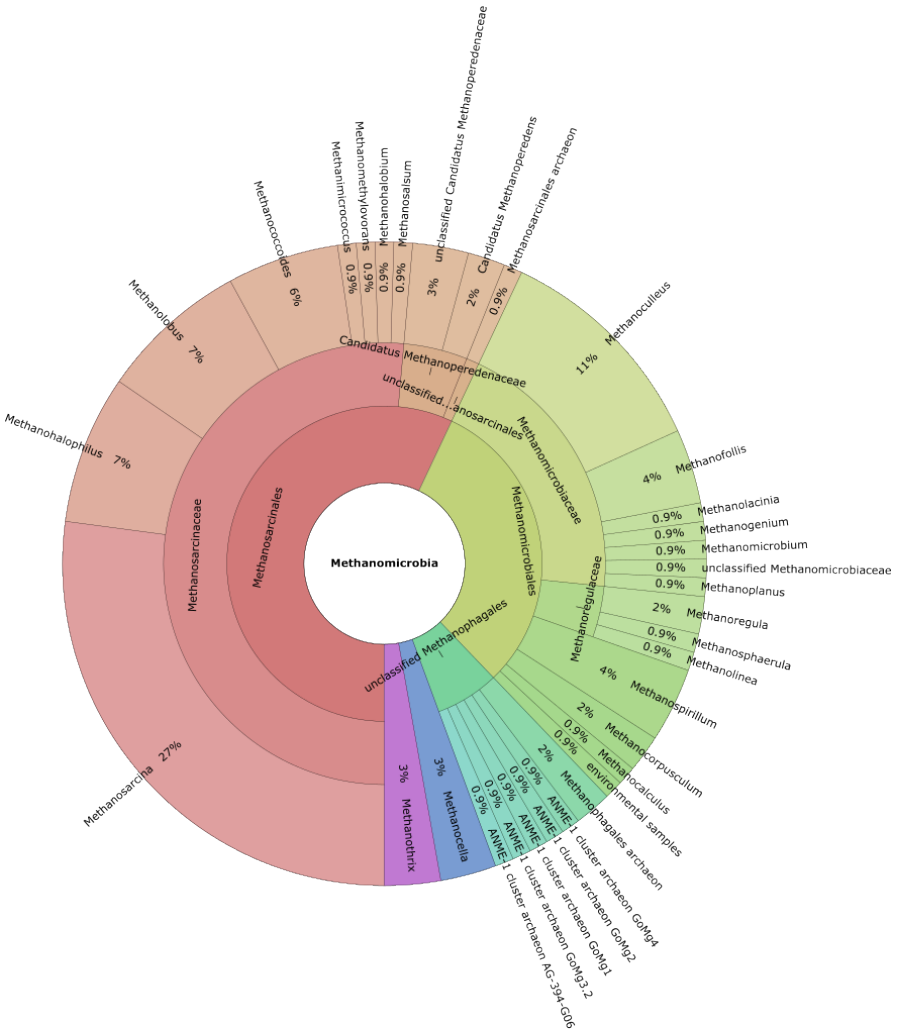
Annexe 6 : Script Bash du programme scafos_3.sh

```
#!/bin/bash

name=$(echo "$1" | tr A-Z a-z)
NAME=$(echo ${name^})

cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}/mafft/

echo "DB processes: $NAME"

# Suppression des gènes bactériens (contaminations) si fichiers de petite taille
rm -f b*

# Suppression des alignements vides
find . -type f -empty -delete

# Ali2phylip
for FILE in *.fasta ; do ali2phylip.pl --bmge-mask=loose --ali --min=0.3 --max=0.5
$FILE; done

mkdir ../ali2phylip
mv *ali ../ali2phylip/ && cd ../ali2phylip

sed -i 's/\#//g' *.ali                # retire les lignes commançant par #
sed -i "/^[ \t]*$/d" *.ali            # retire les lignes vides

# Remplacer les espaces (entre genre et espèce) par des _
perl -i -nle 'if (m/^>.*/) {s/ /_/} ; print' *ali

# Scafos 3
cd ~/thesis_Master2/phylo_ribosomes_DB_SLPs/${NAME}/

scafos.pl in=ali2phylip/ out=scafos3/ otu=${NAME}.otu format=fp g=25

cd scafos3
mv scafos3.fasta scafos3.ali
```

Annexe 7 : Diagrammes Krona illustrant la diversité taxonomique des génomes archéens sélectionnés par ToRQuEMaDA

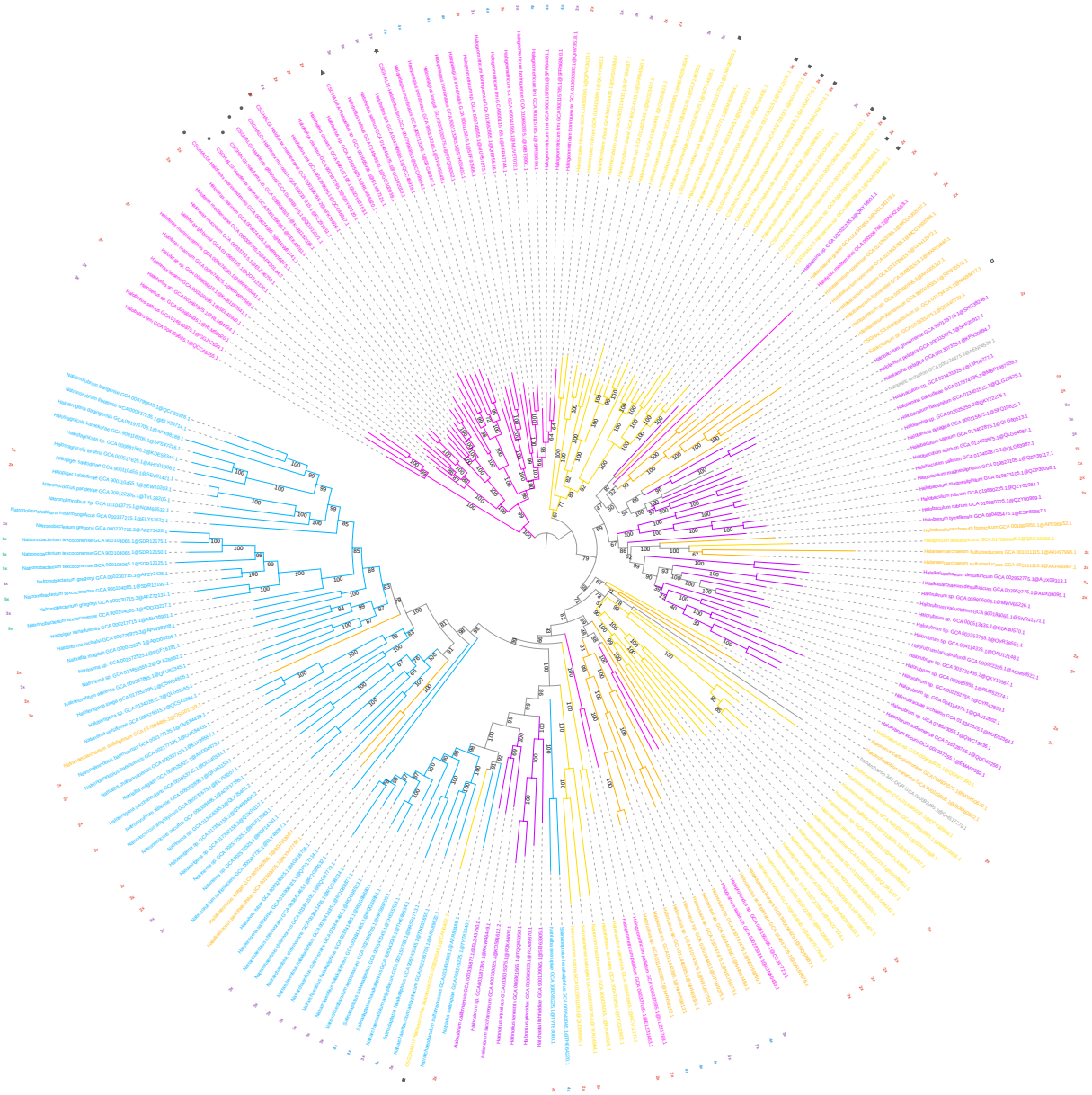




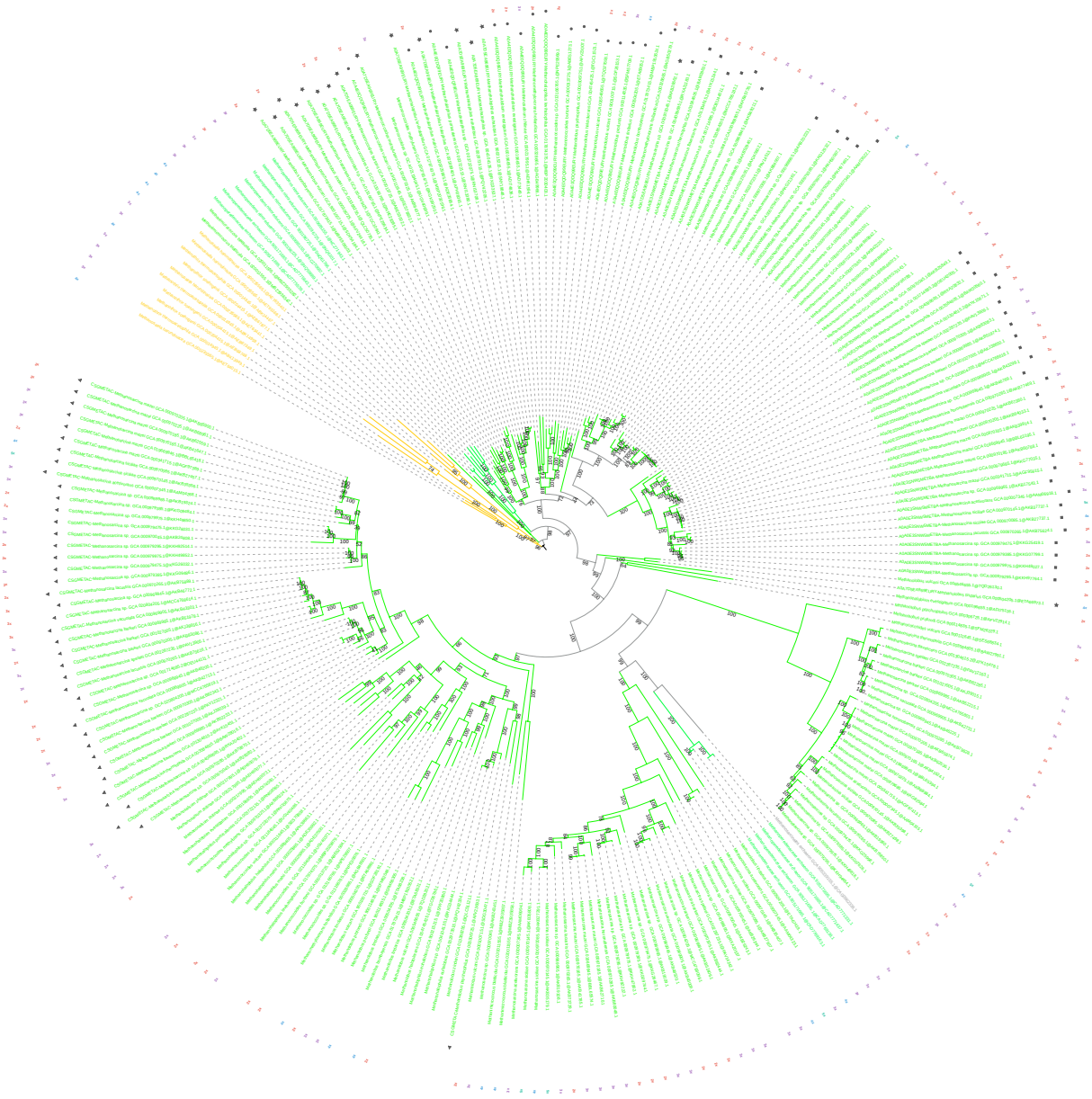

Methanomicrobia
Methanosarcinales
Methanosarcinaceae
Methanosarcina 27%
Methanothrix 3%
Methanophagales
unclassified
Methanocella 3%
ANME-1 cluster archaeon AG-394-G06
ANME-1 cluster archaeon GoMg3.2
ANME-1 cluster archaeon GoMg1
ANME-1 cluster archaeon GoMg2
ANME-1 cluster archaeon GoMg4
Methanophagales archaeon
environmental samples
Methanomicrobiales
Methanomicrobiaceae
Methanoregulaceae
Methanoculleus 11%
Methanofollis 4%
Methanolacinia 0.9%
Methanogenium 0.9%
Methanomicrobium 0.9%
unclassified Methanomicrobiaceae 0.9%
Methanoplanus 0.9%
Methanoregula 2%
Methanosphaerula 0.9%
Methanolinea 0.9%
Methanospirillum 4%
Methanocorpusculum 2%
Methanocalculus 0.9%
unclassified...anosarcinales
Candidatus Methanoperedenaceae
Methanococcoides 6%
Methanolobus 7%
Methanohalophilus 7%
Methanimicrococcus 0.9%
Methanomethylovorans 0.9%
Methanohalobium 0.9%
Methanosalsum 0.9%
unclassified Candidatus Methanoperedenaceae 3%
Candidatus Methanoperedens 2%
Methanosarcinales archaeon 0.9%


Methanococcales
Methanococcaceae
Methanocaldococcaceae
Methanococcus 57%
Methanothermococcus 8%
Methanocaldococcus 24%
Methanotorris 5%
Methanofervidicoccus 5%

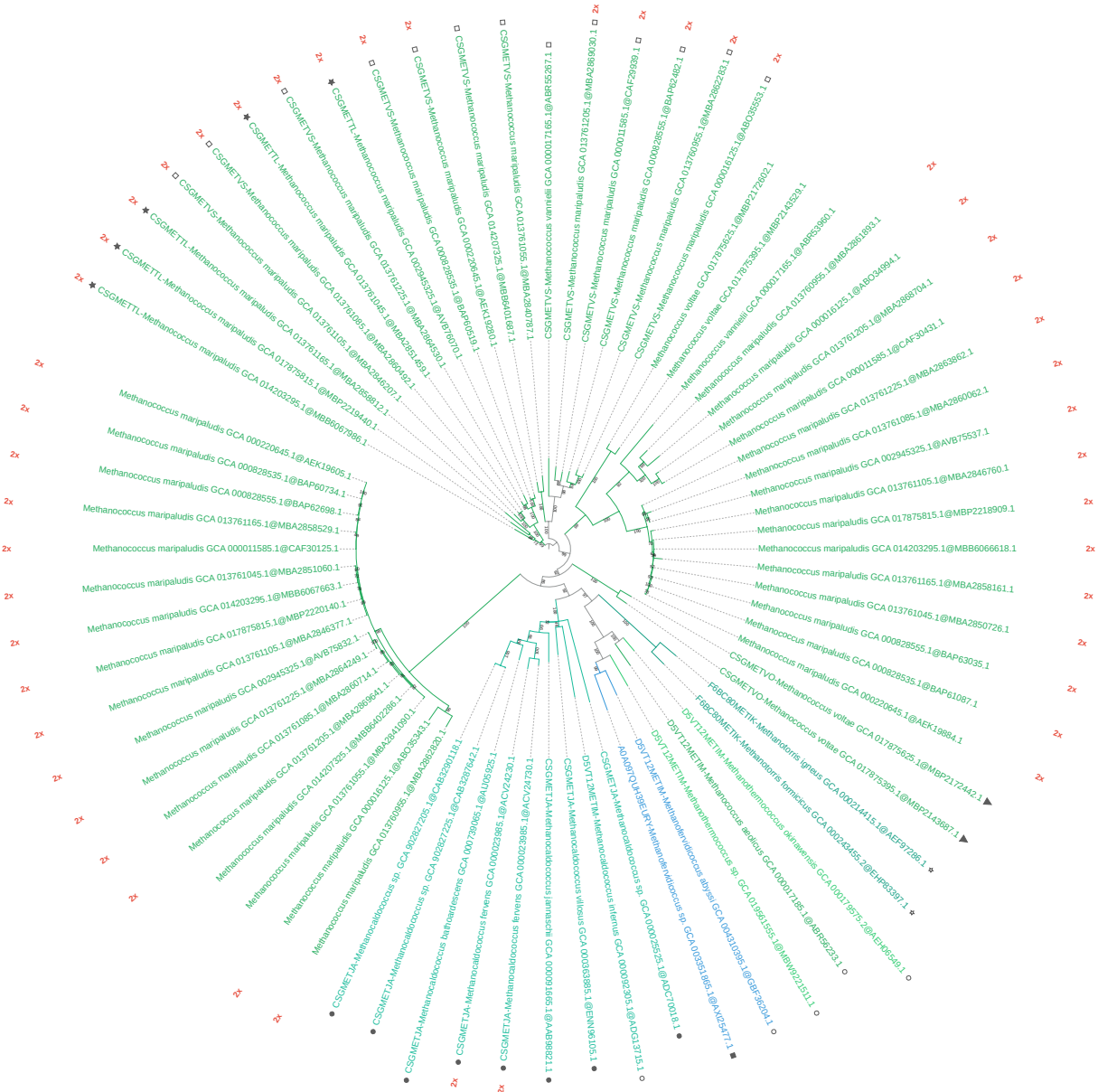
Annexe 8 : Arbres phylogénétiques des SLPs trouvées à l'aide des profils HMM



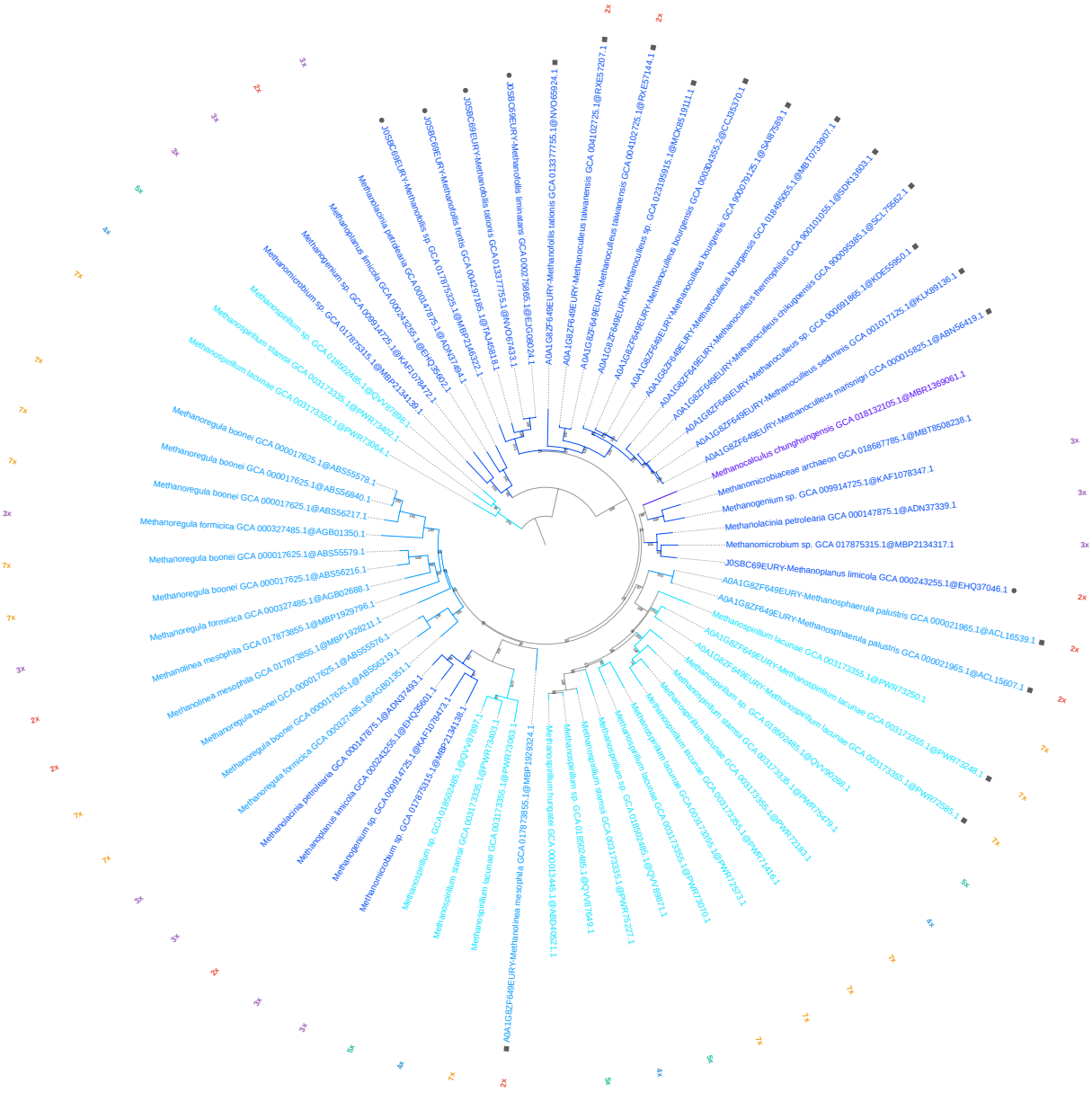
Annexe 8.1 : Arbre phylogénétique final des protéines annotées du cluster12-19 *(222 seqs x 950 sites ; ML LG4X). Chaque couleur représente une famille différente, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur le groupe des Haloferacaceae.*



Annexe 8.2 : Arbre phylogénétique final des protéines annotées du cluster15-17 *(252 seqs x 902 sites ; ML LG4X). Chaque couleur représente une famille différente, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur le groupe des Methanotrichaceae.*



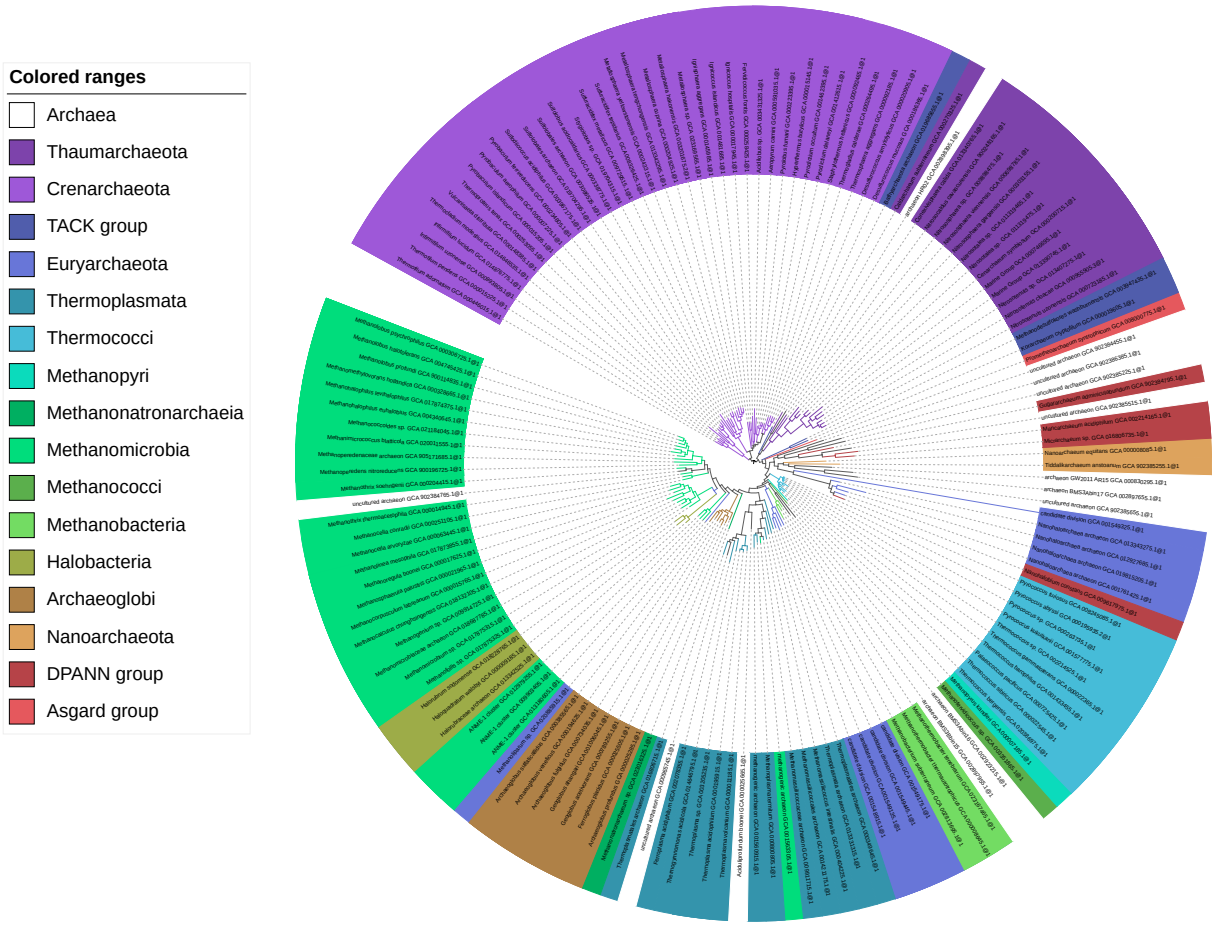
Annexe 8.3 : Arbre phylogénétique final des protéines annotées du cluster32-39 *(252 seqs x 902 sites ; ML LG4X). Chaque couleur représente un genre différent, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur un groupe de Methanococcus.*



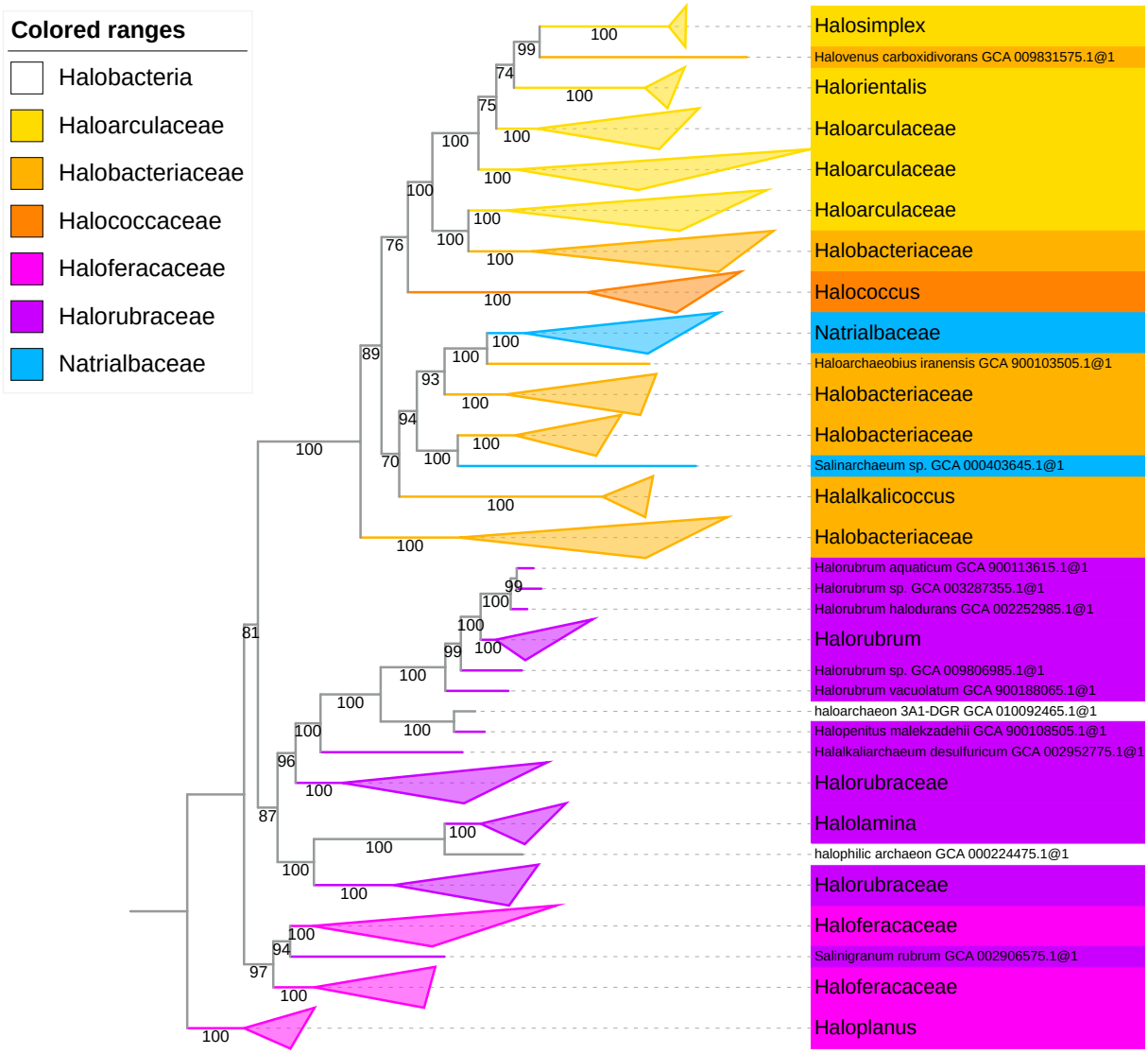
Annexe 8.4 : Arbre phylogénétique final des protéines annotées du cluster43-44 *(64 seqs x 437 sites ; ML LG4X). Chaque couleur représente une famille différente, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur un groupe de Methanospirillum de 3 feuilles.*

Annexe 8.4 : Arbre phylogénétique final des protéines annotées du cluster45-47 *(33 seqs x 477 sites ; ML LG4X). Chaque couleur représente une famille différente, chaque symbole correspond à une annotation distincte. Les indications à droite des branches indiquent les génomes redondants. L'enracinement a été fait sur un groupe de Methanospirillum de 3 feuilles.*

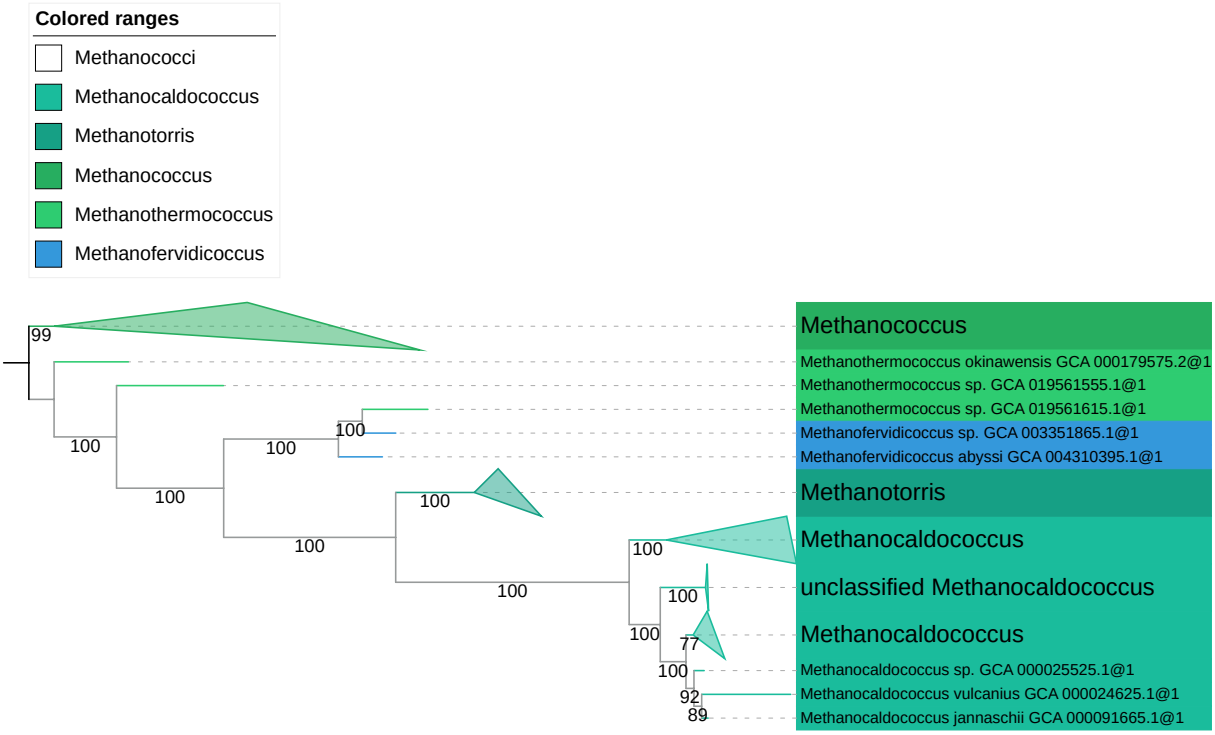
Annexe 9 : Arbres phylogénétiques des génomes sélectionnés par ToRQuEMaDA



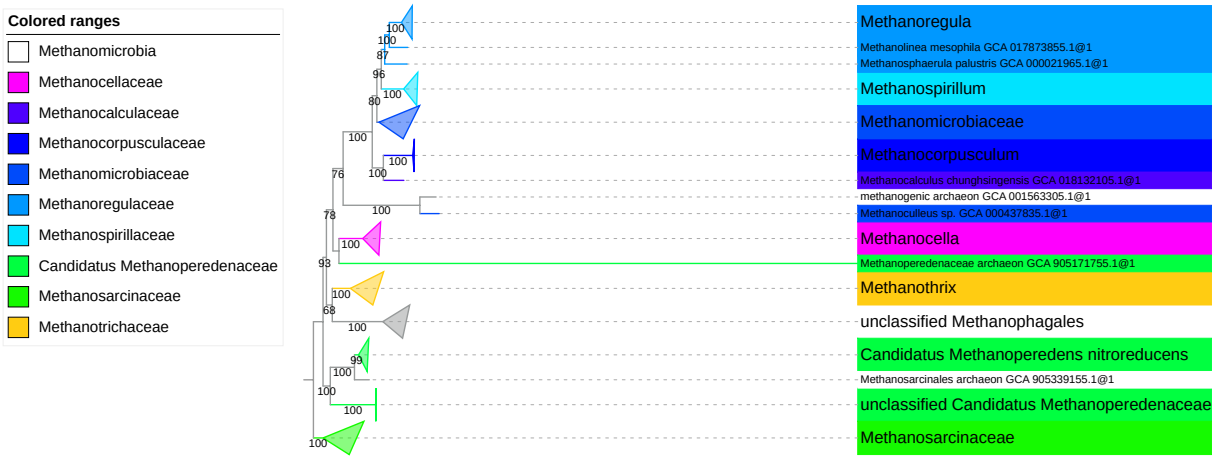
Annexe 9.1 : Arbre phylogénétique basé sur les protéines ribosomiques des séquences d'Archaea *sélectionnées par ToRQuEMaDA (150 seqs x 6956 sites ; ML LG4X). Chaque couleur représente un phylum différent. L'enracinement a été fait sur les Crenarchaeota.*



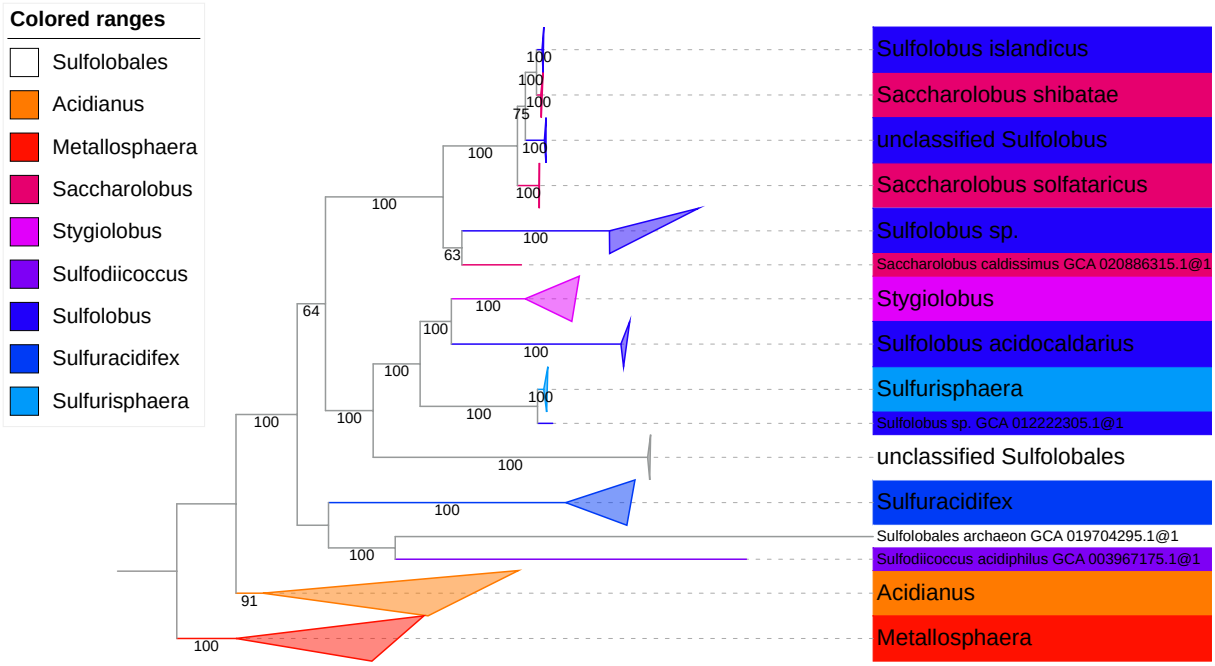
Annexe 9.2 : Arbre phylogénétique basé sur les protéines ribosomiques des séquences d'Halobacteria *sélectionnées par ToRQuEMaDA (185 seqs x 6300 sites ; ML LG4X). Chaque couleur représente une famille différente. L'enracinement a été fait sur le genre Haloplanus.*



***Annexe 9.3 : Arbre phylogénétique basé sur les protéines ribosomiques des séquences de Methanococci** sélectionnées par ToRQuEMaDA (37 seqs x 6495 sites ; ML LG4X). Chaque couleur représente un genre différent. L'enracinement a été fait sur le genre Methanococcus.*



***Annexe 9.4 : Arbre phylogénétique basé sur les protéines ribosomiques des séquences de Methanomicrobia** sélectionnées par ToRQuEMaDA (107 seqs x 6434 sites ; ML LG4X). Chaque couleur représente une famille différente. L'enracinement a été fait sur les Methanosarcinaceae.*



Annexe 9.5 : Arbre phylogénétique basé sur les protéines ribosomiques des séquences de Sulfolobales *sélectionnées par ToRQuEMaDA (45 seqs x 7779 sites ; ML LG4X). Chaque couleur représente un genre différent. L'enracinement a été fait sur les Metallosphaera.*