



Université de Rennes 1

Master 1 Bio-informatique

Rousseau Jérémy

**Annotation structurale de séquences protéiques issue
de métagénomés marins par recherche de profils
HMM, et évaluation d'une approche de prédiction
structurale par deep learning**

Museum National d'Histoire Naturelle

Laboratoire d'accueil : Institut de Systématique, Évolution, Biodiversité (ISYEB)

Équipe : Atelier de Bio-Informatique (ABI)

Responsables de stages :

- Lucie Bittner (Maître de conférence)
- Mathilde Carpentier (Maître de conférence)

2021-2022

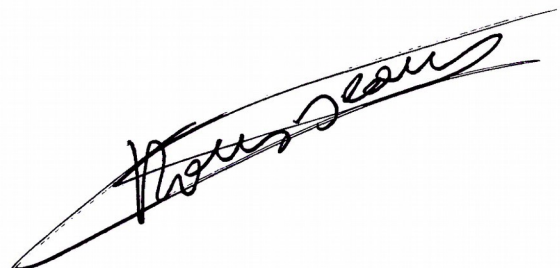
ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) ROUSSEAU Jérémy
Etudiant (e) en Master 1 de Bio-Informatique

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature



Remerciements

Je tiens tout d'abord à remercier vivement mes responsables de stage, Mme Mathilde Carpentier et Mme Lucie Bittner pour m'avoir permis d'effectuer ce stage au sein de leur équipe. Je les remercie également pour les conseils qu'elles m'ont apportées tout au long de mon stage ainsi que leurs aides dans la rédaction de mon rapport.

Je remercie également M. Nicolas Benoit et M. Romuald Bouffet de l'unité de service SACADO, pour m'avoir aidé dans la prise en main de leur cluster de calcul. Je les remercie également pour l'aide apportée dans l'installation d'AlphaFold mais aussi dans les moments difficiles lors de son utilisation.

Je tiens également à remercier M. Faycal Allouti, responsable technique du PCIA, pour m'avoir aidé à comprendre l'utilisation de son cluster.

Merci également Mme. Amandine Blin, ingénieure statisticienne au Pôle Analyse de Données, pour m'avoir conseillée sur les tests statistiques.

Merci aussi à M. Théo Jamay pour m'avoir aidé dans la compréhension d'*hmmsearch*.

Je remercie aussi ma mère, Ghislaine Rousseau, ainsi que M. Loïc Tello y Vazquez, pour avoir pris le temps de relire mon rapport.

Pour finir, je souhaite remercier Mme Martine Boccara, Mme Karen Salazar, M. Hugo Talibart, Mme Pauline Turk (étudiante en master de bio-informatique), M. Gaspard Rhim (étudiant en master BEE) ainsi que toute l'équipe de l'ABI pour leur accueil chaleureux au sein de cette équipe.

Sommaire

Remerciements.....	1
I. Introduction.....	4
II. Matériel et Méthodes.....	6
A. Jeu de données.....	6
B. Matériels : Machines et Clusters et langages de programmation.....	7
1. Machines et Clusters.....	7
2. Langages de programmation.....	8
C. Recherche des profils HMM avec le logiciel HMMER.....	8
1. Définition d'un profil HMM.....	8
2. HMMER - hmmsearch.....	8
3. Banque de données CATH / <i>Gene3D</i>	8
4. Présentation de l'analyses et des paramètres.....	9
D. Prédictiones structurales.....	9
1. CD-HIT et SPOT-Disorder-Single.....	10
2. <i>AlphaFold</i>	10
3. YAKUSA.....	11
E. Analyses statistiques et représentations graphiques.....	11
III. Résultats.....	11
A. Proportion des annotations & distribution des annotations CATH.....	11
B. Distribution de la taille des séquences annotées / non annotées.....	13
C. Distribution des annotations en fonction du <i>kingdom</i> - <i>phylum</i>	13
D. Prédictiones des structures.....	15
1. <i>AlphaFold</i>	15
2. Analyse des résultats de <i>AlphaFold</i>	15
IV. Discussion.....	16
V. Conclusion.....	18
VI. Bibliographie.....	20
Informations complémentaires.....	25
Ressources informatiques.....	25
Disponibilités des codes.....	25
Annexes.....	26

Annexe 1 : Présentation de la structure d'accueil.....	26
Annexe 2 : Bilan du stage.....	28
Annexe 3 : Configuration des machines et clusters.....	29
Annexe 4 : Données du test de χ^2	30
Annexe 5 : Test de corrélation de Pearson.....	31
Annexe 6 : Structures tridimensionnelles.....	33
Annexe 7 : AlphaFold – alignements multiples.....	34
Annexe 8 : AlphaFold – prédiction structurale.....	35
Annexe 9 : Définitions.....	36
Résumé.....	37
Abstract.....	38

I. Introduction

Les micro-organismes sont invisibles à l'œil nu (*i.e.*, généralement < 0.1 mm) et sont très abondants dans les océans, où leur biomasse est estimée entre 10^4 à 10^6 cellules par litre (Sunagawa *et al.*, 2015). Ils interviennent dans de très nombreux cycles biogéochimiques et impactent notre quotidien, *e.g.*, ils produisent notamment près de 50% de l'oxygène que nous respirons et ils contribuent à la fixation de carbone atmosphérique, dont une partie est exportée vers les grands fonds (Falkowski *et al.*, 1998; Field *et al.*, 1998). Dans les océans, de très nombreux micro-organismes sont planctoniques, *i.e.*, ils dérivent passivement au gré des courants, et ils correspondent à des archées, des bactéries, des virus ou encore des eucaryotes unicellulaires.

Nos connaissances sur les différentes espèces de micro-organismes marins ainsi que leurs fonctions dans les écosystèmes restent limitées car la majorité des lignées reste non cultivables, et ceci est d'autant plus vrai pour les lignées d'eucaryotes unicellulaires (Sibbald & Archibald 2017; Carradec *et al.* 2018). Cette dernière décennie, avec l'utilisation grandissante du séquençage moléculaire à haut-débit, de nombreux projets ont contribué à explorer les communautés de micro-organismes planctoniques, et notamment eucaryotes, révélant ainsi leur grande diversité environnementale (Bittner *et al.*, 2018). L'expédition Tara Océans 2009-2013 a permis notamment d'échantillonner 210 stations provenant de 20 régions géographiques différentes, et de se pencher sur les lignées eucaryotes (Pesant *et al.*, 2015). À partir des échantillons prélevés aux différentes stations et à différentes profondeurs, *i.e.* en surface et dans la colonne d'eau au niveau de la zone la plus riche en chlorophylle, les scientifiques ont extrait puis séquencé l'ADN et l'ARN des communautés eucaryotes (Alberti *et al.*, 2017). Ainsi Carradec *et al.* (2018), à partir des séquences courtes ou *reads* obtenues à partir du séquençage des ARNs (métatranscriptomique) ont reconstruit avec des algorithmes d'assemblage *de novo* plus de 116 millions d'unigènes¹ créant ainsi un catalogue de transcrits eucaryotes. À partir des *reads* d'ADN obtenus pour ces mêmes échantillons, Delmont *et al.* (2022) ont reconstruit 683 "espèces métagénomiques" (*Metagenome-Assembled Genome* MAGs) à partir du clustering des reads sur leurs profils d'abondance, puis de l'assemblage *de novo* effectué sur ces mêmes clusters de *reads*. Les MAGs correspondent à des communautés de micro-organismes qui co-abondent et sont interprétés le plus souvent comme des équivalents d'espèces (d'où le terme d'espèces métagénomiques) ou d'holobiontes (hôte et symbionte(s)). Pour cette même étude (Delmont *et al.*, 2022), des cellules d'eucaryotes

lgènes ou morceaux de gène, car ils sont parfois plus longs ou plus courts qu'un gène, un gène étant défini ici par une séquence incluse entre un codon *start* et un codon *stop*

unicellulaires ont été isolées puis séquencées de manière individuelle (*Single Cell Amplified Genomes* SAGs), contribuant à fournir des génomes de référence d'individus issus des communautés naturelles. Cependant que ce soit au niveau des transcrits reconstruits à partir de l'ensemble de la communauté (Carradec *et al.* 2018) ou au niveau des gènes présents dans les MAGs ou les SAGs (Delmont *et al.*, 2022), près de 50% des séquences sont non annotées fonctionnellement. Cette proportion de séquences non annotées est cohérente avec des études précédentes (Perdigão *et al.*, 2015), et révèle ainsi ici dans le cadre de l'étude des métagénomes eucaryotes, plus de 50 millions de séquences inconnues fonctionnellement. Par comparaison, pour les écosystèmes marins planctoniques, le pourcentage de séquences protéiques non annotées chez les bactéries et les archées est de 40% à 90% (Sunagawa *et al.*, 2015) et chez les virus ce pourcentage peut atteindre 95% (Reyes *et al.*, 2012).

La structure d'une protéine reste plus conservée que sa séquence ; Schaefer and Rost (2012) ont montré que malgré la mutation de 50% à 80% des acides aminés d'une séquence, la structure restait la même. L'utilisation de l'information structurale offre donc des perspectives intéressantes afin d'augmenter l'annotation des génomes et des métagénomes. Par ailleurs les 191 565 structures de protéines présentes dans la RCSB PDB (<https://www.rcsb.org/>, Berman *et al.*, 2000) sont regroupées en seulement 5 481 superfamilles dans la classification structurale CATH (<https://www.cathdb.info/>, Sillitoe *et al.*, 2021) et ce nombre ne croît pratiquement plus depuis 10 ans. Ceci semble indiquer que nous connaissons déjà pratiquement tous les repliements des protéines globulaires stables et que ce nombre est étonnamment faible. Néanmoins, il est nécessaire que cette hypothèse soit prouvée et ceci est maintenant possible grâce aux nouvelles méthodes de prédiction structurale par *deep learning* (Jumper *et al.*, 2021) qui ont récemment révolutionné la bioinformatique structurale. Ces méthodes vont nous permettre d'avoir accès à la structure des protéines que nous ne pouvons pas annoter structurellement avec des outils "classiques" et ainsi pouvoir déterminer leurs fonctions, dans le cas du protéome humain, AlphaFold a permis de couvrir 98.5% des protéines.

Ainsi, l'objectif de ce stage est à la fois de prédire la structure des protéines afin d'annoter fonctionnellement les protéines des métagénomes des communautés planctoniques eucaryotes avec des méthodes classiques puis de nouvelles méthodes, mais aussi d'explorer la diversité des repliements prédits. Nous avons cherché à répondre aux questions suivantes : Que peut apporter la prédiction des structures pour l'annotation fonctionnelle des données métagénomiques ? Les nouvelles méthodes de prédiction structurale par *deep learning* sont-elles applicables à l'échelle de plusieurs millions de protéines ? Y a-t-il des repliements

toujours inconnus dans les protéines issues des métagénomes ?

Dans un premier temps, nous avons utilisé la méthode “classique” de comparaison des séquences à annoter structurellement à une banque de profils HMM (Hidden Markov Model), mais l’originalité est que ces profils ont été construits à partir de familles structurales ; ils se trouvent dans la banque *Gene3D* (Lewis *et al.*, 2018) issu de la classification CATH (version 4.3, <https://www.cathdb.info/>, Sillitoe *et al.*, 2021).

Dans un second temps, nous avons utilisé les méthodes récentes de prédiction de structure à partir de la séquence en utilisant des réseaux neurones profonds. Nous avons utilisé AlphaFold, (Jumper *et al.*, 2021), développé par la société DeepMind. Les évaluations CASP (*Critical Assessment of protein Structure Prediction*, <https://predictioncenter.org/>) ont pour objectif de tester différentes méthodes de prédictions structurales ; lors des évaluations CASP13 (en 2018) et CASP14 (en 2020), AlphaFold est arrivé premier. Il a fourni les meilleures prédictions pour la majorité des séquences testées, ce qui en fait l’outil de prédiction structurale le plus performant. Nous avons ensuite déterminé la structure et éventuellement la fonction des protéines à partir de leur structure tridimensionnelle. L’un des objectifs est aussi de déterminer si AlphaFold peut être appliqué sur un jeu de données métagénomiques, comme il a déjà été utilisé sur le génome humain (Tunyasuvunakool *et al.*, 2021).

II. Matériel et Méthodes

A. Jeu de données

Les séquences sont issues du jeu de données de Delmont *et al.* (2022), disponible : <http://www.genoscope.cns.fr/tara/>. La figure 1 présente les différents sites d’échantillonnage ainsi que le nombre d’échantillons de métagénomes disponibles pour chaque région géographique.

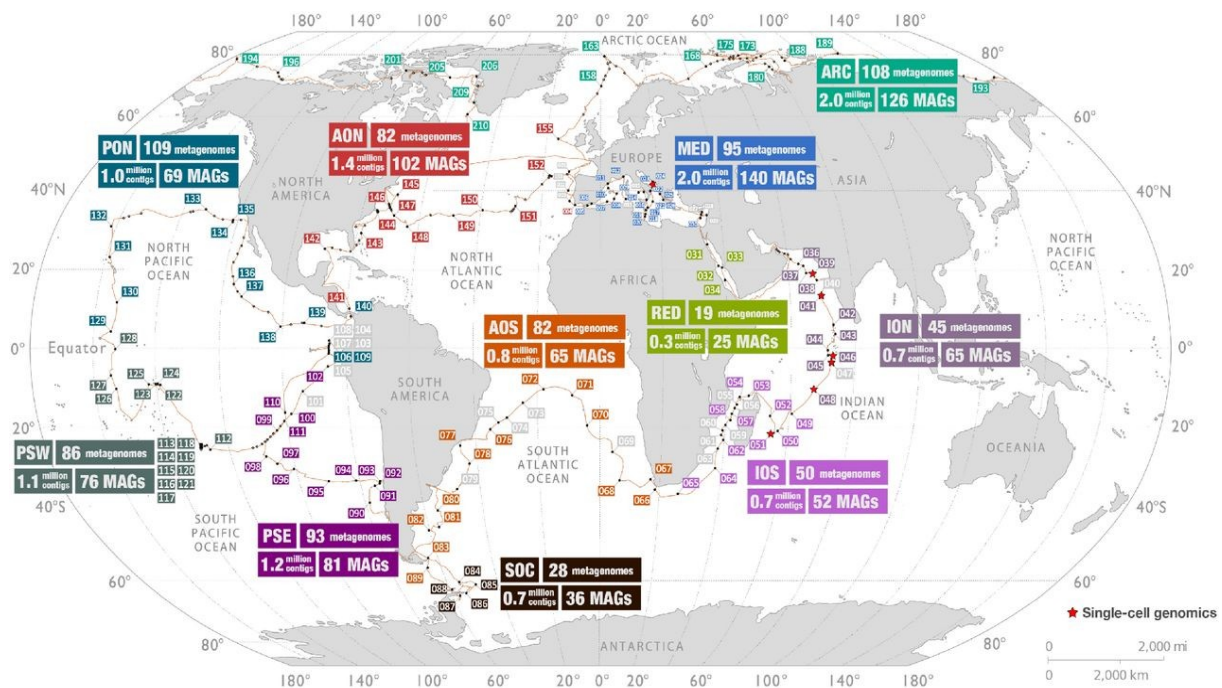


Figure 1 : Carte d'échantillonnage des 210 stations de l'expédition Tara Océans 2009-2013. Les stations ont été regroupées par régions géographiques : **ARC** : Arctic Ocean ; **MED** : Mediterranean Sea ; **RED** : Red Sea ; **ION** : Indian Ocean North ; **IOS** : Indian Ocean South ; **SOC** : Southern Ocean ; **AON** : Atlantic Ocean North ; **AOS** : Atlantic Ocean South ; **PON** : Pacific Ocean North ; **PSE** : Pacific South East ; **PSW** : Pacific South West (Delmont *et al.*, 2022).

Quatre classes de tailles d'organismes ont été échantillonnées (Carradec *et al.*, 2018) : 0.8 - 5 μm (pico-plancton), 5 - 20 μm (nano-plancton), 20 - 180 μm (micro-plancton), 180 - 2000 μm (meso-plancton). Les échantillons provenant de tout le globe, les espèces et les conditions environnementales sont très variées. Le jeu de données utilisé est composé d'un unique fichier (au format fasta) qui contenait 10 207 435 séquences de protéines correspondant aux gènes reconstruits pour 683 MAGs et 30 SAGs.

B. Matériels : Machines et Clusters et langages de programmation

1. Machines et Clusters

Lors du stage, en fonction des analyses à effectuer, différentes machines ont été utilisées, trois appartenant à l'Atelier de Bio-Informatique (<https://bioinfo.mnhn.fr/abi/presentation.FR.html>) (Annexe 3.A) et deux clusters de calcul. Le premier cluster de calcul utilisé est celui de la Plateforme Calcul Intensif et Algorithme (PCIA) (<http://www.ums2700.mnhn.fr/pcia/acces>) ; il appartient au Muséum National d'Histoire Naturelle de Paris. Il est composé de deux partitions (type_1 et type_2) (Annexe 3.B) et utilise le système de gestion de tâche SLURM. Le deuxième cluster de calcul utilisé est celui du Service d'Aide au Calcul Intensif et à l'Analyse de Données (SACADO) (<https://sacado.sorbonne-universite.fr/>) appartenant à la Sorbonne Université. Il est composé de trois partitions (alpha, beta, gamma) (Annexe 3.B) et utilise le système de gestion de tâche

PBS.

2. *Langages de programmation*

Les langages de programmations utilisés ici étaient : Python2 (version 2.7.17, Van Rossum, 1995), Python3 (version 3.9.7, Van Rossum and Drake, 2009), R (version 4.2.0, R Core Team, 2022) et Bash.

C. Recherche des profils HMM avec le logiciel HMMER

1. *Définition d'un profil HMM*

Les profils HMM sont des modèles qui décrivent une probabilité de distribution des acides aminés dans une séquence (Eddy, 1996; EMBL-EBI, n.d.). Ils sont construits à partir d'alignements multiples, ils possèdent ainsi des informations pour chaque position dans l'alignement, notamment sur la conservation de chaque acide aminé dans la colonne. De plus, ils prennent en compte des pénalités de *gap* qui dépendent de la position.

2. *HMMER - hmmsearch*

Afin de pouvoir rechercher les profils HMM correspondant à nos séquences nous avons utilisé le logiciel HMMER (version 3.3.2, <http://hmmer.org/>), plus précisément le programme *hmmsearch* (HMMER 3.3.2 (Nov 2020); <http://hmmer.org/>) qui permet de comparer des profils avec une banque de données de séquences. Celle utilisée ici est la banque de profils HMM *Gene3D* (Lewis *et al.*, 2018).

3. *Banque de données CATH / Gene3D*

La classification structurale utilisée pour cette analyse est la classification CATH (version 4.3 : <https://www.cathdb.info>, Sillitoe *et al.*, 2021). CATH est une classification des 191 565 structures protéiques présentes dans la RCSB PDB (<https://www.rcsb.org/>, Berman *et al.*, 2000). La classification CATH est hiérarchisée en quatre niveaux (table 1).

Table 1 : Hiérarchie de la classification CATH (version 4.3 : <https://www.cathdb.info>, Sillitoe *et al.*, 2021)

Classification	Définition
Class (C)	classement en fonction du contenu en structures secondaires (hélices alpha, feuillets beta, mixte alpha/beta, quelques structures secondaires)
Architecture (A)	classement en fonction de la forme globale des structures secondaires dans l'espace en 3 dimensions
Topology (T)	(ou <i>fold family</i>), classement qui reprend celui de l' <i>Architecture</i> en y ajoutant les informations concernant la connexion des structures secondaires
Homologous superfamily (H)	(ou <i>superfamily</i>), regroupe les domaines protéiques qui ont une forte probabilité d'avoir un ancêtre commun (ils peuvent être considérés comme homologue)

Les données de la classification CATH (version 4.3 : <https://www.cathdb.info>, Sillitoe *et al.*, 2021) peuvent être téléchargées à partir de l'adresse FTP : <ftp://orengoftp.biochem.ucl.ac.uk>.

La banque de profils HMM de *Gene3D* (Lewis *et al.*, 2018) est construite à partir de la classification CATH ; CATH contient 5 481 superfamilles et *Gene3D* contient 151 013 797 prédictions de domaines CATH. *Gene3D* permet de prédire les domaines structuraux pour des millions de séquences de protéines. C'est cette banque de données qui permet de rechercher les profils HMM. Elle peut être téléchargée via l'adresse FTP : <ftp://biochem.ucl.ac.uk/gene3d/>.

4. Présentation de l'analyses et des paramètres

L'objectif de cette première analyse est de chercher des profils HMM pour chaque séquence, puis pour chaque séquence ayant un profil HMM, de les annoter structuellement en lui assignant une ou plusieurs superfamilles. Nous avons utilisé les trois programmes proposés par *Gene3D* (table 2), avec les paramètres conseillés :

([ftp://orengoftp.biochem.ucl.ac.uk/gene3d/v21.0.0/gene3d_hmmsearch/ README_scan.txt](ftp://orengoftp.biochem.ucl.ac.uk/gene3d/v21.0.0/gene3d_hmmsearch/README_scan.txt)).

Table 2 : Présentation des programmes utilisés pour l'annotation structurale des protéines à partir de la banque de données *Gene3D*

Programmes	Descriptions
<i>hmmsearch</i> HMMER 3.3.2 (Nov 2020)	Permet de rechercher les profils HMM. Paramètres : "-Z 10 000 000" indique le nombre total de cibles dans notre recherche ; la e-value a été fixée à 0.001
cath-resolved-hit	Il peut y avoir plusieurs HMM similaires à une même séquence. Ce script permet de résoudre les conflits en conservant les meilleurs hits non chevauchants. (https://cath-tools.readthedocs.io/en/latest/tools/cath-resolve-hits/). --worst-permissible-bitscore=25 permet d'ignorer tous les hit avec bitscore inférieur à 25 et --min-dc-hmm-coverage=80 utilisé pour traiter la sortie de <i>hmmsearch</i> .
assign-cath-superfamilies.py	Script Python, permet d'assigner à chaque domaine détecté une superfamille de la classification CATH

Ces trois programmes ont été exécutés sur les 10 207 435 séquences de notre jeu de données. De plus, afin d'accélérer les calculs, le jeu de données a été séparé en deux afin d'être exécuté sur les clusters du PCIA et de SACADO.

D. Prédiction structurales

Comme dit précédemment, certaines séquences ne sont similaires à aucun des profils HMM. Il faut donc utiliser des outils afin de pouvoir déterminer leur structure et éventuellement leur fonction, nous avons donc décidé d'utiliser AlphaFold (Jumper *et al.*, 2021).

Le nombre de séquences à analyser avec AlphaFold est d'environ 6 124 059. La durée

normale d'analyse d'une séquence par AlphaFold étant de plus de 2 heures, nous avons cherché à la fois à comprendre et diminuer les différentes étapes de AlphaFold mais aussi à réduire le nombre de séquences à analyser. Pour cela nous avons voulu regrouper les séquences similaires et mettre de côté les protéines désordonnées pour lesquelles AlphaFold ne peut prédire de structure (Tunyasuvunakool *et al.*, 2021). Nous avons pour cela utilisé deux méthodes (Table 3).

Table 3 : Présentation des outils utilisés pour préparer l'analyse avec AlphaFold (Jumper *et al.*, 2021)

Outils	Utilisations	Citations
CD-HIT	Regroupement des séquences en fonction d'un pourcentage d'identité	(version 4.8.1, http://weizhong-lab.ucsd.edu/cd-hit/ , Fu <i>et al.</i> , 2012; Li and Godzik, 2006)
SPOT-Disorder-Single	Détermine les résidus désordonnés dans une séquence protéique	(https://sparks-lab.org/ , Hanson <i>et al.</i> , 2018)

1. *CD-HIT et SPOT-Disorder-Single*

Le pourcentage d'identité choisi pour le regroupement avec CD-HIT (Fu *et al.*, 2012; Li and Godzik, 2006) est de 60%. Ce premier programme a permis de diminuer le nombre de séquences à analyser à 4 054 195. De plus, avec SPOT-Disorder-Single (Hanson *et al.*, 2018), nous avons conservé les séquences qui avaient au moins 60% de leurs résidus qui étaient ordonnés, ce qui nous a permis de ne conserver 3 229 526 séquences pour l'analyse avec AlphaFold.

2. *AlphaFold*

AlphaFold est un algorithme développé par DeepMind utilisant plusieurs alignements afin de préparer la prédiction structurale puis un réseau de neurones afin de prédire la structure tridimensionnelle d'une protéine (Jumper *et al.*, 2021) à partir de sa séquence. Il possède deux étapes, chacune prenant environ 1h. La première consiste à rechercher dans plusieurs banques de séquences toutes les protéines similaires à la protéine dont nous voulons déterminer la structure, puis à réaliser les alignements multiples des séquences trouvées. Les banques sont *UniRef90* (≈59Go) (Suzek *et al.*, 2015), *MGnify* (≈64Go) (Mitchell *et al.*, 2020), *PDB70* (≈56.3Go) (Steinegger *et al.*, 2019a), *BFD* (≈1.7To) et *small BFD* (≈17Go) (Jumper *et al.*, 2021; Steinegger *et al.*, 2019b; Steinegger and Söding, 2018), *Uniclust30* (≈86Go) (Mirdita *et al.*, 2017), *UniProt* (≈98.3Go) (The UniProt Consortium, 2021), *PDB mmcif* (206Go) (Westbrook *et al.*, 2022) *PDB seqres* (≈0.2Go) (Zardecki *et al.*, 2022). Par ailleurs, les logiciels utilisés pour réaliser les alignements sur chaque banques de données sont : *jackhmmer* (HMMER 3.3.2 (Nov 2020); <http://hmmer.org/>) pour *MGnify* et *UniRef90*,

HHsearch (Steinegger *et al.*, 2019a) pour la PDB70 et *HHblits* (Steinegger *et al.*, 2019a) sur la BFD.

La deuxième étape utilise les résultats de la première étape afin d'effectuer la prédiction structurale. Nous avons analysé en détail le script d'AlphaFold et l'avons séparé les 2 parties en 2 scripts afin d'accélérer les prédictions. Ainsi, les recherches sur les banques et les alignements et la prédiction structurale par réseaux de neurones ont été effectuées sur la partition gamma du cluster de SACADO car il possède des GPU pouvant être exploités par AlphaFold.

De plus, nous avons utilisé l'option *reduced_dbs*, elle permet d'utiliser une version réduite de la BFD (Jumper *et al.*, 2021; Steinegger *et al.*, 2019; Steinegger and Söding, 2018) ; le volume de cette banque de données est 17 Go au lieu d'environ 1.7 To. AlphaFold ne prend qu'un fichier contenant une unique séquence fasta en entrée.

3. YAKUSA

AlphaFold ne fournit pas une annotation structurale de la protéine ; il permet simplement d'obtenir la structure tridimensionnelle de la protéine dans le format PDB, il faut utiliser un autre logiciel afin de comparer la structure 3D avec celles déjà connues pour en déduire des informations. Nous avons utilisé pour cela le programme YAKUSA (Carpentier *et al.*, 2005) qui permet de scanner une banque de structures. Ce programme attribue un z-score à chaque structure, si le z-score est inférieur à 6-7, alors il se peut que le repliement de la structure soit nouveau.

Le stage étant assez court, nous n'avons pas pu lancer l'analyse complète sur la totalité des protéines identifiées. Nous avons choisi de nous restreindre à un échantillon de 93 protéines ordonnées afin d'obtenir de premiers résultats et savoir la faisabilité de ces prédictions.

E. Analyses statistiques et représentations graphiques

Toutes les analyses et statistiques et les représentations graphiques ont été réalisées avec R (version 4.2.0, R Core Team, 2022) et le package ggplot2 (version 3.3.6, Wickham, 2016) pour les graphiques. Certains packages présents dans le package Tidyverse (version 1.3.1, Wickham *et al.*, 2019) ont également été utilisés.

Pour les analyses statistiques effectuées, le risque alpha a été placé à 5%.

III. Résultats

A. Proportion des annotations & distribution des annotations CATH

Les premiers résultats que l'on a pu obtenir à la suite de la recherche des profils HMM

à l'aide de HMMER (version 3.3.2, <http://hmmer.org/>), concerne la proportion des séquences annotées (séquences connues) et des séquences non annotées (séquences inconnues) avec les HMM de *Gene3D*. Nous avons déterminé que la proportion des séquences annotées était d'environ 40.02% et celle en séquences non annotées était d'environ 59.98%, ces proportions prennent en compte les SAGs ; sans les SAGs (*Single Cell Amplified Genome*) (uniquement les MAGs (*Metagenome-assembled genome*)), la proportion est de 40.08% de séquences annotées et 59.92% de séquences non annotées. La figure 2 permet de visualiser la distribution du nombre de séquences en fonction de l'identifiant de la superfamille CATH, seules les 50 premières familles possédant le plus de séquences ont été représentées.

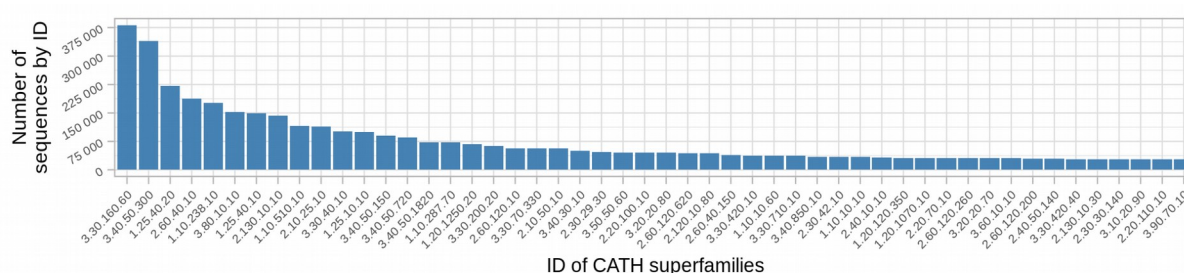


Figure 2 : Distribution du nombre de séquences annotées pour chaque identifiant de la classification CATH identifié. Représentation des 50 premières superfamilles

Il y a une décroissance forte du nombre de séquences en fonction des identifiants. Les deux superfamilles avec le plus de séquences sont 3.30.160.60 (*Zinc Finger*, 381 575 séquences) et 3.40.50.300 (*P-loop containing nucleotide triphosphate hydrolases*, 340 066 séquences).

Nous avons comparé les annotations structurales faites avec *Gene3D* (Lewis *et al.*, 2018) et les annotations fonctionnelles faites avec la base de données Pfam (Mistry *et al.*, 2021) effectuées par (Delmont *et al.*, 2022). Les taux d'annotation sont respectivement de 40% et 43%. 32% des séquences sont à la fois annotées par *Gene3D* et par Pfam (figure 3) et donc *Gene3D* permet d'ajouter 813 357 séquences annotées supplémentaires aux 4 485 410 séquences déjà annotées avec Pfam.

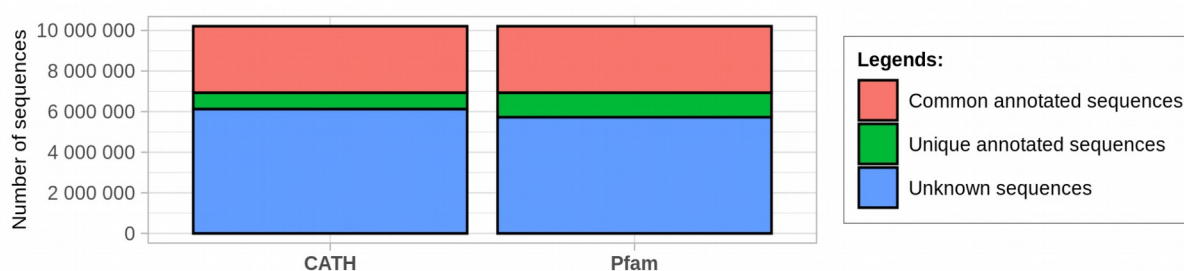


Figure 3 : Comparaison du nombre de séquences annotées avec la banque de données *Gene3D* (Lewis *et al.*, 2018) et Pfam (Delmont *et al.*, 2022; Mistry *et al.*, 2021) (jeu de données initial : 10 207 435 séquences)

B. Distribution de la taille des séquences annotées / non annotées

Nous avons analysé la longueur des séquences annotées et non annotées avec *Gene3D* (figure 4). Les séquences avec une taille supérieure à 1 500 acides aminés n'apparaissent pas sur le graphique car le nombre de séquences était très faible et cela permet de mieux visualiser les données concernant les séquences de petites tailles.

Sur la figure 4, nous pouvons constater que la majorité des séquences ont une taille inférieure à 1 000 acides aminés. Il y a 9 747 332 séquences avec une taille inférieure à 1 000 acides aminés et 461 490 séquences avec une taille supérieure à 1 000 acides aminés. De plus, la médiane des tailles des séquences annotées est de 376 acides aminés alors que celle des séquences non annotées est de 131 acides aminés ; les séquences de petite taille sont majoritairement non annotées.

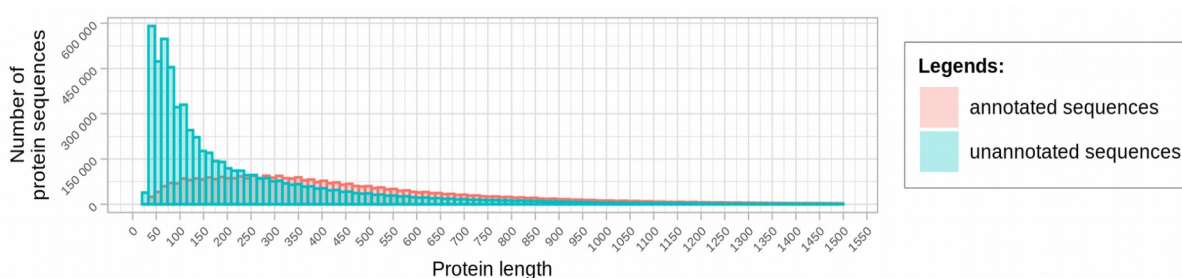


Figure 4 : Distribution du nombre de séquences en fonction de la taille (en acides aminés) pour les séquences annotées et les séquences non annotées. Les séquences avec une taille supérieure à 1 500 acides aminés ont été supprimées.

C. Distribution des annotations en fonction du *kingdom* - *phylum*

Nous nous sommes demandé si l'annotation structurale des séquences était dépendante de la classification des lignées dont elles sont issues (notamment les rangs taxonomiques au niveau du *kingdom* et du *phylum*). Delmont *et al.* (2022) ont également fourni une assignation taxonomique pour chaque MAG et SAG (tableau *Table_S02_genome-resolved metagenomics_and_SAGs* disponible sur

https://www.genoscope.cns.fr/tara/localdata/data/SMAGs-v1/Supplemental_Tables.zip).

Nous avons croisé les informations de taxonomie avec celle d'annotation de structure. Pour cela, nous avons réalisé un test de χ^2 (H_0 : l'annotation structurale est indépendante du *kingdom* / *phylum* ; H_1 : l'annotation structurale est dépendante du *kingdom* / *phylum*). En résultat, nous avons mis en évidence que l'annotation structurale des séquences était dépendante du *kingdom* et du *phylum* (p-value < 0.05), les données sont présentées dans l'Annexe 4.A.

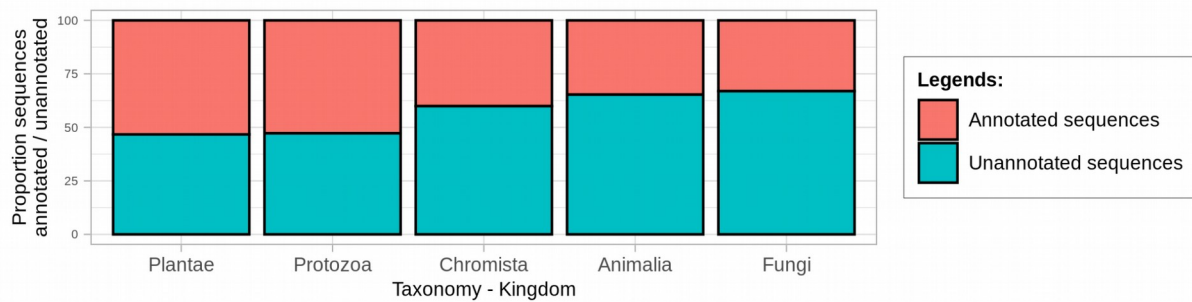


Figure 5 : Proportion des séquences annotées et des séquences non annotées pour chaque *Kingdom*. Nombre de séquences totales : **Animalia** : 3 803 641, **Chromista** : 4 209 409, **Fungi** : 28 682, **Plantae** : 528 010, **Protozoa** : 1 049 916.

De plus, la [figure 5](#) nous permet de constater que ce sont les Plantae et les Protozoa qui possèdent le plus de séquences annotées (environ 50%). Quant aux Animalia, Chromista et Fungi, le pourcentage de séquences non annotées est d'environ 60%.

Il est également intéressant de regarder la présence des différentes superfamilles trouvées en fonction des phylums. Nous pouvons constater, avec la [figure 6.A](#), que nous trouvons au minimum 25% de superfamilles dans un phylum, c'est notamment le cas des Echinodermata ; de plus, nous pouvons voir que certains phylums possèdent environ 70%-75% des superfamilles. C'est par exemple le cas des arthropoda pour les animalia, des miozoa pour les chromista, des choanozoa pour les protozoa ou encore des chlorophyta pour les plantae. Nous nous sommes ensuite intéressés à la corrélation entre le pourcentage de séquences annotées et le pourcentage de superfamilles détectées pour chaque phylum. Un test de corrélation de Pearson nous permet d'affirmer qu'il n'y a pas de corrélation entre les deux variables ($p\text{-value} = 0.8077$) (H_0 : Les variables ne sont pas corrélées, le coefficient de corrélation est égal à 0 ; H_1 : les variables sont corrélées, le coefficient de corrélation est différent de 0) ([Annexe 5.A](#)) ; ce qui est confirmé par l'observation graphique ([Annexe 5.B](#)).

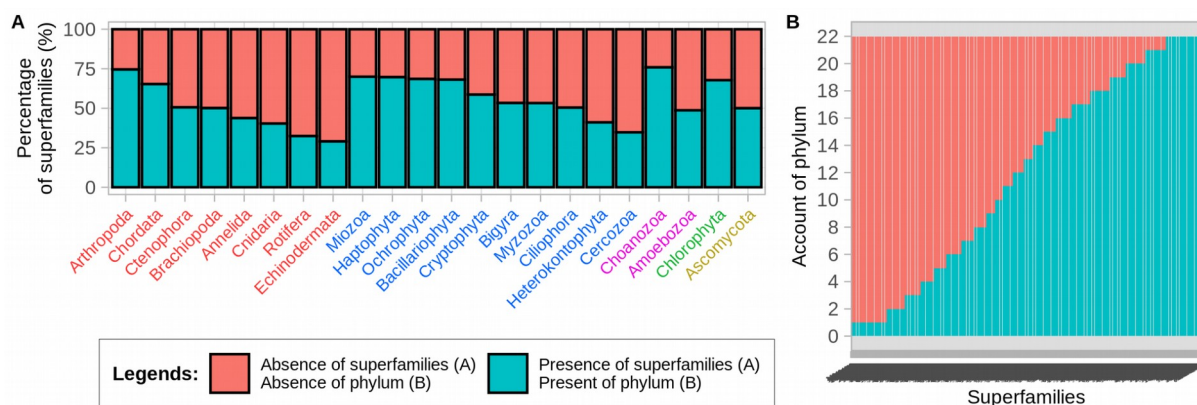


Figure 6 : (A) Proportion de la présence et de l'absence en superfamille pour chaque phylum. Les couleurs l'axe des abscisses correspondent au kingdom : **Rouge** : Animalia ; **Bleu** : Chromista ; **Violet** : Protozoa ; **Vert** : Plantae ; **Jaune** : Fungi. (B) Nombre de phylums pour chaque superfamille.

Par ailleurs, avec la [figure 6.B](#), nous pouvons constater que 346 superfamilles sont

retrouvées dans tous les *phylums* ; c'est notamment le cas des superfamilles citées dans la partie III.A (3.40.50.300, 3.30.160.60). À l'inverse, 344 superfamilles ne sont retrouvées que dans un seul *phylum* (1.10.10.1150 : *Coenzyme PQQ synthesis protein D* ; 1.10.10.1550 : *ROS/MUCR transcriptional regulator protein*). Par ailleurs, les superfamilles 1.10.10.1150 et 1.10.10.1550, ont été assignées respectivement à 2 et 6 séquences dans le jeu de données initial.

D. Prédiction des structures

1. AlphaFold

Le temps ne nous a pas permis d'exécuter AlphaFold sur tout le jeu de données mais nous avons pu estimer les temps de calculs et donc la faisabilité d'une application à aussi large échelle (table 4). Nous n'avons les résultats que pour les plus de 6 millions de protéines non annotées. Le temps de calcul de la première partie (recherche sur banque) a été divisé par 3 (20 min au lieu d'une heure) et le temps de la seconde partie est plutôt à environ 100 min. Avec ces temps de calculs, il est donc possible de prédire la structure de 12 protéines par jour et par cœur. Nous avons environ 3 millions de protéines a priori globulaires et différant à plus 60% mais en l'état, il n'est pas raisonnable de prédire leur structure: il faudrait 500 cœurs et 500 jours. Par contre, il est tout à fait possible de prédire des protéomes entiers : 20 000 structures peuvent être prédites en 20 jours et 80 cœurs.

Table 4 : Temps d'exécution de chaque parties d'AlphaFold, les temps de calculs ont été déterminé sur les deux machines utilisé pour AlphaFold, la machine 3 de l'ABI et le cluster de SACADO (<https://sacado.sorbonne-universite.fr/>)

	Temps de calcul - première partie	Temps de calcul - deuxième partie
Machine 3 - ABI	≈20 minutes	≈105 minutes
Cluster - SACADO	≈30 minutes	≈70 minutes

2. Analyse des résultats de AlphaFold

Afin d'avoir un premier aperçu des résultats d'AlphaFod, nous avons prédit la structure de 93 protéines (Jumper *et al.*, 2021). Le temps a été un facteur limitant mais il sera tout à fait possible de poursuivre les prédictions par la suite. Nous avons ensuite utilisé YAKUSA pour comparer les structures aux structures de la PDB (Carpentier *et al.*, 2005) ; ceci nous permet de déterminer si les repliements des protéines sont nouveaux, ou similaires à

une ou des structures déjà présentes dans RCSB PDB (<https://www.rcsb.org/>, Berman *et al.*, 2000). Si le z-score est supérieur à 6 ou 7 alors nos *matches* sont susceptibles d’être significatifs. Ainsi, nous avons au moins 27 protéines (30%) qui ont potentiellement des repliements nouveaux mais il faudra les analyser plus en détail pour le confirmer (table 5). Les structures de MAG TARA_SOC_28_MAG_00063_000000007425.10.1 et de TARA_SOC_28_MAG_00063_000000008836.1.3 sont présentées Annexe 6 à titre d’exemple.

Table 5 : Nombre de protéines en fonction du z-score

z-score < 6	6 < z-score < 7	z-score > 7
27 protéines	27 protéines	39 protéines

IV. Discussion

Ce stage avait pour objectif d’explorer un jeu de données métagénomiques contenant 10 207 435 séquences de protéines, issues de l’expédition Tara Océans 2009-2013 provenant de 20 régions géographiques (210 stations) (Pesant *et al.*, 2015). La recherche des profils HMM avec le programme *hmmsearch* (HMMER 3.3.2 (Nov 2020); <http://hmmer.org/>) nous a permis d’annoter structurellement 40.02% des séquences, à l’aide de la banque de profils HMM construits à partir de familles structurales *Gene3D* (Lewis *et al.*, 2018). Ce résultat est cohérent avec les études précédentes ; en effet, Delmont *et al.* (2022) avait également pu annoter fonctionnellement environ 40% de jeu. L’annotation fonctionnelle des séquences de Delmont *et al.* (2022) a été réalisée avec la base de données Pfam (Mistry *et al.*, 2021). En comparant nos annotations CATH avec celles de Pfam, nous avons montré que CATH permettait de rajouter 813 357 annotations. Pour de prochaines études, il peut être intéressant d’utiliser à la fois CATH et Pfam afin d’obtenir un maximum d’annotations structurales ou fonctionnelles pour les séquences du jeu de données. De plus, il peut également être intéressant d’utiliser une autre banque de classification des structures de protéines, SCOPe (*Structural Classification of Proteins - extended*) (<https://scop.berkeley.edu/>, Fox *et al.*, 2014), afin de voir si elle permet d’apporter de nouvelles informations en plus de CATH et Pfam. Ainsi, en annotant un maximum des séquences à l’aide des différentes banques de données, le nombre de séquences sans annotation structurale ou fonctionnelle pourrait diminuer, avant même l’utilisation d’outil de *deep learning* tel qu’AlphaFold (Jumper *et al.*, 2021).

Il existe un lien significatif entre proportion de séquences annotées/non annotées, et les kingdoms Animalia, Chromista et Fungi présentent la plus grande proportion de séquences non annotées (~60%).

Par ailleurs, lors de recherche des profils HMM, les deux identifiants de superfamille les plus souvent identifiés sont l'identifiant 3.30.160.60 (*Zinc Finger*) et l'identifiant 3.40.50.300 (*P-loop containing nucleotide triphosphate hydrolases*), et ils sont retrouvés dans les 22 *phylums* eucaryotes ici étudiés. Certaines superfamilles ne sont au contraire retrouvées que dans un seul phylum, c'est le cas notamment des identifiants 1≈.10.10.1150 (*Haptophyta, Coenzyme PQQ synthesis protein D*) et 1.10.10.1550 (*Arthropoda, ROS/MUCR transcriptional regulator protein*). Lors de prochaines analyses, il sera intéressant de regarder la distribution des superfamilles en fonction de la taxonomie à divers rangs mais également en fonction des conditions environnementales (e.g., température, salinité, lumière, concentration en nutriments) ; nous pourrions ainsi déterminer si la présence de certaines structures pourrait être corrélée à certaines conditions environnementales ou encore si elles peuvent être retrouvées dans tous les écosystèmes marins. Nous pouvons poser les hypothèses : les superfamilles présentes dans tous les phylums peuvent être retrouvées dans la majorité des 20 régions géographiques, quant aux superfamilles retrouvées dans un seul *phylum* peuvent être retrouvées que dans des conditions environnementales précises.

Nous avons également pu constater que la majorité des séquences non annotées était de petite taille, avec une médiane de 131 acides aminés. Cette importante proportion en séquences non annotées pour les fragments peut venir du fait que les domaines sont tronqués, et donc que les profils HMM ne permettent de les identifier.

Les méthodes "classiques" ne nous ont pas permis d'annoter structurellement la totalité des séquences, plus de 50% des séquences de chaque phylum étaient non annotées, c'est pour cela que nous avons décidé d'utiliser AlphaFold ([Jumper et al., 2021](#)). Dans ce programme, comme précisé dans la partie Matériels et Méthodes, nous avons pu identifier deux parties, la première réalise plusieurs alignements multiples et la deuxième la prédiction structurale. La séparation en deux parties a été réalisée seul et nous avons utilisé une version réduite des bases de données de séquences, mais dans l'état actuel d'AlphaFold ne peut pas être exécuté sur de gros jeux de données métagénomiques de plusieurs millions de séquences. Nous avons plusieurs pistes pour tenter de le rendre plus rapide. Pour la première partie, il faudrait tester s'il est possible de se passer de certaines banques sans altérer les résultats. En effet, nous savons par exemple qu'il est très peu probable de trouver une protéine de la PDB similaire aux séquences non annotées. Il semblerait aussi judicieux de découper les séquences à annoter en séquences réellement uniques (sans aucune séquence similaire dans d'autres banques) et en séquences ayant des séquences similaires dans des banques. Il est probable que les performances d'AlphaFold avec les premières séquences soient beaucoup moins bonnes et

que dans ce cas il faille utiliser d'autres méthodes de prédiction dites *de novo* comme Rosetta (<https://www.rosettacommons.org/software>, Hanson *et al.*, 2018; Humphreys *et al.*, 2021). Nous avons réalisé un clustering des séquences de notre jeu de données. Il serait aussi intéressant de pouvoir donner à AlphaFold toutes les séquences d'un même cluster sous forme d'un alignement, ce qui permettrait peut-être de se passer de certaines recherches sur banques faites actuellement. Afin de l'accélérer la première partie, il faudrait pouvoir donner un fichier multifasta en entrées afin que les programmes d'alignements (*jackhmmer* (HMMER 3.3.2 (Nov 2020); <http://hmmer.org/>) et *HHsearch* / *HHblits* (Steinegger *et al.*, 2019a)) puissent traiter la totalité des séquences. Par ailleurs, *HHsearch* et *HHblits* peuvent déjà être employés pour des projets en métagénomique (Steinegger *et al.*, 2019a) ; ainsi, si un fichier multifasta peut être donné comme entrée pour la première partie, ça pourrait permettre d'obtenir les fichiers *pickle* plus rapidement. Concernant la partie permettant la prédiction structurale, une piste qui pourrait être explorée serait de diminuer la qualité de la prédiction afin d'accélérer AlphaFold. De plus, l'utilisation d'un cluster de calcul plus performant, notamment avec plus de GPU, pourrait permettre d'accélérer AlphaFold ; dans le cas du génome humain, le nombre de GPU utilisés par jour s'élève à 930 (NVIDIA Tesla V100) (Tunyasuvunakool *et al.*, 2021). Enfin, un cluster de calcul possédant un SSD avec une capacité de stockage important afin de pouvoir stocker toutes les banques de données nécessaires pour AlphaFold ; la partition gamma du cluster de calcul de SACADO ne possédant que 2 noeuds et 12 CPU, il n'est pas possible d'exécuter plusieurs fois AlphaFold.

V. Conclusion

Pour conclure, avec des outils bioinformatiques, comme HMMER (version 3.3.2, <http://hmmer.org/>), nous avons pu annoter structurellement environ 40% de nos séquences, en utilisant la banque de profils HMM *Gene3D* (Lewis *et al.*, 2018). L'utilisation d'annotations fonctionnelle effectuées avec une banque de données telle que Pfam, pourrait permettre de diminuer la proportion d'inconnues. Cependant au moins 50% des données restent non annotées structurellement et ou fonctionnellement. De plus, les séquences non annotées étaient majoritairement de petite taille, avec une médiane de 131 acides aminés ; ce qui montre que nous avons des difficultés à annoter les séquences de petites tailles. Afin de pouvoir déterminer la structure tridimensionnelle et d'annoter structurellement les séquences inconnues, nous avons utilisé AlphaFold (Jumper *et al.*, 2021). Cependant, c'est un outil qui ne permet pas pour l'instant de traiter des jeux de données métagénomiques ; un premier travail afin de mieux le comprendre et de l'accélérer a été fait. D'autres travaux pourront être

entrepris afin de l'optimiser pour des jeux de données métagénomiques. Par ailleurs, des analyses de la biogéographie des superfamilles de protéines pourraient permettre d'investiguer quels seraient les conditions déterminantes dans la sélection de la structure des protéines. Une autre étude pourrait consister à comparer le nombre de séquences qui ont été annotées fonctionnellement avec la taille des génomes ou encore celle des cellules.

VI. Bibliographie

- Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.-M., Belser, C., Bertrand, A., Cruaud, C., Da Silva, C., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquell, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Acinas, S.G., Royo-Llonch, M., Cornejo-Castillo, F.M., Logares, R., Fernández-Gómez, B., Bowler, C., Cochrane, G., Amid, C., Hoopen, P.T., De Vargas, C., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M.E., Stepanauskas, R., Sullivan, M.B., Brum, J.R., Duhaime, M.B., Poulos, B.T., Hurwitz, B.L., Pesant, S., Karsenti, E., Wincker, P., 2017. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* 4, 170093. <https://doi.org/10.1038/sdata.2017.93>
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Bittner, L., Guidi, L., Chaffron, S., Eveillard, D., 2018. Les microbiomes de l’océan: une démarche à haut débit pour une compréhension globale et systémique, in: Palka, L. (Ed.), *Microbiodiversité: Un Nouveau Regard*. éditions matériologiques, France, p. 26.
- Carpentier, M., Brouillet, S., Pothier, J., 2005. YAKUSA: A fast structural database scanning method. *Proteins Struct. Funct. Bioinforma.* 61, 137–151. <https://doi.org/10.1002/prot.20517>
- Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D.J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M.B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler, C., Wincker, P., 2018. A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 373. <https://doi.org/10.1038/s41467-017-02342-1>
- Danchin, A., 2000. A brief history of genome research and bioinformatics in France. *Bioinformatics* 16, 65–75. <https://doi.org/10.1093/bioinformatics/16.1.65>
- Delmont, T.O., Gaia, M., Hinsinger, D.D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A.M., Kourlaiev, A., d’Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud,

- C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., Sunagawa, S., Acinas, S.G., Bork, P., Karsenti, E., Bowler, C., Sardet, C., Stemmann, L., de Vargas, C., Wincker, P., Lescot, M., Babin, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Jaillon, O., Kandels, S., Iudicone, D., Ogata, H., Pesant, S., Sullivan, M.B., Not, F., Lee, K.-B., Boss, E., Cochrane, G., Follows, M., Poulton, N., Raes, J., Sieracki, M., Speich, S., de Vargas, C., Bowler, C., Karsenti, E., Pelletier, E., Wincker, P., Jaillon, O., 2022. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 2, 100123. <https://doi.org/10.1016/j.xgen.2022.100123>
- Eddy, S.R., 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* 6, 361–365. [https://doi.org/10.1016/S0959-440X\(96\)80056-X](https://doi.org/10.1016/S0959-440X(96)80056-X)
- EMBL-EBI, n.d. What are profile hidden Markov models? | Pfam. URL <https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/> (accessed 6.16.22).
- Falkowski, P.G., Barber, R.T., Smetacek, V., 1998. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* 281, 200–206. <https://doi.org/10.1126/science.281.5374.200>
- Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* 281, 237–240. <https://doi.org/10.1126/science.281.5374.237>
- Fox, N.K., Brenner, S.E., Chandonia, J.-M., 2014. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* 42, D304–D309. <https://doi.org/10.1093/nar/gkt1240>
- Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Hanson, J., Paliwal, K., Zhou, Y., 2018. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J. Chem. Inf. Model.* 58, 2369–2376. <https://doi.org/10.1021/acs.jcim.8b00636>
- Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., Stancheva, V.G., Li, X.-H., Liu, K., Zheng, Z., Barrero, D.J., Roy, U., Kuper, J., Fernández, I.S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E.C., Biggins, S., Keeney, S., Miller, E.A., Fromme, J.C.,

- Hendrickson, T.L., Cong, Q., Baker, D., 2021. Computed structures of core eukaryotic protein complexes. *Science* 374, eabm4805. <https://doi.org/10.1126/science.abm4805>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Lewis, T.E., Sillitoe, I., Dawson, N., Lam, S.D., Clarke, T., Lee, D., Orengo, C., Lees, J., 2018. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* 46, D435–D439. <https://doi.org/10.1093/nar/gkx1069>
- Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., Steinegger, M., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* 45, D170–D176. <https://doi.org/10.1093/nar/gkw1081>
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. <https://doi.org/10.1093/nar/gkaa913>
- Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., 2020. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* 48, D570–D578. <https://doi.org/10.1093/nar/gkz1035>
- Perdigão, N., Heinrich, J., Stolte, C., Sabir, K.S., Buckley, M.J., Tabor, B., Signal, B., Gloss, B.S., Hammang, C.J., Rost, B., Schafferhans, A., O'Donoghue, S.I., 2015. Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* 112, 15898–15903. <https://doi.org/10.1073/pnas.1508380112>
- Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., Searson, S., 2015. Open science

- resources for the discovery and analysis of Tara Oceans data. *Sci. Data* 2, 150023. <https://doi.org/10.1038/sdata.2015.23>
- R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., Gordon, J.I., 2012. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* 10, 607–617. <https://doi.org/10.1038/nrmicro2853>
- Schaefer, C., Rost, B., 2012. Predict impact of single amino acid change upon protein structure. *BMC Genomics* 13, S4. <https://doi.org/10.1186/1471-2164-13-S4-S4>
- Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S.D., Berka, K., Varekova, I.H., Svobodova, R., Lees, J., Orengo, C.A., 2021. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* 49, D266–D273. <https://doi.org/10.1093/nar/gkaa1079>
- Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J., 2019a. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 473. <https://doi.org/10.1186/s12859-019-3019-7>
- Steinegger, M., Mirdita, M., Söding, J., 2019b. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16, 603–606. <https://doi.org/10.1038/s41592-019-0437-4>
- Steinegger, M., Söding, J., 2018. Clustering huge protein sequence sets in linear time. *Nat. Commun.* 9, 2542. <https://doi.org/10.1038/s41467-018-04964-5>
- Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmiento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P., 2015. Structure and function of the global ocean microbiome. *Science* 348, 1261359. <https://doi.org/10.1126/science.1261359>
- Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., 2015. UniRef clusters: a

- comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. <https://doi.org/10.1093/bioinformatics/btu739>
- The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G.J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S.A.A., Potapenko, A., Ballard, A.J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A.W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., Hassabis, D., 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596. <https://doi.org/10.1038/s41586-021-03828-1>
- Van Rossum, G., 1995. *Python Reference Manual*. Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.
- Van Rossum, G., Drake, F.L., 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
- Westbrook, J.D., Young, J.Y., Shao, C., Feng, Z., Guranovic, V., Lawson, C.L., Vallat, B., Adams, P.D., Berrisford, J.M., Bricogne, G., Diederichs, K., Joosten, R.P., Keller, P., Moriarty, N.W., Sobolev, O.V., Velankar, S., Vonnrhein, C., Waterman, D.G., Kurisu, G., Berman, H.M., Burley, S.K., Peisach, E., 2022. PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology. *J. Mol. Biol., Computation Resources for Molecular Biology* 434, 167599. <https://doi.org/10.1016/j.jmb.2022.167599>
- Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.Da., François, R., Grolemund, G., Hayes, A., Henry, L.A., Hester, J.J., Kuhn, M., Pedersen, T.Q., Miller, E., Bache, S., Müller, K., Ooms, J.C.L., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C.O., Woo, K.H., Yutani, H., 2019. Welcome to the Tidyverse. *J. Open Source Softw.* <https://doi.org/10.21105/joss.01686>
- Zardecki, C., Dutta, S., Goodsell, D.S., Lowe, R., Voigt, M., Burley, S.K., 2022. PDB-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Sci.* 31, 129–140. <https://doi.org/10.1002/pro.4200>

Informations complémentaires

Ressources informatiques

Ce travail a obtenu l'accès aux ressources HPC de la plateforme SACADO MeSU de Sorbonne-Université. Il a également obtenu l'accès aux ressources de la Plateforme de Calcul Intensif et Algorithmique PCIA, Muséum National d'Histoire Naturelle, Centre National de la Recherche Scientifique, UAR 2700 2AD, CP 26, 57 rue Cuvier, F-75231 Paris Cedex 05, France

Disponibilités des codes

Les codes utilisés sont disponibles sur : https://github.com/jroussea/Stage_MNHN.git

Annexes

Annexe 1 : Présentation de la structure d'accueil

<https://bioinfo.mnhn.fr/abi/presentation.FR.html>

Bref historique

L'Atelier de BioInformatique (ou ABI) est une structure ouverte regroupant des enseignants-chercheurs, chercheurs et étudiants désirant travailler à l'interface entre Biologie, Informatique et Mathématiques dans un environnement multidisciplinaire.

Cette structure fondée en 1986 à l'Institut Curie a rejoint l'UFR des Sciences de la Vie de l'UPMC en 1996 puis le Muséum National d'Histoire Naturelle en 2015. L'ABI est l'une des structures ayant accompagné l'essor de la bioinformatique en France (Danchin, 2000). Ainsi, parmi les nombreux chercheurs qui ont pris part à la vie de l'ABI, certains aujourd'hui sont devenus des référents de la bioinformatique dans plusieurs grands organismes de recherche : à l'INRIA, au CNRS et au CEA.

Situation actuelle

L'ABI est un lieu de travail, de formation, et d'échanges pour les chercheurs, enseignants-chercheurs et étudiants. Il accueille des enseignants-chercheurs et des chercheurs souhaitant traiter leur sujet de recherche par l'approche bioinformatique mais il accompagne également des reconversions vers la bioinformatique. A la différence d'une plateforme ou d'une unité de service, l'Atelier apporte bien plus qu'un travail « à façon » en permettant à chacun de partager ses connaissances, de faire part de ses idées et surtout de les confronter dans un environnement multidisciplinaire. Conformément à son objectif de formation, l'ABI est le lieu d'accueil de nombreux étudiants. Ces stagiaires, doctorants et post-doctorants provenant de divers laboratoires ou formations bénéficient de l'environnement multidisciplinaire de l'ABI pour développer l'aspect bioinformatique de leur travail de recherche. Le travail est mené en étroite collaboration avec leur laboratoire d'appartenance.

L'ABI existe aujourd'hui sous la forme d'une équipe de recherche de l'ISYEB au MNHN (UMR 7205). Les travaux de l'ABI donnent lieu à des publications dans des revues internationales ou dans des actes de conférences à comité de lecture ainsi qu'à plusieurs logiciels mis en ligne.

Les membres permanents actuels de l'ABI sont principalement des enseignants-chercheurs de l'Université Pierre et Marie Curie. Trois membres associés actuellement à l'ABI proviennent d'autres universités (UVSQ et Université Paris 13) ou organismes de recherche (INRA).

L'ABI est fortement investi dans la formation universitaire. L'ABI a eu, et possède toujours, une place centrale dans la mise en place et la responsabilité des enseignements d'informatique et de bioinformatique de l'UPMC.

Annexe 2 : Bilan du stage

Ce stage à l'Atelier de Bio-Informatique a tout d'abord été pour moi une première expérience dans le milieu de la recherche. Au sein de l'ABI, j'ai pu travailler sur un sujet passionnant concernant l'annotation structurale de protéines issus de métagénomiques. En plus de découvrir la recherche, il m'a permis de mieux comprendre les protéines : leur diversité, leur répartition.

Afin de répondre à différentes questions, j'ai découvert et utilisé un outil, HMMER, qui permet d'annoter structurellement les protéines en utilisant la banque de données CATH/*Gene3D*. J'ai également utilisé l'un des meilleurs outils de prédictions structurales : AlphaFold. J'ai également dû utiliser plusieurs autres outils bioinformatiques afin d'étudier le jeu de données

Cependant, il reste de nombreuses questions, que ce soit notamment pour les outils bioinformatiques (accélération d'AlphaFold) ou biologiques (répartitions spatiales des superfamilles de protéines, les facteurs environnementaux qui influencent le plus la structure tridimensionnelle des protéines). Avec un peu plus de temps ces différentes questions auraient pu être abordées.

Le défi de ce stage a été de traiter un jeu de données métagénomique avec plus de 10 200 000. Il a donc fallu trouver le moyen d'automatiser et paralléliser les analyses. Certains scripts, qui permettent d'analyser les résultats, écrits en Bash, pourraient être accélérés en utilisant Awk.

Annexe 3 : Configuration des machines et clusters

Annexe 3.A : Configuration des machines appartenant à l'Atelier de Bio-Informatique

Configurations	Machine 1	Machine 2	Machine 3
Utilisation	Quotidienne - machine de travail principale	Calculs - exécution de programmes	AlphaFold
Système d'exploitation	Ubuntu 18.04.6 LTS	Ubuntu 20.04.4 LTS	Debian 4.19.181-1
RAM	15.4 Go	125.6 Go	93.1 Go
Processeurs	Intel® Core™ i5-10210U CPU @ 1.60GHz (x8)	Intel® Xeon® Silver 4214R CPU @ 2.40GHz (x48)	
Cartes graphiques (GPU)	Intel® UHD Graphics (CML GT2)	Nvidia RTX A5000/PCIe/SSE2	AMD® Radeon™ Pro WX 3200

Annexe 3.B : Caractéristiques des clusters de calculs de SACADO (<https://sacado.sorbonne-universite.fr/>) et du PCIA (<http://www.ums2700.mnhn.fr/pcia/presentation>)

Caractéristiques	Cluster de calcul SACADO			Cluster de calcul PCIA	
Partition	Partition Alpha	Partition Beta	Partition Gamma	Partition Type_1	Partition Type_2
Utilisation	Calculs - exécution de programmes		AlphaFold	Calcul - exécution de programme	
Système de gestion des tâches	PBS (Portable Batch System)			SLURM (Simple Linux Resources Management)	
Noeud	1	144	2	4	29
CPU par noeud	Intel® Xeon® E5-4650L (x1024)	Intel® Xeon® E5-2670v3 (x24)	Intel® Xeon® E5-2620v3 (x12)	16	28
RAM par noeud	16 To	128 Go	256 Go	500 Go	250 Go
Cartes graphiques (GPU)	Non		Nvidia Quadro K5200 (x2)	Non	

Annexe 4 : Données du test de χ^2

Annexe 4 : Nombre total de séquences annotées et non annotées pour chaque phylum présent dans le jeu de données. Données utilisées dans pour un test de χ^2

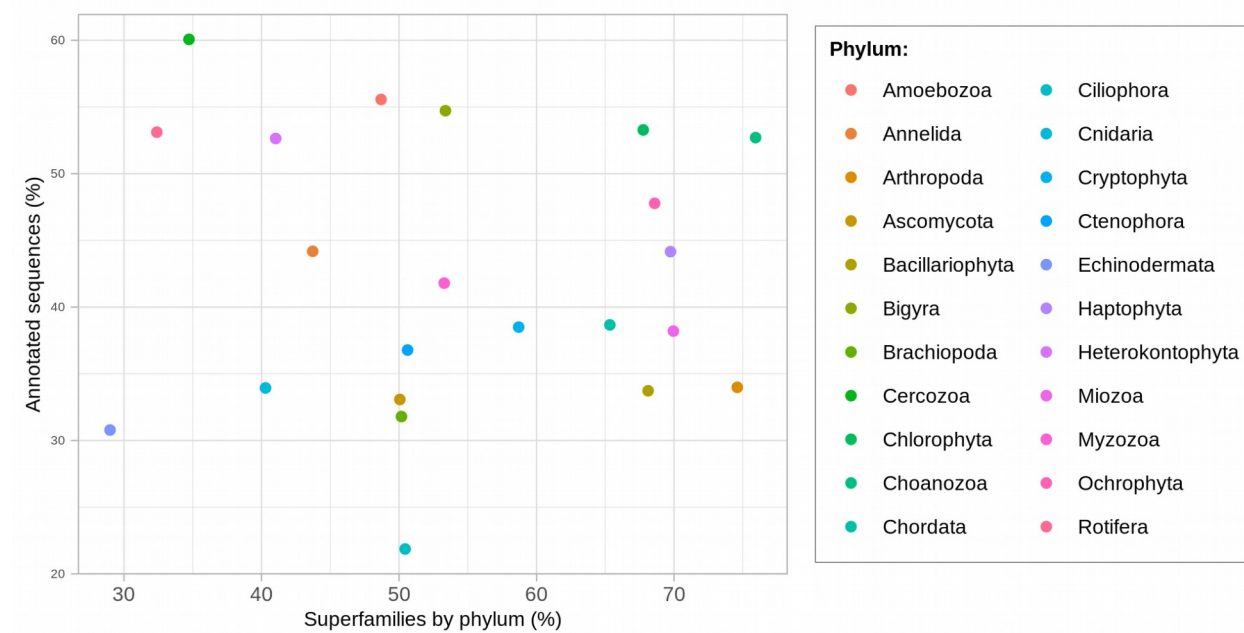
Taxonomie <i>Kingdom</i>	Taxonomie <i>Phylum</i>	Nombre total de séquences annotées	Nombre total de séquences non annotées
Animalia	Annelida	9 713	12 271
Animalia	Arthropoda	1 058 075	2 055 336
Animalia	Brachiopoda	15 714	33 708
Animalia	Chordata	197 987	314 070
Animalia	Cnidaria	9 804	19 083
Animalia	Ctenophora	18 982	32 627
Animalia	Echinodermata	6 257	14 068
Animalia	Rotifera	3 158	2 788
Chromista	Bacillariophyta	296 946	583 537
Chromista	Bigyra	18 709	15 481
Chromista	Cercozoa	4 527	3 009
Chromista	Ciliophora	64 393	230 104
Chromista	Cryptophyta	67 383	107 647
Chromista	Haptophyta	717 090	906 940
Chromista	Heterokontophyta	5 047	4 541
Chromista	Miozoa	202 105	326 968
Chromista	Myxozoa	17 297	24 089
Chromista	Ochromophyta	293 172	320 424
Fungi	Ascomycota	9 487	19 195
Plantae	Chlorophyta	281 316	246 694
Protozoa	Amoebozoa	10 227	8 179
Protozoa	Choanozoa	543 662	487 848

Annexe 5 : Test de corrélation de Pearson

Annexe 5.A : Tableau de données utilisé pour le test de corrélation de Pearson.

Phylum	Pourcentage du nombre de superfamilles	Pourcentage du nombre de séquence annotées
Amoebozoa	48.69745%	55.56340%
Annelida	43.72410%	44.18213%
Arthropoda	74.60036%	33.98443%
Ascomycota	50.05921%	33.07649%
Bacillariophyta	68.11723%	33.72535%
Bigyra	53.37478%	54.72068%
Brachiopoda	50.17762%	31.79556%
Cercozoa	34.72469%	60.07166%
Chlorophyta	67.76199%	53.27854%
Choanozoa	75.93250%	52.70545%
Chordata	65.33452%	38.66503%
Ciliophora	50.44405%	21.86542%
Cnidaria	40.29011%	33.93914%
Cryptophyta	58.70337%	38.49797%
Ctenophora	50.62167%	36.78041%
Echinodermata	28.98165%	30.78475%
Haptophyta	69.74541%	44.15497%
Heterokontophyta	41.03020%	52.63872%
Miozoa	69.95263%	38.19983%
Myxozoa	53.28597%	41.79433%
Ochromyxa	68.59088%	47.77932%
Rotifera	32.38603%	53.11134%

Annexe 5.B : Graphique de corrélation entre le pourcentage de séquences annotées et le pourcentage de superfamilles CATH détectées dans chaque phylum.

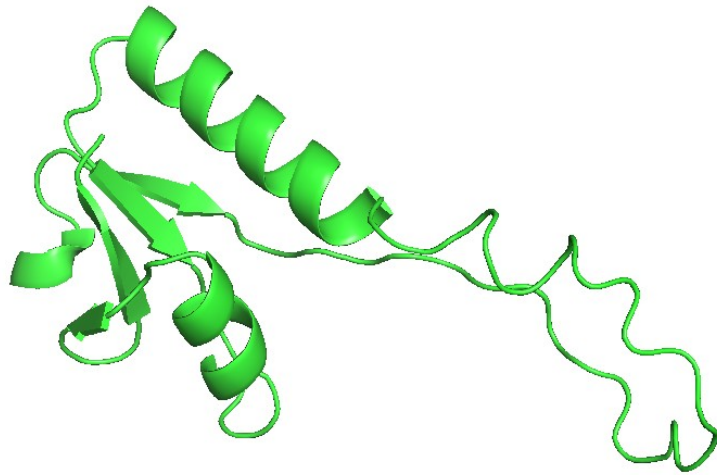


Annexe 6 : Structures tridimensionnelles

Annexe 6 : Structure tridimensionnelle deux protéines prédite avec AlphaFold ([Jumper et al., 2021](#))

Protéine 1 : TARA_SOC_28_MAG_00063_000000007425.10.1 (Plantae / Choanozoa)
z-score : 5.73 ; pLDDT : 85.37

Structure inconnue

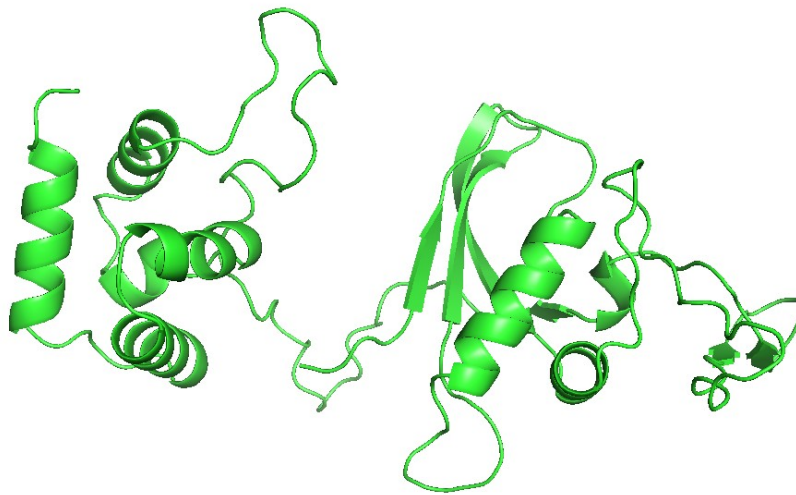


Protéine 2 : TARA_SOC_28_MAG_00063_000000008836.1.3 (Plantae / Choanozoa)
z-score : 19.03 ; pLDDT : 89.95

Structure connue

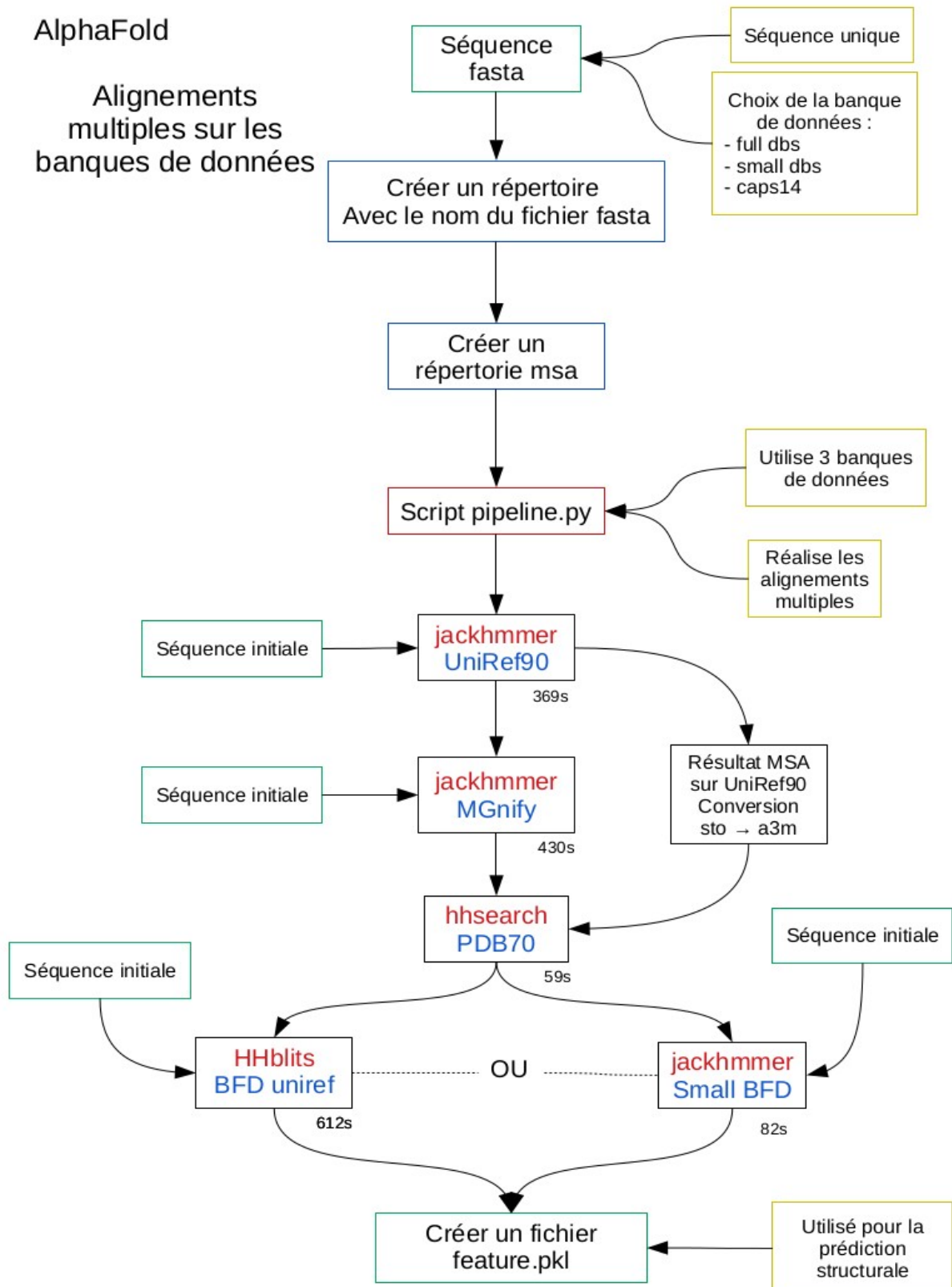
Fonction : Protéine de liaison

Identifiant PDB : 6CDD



Annexe 7 : AlphaFold – alignements multiples

Annexe 7 : Schéma représentant la succession des différents outils et banques de données utilisées par AlphaFold pour réaliser les alignements multiples (correspond à la première partie).

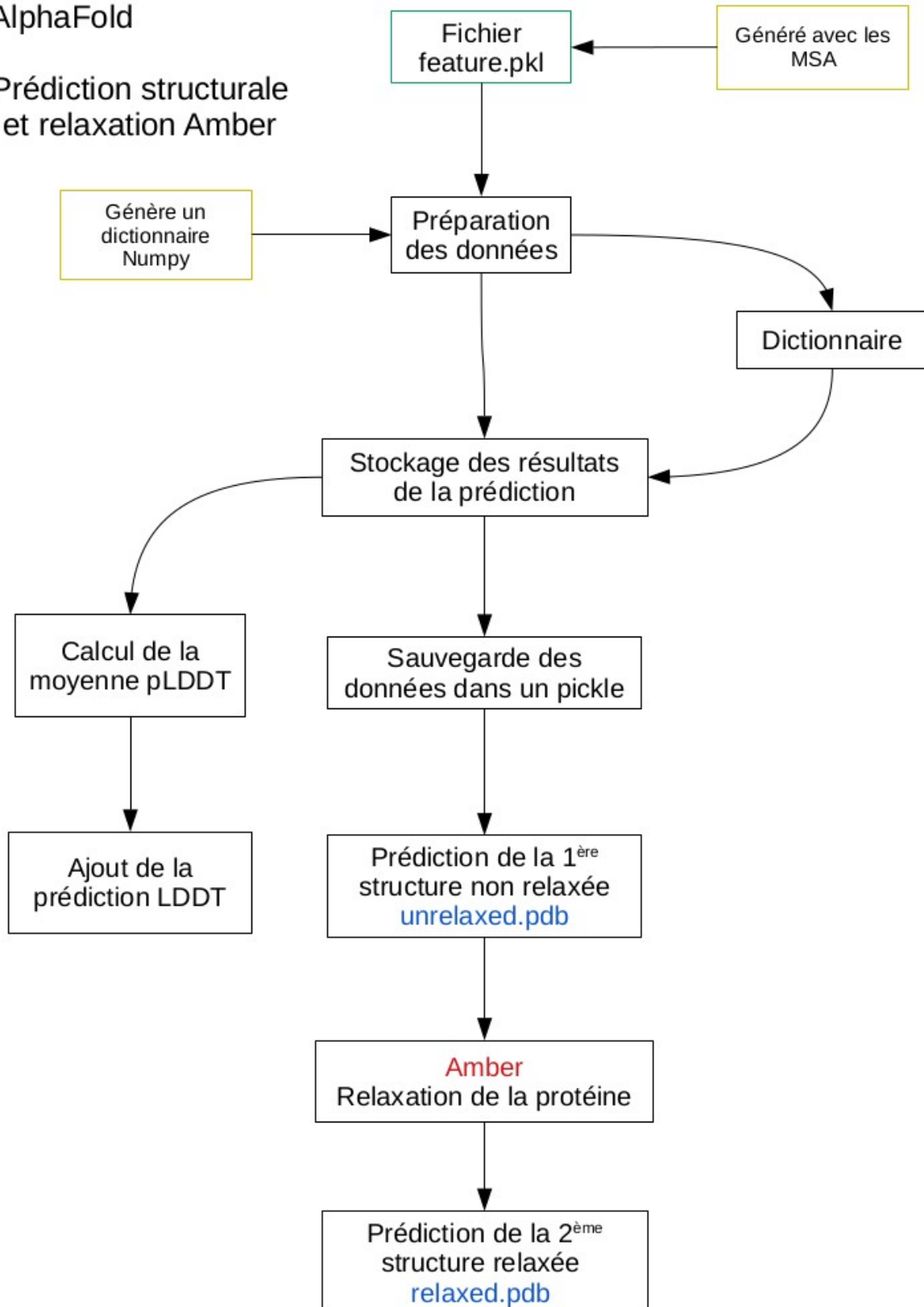


Annexe 8 : AlphaFold – prédiction structurale

Annexe 8 : Schéma des différentes étapes d'AlphaFold pour la prédiction structurale (correspond à la deuxième partie)

AlphaFold

Prédiction structurale et relaxation Amber



Annexe 9 : Définitions

- ❖ **HMM** : Hidden Markov Model
- ❖ **MAGs** : Metagenome-Assembled Genome
- ❖ **SAGs** : Single Cell Amplified Genome
- ❖ **Unigènes** : Gène ou morceaux de gène, car ils sont parfois plus longs ou plus courts qu'un gène. Ici un gène est défini par une séquence incluse entre un codon start et un codon stop

Résumé

Annotation structurale de séquences protéiques issue de métagénomes marins par recherche de profils HMM, et évaluation d'une approche de prédiction structurale par *deep learning*

Les micro-organismes eucaryotes sont abondants et primordiaux dans le fonctionnement des écosystèmes marins. Leurs génomes et transcriptomes présentent en moyenne près de 50% de séquences sans annotation fonctionnelle. Le but de ce travail a pour but d'annoter fonctionnellement des séquences à partir d'outils "classiques" et d'évaluer si l'annotation structurale apporte de nouvelles informations. À partir d'un fichier fasta contenant 10 207 435 séquences protéiques issues de l'expédition Tara Océans 2009-2013, avec le logiciel HMMER, nous avons cherché à les annoter en utilisant les profils HMM présent dans la classification CATH/*Gene3D*. Avec cette première approche, environ 40% des séquences ont pu être annotées à une superfamille de protéines ; l'utilisation de Pfam, en plus de CATH/*Gene3D*, permet d'annoter 50 % des séquences. La majorité des séquences restant non annotées sont de petite taille (inférieur à 100 acides aminés). Afin de pouvoir annoter les séquences pour lesquelles aucune superfamille n'a été attribuée, nous avons testé une approche de *deep learning* (AlphaFold) sur une sous-partie des séquences non annotées. Deux étapes ont permis de réduire le nombre de séquences à analyser : CD-HIT (regroupement des séquences) puis SPOT-Disorder-Single (recherche des résidus désordonnés). AlphaFold a été utilisé sur 93 séquences en raison du temps de calcul élevé. Avec cette méthode, nous pouvons annoter la structure des protéines avec celle présente dans RCSB PDB, grâce au programme YAKUSA. Il y a encore des difficultés pour appliquer AlphaFold, mais il nous permet d'apporter des informations supplémentaires à celles de programme comme HMMER.

Mots clés : HMMER ; AlphaFold ; CATH/*Gene3D* ; génomique environnementale ; structure tridimensionnelle

Abstract

Structural annotation of protein sequences from marine metagenomes by HMM profile search, and evaluation of a structural prediction approach by *deep learning*

Eukaryotic microorganisms are abundant and essential for the functioning of marine ecosystems. Their genomes and transcriptomes present on average nearly 50% of sequences without functional annotation. The aim of this work is to functionally annotate sequences using "classical" tools and to evaluate if structural annotation brings new information. From a fasta file containing 10,207,435 protein sequences from the Tara Oceans 2009-2013 expedition, with the HMMER software, we sought to annotate them using the HMM profiles present in the CATH/*Gene3D* classification. With this first approach, about 40% of the sequences could be annotated to a protein superfamily; the use of Pfam, in addition to CATH/*Gene3D*, allows to annotate 50% of the sequences. The majority of the remaining unannotated sequences are small (less than 100 amino acids). In order to be able to annotate the sequences for which no superfamily has been assigned, we tested a *deep learning* approach (AlphaFold) on a subpart of the unannotated sequences. Two steps were used to reduce the number of sequences to be analyzed: CD-HIT (clustering of sequences) and SPOT-Disorder-Single (search for disordered residues). AlphaFold was used on 93 sequences because of the high computation time. With this method, we can annotate the protein structure with the one present in RCSB PDB, thanks to the YAKUSA program. There are still difficulties to apply AlphaFold, but it allows us to bring additional information to those of programs like HMMER.

Keywords : HMMER ; AlphaFold ; CATH/*Gene3D* ; environmental genomics; three-dimensional structure