Évolution de la thermophilie au sein de la lignée des annélides polychètes Alvinellidae

Pierre-Guillaume Brun

Sorbonne Université

École doctorale 227
Sciences de la nature et de l'Homme : évolution et écologie
UMR7144 Adaptation et Diversité en Milieu Marin
Equipe Dynamique de la diversité marine - Station Biologique de Roscoff

Evolution de la thermophilie au sein de la lignée des annélides polychètes Alvinellidae

Par Pierre-Guillaume BRUN

Thèse de doctorat de biologie évolutive

Sous la direction de Jean MARY

Présentée et soutenue publiquement le 29 avril 2024

Devant un jury composé de :

Sarah SAMADI, Professeure	MNHN	Présidente
Céline BROCHIER-ARMANET, Professeure	Université Claude Bernard Lyon	Rapportrice
Bruno FRANZETTI, Directeur de recherche	CNRS IBS Grenoble	Rapporteur
Jean MARY, Maître de Conférences	Sorbonne Université	Directeur de thèse
Didier JOLLIVET, Directeur de recherche	CNRS Station Biologique de Roscoff	Invité
Elodie LAINE, Professeure	Sorbonne Université	Invitée

Remerciements.

Je tiens en premier lieu à remercier mes directeurs de thèse, Jean Mary et Didier Jollivet, pour m'avoir accompagné tout au long de ces trois années (et même un peu plus!). Vous m'avez épaulé sur ce projet tout en me laissant la liberté d'explorer des questions plus périphériques mais tout aussi passionnantes. Il me reste encore beaucoup à apprendre, mais il faut bien commencer quelque part!

Je remercie également les membres du comité de thèse, Frédéric Partensky, Mirjam Czjzek et Manolo Gouy, pour les conseils prodigués et les encouragements ainsi que les membres du jury de thèse, Sarah Samadi, Céline Brochier-Armanet, Bruno Franzetti et Elodie Laine, pour la relecture et la critique de mon travail qui profitera grandement d'un point de vue extérieur.

Merci à tous les collaborateurs du projet, en particulier Anne-Sophie Le Port pour la patience et le travail déployé avec ces trop nombreux litres de culture bactérienne, Lionel Cladière pour m'avoir formé sur la purification de protéines, et Murielle Jam pour avoir pris le temps de corriger mes bêtises avec "délicatesse", bien que je pense rester éloigné de ces machines diaboliques dorénavant.

Enfin, un grand merci à Marie-Noëlle Le Meur et France Hall pour leur disponibilité lors de mes nombreux allés-retours les bras remplis de bidons en déséquilibre, et d'avoir rendu possible mon utilisation non raisonnée de tous les récipients gradués qui me tombaient sous la main.

Parce qu'il y a un monde plus loin que le bout du monde, un grand merci à Sébastien Brûlé et Magali Dumont-Nicaise pour leur réactivité et leur expertise. J'ai été très bien accueilli et ai pu découvrir de nouvelles techniques. Je garde un petit béguin pour le spectromètre de masse, c'est fabuleux.

Merci à Stéphane Hourdez pour m'avoir introduit au monde de l'hydrothermal en master. Tenté qu'on fait un détour par le master, merci à Tom Weihmann. Vielen Dank, Tom. Ich hatte eine sehr gute Zeit während des Masterpraktikums und danke dir für deine Unterstützung mit meiner Bewerbung für diese Doktorat. Entschuldigung für die Kapitel auf Frenzisch, aber ich bin glücklich über die Arbeit die sich eher in Richtung der molekularen Biomecanik bewegt, und das ist auch ein bisschen dein Verdienst! Ich lese weiter deine Veröffentlichungen. Bis dann!

~

Passons aux amis! Désolé de faire si court, mais je soutiens dans trois heures. Pédagie! Pour le soutien durant ces dernières semaines, ton rire agaçant et le temps passé en ta compagnie. Je serai amende honorable!

Louise Fouqueau PhD et Yasmine Lund-Ricard, beaucoup de discussions et conversations, franches R.I.G.O.L.A.D.E.S. J'espère que nos chemins se recroiseront souvent malgré la boycotte les uns et les autres.

Romeo Stetsenko, la force tranquille. Toujours curieux et enthousiaste, un vrai geek qu'on veut compter parmi ses amis.

Alex le Moan. Que dire d'Alex? Son charme, son magnétisme, son esprit. Je n'ai pas le temps de développer, mais merci de m'avoir supporté dans mes radotages scientifiques, de m'avoir montré les triangles, et ouvert les portes de cette belle terre bretonne imprégnée de légendes et imbibée de bière.

Vincent Monchi, les jeux, les soirées, les discussions. Je te souhaite vraiment de réussir et au plaisir de se revoir baignés de gloire intellectuelle.

Louison Dufour pour son attitude impayable "Ok, on fait comme ça". Antonin Chevenier, petit farceur qui m'a enlevé quelques épines du pied.

Coral & Carl, late addition to the crew but a good one. We'll see very soon for Shlingui's un-burial and some Vampire week-ends. Can't wait!

Etienne et Yacine, qui ont "retourné la baie d'Halong". Nuit mémorable au moment où j'avais besoin de vider les batteries.

PG Blond, mon alter ego plus aventurier, bien que moins beau. A cet été sur ton bateau! Le Fondoir de l'Océan. Paul Mulot, avec qui il n'y a jamais de problème, Solène Marcel avec qui il faut toujours rester sur le qui-vive Merci pour tout le temps passé ensemble, tous les coups de mains présents et -bisdé- les nombreux à venir!

Jérome Coudret, Claire Daguin, et bien sûr Thierry Comlet et Thomas Broquet pour les échanges scientifiques mais surtout amicaux, entre Björk et Supertramp. Morgane Guillem également, même si elle passe maintenant sa vie en mer, qui reste un membre de ce bout du couloir avec qui j'ai passé tant de bons moments.

Pour finir, merci au Snoop et aux membres de la loge maçonnique de Quimper (en particulier Ed, Dudu, Alex et P-E). Je suis heureux d'avoir été accepté dans cette grande famille. On les aura, ces Reptiliens!

Merci à tous,
Pierre-Guillaume

Titre : Evolution de la thermophilie au sein de la lignée des annélides polychaetes Alvinellidae.

Mots clefs : thermophilie ; sources hydrothermales ; Alvinellidae ; Evolution moléculaire ; ASR ; Biologie structurelle

Résumé : Les Alvinellidae (annélides polychètes terebellides) constituent une famille d'espèces endémiques des sources hydrothermales profondes, dispersées entre l'océan Pacifique et Indien. Depuis leur découverte avec l'espèce emblématique *Alvinella pompejana*, le ver de Pompéi, ces animaux ont suscité l'intérêt de la communauté scientifique. En effet, si les sources hydrothermales constituent des environnements réputés extrêmes (gradients de température, absence de photosynthèse, anoxie du milieu, présence de divers métaux et sulfides issus de la percolation du fluide hydrothermale dans la croûte basaltique, pH acide), les Alvinellidae sont parvenus à coloniser des niches écologiques variées et montrent une grande diversité morphologique, physiologique et génétique, inter et intra-espèces. Dans le cadre de cette thèse, nous nous sommes plus particulièrement intéressés aux adaptations permettant à ces vers de faire face à des régimes thermiques contrastés. *A. pompejana*, par exemple, est thermophile, survivant à des températures proches de 50°C. D'autres espèces en revanche, comme *Paralvinella grasslei*, sont considérées psychrophiles, vivant à distance des cheminées hydrothermales à des températures entre 10 et 25°C. Plus spécifiquement, nous avons étudié l'acquisition de la thermophilie/psychrophilie au cours de l'évolution de la lignée, en essayant de répondre à la question du phénotype thermique de l'ancêtre des Alvinellidae. Pour cela, nous avons établi la phylogénie moléculaire des Alvinellidae, sur la base des données moléculaires transcriptomiques récupérées pour onze des quatorze espèces de la famille au cours de plusieurs campagnes scientifiques. Ce premier résultat amène à conclure à un ancêtre datant de la fin du Crétacée (entre 60 et 90 millions d'années), déjà présent dans les sources hydrothermales du Pacifique Est. La radiation des Alvinellidae à cette époque a été rapide, en quelques millions d'années, aboutissant à l'apparition de plusieurs espèces présentant de forts taux de tri incomplet de lignée et d'introgression interspécifique. Les résultats de cette phylogénie nous ont permis d'établir le modèle permettant de construire des propositions statistiques de protéines appartenant aux ancêtres de la lignée. Trois protéines ont été choisies, à savoir la malate déshydrogénase cytosolique, la superoxyde dismutase Cu/Zn et une hémoglobine intracellulaire, pour être reconstruites, exprimées et expérimentalement caractérisées. En effet, pour des organismes ectothermes comme les Alvinellidae, il est attendu que les protéines des espèces thermophiles soient en moyenne plus stables thermiquement que les protéines issues des espèces psychrophiles. Ces reconstructions ancestrales nous ont permis de conclure que l'ancêtre de la lignée était un ver déjà adapté aux environnements chauds, et que la psychrophilie de certaines espèces de la lignée est un caractère dérivé acquis plus récemment. Enfin, dans une dernière partie, je me suis intéressé à l'optimisation des modèles de reconstruction des séquences protéiques ancestrales. Ces modèles sont basés sur la diversité des séquences contemporaines et leurs relations phylogénétiques. J'ai essayé d'implémenter ces approches en utilisant deux types d'informations supplémentaires : celles liées aux événements d'insertions/délétions de séquence, et celles concernant l'évolution de la structure secondaire des protéines et la variabilité temporelle des fréquences attendues des résidus aux différentes positions des protéines. Je montre que l'introduction de ces deux derniers types de paramètres dans les méthodes ASR est bénéfique et aboutit à des modèles ayant de meilleures vraisemblances. Toutefois, l'optimisation de ces modèles, nécessairement probabilistes, ne garantit pas un meilleur résultat pour l'expérimentateur, et les limites de ces modèles à estimer l'incertitude des séquences ancestrales inférées sont discutées.

Title : Evolution of thermophily in the Alvinellidae (Annelida : Polychaeta).

Keywords : Thermophily ; Hydrothermal vents ; Alvinellidae ; Molecular evolution ; Structural biology

Abstract : The Alvinellidae (Annelida : Terebelliformiaterebellid) are a species family endemic to deep hydrothermal vents from the Pacific and Indian Oceans. Since the discovery of the emblematic species *Alvinella pompejana*, the Pompeii worm, these animals have aroused the interest of the scientific community. Although hydrothermal vents are extreme environments (strong temperature gradients, absence of photosynthesis, anoxy, presence of various metals and sulphides due to the percolation of hydrothermal fluid into the basaltic crust, acid pH), the Alvinellidae have managed to colonise a variety of ecological niches and show great morphological, physiological and genetic diversity, both between and within species. In this thesis, we were notably interested in the adaptations that enable these worms to cope with contrasting thermal regimes. *A. pompejana*, for example, is thermophilic, surviving at temperatures close to 50°C. Other species, however, such as *Paralvinella grasslei*, are psychrophilic, living further from hydrothermal chimneys at temperatures between 10 and 25°C. More specifically, we studied the acquisition of thermophilia/psychrophilia during the evolution of the lineage, in an attempt to characterize the thermal phenotype of the ancestor of the Alvinellidae. To this end, we have established the molecular phylogeny of the Alvinellidae, based on molecular transcriptomic data recovered for eleven of the fourteen species in the family during several scientific campaigns. This initial result points to an ancestor dating from the end of the Cretaceous (between 60 and 90 million years ago), already present in the hydrothermal vents of the eastern Pacific. The radiation of the Alvinellidae was a quick event, within a few million years, resulting in several species with high rates of incomplete lineage sorting and showing traces of high interspecific introgression. The results of this phylogeny enabled us to establish a model to construct statistical proposals of proteins belonging to the ancestors of the lineage. Three proteins were chosen, namley the cytosolic malate dehydrogenase, the Cu/Zn superoxide dismutase and an intracellular hemoglobin, for reconstruction, expression and experimental characterisation. For ectothermic organisms such as the Alvinellidae, proteins from thermophilic species are expected to be on average more stable at high temperatures compared to their counterparts from psychrophilic species. These ancestral reconstructions allowed us to conclude that the ancestor of the lineage was a worm that was already adapted to warm environments, and that psychrophily of modern-day alvinellid species is a derived character acquired more recently. Finally, I looked at the optimisation of models for reconstructing ancestral protein sequences. These models are based on the diversity of contemporary sequences and their phylogenetic relationships. I tried to implement these approaches using two types of additional information : those linked to sequence insertion/deletion events, and those regarding the evolution of secondary structures of proteins and temporal variability of the expected frequencies of residues at different protein positions. I show that the introduction of these last two types of parameters into ASR methods is beneficial and leads to models with better likelihoods. However, the optimisation of these models, which are necessarily probabilistic, does not guarantee a better result for the experimenter, and the limits of these models to estimate the uncertainty of the inferred ancestral sequences are discussed.

Table des matières

Liste des figures

Liste des tableaux

Chapitre I

Introduction

I.1 Les Alvinellidae, une famille de polychètes marins endémiques des sources hydrothermales des océans Pacifique et Indien

I.1.1 Le milieu hydrothermal, un environnement extrême et très fluctuant.

Les sources hydrothermales constituent des environnements hors du commun et ont été découvertes au niveau de la faille des Galapàgos grâce au sous-marin Alvin en 1977 (P. Lonsdale, 1977). Depuis lors, des centaines de sites hydrothermaux actifs ont été répertoriés (Beaulieu et al., 2013), au niveau des bassins arrière-arc ou des failles médio-océaniques. Situés entre 1000 et 4000 mètres de profondeur, ces systèmes sont différents de par leur âge (généralement de quelques millions d'années *vs.* plusieurs dizaines de millions d'années, Maria Seton et al., 2020 ; Boulart et al., 2022 ; Artemieva, 2023), mais sont systématiquement caractérisés par la formation et le fonctionnement d'une chambre magmatique quelques kilomètres sous le plancher basaltique liés à la décompression du plancher océanique (Barnes, 1991). La croûte sous-jacente fracturée, illustrée en figure I.1, permet la circulation d'eau de mer profonde qui, au contact des chambres, est expulsée à des températures très élevées (jusqu'à 400°C), devient acide (pH compris entre 2 et 4), anoxique (réduction de O_2 par les sulfures et le Fe II) et riche en CO_2 (entre 3,5 et 6 mM), chargée de sulfures d'hydrogène (concentration en H_2S supérieure à 300 μM) et de métaux (concentrations de plusieurs ordres de grandeur supérieurs aux conditions côtières et d'estuaires) (Le Bris et Françoise Gaill, 2007 ; Stéphane Hourdez et Jollivet, 2020). Selon le débit du fluide et son mélange sous la surface, les émissions peuvent être soit diffuses et lentes, étendues sur de grandes zones, soit concentrées et intenses (plusieurs mètres par seconde). Dans ce cas, le fluide se charge de métaux depuis la croûte basaltique, et précipite sous la forme de fumeurs noirs (cheminées hydrothermales), édifices de sulfures polymétalliques composés majoritairement de fer, sulfures de zinc et cuivre, selon la nature de la

roche basaltique (Le Bris et Françoise Gaill, 2007). Le mélange de ce fluide avec l'eau des fonds marins est chaotique dans le temps et l'espace, entraînant des variations importantes des propriétés physico-chimiques de l'eau en quelques minutes et à quelques centimètres d'écart (Le Bris et Françoise Gaill, 2007 ; Stéphane Hourdez et Jollivet, 2020). En outre, l'activité volcanique et tectonique propre à ces milieux rend les sources très éphémères, et certains sites peuvent apparaître et être détruits en quelques décennies (Le Bris et Françoise Gaill, 2007).

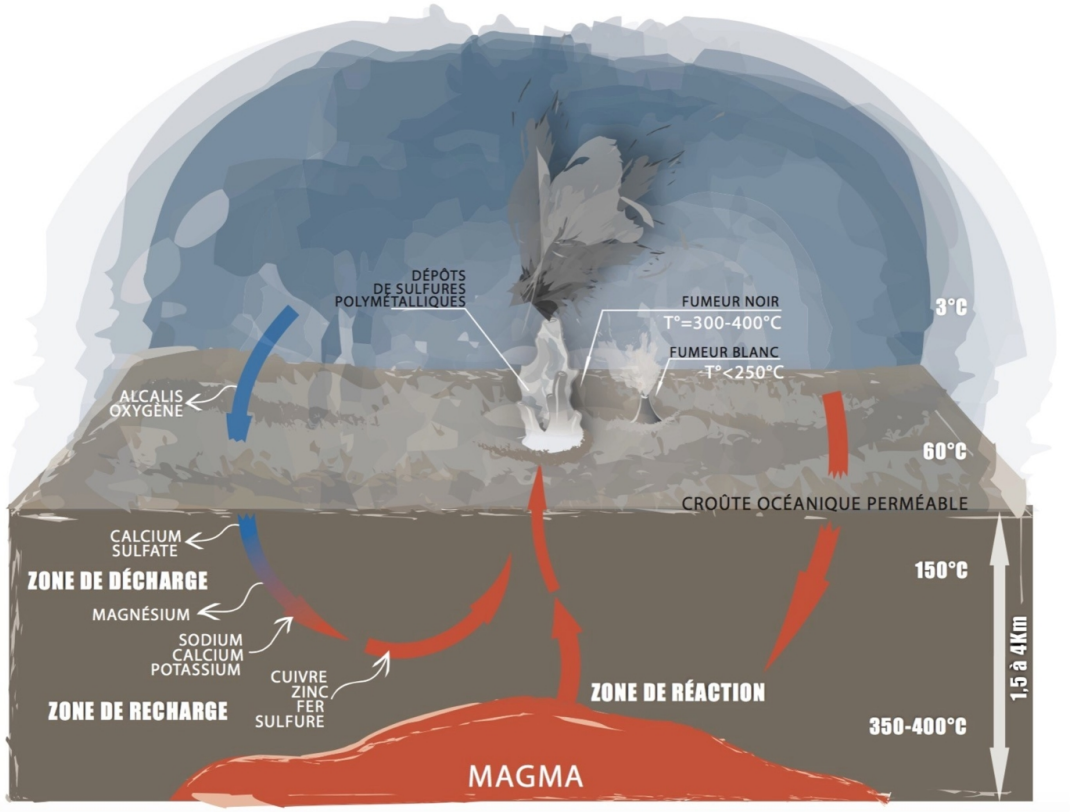


Figure I.1 – Echanges entre l'eau de mer profonde et la croûte océanique au niveau d'une dorsale océanique. L'eau de mer, froide, percole au sein des fractures de la croûte océanique et se charge en divers éléments minéraux et métalliques (cuivre, zin, fer, souffre). L'eau, anoxique et chauffée par les chambres magmatiques sous la dorsale, est ensuite émise au niveau des sources hydrothermales. Illustration Capsule Graphik 2014.

Au contraire des plaines abyssales, les sources hydrothermales sont réputées pour abriter une biomasse importante d'espèces chimiosynthétiques (Hourdez et Weber, 2005) situées au delà de la zone photique où la photosynthèse est possible (200 premiers mètres de la colonne d'eau). Ces espèces sont considérées abondantes, mais faiblement diversifiées, ce qui s'expliquerait par l'effet antagoniste entre des conditions extrêmes (anoxie, concentration élevées en sulfures et métaux, gradients importants de température) et une production primaire riche, basée sur l'oxydation de nutriments réduits (CH_4, H_2S) par des bactéries chimiosynthétiques (Tunnicliffe, Juniper et Sibuet, 2003). Ainsi, 95% des espèces des sources hydrothermales sont endémiques de ces milieux extrêmes.

Preuve des conditions extrêmes rencontrées par certains organismes au niveau des sources hydrothermales, certains procaryotes décrits au niveau des fumeurs noirs peuvent croître à des températures jusqu'à 122°C, ce qui les classe parmi les organismes les plus thermophiles décrits à ce jour (Takai et al., 2008). Concernant les animaux, toutefois, la plupart des espèces des sources (gastéropodes, moules, vers) vivent à proximité des émissions de fluide chaud et préfèrent des températures relativement faibles entre 5 et 15°C (Ravaux et al., 2013), le gradient thermique entre fluide chaud et eau de fond froide étant de l'ordre du décamètre. Certaines espèces peuvent même être considérées comme véritablement froides sténothermes, dites psychrophiles, évoluant à des températures inférieures à 4°C. L'étude enzymatique de la malate déshydrogénase ainsi que l'affinité de l'hémocyanine pour l'oxygène d'un ensemble d'espèces hydrothermales démontrent cependant que les espèces hydrothermales même froides sont globalement moins sensibles aux variations importantes de température, comparées à des espèces proches des milieux marins profonds (Dahlhoff et Somero, 1991 ; Sanders, Arp et Childress, 1988 ; Lallier et Truchot, 1997 ; Chausson, Bridges et al., 2001 ; Chausson, Sanglier et al., 2004). Ces résultats suggèrent un caractère eurythermal plus prononcé propre aux espèces hydrothermales, qui seraient capable de soutenir des températures plus élevées pendant de courts laps de temps.

En effet, les émissions de fluide hydrothermal imposent à ces espèces des conditions environnementales fluctuantes, en particulier pour les espèces vivant au plus prêt des émissions (Monaco et Prouzet, 2015). Les conditions chimiques réelles auxquelles sont confrontés les organismes hydrothermaux sont par ailleurs difficiles à caractériser précisément. La figure I.2 illustre plusieurs paramètres physico-chimiques mesurés par Le Bris, Zbinden et Françoise Gaill, 2005 au niveau des tubes d'une colonie d'*Alvinella pompajena*. Le dispositif est constitué d'un thermomètre associé à une sonde pH, ce qui permet de déterminer si le tube est intègre ou s'il est détérioré, entraînant des fuites entre le milieu extérieur et le tube. Les auteurs ont constaté que les conditions de pH et de température mesurées à l'intérieur des tubes et dans la matrice déviaient d'un processus conservatif entre l'eau de mer et le fluide émis depuis la roche, "end-fluid member". Ceci démontre que dans le cas de cette espèce, le micro-environnement dans lequel baigne *A. pompejana* n'est pas directement le résultat du mélange des deux fluides. L'eau issue de la percolation hydrothermal représente moins de 10% du mélange, et les températures importantes relevées à l'intérieur du tube seraient plutôt le fait de conduction thermique. Outre ces micro-environnements, la composition des deux fluides est elle-même fortement variable. Ainsi, la saturation des espèces chimiques (fer, zinc) et les conditions oxydo-réductrices du milieu influent sur les concentrations en métaux libres dans l'eau par rapport à leur forme complexée, mais dépend principalement de la composition de la roche basaltique qui nourrit le fluide hydro-

thermal. Réciproquement, les concentrations en H_2S sont modulées par les concentrations en fer, et les équilibres chimiques prédisent une baisse des quantités de H_2S au profit de HS^- aux sites où le fer est le plus abondant (Le Bris et Françoise Gaill, 2007). De la même manière, le fluide hydrothermal riche en sulfure d'hydrogène tend à être plus anoxique, du fait de la réaction spontanée entre les deux espèces : $2\,H_2S + 3\,O_2 \longrightarrow 2\,H_2O + 2\,SO_2$ (Johnson et al., 1986 ; Stéphane Hourdez et Lallier, 2007). Par conséquent, certains paramètres physico-chimiques du milieu corrèlent avec les émissions de fluide hydrothermal, et les espèces vivant au plus proche des émissions, sur les cheminées, subissent des variations importantes de température, d'oxygène et de stress oxydatif qu'il est difficile de modéliser sans mesure directe.

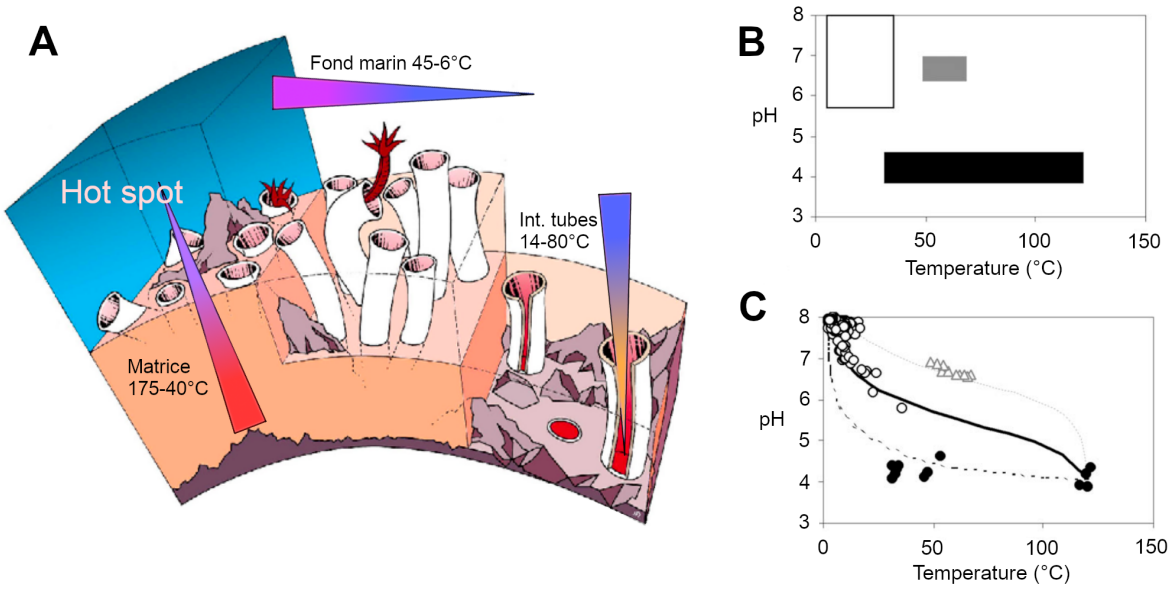


FIGURE I.2 – Mesures physico-chimiques de l'eau au niveau d'une colonie d'*A. pompejana*. (A) coupe de la colonie, montrant les variations importantes de température selon les points de mesure. (B) pH et température au dessous de l'ouverture des tubes (carré blanc), dans un tube (carré gris), dans la matrice autour des tubes (carré noir). (C) Mesures reportées avec une simulation géochimique d'un processus de mélange conservatif entre le fluide hydrothermal et l'eau de mer, ligne noire. Adapté de LE BRIS, ZBINDEN et Françoise GAILL, 2005.

I.1.2 Les Alvinellidae : un modèle d'étude exceptionnel des adaptations aux conditions extrêmes

Les Alvinellidae constituent une famille de quatorze espèces de polychètes endémiques des sources hydrothermales profondes des océans Pacifique et Indien, comme illustré en figure I.3. Elles sont toutes associées à la circulation hydrothermale, soit au niveau des émissions diffuses de fluide, soit directement sur les murs des cheminées. Ces espèces, pour la plupart, construisent des tubes ou des cocons qu'elles ne quittent que rarement, composés d'un mucus de polysaccharides et de glycoprotéines (Jollivet et Stéphane Hourdez, 2020 ; Han et al., 2021). Il est en outre probable que d'autres espèces soient encore à découvrir, notamment dans l'océan Indien où certaines populations d'Alvinellidae géographiquement éloignées ont été observées, sans toutefois avoir été décrites au niveau taxonomique (Nakamura et al., 2012). Les espèces actuellement décrites sont dispersées entre le Pacifique Est (ride Juan de Fuca et dorsale Est-Pacifique), les bassins arrière-arc du Pacifique Sud-Ouest et nord-est, ainsi que sur la dorsale indienne au niveau du site Wocan. Ces ensembles sont par conséquent isolés les uns des autres. En effet, les anciennes dorsales de Kula et au nord de la Nouvelle Guinée, qui ont été proposées pour avoir joué un rôle important dans la dispersion des espèces entre le Pacifique Est et Ouest (Hessler et P. F. Lonsdale, 1991 ; Bachraty, Legendre et Desbruyères, 2009), ne sont plus actives depuis environ 40 millions d'années (Smith, 2003), environ à la même période que l'ouverture de la mer de Chine et de la fermeture des océans Pacifique et Indien (Parker et Gealey, 1985 ; Moalic et al., 2012).

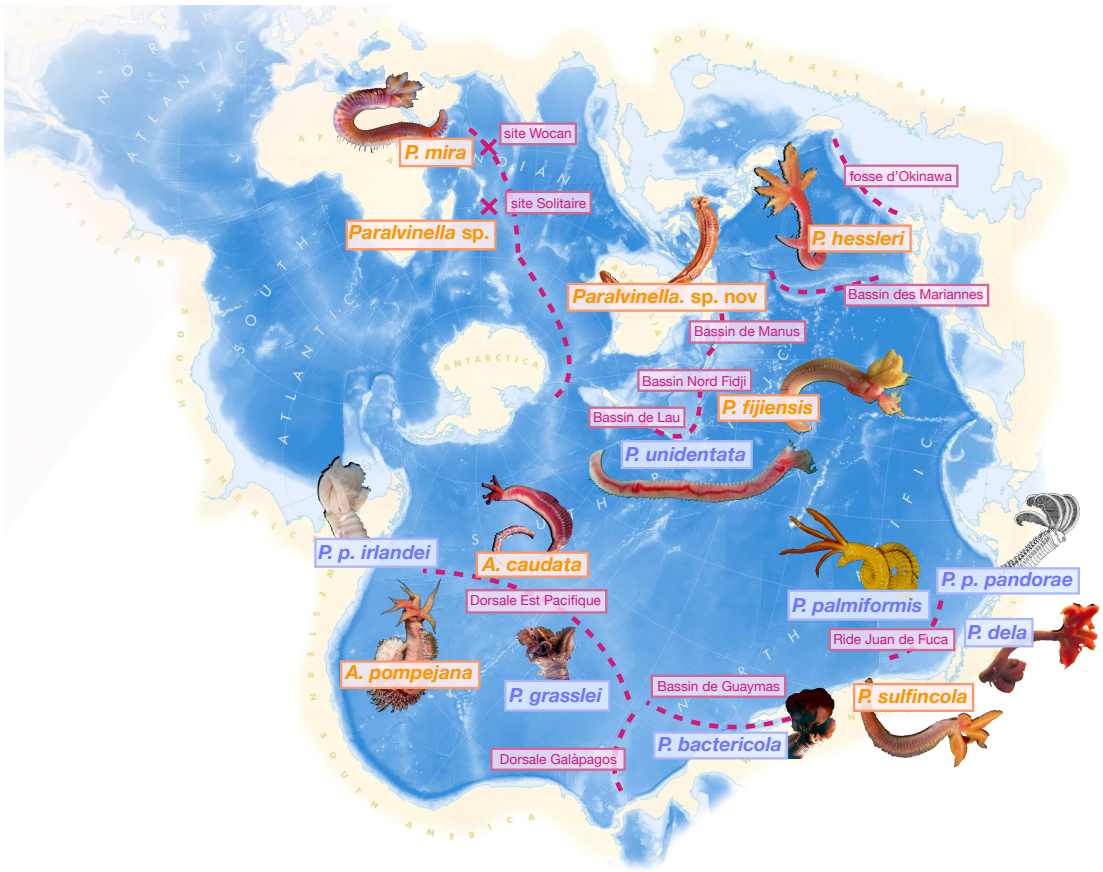


FIGURE I.3 – Répartition des espèces d'Alvinellidae entre les océans Pacifique et Indien. crédits photos : K. Onthank, Han et al. 2021, Briand IFREMER, Dugornay IFREMER, Deep-sea photography, S. Tsuchida, F. Pleijel, V. Tunnicliffe.

La famille des Alvinellidae est inclue au sein des Terebelliformia (voir figure I.4). Il s'agit d'une famille soeur des Ampharetidae, un large groupe de polychètes marins ayant également colonisé le milieu hydrothermal à de multiple reprises (Eilertsen et al., 2017). La vue actuelle est que les Alvinellidae forment bel et bien une famille séparée des Ampharetidae (Rousset, Rouse et al., 2003 ; Stiller, Tilic et al., 2020). Toutefois, l'hypothèse que les Alvinellidae serait une sous-famille des Ampharetidae ne peut pas être tout à fait écartée (Eilertsen et al., 2017, et la ré-analyse de certaines données moléculaires utilisées par Stiller, Tilic et al., 2020 donnent des résultats plus ambigus, voir chapitre 1). En effet, ces deux familles d'espèces partagent de nombreux traits morphologiques, dont la présence de tentacules buccaux filiformes et rétractiles ainsi que la présence de tores uncinigères ventraux et de soies dorsales simples. Ces critères avaient amené Desbruyères et Laubier à classer initialement *A. pompejana* et *Alvinella caudata* (cette dernière étant alors identifiée comme une forme ontogénique jeune de *A. pompejana*) parmi les Ampharetidae lors de leur découverte (Desbruyères et Laubier, 1980). Les auteurs, toutefois, considéraient déjà *A. pompejana* comme un Ampharetidae "aberrant" du fait de la fusion du *prostomium* et du *peristomium*, des modifications des notopodes sur les quatrième et cinquième segments, de la présence de notopodes et de soies dorsales sur tous les segments de l'individu jusqu'à son extrémité ainsi que de l'apparition des tores uncinigères au dixième segment seulement. La diversité morphologique très importante connue au sein des Ampharetidae montre cependant d'autres exemples de formes aberrantes (Day, 1964 ; Desbruyères et Laubier, 1980) et relativement peu de données moléculaires haut-débit (comme des transcriptomes) sont pour l'instant disponibles. Les Alvinellidae forment bien une famille d'espèces monophylétique, mais leur relation avec les Ampharetidae n'est pas tout à fait élucidé.

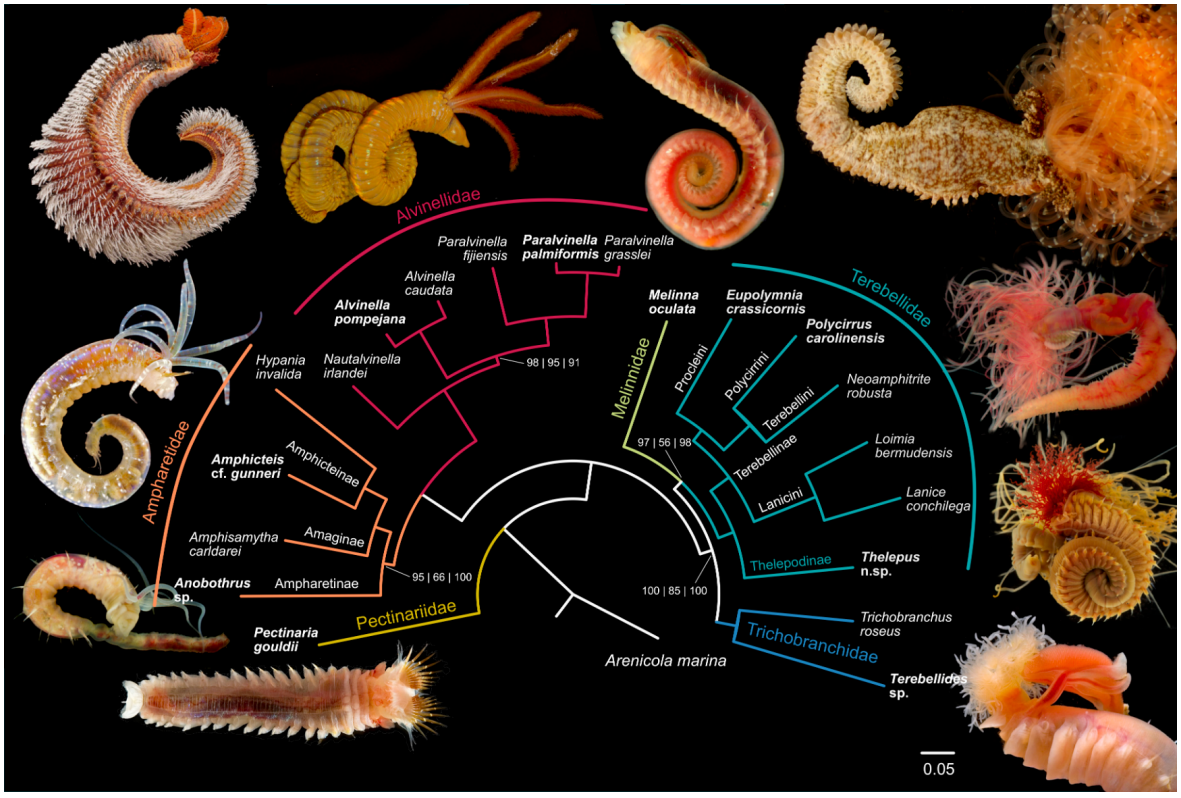


FIGURE I.4 – Phylogénie des Terebelliformia, tirée de STILLER, TILIC et al., 2020. Les espèces en gras sont illustrées par des photographies.

De manière similaire aux autres espèces hydrothermales, les Alvinellidae présentent de nombreuses d'adaptations spectaculaires pour faire face aux conditions extrêmes rencontrées dans ces environnements. Ces adaptations sont observables tant au niveau moléculaire qu'au niveau morphologique : augmentation de la surface des branchies et forte affinité pour l'oxygène des pigments respiratoires en réponse à l'hypoxie des milieux profonds, efficacité des voies métaboliques de détoxification des sulfides et des dérivés réactifs de l'oxygène (ROS), expression de protéines thermostables ainsi que de protéines chaperones et heat-shock en réponse aux conditions dénaturantes des hautes températures (Jollivet, Desbruyères et al., 1995 ; Le Bris et Françoise Gaill, 2007 ; Dilly et al., 2012).

Le ver de Pompéi, *A. pompejana*, est l'espèce ayant suscité l'intérêt le plus fort auprès de la communauté scientifique. Il s'agit du premier Alvinellidae découvert en 1979 sur le segment 21°N de la dorsale Est-Pacifique (EPR) accompagné d'une espèce proche alors assujettie à une forme juvénile de la même espèce, *A. caudata* (Desbruyères et Laubier, 1980). *A. pompejana* a été initialement décrite comme vivant à des températures supérieures à 60°C, subissant des pics de température avoisinant les 80°C (Stéphane Hourdez et Jollivet, 2020). Toutefois, les conditions réelles auxquelles est soumise *A. pompejana* sont probablement moins extrêmes. En effet, ce ver vit dans des tubes qu'il sécrète directement à la surface des cheminées hydrothermales, constitués à plus de 50% de protéines stables en conditions acides jusqu'à des températures de 100°C, ce qui constitue une première barrière permettant à ce ver de contrôler la température de son microenvironnement (voir fig. I.2 et photographie I.5). Ainsi, par l'étude de la conservation du pH et des concentrations en magnésium, on estime que l'eau de mer constitue entre 70 et 100% de l'eau trouvée à l'intérieur des tubes, tandis que l'eau circulant entre les tubes provient à 95% du fluide hydrothermal chaud et acide. Les colonies du ver de Pompéi agiraient comme des échangeurs thermiques au niveau des cheminées hydrothermales, et les fortes températures mesurées à l'intérieur des tubes sont attribuables à la conduction de chaleur à travers la roche de la cheminée (Le Bris, Zbinden et Françoise Gaill, 2005).

Figure I.5 – Photographie d'*A. pompejana* au niveau de la dorsale EPR. Sur la gauche de l'individu, on peut voir un tube inhabité à la texture parcheminée. L'animal quitte rarement son tube, mais il le ventile régulièrement en exposant sa partie antérieure. Il se nourrit probablement des protéobactéries ϵ symbiotiques qui croissent pour former les filaments blancs présents sur son dos. La vidéo complète est disponible sur le site du Monterey Bay Aquarium Research Institute.

En conséquence, les températures maximales que peut soutenir cet animal au delà de quelques minutes sont moindres. Ravaux et al., 2013 ont démontré qu'une exposition à plus de 50°C durant deux heures en aquarium pressurisé était déjà létale pour le ver. A l'issue de ce stress, les individus présentaient de graves dommages tissulaires et une surexpression de protéines heat-shock (HSP70). Néanmoins, l'*optimum* thermique du ver de Pompéi reste au delà de 42°C, ce qui le situe parmi les animaux les plus thermotolérants décrits à ce jour et proche de la limite UTL (upper thermal limit) estimée à 50°C pour les métazoaires (Pörtner, 2002). Alors que certaines espèces d'Alvinellidae telles *A. pompejana* ou *Paralvinella sulfincola* vivent entre 40 et 50°C, en association avec les cheminées hydrothermales (Le Bris et Françoise Gaill, 2007 ; Girguis et Lee, 2006), d'autres espèces comme *Paralvinella palmiformis* ou *Paralvinella grasslei* évitent d'être confrontées à des températures si élevées et préfèrent la vie loin des parois des cheminées, entre 3 et 30°C (Le Bris et Françoise Gaill, 2007 ; Girguis et Lee, 2006). Par conséquent, les Alvinellidae regroupent des espèces hydrothermales dont les températures de vie s'étendent sur l'ensemble du spectre envisagé pour la vie animale, et récapitulées dans le tableau I.1. Cette particularité permet une étude comparative des adaptations à la température au sein de cette famille de métazoaires le long de contrastes phylogénétiquement indépendants (Pörtner, 2002 ; Girguis et Lee, 2006 ; Jollivet, Mary et al., 2012).

Outre des régimes variés de température, les espèces d'Alvinellidae sont également exposées aux autres singularités du milieu hydrothermal. L'anoxie des sources est associée à un élargissement des branchies des Alvinellidae, traversées par un réseau de vaisseaux intra-épidermiques séparés de l'eau de mer par une distance très réduite entre 3 et 10 μm (Jouin et Françoise Gaill, 1990). Ces vers possèdent également un organe d'échange de gaz interne hypertrophié comparé à d'autres Terebellidae, contenant de nombreux erythrocytes qui permettent des échanges entre le système vasculaire et le compartiment cœlomique (Stéphane Hourdez, Lallier et al., 2000). En effet, les hémoglobines extracellulaires vasculaires et les hémoglobines intracellulaires cœlomiques possèdent des affinités très similaires et fortes pour l'oxygène, et cet organe d'échange est interprété comme ayant un rôle dans le stockage de l'oxygène puis son relargage lors d'épisodes d'anoxie afin de maintenir l'oxygénation de la tête et du cerveau. La forte affinité pour l'oxygène ($P_{50} < 0.3 mmHg$ pour *A. pompejana* et *A. caudata* contre des P_{50} comprises entre 1 et 100 $mmHg$ pour des espèces côtières) est contrebalancée par un effet Bohr très prononcé de l'hémoglobine (baisse de l'affinité pour l'oxygène à pH faible) permettant la dissociation de l'oxygène au niveau des tissus (Stéphane Hourdez et Jollivet, 2020 ; Jollivet et Stéphane Hourdez, 2020). Ces pigments peuvent donc *a priori* constituer une signature moléculaire de l'adaptation à l'hypoxie et potentiellement aux conditions des sources hydrothermales. D'autres enzymes, cette fois impliquées dans le métabolisme anaérobie, pourraient également présenter des activités différentes entre les Alvinellidae et d'autres espèces hydrothermales ou côtières comme la phosphofructokinase (Hand et Somero, 1983). En effet, chez *A. pompejana*, l'activité de la cytochrome c oxydase, bien que relativement faible, est présente, ce qui indique que les métabolismes aérobie et anaérobie sont tout deux présents chez l'espèce (Hand et Somero, 1983). Ceci pourrait être la conséquence de phases anoxiques prolongées du milieu qui engendrent une baisse de l'activité mitochondriale. Ces résultats nécessitent tout de même d'être nuancés et confirmés, car la réponse cellulaire au stress hypoxique ou hyperoxique dans les branchies d'*A. pompejana* est relativement faible, sans montrer de spécificité forte vers une modulation des métabolismes aérobique ou anaérobique Mary et al., 2010.

Enfin, les concentrations importantes en H_2S ainsi qu'en divers métaux (cuivre, fer, zinc, cadmium, mercure, manganèse) dans le fluide émis par les cheminées induisent un stress oxydatif important pour les espèces hydrothermales, à l'origine de la création de ROS qui peuvent engendrer des dommages à l'ADN, aux protéines et aux lipides constitutifs des membranes. (Stéphane Hourdez et Jollivet, 2020 ; Jiang et al., 2016). Chez le ver de Pompéi, de fortes concentrations de métallothioneines, protéines ayant une forte affinité pour le cuivre, le zinc, le cadmium et le mercure, ont été relevées dans l'épiderme et les tissus digestifs (Desbruyères, Chevaldonné et al., 1998). Ces protéines pourraient servir à la séquestration de ces métaux. En outre, comparativement aux annélides des côtes, une forte activité de la superoxide dismutase, qui convertit l'anion superoxyde $O_2^{\cdot -}$ en H_2O_2, a été relevée chez *A. pompejana* et *P. grasslei* en particulier là encore dans les tissus digestifs. Au contraire, l'activité tissulaire de la catalase, une peroxidase qui convertit H_2O_2 en eau, apparaît très faible chez ces deux espèces (Marie et al., 2006 ; Genard et al., 2013). Ces deux constats peuvent sembler contradictoires, et une étude élargie de l'activité des autres peroxydases au sein des Alvinellidae ainsi que d'autres enzymes impliquées dans l'équilibre Redox cellulaire (comme la malate dehydrogenase dans le cas du métabolisme anaérobie chez plusieurs organismes marins Fields et Quinn, 1981 ; Lazou et al., 1987 ; Dahlhoff et Somero, 1991) serait très instructive. En outre, ces études ont été effectuées sur l'activité tissulaire totale d'Alvinellidae, mais les performances catalytiques des enzymes du stress oxydatif elles-mêmes n'ont pas été comparées entre espèces hydrothermales et non hydrothermales. Cette caractérisation serait nécessaire pour savoir si l'activité de ces protéines peut être utilisée comme proxy pour déterminer les concentration en sulfure d'hydrogène ou en métaux de l'environnement (Jollivet et Stéphane Hourdez, 2020 ; Le Bris et Françoise Gaill, 2007).

La caractérisation de l'environnement par la mesure de l'activité physiologique des espèces prend également du sens si l'on considère que la mesure directe des conditions du milieu *in situ* reste un défi. En outre, établir le lien entre les conditions environnementales et l'exposition réelle des animaux n'est pas trivial non plus. Par exemple, les quantités en métaux ou sulfures mesurées aux différents sites hydrothermaux ne sont pas nécessairement fortement corrélées avec les quantités retrouvées dans les tissus des Alvinellidae, et pourraient dépendre de mécanismes propres aux tissus des différentes espèces (Martineu et al., 1997 ; Genard et al., 2013). En particulier, les tubes ou cocons secrétés par la plupart des Alvinellidae, outre le fait de permettre le contrôle de la nature de l'eau directement au contact des vers, pourraient également permettre le piégeage des métaux et la régulation de l'oxygène grâce à leur ventilation avec de l'eau de mer du fond avoisinante (Le Bris, Zbinden et Françoise Gaill, 2005). L'association avec des bactéries (épibionte de protéobactéries ϵ filamenteuse pour les espèces *Alvinella* ainsi qu'avec des bacilles sur la surface interne des tubes de *Paralvinella*) est également suggérée comme un moyen de détoxifier les sulfures environnants (Jollivet et Stéphane Hourdez, 2020 ; Le Bris et Françoise Gaill, 2007).

Enfin, la compréhension de la physiologie globale des Alvinellidae doit prendre en compte la phase larvaire de ces organismes, assez peu décrite encore à ce jour. En effet, en se bornant à ne considérer que les stades adultes de ces espèces considérées chaudes ou froides, les conditions environnementales et adaptations moléculaires associées aux stades précoces de développement sont ignorées. Les larves de certaines espèces telles *P. palmi-*

formis et *A. pompejana* sont probablement lécithotrophes (contenant leur propre réserve de nutriments), ce qui suggère une capacité de dispersion importante entre sites hydrothermaux (McHugh, 1989 ; Zal et al., 1995). Au contraire, McHugh, 1989 ont proposé que *Paralvinella pandorae pandorae* couve les juvéniles qui ont une capacité limité de dispersion en dehors du site hydrothermal. Dans le cas d'*A. pompejana*, il est montré que la colonisation de nouveaux substrats s'opère par l'intermédiaire de juvéniles et d'adultes migrant depuis une colonie adjacente. En effet, les embryons jeunes se développent à des températures de 5 à 10°C, et des températures de 20°C leur sont létales, ce qui prouve que ces stades vivent à l'écart des colonies localisées sur les cheminées hydrothermales (Pradillon et al., 2005). Les larves pourraient être plus sensibles aux variations soudaines de l'environnement, notamment les périodes d'hypoxie induites par les émissions de fluide. Dans ce cas, le développement larvaire à l'écart des cheminées serait un moyen de stabiliser les conditions de l'environnement (Pörtner, 2002). Cela implique que la physiologie de ces vers doit présenter une nécessaire eurythermie et une grande flexibilité des *optima* thermiques de fonctionnement des flux métaboliques, possiblement par des expressions différentielles de variants enzymatiques (allélisme ou duplication de gènes). Par exemple, chez *A. pompejana*, l'allozyme 90 de la phosphoglucomutase est plus thermostable de 3°C par rapport à l'allozyme 100 et sa fréquence est plus élevée au niveau des populations jeunes lors de la colonisation des nouveaux édifices beaucoup plus chauds (Bioy et al., 2022). Dès lors, les attendus moléculaires que l'on peut avoir dans l'étude de ces espèces sont grandement complexifiés.

Espèce	Température de vie	Localisation
Alvinella pompejana	40-50°C (Ravaux et al., 2013)	Parois des cheminées : Dorsale Est Pacifique, Dorsale Pacifique-Antarctique, Bassin de Guaymas, Dorsale des Galàpagos
Alvinella caudata	Chaud	Parois des cheminées : Dorsale Est Pacifique, Dorsale Pacifique-Antarctique, Dorsale des Galàpagos
Paralvinella pandorae pandorae	2-5°C (*)	Plancher : Ride Juan de Fuca
Paralvinella pandorae irlandei	2-5°C (*)	Plancher : Dorsale Est Pacifique, Dorsale Pacifique-Antarctique
Paralvinella unidentata	Froid	Plancher : Bassin de Lau, Bassin Nord Fidji, Arc de subduction de Vanatu
Paralvinella palmiformis	20 - 35°C (Girguis et Lee, 2006 ; Dilly et al., 2012)	Plancher : Ride Juan de Fuca
Paralvinella grasslei	10-30°C (Cottin et al., 2008)	Plancher : Dorsale Est Pacifique, Bassin de Guaymas
Paralvinella mira	Chaud	Sites hydrothermaux de Wocan et Daxi
Paralvinella hessleri	Chaud	Parois des cheminées : Fosse d'Okinawa, Bassins de Manus et des Mariannes
Paralvinella sulfincola	40-50°C (Girguis et Lee, 2006)	Parois des cheminées : Ride Juan de Fuca
Paralvinella dela	Froid	Ride Juan de Fuca, sites Middle Valley, Cleft, Endeavour
Paralvinella sp. nov.	Chaud	Bassin de Manus
Paralvinella fijiensis	Chaud	Parois des cheminés : Bassins de Lau et Nord Fidji
Paralvinella bactericola	Froid	Plancher : Bassin de Guaymas

Table I.1 – Caractéristiques thermiques des Alvinellidae et localisation. Les espèces marquées de (*) ont également été observées dans des colonies à des températures plus élevées autour de 25°C (communication personnelle, D. Jollivet). (Desbruyères et Laubier, 1986 ; Jollivet et Stéphane Hourdez, 2020)

I.2 Diversité des mécanismes impliqués dans les adaptations du vivant aux températures extrêmes

Dans cette thèse, nous nous sommes spécifiquement intéressés aux adaptations moléculaires (stabilité des protéines) permettant aux Alvinellidae de s'adapter à une gamme large de température, aussi bien à un niveau inter- qu'intraspécifique considérant leur caractère souvent eurytherme. Quelques grands principes de l'adaptation à la température peuvent d'abord être dressés pour expliquer ce qui différencie la thermotolérance de certaines espèces eucaryotes de la véritable thermophilie (voir hyperthermophilie) de certains procaryotes.

I.2.1 Principes généraux des adaptations à la température

Comme nous l'avons vu à travers l'exemple des sources hydrothermales, la vie s'est développée sur toute une gamme de températures, des émissions de fumeurs noirs à plus de 100°C jusqu'aux plaines abyssales dont la température est inférieure à 4°C (Ravaux et al., 2013 ; Stéphane Hourdez et Jollivet, 2020). Dans le spectre chaud, l'UTL (Upper Thermal Limit pour limite supérieure de température permettant la vie) pour les eucaryotes est estimée en dessous de 55°C. Les eucaryotes les plus thermotolérants connus en milieu ducalcicole étant les ostracodes *Potamocypris* sp. (dans les sources chaudes du parc de Yellowstone, Wickstrom et Castenholz, 1973), en milieu marin le ver *A. pompejana* vraisemblablement en dessous de 50°C, et en milieu terrestre les fourmis du désert *Cataglyphis bombycina* et *C. bicolor* qui montrent des dysfonctionnements à 55°C (Gehring et Wehner, 1995). Le record étant détenu par le champignon *Chaetomium thermophilum* qui peut croître à 55°C (La Touche, 1950). Au contraire, les organismes considérés comme thermophiles *stricto sensu* vivent au delà de 60°C, et sont tous des procaryotes.

De l'autre côté du spectre, concernant la limite basse de température, certaines espèces animales peuvent vivre à des températures très froides et peuvent être considérées comme de véritables psychrophiles. On peut citer l'exemple du bivalve Antarctique *Limopsis marionensi* dont la tolérance à la température baisse jusqu'à se situer entre -1,5 et 2°C (Pörtner et al., 1999), ce qui en fait un véritable psychrophile sténotherme dont la gamme possible de températures de vie est très étroite.

On peut alors s'interroger sur les raisons de ces limites de température de vie, et sur la nature des mécanismes moléculaires qui ont été sélectionnés par l'évolution pour y faire face.

Une première observation est que les organismes thermophiles *stricto sensu* (température de croissance supérieure à 60°C) et hyperthermophiles (température supérieure à 80°C) sont tous procaryotes, et en particulier des archées. Une limite à la croissance à des températures équivalentes chez les eucaryotes serait liée au fonctionnement des mitochondries qui seraient incapables de fournir les besoins de l'organisme via le métabolisme oxydatif à haute température (arrêt du fonctionnement de la chaîne de phosphorylation oxydative lié à la baisse de potentielle de la membrane mitochondriale). Ceci entraînerait l'augmenta-

tion du stress oxydatif et l'hypoxie croissante de l'organisme (Jarmuszkiewicz et al., 2015 ; Zukiene et al., 2010 ; Pörtner, 2002). L'abaissement de la température a également pour conséquence de ralentir le métabolisme des mitochondries (Pörtner, 2002). Ainsi, les organismes complexes, pour lesquels les besoins métaboliques sont élevés et qui ne peuvent pas survivre longtemps sur la seule base d'un métabolisme anaérobie et/ou d'un métabolisme aérobie ralenti, sont plus rapidement limités par les températures extrêmes.

L'augmentation de la température au delà de 100°C provoquerait des processus physico-chimiques délétères pour la survie d'organismes eucaryotes comme procaryotes. Ainsi, l'hydrolyse spontanée de l'ATP est beaucoup plus importante entre 110 et 140°C y compris à hautes pressions jusqu'à 220 atmosphères (Leibrock, Bayer et Lüdemann, 1995), et la sphère d'hydratation des protéines disparaît dans ces mêmes gammes de températures (Jaenicke et Böhm, 1998). Ces mécanismes sont universels, aussi les organismes les plus thermophiles décrits à ce jour, comme l'archée *Pyrococcus abyssi* des sources hydrothermales du bassin Nord Fidji, ont des températures optimales de croissance situées légèrement au-dessus de 100°C à 200 atmosphères (Erauso et al., 1993).

Dans le cadre de cette thèse, nous avons étudié la stabilité de certaines protéines d'Alvinellidae issus de différents régimes de température de vie. En effet, les protéines sont caractérisées en général par une conformation tri-dimensionnelle, dite native dans la cellule. Cette conformation est liée à la fonction de la protéine, et elle est théoriquement déductible de la structure primaire de la protéine, c'est-à-dire de sa séquence en acides aminés (Matsumura, Yasumura et Aiba, 1986). Toutefois, la stabilité de la protéine native dépend de paramètres intrinsèques à la protéine (dont sa séquence), mais aussi de la température. En acceptant un modèle à deux états, natif (N) et dénaturé (D) La stabilité est traduite par l'équation modifiée de Gibbs-Helmholtz (Razvi et Scholtz, 2006) :

$$\Delta G^o(T) = \Delta H_m \times (1 - \frac{T}{T_m}) - \Delta C_p \times (T_m - T + T \times ln(\frac{T}{T_m}))$$

Cette équation relie la différence d'énergie libre entre les formes dénaturée et native ($\Delta G^o(T)$) de la protéine avec la température à laquelle la dénaturation est étudiée, selon une modèle de dénaturation à deux états

$$\mathrm{N} \rightleftharpoons \mathrm{D} \tag{I.1}$$

$\Delta G^o(T)$ dépend également de plusieurs paramètres thermodynamiques : la différence d'enthalpie (ΔH_m) et la différence de capacité calorifique (ΔC_p) mises en jeu pendant la réaction de dénaturation à T_m, qui correspond à la température de demi-dénaturation pour laquelle la moitié des protéines sont dans leur forme native. Ces paramètres dépendent de la séquence primaire de la protéine, en interaction avec le milieu. Une représentation de cette équation est donnée en figure I.6. Si $\Delta G^o(T)$ prédit est positif, alors la réaction de dénaturation est défavorable et tirée vers un équilibre thermodynamique où la protéine est majoritairement dans sa forme native. En simplifiant, la température intracellulaire doit par conséquent correspondre à une gamme pour laquelle l'ensemble du protéome est stable et fonctionnel. Cette libération d'énergie associée à la dénaturation de la protéine provient (Jaenicke et Böhm, 1998 ; Scandurra et al., 1998) :

1. d'interactions entre résidus à très courte distance (interactions de Van der Waals) ;

2. de la libération des molécules d'eau des surfaces hydrophobes enfouies dans la protéine ;
3. de la formation de liaisons hydrogènes et d'interactions électrostatiques ;
4. de la proportion d'hélices α et de feuillets β ainsi que la formation de groupes de résidus chargés (interactions ioniques) dans le cœur de la protéine.

Le moteur principal du repliement correspond principalement à la forte création d'entropie $\Delta S = C_p \times ln(T/T_m)$ conséquence de la constitution du coeur hydrophobe de la protéine et donc à l'effet solvant (Jaenicke et Böhm, 1998). De plus l'équation montre que l'effet relatif de l'entropie par rapport à l'enthalpie est de plus en plus important à mesure que T augmente au-dessus du T_m, alors que la variation d'énergie libre est dominée par la différence d'enthalpie autour de T_m. Toutefois, l'ordre de grandeur de la différence d'enthalpie pour le repliement d'une protéine est de quelques centaines de $kJ.mol^{-1}$ contre quelques $J.mol^{-1}$ concernant la différence d'entropie (Nicholson, W. J. Becktel et Matthews, 1988 ; Razvi et Scholtz, 2006). Chez les organismes thermophiles, les protéines ont généralement une plus grande flexibilité de leur structure native comparée à leurs homologues mésophiles prises à leur température physiologique respective, ce qui a pour conséquence une réduction de la différence d'entropie lors de la dénaturation chez les thermophiles (Karshikoff, Nilsson et Ladenstein, 2015). En effet, l'entropie de la molécule est liée aux degrés de liberté de mouvement des formes natives et dénaturés. L'augmentation de la flexibilité des protéines issus des organismes thermophiles implique moins d'énergie libérée lors de leur dénaturation et la réaction de dénaturation est moins favorable à température physiologique.

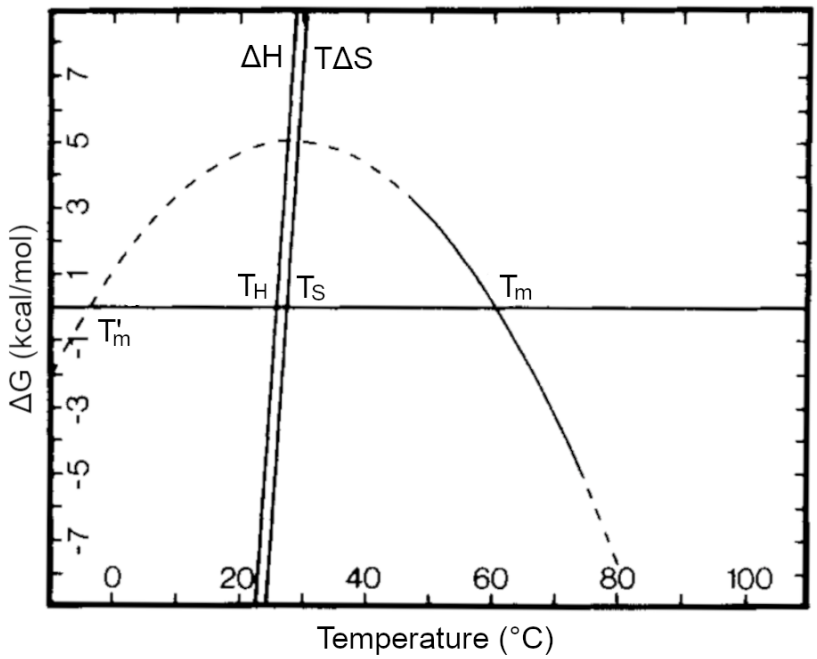


FIGURE I.6 – Courbe de stabilité théorique d'une protéine hypothétique selon la température. Le domaine où ΔG est positif correspond aux températures pour lesquelles la réaction de dénaturation n'est pas favorable et où la protéine est majoritairement dans sa forme native. T_m et T'_m correspondent aux températures de demi-dénaturation chaude et froide. Figure tirée de Wayne J. BECKTEL et SCHELLMAN, 1987.

Ainsi, si les paramètres thermodynamiques d'une protéine sont connus dans l'environnement cellulaire, il est en principe possible de déterminer la proportion de la population de protéines correctement repliées en forme native *via* la relation :

$$\Delta G(T) = \Delta G^o(T) + R \times T \times ln(K)$$

où R est la constate des gaz parfaits. A l'équilibre thermodynamique, $\Delta G(T) = 0$, et l'on peut en déduire K qui correspond à la part de la population de protéines en forme dénaturée divisée par la fraction de protéines natives. Dans les conditions physiologiques, $K << 1$ avec des valeurs comprises entre (10^{-4} à 10^{-11}). Pour une protéine, les variants de conformation de la forme dénaturée sont très nombreux, mais cette simplification du problème en considérant deux états est généralement valide statistiquement (Serohijos et Eugene I Shakhnovich, 2014 ; Sikosek et Chan, 2014). En conséquence, plus la fraction de protéines natives est importante à la température de vie de l'organisme, plus $\Delta G(T)$ est élevé.

On pourrait alors s'attendre à ce que les protéines évoluent vers des stabilités élevées. Pourtant, les mesures mentionnées dans la littérature correspondent en général à une stabilité relativement faible, qualifiée de "marginale", entre 5 et 15 kcal/mol pour des protéines globulaires d'une centaine d'acides aminés (Taverna et Goldstein, 2002 ; Goldstein, 2011 ; Serohijos et Eugene I Shakhnovich, 2014). Ce constat vaut autant pour les protéines issues d'organismes mésophiles que pour les organismes thermophiles (Jaenicke et Böhm, 1998 ; Serohijos et Eugene I Shakhnovich, 2014). Ceci ne correspond à l'énergie que de quelques liaisons hydrogènes, alors même qu'une protéine présente en moyenne plusieurs centaines de ce type d'interactions non covalentes dans sa forme native. En outre, la température correspondant à la valeur maximale de $\Delta G^o(T)$ (sommet de la parabole de la figure I.6) est systématiquement inférieure à la température optimale de croissance des organismes (Jaenicke et Böhm, 1998). Par conséquent les protéines ne sont pas non plus à leur stabilité maximale dans la cellule. L'observation d'une stabilité marginale pour les protéines peut s'expliquer de deux manières, en partant d'une approche fonctionnelle ou par une approche thermodynamique et stochastique.

Dans le cas d'une explication fonctionnelle, la stabilité marginale est sélectionnée par l'évolution car il existe un compromis entre la rigidité et la fonction des protéines. Dans ce cas, les répertoires de protéines des mésophiles et de leurs homologues chez les organismes thermophiles doivent présenter dans leur environnement respectif une flexibilité équivalente qui n'altère pas leur fonctionnalité. Les protéines ne doivent pas être considérées comme des entités rigides ou flexibles, mais on doit prendre en compte l'ajustement de la flexibilité dans les différentes parties de la protéine (Karshikoff, Nilsson et Ladenstein, 2015). La nécessité de conserver une certaine flexibilité peut s'expliquer pour les enzymes par la nécessité de changements de conformation, ou des interactions avec des ligands pour lesquelles les différences d'énergies d'association/dissociation ou de bascule entre des conformations différentes doivent rester relativement faibles pour être réversibles (Karshikoff, Nilsson et Ladenstein, 2015). Beaucoup de protéines ne correspondent pas au modèle classique de repliement de l'entonnoir énergétique, pour lequel depuis une multitude de formes dénaturées, la protéine se replie vers un minimum de stabilité de plus en plus étroit jusqu'à une conformation unique en passant par des stades intermédiaires (Modèle de globules fondus). Ces protéines montrent des profils énergétiques bi- ou poly-stables (plusieurs *optima* avec des énergies libres similaires), ou des régions intrinsèquement non

structurées très flexibles, ce qui concerne environ un tiers des résidus de protéines dans le vivant Sikosek et Chan, 2014. Sur des mutants de kanamycin nucleotidyltransferase, Matsumura, Yasumura et Aiba, 1986 ont montré que les mutants plus thermostables, ayant un *optimum* de température de fonctionnement supérieur de 10°C par rapport à la protéine sauvage, ont une baisse relative d'activité de 20% à 37°C. Dans ce cas, l'augmentation de la stabilité de la protéine se fait au détriment de son activité à une température plus faible démontrant un compromis entre stabilité et activité. Les auteurs notent en outre que l'énergie d'activation de l'enzyme mutante est similaire à celle de la protéine sauvage, environ 11 kcal/mol. A leur *optimum* thermique de fonctionnement, les protéines ont donc une activité similaire. Un exemple extrême a été mis en évidence par Bouvignies et al., 2011, à partir du lysozyme du phage T4. Une conformation, dite excitée (non stable et transitoire, 1 ms d'existence), a été caractérisée dans la population de ces protéines. Cette forme correspond à 3% de la population totale, ce qui montre une stabilité marginale particulièrement faible de cette protéine (environ 2 kcal/mol à 25°C). La fonction biologique de la population excitée n'est pas connue (seule la forme native se lie à son ligand), mais une implication évolutive possible est que ces protéines, de part leur équilibre entre plusieurs conformations simultanément, ont également une variabilité et une adaptabilité potentiellement très forte. Ceci est particulièrement vrai pour les protéines de la capsule du virus, contraintes d'évoluer rapidement pour conserver leur capacité à infecter leur hôte. Le pendant de cela est que pour ces protéines, les modèles de dénaturation à deux états sont une simplification grossière et la compréhension de la dynamique évolutive est beaucoup plus complexe, nécessitant une capacité de modélisation des états de repliement intermédiaires qui n'existe pas actuellement.

La deuxième explication est purement thermodynamique et ne fait pas intervenir directement la fonction de la protéine. Selon cette hypothèse, la stabilité de la protéine est fondamentalement reliée à son abondance cellulaire. Une protéine plus abondante doit être plus stable, car les formes dénaturées peuvent s'agréger (modèle amyloïde) et être toxiques pour la cellule (misfolding avoidance hypothesis, Serohijos et Eugene I Shakhnovich, 2014 ; Razban, 2019). Ainsi, la sélection agirait uniquement dans le sens de la stabilité accrue des protéines, et la stabilité marginale ne serait que le reflet de l'équilibre entre la sélection et la dérive génétique, qui tend à déstabiliser les protéines comme illustré en figure I.7. En partant de la relation de Boltzmann et Maxwell, Serohijos et Eugene I Shakhnovich, 2014 relient la stabilité des protéines (donc l'énergie libre entre les formes dénaturée et native) avec la proportion de protéines correctement repliées dans le cytoplasme. Selon l'hypothèse que les formes mal repliées sont toxiques pour la cellule et diminuent le rendement métabolique, ils établissent que l'effet d'une mutation dans la protéine impacte la valeur sélective (s) de la protéine selon l'équation

$$s \approx Ae^{\beta \Delta G_{pre_mutation}}(1 - e^{\beta \Delta \Delta G})$$

où A correspond à l'abondance de la protéine dans le cytoplasme, $\beta = 1/k_B T$ fait intervenir la constante de Boltzmann et la température, et $\Delta\Delta G$ mesure la différence d'énergie libre du repliement entre la protéine incluant la nouvelle mutation et la protéine sauvage prémutation.

Finalement, les protéines se situent dans un régime tel que $N|s| \approx 1$ où les mutations stabilisatrices et déstabilisatrices ont une probabilité égale de se fixer (voir figure I.7A.). Au

delà, la sélection aura pour effet de favoriser la fixation de mutations stabilisatrices, et en deçà, la dérive, biaisée en faveur de mutations globalement déstabilisatrices, jouera en sens inverse (voir figure I.7B.).

Les auteurs prédisent alors l'énergie libre attendue pour une protéine dans un certain contexte selon l'équation

$$\Delta G \propto -k_B T ln N_e - k_B T ln A + k_B T ln(\frac{1}{k_B T} \frac{\Delta\Delta G_{sd}^2}{\Delta\Delta G_{mean}})$$

où $\Delta\Delta G_{sd}$ et $\Delta\Delta G_{mean}$ caractérisent la distribution de l'effet thermodynamique moyen d'une mutation aléatoire dans la séquence d'une protéine (correspondant environ à $1 kcal/mol$ et $1,7 kcal/mol$ respectivement). Ils démontrent par exemple que la simple différence de taille de population entre procaryotes et vertébrés (10^8 *vs.* 10^5) suffit à expliquer des protéines en moyenne $6 kcal/mol$ moins stables chez ces derniers. Cette équation a l'intérêt d'intégrer d'autres paramètres pour expliquer la stabilité observée d'une protéine, à savoir son abondance cellulaire et la taille de la population.

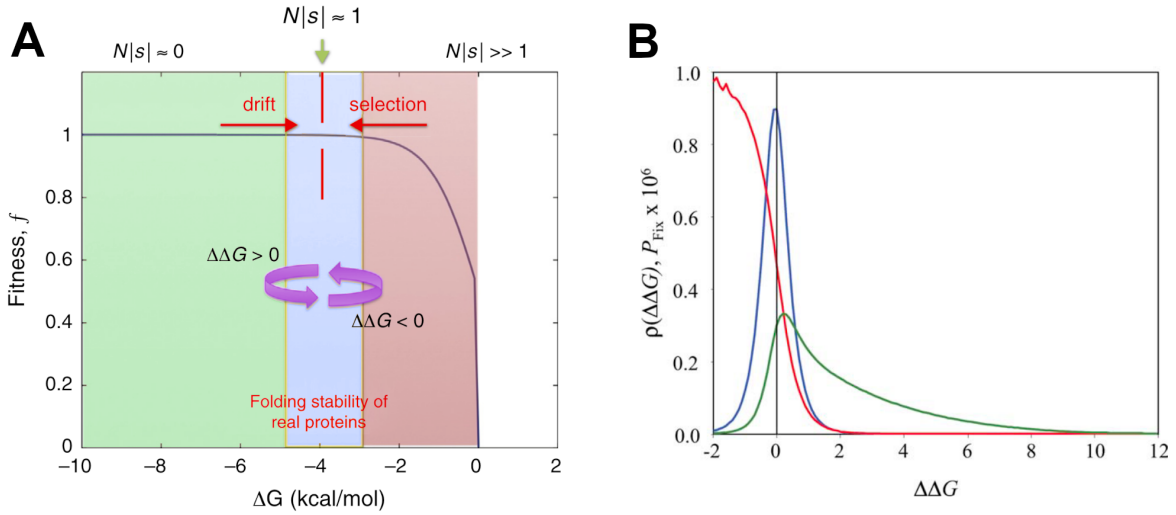


FIGURE I.7 – Effet de la sélection et de la dérive sur la stabilité d'une protéine. (A) domaines de sélection et de dérive. Dans la partie verte, la stabilité de la protéine est très élevée, et l'effet de la sélection est faible par rapport à la dérive. Dans la partie rouge, la protéine est peu stable, les formes mal repliées s'accumulent, et la sélection tend à stabiliser la protéine. (B) Modélisation des forces agissant sur la stabilité. Courbe verte : effet sur la stabilité des mutations aléatoires apparaissant dans la protéine, biaisées dans le sens de la déstabilisation. Courbe rouge : force de la sélection sur la mutation, qui tend à favoriser les mutations stabilisatrices. Courbe bleue : mutations fixées, à l'équilibre entre les mutations aléatoires et l'effet de la sélection. Adapté de SEROHIJOS et Eugene I SHAKHNOVICH, 2014 et GOLDSTEIN, 2011.

Ces deux explications fonctionnelles et thermodynamiques ne sont pas incompatibles. Par exemple, les prédictions des modèles thermodynamiques se placent en général dans l'hypothèse simplificatrice que l'état natif de la protéine est unique et lui confère sa fonction. En revanche, dans le cas de protéines intrinsèquement non structurées ou avec plusieurs *optima* de conformation, une moindre stabilité sur la protéine n'aboutit pas nécessairement à des conformations alternatives toxiques avec des effets délétères sur la valeur sélective. Aussi, la plupart des mutations apparaissant aléatoirement tendent à déstabiliser la protéine (en moyenne, une mutation stabilise une protéine entre 0.2-0.4 kcal/mol ou la déstabilise entre 1.7-2.1 kcal/mol, Sikosek et Chan, 2014), ce qui vaut aussi pour les mutations impliquées directement dans la fonction de la protéine, par exemple dans le site catalytique. Par conséquent, une mutation qui confère une nouvelle fonction à la protéine tend à déstabiliser la protéine dans des proportions similaires à une mutation aléatoire (Tokuriki et al., 2008), et dans ce cas la pression de sélection qui s'opère sur la protéine est complexe. La sélection sur certains acides aminés aminés conservés pour la fonction de la protéine implique ainsi une sélection stabilisatrice compensatrice sur d'autres acides aminés de la protéine, y compris dans des régions de la protéines qui *a priori* ne sont pas impliquées dans la fonction de la protéine.

Enfin, si la sélection sur la stabilité ne fait pas de doute, la question initiale est de savoir si la stabilité *marginale* est un caractère sélectionné. Les simulations de Goldstein, 2011 prédisent en effet que la stabilité marginale n'est qu'une résultante statistique du fait que la plupart des séquences possibles qui encodent une structure sont peu stables, et qu'en conséquence, plus la protéine est stable, moins elle a de chance de trouver aléatoirement une alternative plus stable dans une proportion suffisamment bénéfique pour être sélectionnée. Toutefois, même si ce cas est peu probable, il n'est pas impossible. Ainsi, selon ce modèle, si la moyenne de la stabilité des protéines est environ de 9 kcal/mol, la queue de la distribution prédit des protéines aussi stables que 118 kcal/mol. Cette valeur n'est jamais observée et probablement due à des simplifications du modèle. Les mesures expérimentales montrent que les protéines observées les plus stables, y compris lorsqu'elles sont obtenues par mutation dirigée pour augmenter leur stabilité, n'excèdent pas 15 kcal/mol (Sikosek et Chan, 2014). Il est donc possible que malgré une stabilité marginale apparente, les protéines soient en fait proches de leur maximum de stabilité théorique, ce que Sikosek et Chan résument par "marginally stable, but nearly maximally stable".

Quoiqu'il en soit, on remarque qu'aussi bien dans l'explication fonctionnelle que dans l'explication thermodynamique, la température est un paramètre qui intervient dans la prédiction de la stabilité de la protéine. C'est un résultat important car il implique que la température de vie de l'organisme est corrélée à la stabilité de ses protéines. En effet, on aurait pu penser que la température imposait uniquement une limite basse à la stabilité d'une protéine, mais pas de limite haute aux stabilités potentiellement observées. Or dans ces deux explications la température devrait se traduire par une gamme relativement étroite des stabilités observées.

Dans cette thèse, nous avons utilisé la stabilité de protéines d'espèces d'Alvinellidae, contemporaines et ancestrales, comme proxy de la température de vie des organismes. La justification de l'utilisation de ce proxy et ses limites (stochasticité du processus évolutif, importance relative de la stabilité à haute température et de la flexibilité à basse température

pour des espèces eurythermales) sont cruciales pour interpréter les résultats obtenus.

I.2.2 Des adaptations particulièrement étudiées chez les procaryotes

De très nombreuses études se sont consacrées à comparer la stabilité de protéines provenant d'organismes vivant à différentes températures, notamment des organismes procaryotes réellement thermophiles ou hyperthermophiles (plus de 60°C) avec des procaryotes mésophiles (entre 20 et 40°C) évolutivement proches. En effet, c'est au sein des procaryotes que les écarts de température observés sont les plus importants (Jaenicke et Böhm, 1998).

Il ressort de ces études que l'ensemble des macromolécules des organismes thermophiles subissent des adaptations structurales (lipides particuliers jouant sur la fluidité membranaire (Siliakus, Oost et Kengen, 2017), enroulement positif de l'ADN fréquent chez les hyperthermophiles, Valenti et al., 2011, biais de composition en G+C qui stabilise la conformation des ARN, Deckert et al., 1998 ; Hickey et Singer, 2004) mais aucun mécanisme n'est systématique chez tous les extrêmophiles.

Nous allons nous concentrer spécifiquement sur les adaptations au niveau des protéines. Il a été démontré que la température de croissance des organismes a une influence sur la composition moyenne en acides aminés. Ainsi, si le biais d'utilisation des codons, et par conséquent des acides aminés, dépend en premier lieu du contenu en G+C des génomes (et donc de facteurs non liés à la température impliquant le biais mutationnel et les systèmes de réparation de l'ADN), il existe également un second biais sur l'usage des codons lié directement à la température et répété indépendamment dans différentes lignées, qui tend à favoriser les acides aminés chargés et aromatiques (Hickey et Singer, 2004 ; Jollivet, Mary et al., 2012). La corrélation entre l'usage de ces acides aminés et la température est globalement linéaire entre 7 et 103°C, biais impliquant donc à la fois des organismes psychrophiles, mésophiles et thermophiles (G.-Z. Wang et Lercher, 2010). Ainsi, de façon générale, les protéines issues des organismes (hyper)thermophiles sont enrichies en acides aminés chargés (DEKR) ainsi qu'en acides aminés hydrophobes de grande taille (ILVW) et présentent une proportion moins grande d'acides aminés polaires et non chargés (STNQC) par rapport à leurs homologues mésophiles, comme illustré en figure I.8A. (Jaenicke et Böhm, 1998 ; Deckert et al., 1998).

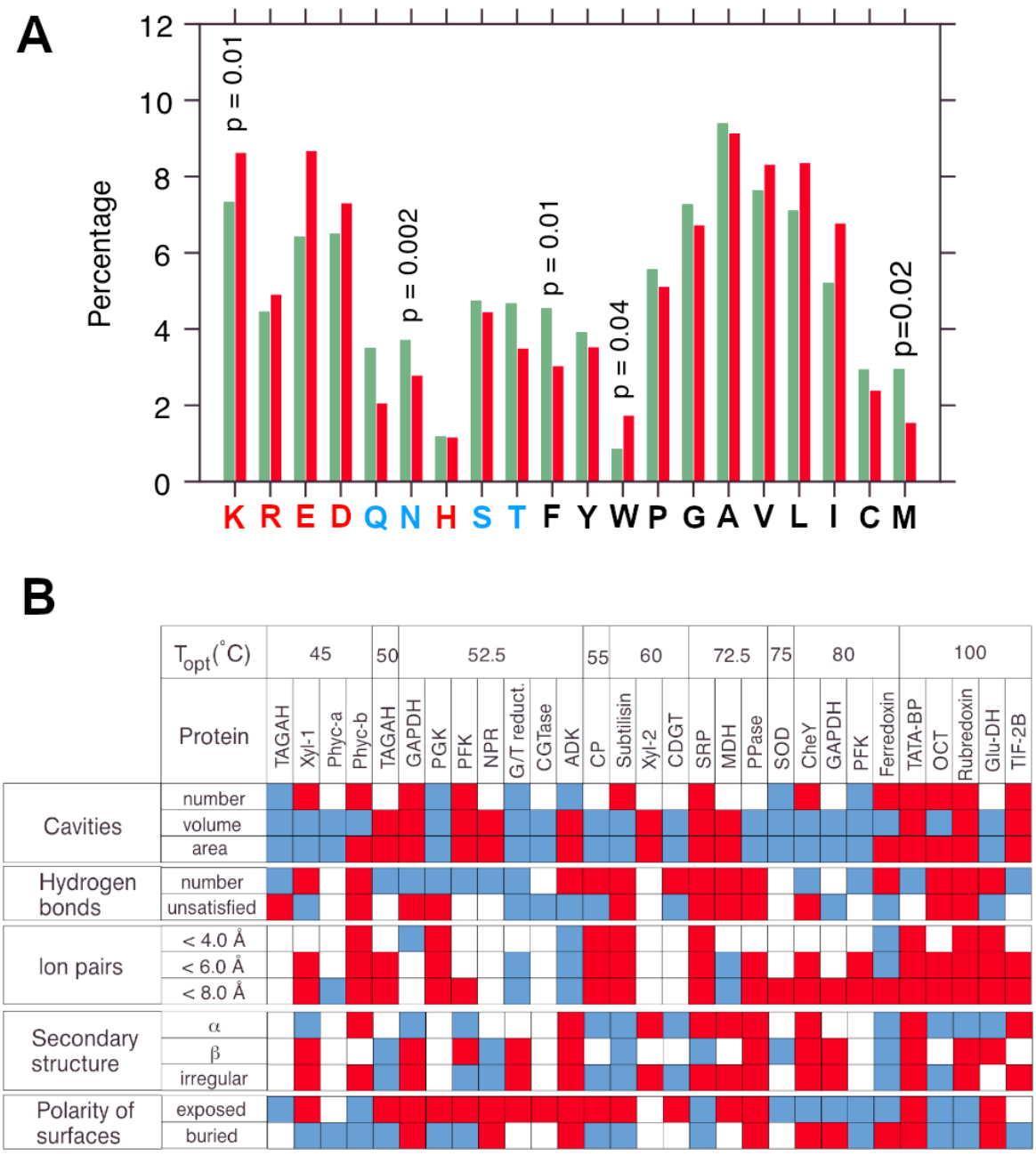


Figure I.8 – Comparaison des structures primaires et secondaires entre protéines de thermostabilités différentes. (A) Utilisation des acides aminés selon la stabilité des protéines. Les acides aminés chargés sont en rouge, les acides aminés polaires en bleu. Les barres verticales rouges correspondent aux fréquences des acides aminés observés chez des protéines d'organismes hyperthermophiles, et les barres vertes aux protéines homologues provenant d'organismes mésophiles. (B) Adaptation au niveau de la structure secondaire des protéines. Une case rouge indique une contribution stabilisatrice chez la protéine thermostable par rapport à son homologue mésophile, et une case bleu un effet déstabilisant. Adapté de Szilágyi et Závodszky, 2000.

Sur la base de ces observations, plusieurs indices ont été développés afin de rendre compte de l'adaptation moléculaire des protéines à la température. Ces indices sont censés corréler avec la stabilité de repliement des protéines, et par extension avec la température de vie des organismes. Parmi les plus usuels, on peut citer les indices IVYWREL (Zeldovich, Berezovsky et Eugene I. Shakhnovich, 2007), EK/HQ et le bias CvP (chargés *vs.* polaires) (Suhre et Claverie, 2003), ces trois derniers indices faisant directement référence à la composition moyenne en acides aminés des protéines, qui favorise les acides aminés chargés et hydrophobes chez les organismes thermophiles. Ces biais en acides aminés refléteraient l'établissement de liaisons hydrogènes et la compacité du cœur hydrophobe des protéines chez les organismes thermophiles. En revanche, Szilágyi et Závodszky, 2000 ont montré que les stratégies permettant aux protéines d'atteindre des stabilités plus élevées sont diverses, comme montré en figure I.8B. Ainsi, il y a une corrélation positive entre la thermostabilité et le nombre de liaisons hydrogènes et de liaisons électrostatiques (qui affectent l'enthalpie des protéines), et une corrélation plutôt positive avec la polarité des résidus exposés et avec la compacité de la protéine (par réduction du nombre et de la taille des cavités), qui affectent les différences de capacité calorifique (Jaenicke et Böhm, 1998 ; Scandurra et al., 1998). En revanche, il n'y a pas de mécanisme obligatoire pour augmenter la stabilité des protéines, et les mêmes phénotypes peuvent s'expliquer par une grande variété d'adaptations. La figure I.8B. montre que la stabilité prédite par ces différents mécanismes peut être due à des mutations stabilisatrices ou déstabilisatrices, y compris lorsqu'on considère des protéines issues des organismes les plus thermophiles étudiées par Szilágyi et Závodszky, 2000. Les différents chemins adaptatifs possibles pour augmenter la stabilité des protéines sont bien illustrés par le concept de *Thermodynamic System Drift* développé par Hart et al., 2014. Pour un même phénotype thermostable, les protéines peuvent explorer un large éventail de mécanismes de stabilisation. La sélection qui s'opère sur la stabilité de la protéine autorise beaucoup de variation dans l'évolution des paramètres thermodynamiques, plus particulièrement l'évolution des différences de capacité calorifique et d'enthalpie entre les formes dénaturée et native, qui respectivement s'expliquent pour l'essentiel par la compacité de la protéine, ainsi que par l'augmentation du nombre de liaisons hydrogènes et électrostatiques . Le terme *drift* fait donc référence à la fluctuation des paramètres thermodynamiques qui pris individuellement sont sans corrélation directe avec le phénotype résultant, et non pas à la dérive définie en biologie évolutive (variation aléatoire de la fréquence des allèles dans la population) (Hart et al., 2014). La préférence pour certains mécanismes de stabilisation par rapport à d'autres pour une même protéine est vraisemblablement liée aux interactions épistasiques entre les mutations. En effet, selon le contexte de la structure tertiaire dans laquelle elle survient, une même mutation n'a pas la même incidence sur la stabilité d'une protéine et par conséquent l'héritage évolutif impose une contrainte sur les mécanismes de stabilisation que la protéine peut immédiatement explorer (Pollock, Thiltgen et Goldstein, 2012). L'indice le plus fiable observé pour identifier des protéines issues d'organismes thermophiles est généralement le nombre d'acides aminés chargés par rapport à l'homologue mésophile comme montré figure I.8B. (Szilágyi et Závodszky, 2000 ; Hickey et Singer, 2004). Ceci s'explique sans doute par le fait que l'établissement de nouvelles liaisons électrostatiques est un chemin plus rapide d'adaptation d'une protéine.

D'autres spécificités que le biais en acide aminés sont également observées chez les protéines des organismes thermophiles. Ainsi, ces protéines ont tendance à être plus courtes que leurs homologues mésophiles avec une réduction de la taille des boucles, feuillet β et

régions intrinsèquement non structurées au profit des hélices α (Szilágyi et Závodszky, 2000 ; Hickey et Singer, 2004). En outre, chez les organismes hyperthermophiles, les protéines adoptant une structure quaternaire peuvent subir de nouvelles modifications (associations originales et fusions) qui permettent la réduction de la surface accessible au solvant par la compacité de la protéine multimérisée (Jaenicke et Böhm, 1998 ; Fraser et al., 2016).

L'adaptation des organismes aux environnements froids a été moins étudiée. Les organismes psychrophiles sont généralement définis comme vivant à moins de 20°C, et peuvent se développer à des températures jusqu'à plusieurs dizaines de degrés sous le zéro (Margesin et Schinner, 1999). Certains indices d'adaptation à la température observés dans les protéines des procaryotes thermophiles tendent à être renversés chez les procaryotes psychrophiles. Ainsi, les acides aminés chargés (acide glutamique, arginine) ainsi que certains acides aminés hydrophobes comme la leucine tendent à être moins présents, au profit d'acides aminés polaires (sérine et thréonine en particulier) et de la glycine (Russell, 2000 ; Metpally et Reddy, 2009). La tendance est cependant moins nette que pour les protéines thermostables, car d'autres acides aminés chargés ou hydrophobes comme l'acide aspartique et l'alanine sont un peu plus représentés également (Metpally et Reddy, 2009). En outre, le biais de mutation n'est pas homogène dans toutes les régions des protéines. Ainsi, c'est dans les boucles exposées que le biais est le plus fort en faveur d'acides aminés petits et polaires et en déficit d'acides aminés chargés. Ces boucles sont en outre plus longues de 2% en moyenne, au détriment des hélices α plus courtes de 2% (Metpally et Reddy, 2009). La contrainte sélective la plus importante pour les enzymes des organismes psychrophiles ne serait pas tant un problème de stabilité qu'un problème d'efficacité catalytique (Gerday et al., 1997). Au contraire des environnements chauds, le facteur limitant devient l'énergie d'activation nécessaire à la réalisation de la réaction enzymatique (Siddiqui et Cavicchioli, 2006). En conséquence, les protéines des procaryotes psychrophiles ont en général des différences d'enthalpie d'activation plus faibles que leurs homologues mésophiles et thermophiles entre l'état initial et l'état de transition enzyme-substrat, grâce à un nombre plus faible de liaisons non covalentes (Van der Waals, électrostatiques, hydrogènes) devant être rompues lors de changement de conformation de la protéine (Siddiqui et Cavicchioli, 2006). En conséquence, les protéines psychrophiles ont des efficacités catalytiques (k_{cat}) jusqu'à dix fois plus élevées que leurs homologues méso et thermophiles, bien que leur thermolabilité soit également plus importante (Saunders et al., 2003 ; Siddiqui et Cavicchioli, 2006). On peut tirer une image différente de l'évolution de la stabilité entre protéines d'organismes thermophiles et psychrophiles, où les protéines des premiers sont sous une contrainte forte de stabilité globale de la structure, tandis que les protéines des seconds sont plus thermolabiles du fait d'un relâchement de la pression de sélection liée à la température, sauf localement, autour du site actif ou des régions impliquées dans les changements de conformation lors de la reconnaissance du ligand.

I.2.3 Des adaptations à la température similaires chez les organismes eucaryotes

Partant des nombreux constats établis chez les procaryotes thermophiles, plusieurs recherches ont tenté de transposer ces résultats chez les eucaryotes les plus thermotolérants,

plutôt par l'étude d'espèces modèles que par des études statistiques globales.

Ainsi, la comparaison de l'expression des gènes du champignon *Chaetomium thermophilum*, dont la température optimale de croissance s'établit entre 50 et 55°C, avec celui de l'espèce fortement apparentée *Chaetomium globosum* vivant à 24°C, révèle l'absence de réponse thermo-induite de certaines protéines (par exemple des chaperonnes ou protéines heat-shock) par rapport à l'espèce mésophile, ce qui argumente en faveur d'une véritable thermotolérance de l'espèce (Bock et al., 2014). Noort et al., 2013 ont effectivement retrouvé plusieurs adaptations similaires chez *C. thermophilum* à celles observées chez les procaryotes thermophiles : réduction de la taille du génome, grâce à la diminution du nombre de gènes codant des protéines, de plus courts introns et régions intergéniques, ainsi qu'un biais dans l'usage des acides aminés comparé à *C. globosum*. Ainsi, les protéines de *C. thermophilum*, *Thermomyces lanuginosus*, *Talaromyces thermophylus*, *Thielavia terrestris* et *Thielavia heterothallica* (ces deux derniers vivant à des températures élevées sur les prairies semi-arides du Nouveau Mexique) montrent un remplacement préférentiel de la lysine par l'arginine, de l'acide aspartique par l'acide glutamique, et de la threonine par l'alanine, et une sur-représentation des prolines comparativement à des espèces mésophiles proches. En outre, on observe également un remplacement préférentiel de la glycine par l'alanine, ce qui stabilise les hélices α, et, chez *C. thermophilum*, une surreprésentation de cystéines qui pourrait contribuer à la stabilité des protéines par des interactions dans le cœur protéique. Ces deux derniers biais sont spécifiques à ces champignons et n'ont pas été relevés chez les procaryotes. Globalement, l'indicateur IVYWREL, établi à l'origine pour les organismes procaryotes (Zeldovich, Berezovsky et Eugene I. Shakhnovich, 2007), montre une bonne corrélation avec les températures de croissance de ces champignons, tiré notamment par l'arginine ainsi que la tyrosine, l'isoleucine et le tryptophane qui pourraient jouer un rôle dans la compacité du cœur hydrophobe des protéines (Noort et al., 2013). Ainsi, certains acides aminés chargés (arginine, acide glutamique) sont favorisés, tandis que des acides aminés polaires (thréonine) sont moins présents. En prenant en compte 457 protéines, Holder et al., 2013 identifient que le biais CvP (acides aminés chargés contre polaires) montre un biais positif important pour l'espèce *C. thermophilum*, en accord avec les résultats sur les protéomes de procaryotes thermophiles pour lesquels le biais en faveur d'acides aminés chargés à la surface de la protéine est récurrent. Ce mécanisme n'est cependant pas le seul permettant d'augmenter la stabilité de la protéine chez les eucaryotes.

Le cas des animaux endothermes est particulièrement intéressant. En effet, la température de vie des vertébrés homéothermes (mammifères et oiseaux) est plus élevée que celle des vertébrés ectothermes (poissons, amphibiens, reptiles). En pratique, les protéines des homéothermes dont les températures internes varient entre 35 et 42°C pourraient être considérées comme au moins thermotolérantes. G.-Z. Wang et Lercher, 2010 montrent qu'on retrouve effectivement un biais dans l'usage des acides aminés des homéothermes par rapport aux ectothermes, qui ne s'explique pas uniquement par les contenus en G+C différents entre les deux groupes. Ainsi, les indicateurs CvP et ERK (E+R+K-D-N-Q-T-S-H-A) corrèlent fortement avec la température interne des organismes, et les ARN ribosomiques des vertébrés endothermes présentent un contenu en G+C plus élevé que celui des ectothermes. Toutefois, bien que l'effet soit significatif, il reste relativement faible, ce qui s'explique peut-être par le fait que les vertébrés ectothermes ne sont pas nécessairement toujours plus froids que les homéothermes et sont vraisemblablement eurythermaux dans une certaine mesure.

Ainsi les indices ne corrèlent pas avec les températures environnementales théoriques des vertébrés ectothermes qui doivent pouvoir s'adapter à des variations de température. En outre, l'effet de la température entre ces deux groupes est difficile à observer rigoureusement, étant donné que les caractères endotherme et ectotherme sont partagés au sein des groupes monophylétiques. Il n'existe par conséquent pas d'espèces sœurs au sein des vertébrés qui vivraient à des températures radicalement différentes. Après contrôle du lien phylogénétique entre les espèces, les fréquences des acides aminés C, D, M, N, Q, S, T corrèlent négativement avec la température de vie des vertébrés, tandis que les acides aminés R et P sont plus représentés. L'indicateur ERK est toujours significativement corrélé avec le caractère endothermique, mais le biais est très faible (G.-Z. Wang et Lercher, 2010).

En conclusion, il apparaît que dans le cas des eucaryotes, bien que la gamme de température de vie soit plus restreinte que celle observée chez les procaryotes, on puisse retrouver dans leurs protéines un certain nombre de signaux d'adaptations à la température. L'indice CvP en particulier ressort comme étant le plus fiable pour discriminer les températures de croissance autant chez les procaryotes que chez les eucaryotes (Holder et al., 2013 ; G.-Z. Wang et Lercher, 2010). C'est un résultat attendu, puisque les bases thermodynamiques du repliement et de la stabilité des protéines sont identiques dans tout le vivant. Toutefois, la question du modèle d'étude se pose, en particulier chez les animaux pour lesquels les températures de vie sont souvent corrélées avec la phylogénie. Ceci rend la mise en évidence d'adaptations strictement liées à la température plus difficile, car il faut s'affranchir des effets confondants liés à l'héritage phylogénétique des espèces (contenu en GC ou en purine).

I.2.4 Des adaptations moléculaires à la température étudiées dans la lignée des Alvinellidae

La famille des Alvinellidae représente un modèle d'étude exceptionnel au sein des animaux pour observer les adaptations à la température. En effet, certains Alvinellidae, tels *A. pompejana* et *P. sulfincola*, font partie des animaux les plus thermotolérants connus actuellement (Girguis et Lee, 2006 ; Ravaux et al., 2013), mais la famille comprend également des espèces proches vivant sur toutes les gammes de températures connues pour les métazoaires (Cottin et al., 2008 ; Jollivet et Stéphane Hourdez, 2020). Cette diversité écologique au sein de la famille permet l'application de contrastes phylogénétiques pour déterminer rigoureusement quels traits sont associés à des adaptations à la température, et quels traits sont seulement dus à l'héritage génétique des espèces. En effet, en tant qu'invertébrés ectothermes, on s'attend à ce que la température de vie de ces organismes engendre des adaptations moléculaires, de la même manière que ce qui a été observé chez les procaryotes psychrophiles et thermophiles et certains eucaryotes vivant à des températures élevées. Comme nous l'avons rappelé, une limitation importante à l'adaptation des eucaryotes et plus spécifiquement des animaux aux températures extrêmes est lié au métabolisme de l'oxygène, et plus spécifiquement à la baisse fonctionnement des mitochondries aux températures élevées (Pörtner, 2002). Il est d'autant plus remarquable que les Alvinellidae aient réussi à coloniser les cheminées hydrothermales, dont l'environnement est à la fois hypoxique et chaud, alors que la plupart des espèces hydrothermales vivent dans des

habitats plus froids et mieux oxygénés (Stéphane Hourdez et Jollivet, 2020).

Ainsi, il a été montré que les protéines d'*A. pompejana* sont globalement plus stables que leurs homologues chez d'autres espèces. On peut donner l'exemple du collagène cuticulaire, stable jusqu'à une température de 46°C chez le ver de Pompéi, soit 17°C de plus que la protéine homologue des annélides des côtes tempérées. Dans cette protéine, les caractéristiques moléculaires connues pour stabiliser le collagène chez les vertébrés sont amplifiées, comme la quantité de proline et de triplets impliquant la glycine dans les hélices (Sicot et al., 2000). Cette température de dénaturation est proche de la limite de vie d'*A. pompejana*, et par comparaison le collagène cuticulaire de *P. grasslei* est stable jusqu'à 35°C, soit 11°C de moins (Francoise Gaill et al., 1995). A noter qu'une fois assemblée en fibriles, la supramolécule de collagène a une température de dénaturation qui peut augmenter de 20°C par rapport à la molécule seule (Mann et al., 1996). La thermostabilité d'autres protéines d'*A. pompejana* a également été comparée à leurs homologues humaines : ainsi l'ADN polymérase $Pol\mu$ est toujours active à 49°C chez *A. pompejana*, comparativement à 43°C chez l'Homme (Kashiwagi et al., 2010), ou encore le facteur d'épissage U2AF65 est 6°C plus stable chez *A. pompejana* (Henscheid et al., 2005). La Superoxyde Dismutase Cu/Zn est également plus stable chez *A. pompejana* que chez son homologue humain (Shin, DiDonato, Barondeau, Hura, Hitomi et al., 2009). Dans le cas de l'hémoglobine géante extracellulaire toutefois, bien que la protéine présente des caractéristiques propres au milieu de vie d'*A. pompejana* (plus haute affinité pour l'oxygène jusqu'à 45°C par rapport à d'autres annélides, effet Bohr prononcé pour relâcher l'oxygène dans les tissus plus acides) ; elle n'est stable qu'en dessous de 50°C, ce qui ne la distingue pas de l'hémoglobine du lombric (Toulmond et al., 1990). Toujours chez *A. pompejana*, les protéines ribosomales sont enrichies en arginine, proline et tyrosine et appauvries en glycine et phénylalanine par rapport à leurs homologues chez les lophotrochozoaires, ce qui corrèle avec des attendus de stabilité établis chez les espèces procaryotes (Jollivet, Mary et al., 2012).

Des différences de stabilité entre les protéines des Alvinellidae ont également été mesurées entre les espèces chaudes et froides de la famille. Jollivet, Desbruyères et al., 1995 ont observé que l'activité résiduelle de trois allozymes (l'aspartate-amino transferase AAT, la glucose-6-phosphate isomerase GPI et la phosphoglucomutase PGM) après incubation à différentes températures distinguait les espèces chaudes *A. pompejana*, *A. caudata*, *P. sulfincola* et *Paralvinella hessleri* et froides *P. palmiformis*, *P. grasslei*, *Paralvinella pandorae irlandei* et *P. p. pandorae*. Ainsi, l'AAT et la GPI issues des espèces chaudes conservent 80% de leur activité résiduelle (T_{80}) entre 46 et 61°C, tandis que les allozymes des espèces froides ont des T_{80} entre 32 et 53°C. La différence est moins marquée en ce qui concerne la PGM, avec des T_{80} entre 26 et 46°C pour toutes les espèces. On note toutefois qu'*A. pompejana*, qui est probablement l'espèce la plus thermotolérante des Alvinellidae (considérant que ses protéines sont systématiquement les plus stables ou parmi les plus stables dans les études comparatives avec d'autres espèces de la famille, y compris dans cette thèse) présente là aussi un variant de PGM le plus stable à 46°C. En se concentrant sur les espèces *P. sulfincola* et *P. palmiformis*, Rinke et Lee, 2009 ont montré que l'activité de l'alanopine déshydrogenase, strombine déshydrogenase, citrate synthase et lactate déshydrogenase étaient maintenues jusqu'à 50-60°C chez *P. sulfincola*, et 35-50°C chez *P. palmiformis*. Les enzymes de *P. palmiformis* étaient 5 à 15°C moins stables que chez *P. sulfincola*. Par conséquent, bien que les espèces froides des Alvinellidae soient vraisemblablement eurythermales sur une gamme

assez large de températures entre 10 et 30°C, des différences d'adaptation notables entre les protéines des deux groupes d'espèces sont déjà mesurables. Un dernier point intéressant à noter est la présence d'allèles avec des stabilités très différentes chez certaines espèces. Jollivet, Desbruyères et al., 1995 rapportent la présente de deux variants de l'AAT chez *P. sulfincola*, ayant des T_{80} de 46 ou 56°C. La variabilité est plus grande encore pour la PGM, avec trois variants chez *P. sulfincola* (T_{80} de 22°C pour l'allèle 78, 24°C pour l'allèle 66 et 36°C pour l'allèle 90) et deux chez *A. pompejana* (23°C pour l'allèle 100 et 46°C pour l'allèle 90). La répartition de ces derniers variants dans les populations sauvages a été étudiée plus en détail pour *A. pompejana*. Les allèles 90 et 100 de la PGM sont répartis environ à 50% de fréquence dans les populations, en revanche l'allèle 90 (le plus stable) est majoritaire sur les nouvelles cheminées hydrothermales les plus chaudes (diffuseurs d'anhydrite), tandis que l'allèle 100 est plus représenté sur les cheminées en cours de refroidissement. D'autres raisons pourraient expliquer cette répartition (déséquilibre de liaison avec d'autres loci sous sélection balancée), néanmoins il faut garder à l'esprit que le polymorphisme important dans les espèces actuelles engendre certainement une variabilité intra-espèce de la tolérance à la température (Ravaux et al., 2013).

Concernant les changements de composition en acides aminés des protéines, des biais similaires à ceux observés chez les procaryotes thermophiles et certains eucaryotes thermotolérants ont été également relevés chez certaines espèces d'Alvinellidae. Ainsi, Fontanillas et al., 2017 ont montré une prédominance de résidus polaires et aliphatiques chez les espèces froides *P. grasslei* et *P. p. irlandei*, au contraire d'une plus grande fréquence de résidus chargés et hydrophobes chez les espèces chaudes *A. pompejana*, *P. sulfincola*, *A. caudata*, et *Paralvinella fijiensis* parmi 423 gènes orthologues. La nature des remplacements n'est toutefois pas équivalente : ainsi chez *P. grasslei*, c'est surtout l'abondance de résidus polaires qui ressort, tandis que *P. p. irlandei* a un nombre plus important de résidus aliphatiques, peut-être dû à des temps différents d'adaptation au froid chez ces deux espèces qui serait plus ancien pour *P. p. irlandei* (Fontanillas et al., 2017). En outre, l'étude du $\frac{d_N}{d_S}$ sur une phylogénie élargie avec d'autres lophotrochozoaires montre que la branche menant à la lignée des Alvinellidae présente des signes de sélection positive. Sur la base de 335 gènes orthologues, 40 codons sont ainsi mis en évidence avec des remplacements préférentiels de résidus chargés par des résidus hydrophobes, et encore plus fortement des résidus polaires par des résidus hydrophobes (Jollivet, Mary et al., 2012). Au contraire, au sein de la lignée, le $\frac{d_N}{d_S}$ moyen des gènes est très faible, entre 0,01 et 0,04 (Fontanillas et al., 2017), suggérant une forte sélection purifiante à l'échelle de la famille (à titre de comparaison, les $\frac{d_N}{d_S}$ moyen au sein des mammifères s'établissent entre 0,2 et 0,25, et celui des huîtres et tuniciers, qui ont des tailles de population et temps de génération similaires aux Alvinellidae, évolue entre 0,10 et 0,20, Jollivet, Mary et al., 2012). Ce $\frac{d_N}{d_S}$ est encore plus faible sur les branches menant aux espèces chaudes. Ces différences sont des arguments en faveur de l'acquisition du caractère thermophile par l'ancêtre de la famille, puis un maintien de ce caractère par une forte sélection purifiante menant aux espèces chaudes contemporaines. Au contraire, la colonisation des zones plus froides des milieux hydrothermaux serait intervenue dans un second temps, comme en témoigne un relâchement sélectif plus important sur les branches menant aux espèces froides contemporaines (Fontanillas et al., 2017). Ces remplacements sont illustrés en figure I.9, où les branches menant aux espèces froides portent une divergence plus importante, avec des remplacements préférentiels d'acides aminés considérés comme associés à une déstabilisation des protéines par rapport aux branches

chaudes moins divergentes de l'ancêtre de la famille.

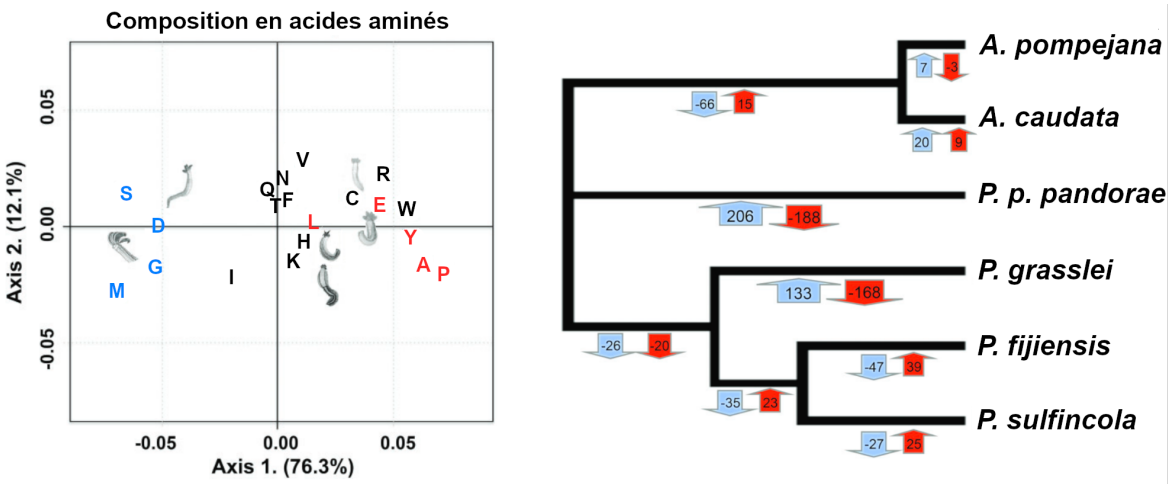


Figure I.9 – Biais en acides aminés et remplacements préférentiels dans la phylogénie des Alvinellidae. Le biais entre les acides aminés associés aux espèces chaudes (PAYLE) et froides (DGMS) explique une parte importante de la variabilité du protéome, indépendamment du biais en GC. Le nombre de mutations vers les acides aminés PAYLE ou DGMS est reporté en rouge ou en bleu sur les branches de la phylogénie des alvinellidae. Adapté de Fontanillas et al., 2017.

Certains biais de composition en acides aminés établis chez les procaryotes thermophiles ont aussi pu être observés chez les Alvinellidae. Le protéome d'*A. pompejana*, après contrôle du biais dû aux différents contenus en G+C, révèle un enrichissement en lysine et arginine (résidus chargés) et en alanine et proline, et une fréquence moindre de valine, méthionine et glycine (Jollivet, Mary et al., 2012). Il est intéressant de noter que l'enrichissement en alanine, s'il n'a pas été relevé chez les procaryotes thermophiles, avait déjà été noté dans le protéome de plusieurs champignons thermophiles dont *Chaetomium thermophilum* (Noort et al., 2013). Le biais en acides aminés chargés par rapport aux polaires, CvP, semble rester l'indicateur le plus robuste pour établir la thermotolérance de *A. pompejana*. Ainsi, selon Holder et al., 2013, il s'agit du seul indicateur qui classe systématiquement les protéines d'*A. pompejana* comme étant plus thermostables. Fontanillas et al., 2017 notent cependant qu'en prenant en compte le transcriptome de six espèces d'Alvinellidae (deux espèces froides, quatre espèces chaudes), les indices EK/QH *vs.* IVYWREL permettent de distinguer les espèces chaudes et froides de la famille. Dans cet article, les auteurs développent un indice de thermotolérance propre aux Alvinellidae, qui sépare les espèces chaudes et froides après contrôle du contenu en G+C des génomes et présenté en figure I.9. Ainsi, la proline, alanine, tyrosine, leucine, acide glutamique (PAYLE) sont enrichis chez les espèces chaudes, au détriment de l'acide aspartique, glycine, méthionine et sérine (DGMS). Cet indicateur est établi sur la base d'un jeu de 423 séquences orthologues provenant de 6 espèces d'Alvinellidae, et il serait intéressant de l'éprouver sur un nombre de séquences plus grand, incluant notamment plus d'espèces de la famille pour tenir compte du contraste phylogénétique entre les espèces (Fontanillas et al., 2017). En effet, au niveau de la famille, ces enrichissements chez les espèces chaudes ne se retrouvent pas nécessairement sur les mêmes gènes orthologues. Ceci suggère que l'héritage phylogénétique des espèces a aussi pu favoriser la fixation de certaines mutations par rapport à d'autres, tout en explorant un paysage mutationnel large qui aboutit à des convergences évolutives phénotypiques similaires (en pratique, le ΔG^o correspondant à la réaction de dénaturation de la protéine, (Fontanillas et al., 2017 ; Hart et al., 2014)).

Aussi, tout changement relatif à la stabilité des protéines, et donc à la composition moyenne en acides aminés, n'est pas lié à des adaptations à la température. D'autres facteurs peuvent intervenir et brouiller le signal, sur certaines protéines particulières ou sous-ensembles du protéome. C'est le cas notamment de la profondeur de vie des espèces (pression hydrostatique) qui peut induire des biais similaires à l'adaptation aux températures froides et affecter de manière différente certaines protéines issues d'espèces hydrothermales profondes par rapport à des espèces proches de surface (Dahlhoff et Somero, 1991 ; Jollivet, Mary et al., 2012). Comme prédit par l'hypothèse d'évitement des formes dénaturées (misfolding avoidance hypothesis, Razban, 2019), l'augmentation de la concentration d'une protéine, comme observée dans le cas des hémoglobines de mammifères sous-marins, tend à augmenter leur stabilité. Ces hémoglobines ont par exemple une charge nette plus élevée des acides aminés exposés, ce qui pourrait se confondre avec le biais CvP lié à la température, probablement pour éviter leur agrégation entre elles ou avec d'autres protéines (Isogai et al., 2018). D'autres mécanismes adaptatifs peuvent aussi augmenter la thermotolérance des animaux : ainsi, contrairement à *A. pompejana*, l'expression de protéines chaperonnes est en permanence élevée chez *P. sulfincola*, ce qui pourrait impliquer que certaines protéines de cet animal se replient dans un micro-environnement relativement différent du cytosol et soient potentiellement moins stables que celles d'*A. pompejana* bien que les deux

espèces vivent à des gammes de température similaires (Dilly et al., 2012).

De ces observations, la question principale à laquelle cette thèse tente de répondre est la suivante :

L'ancêtre commun des Alvinellidae était-il déjà une espèce thermotolérante des sources hydrothermales de l'Océan Pacifique ?

Un certain nombre de questions corollaires sont associées à cette problématique :

1. Quelle est l'histoire évolutive des Alvinellidae ? Quel scénario a vraisemblablement permis la colonisation ancestrale des sources hydrothermales du Pacifique par les Alvinellidae ? Cette histoire souscrit-elle à l'origine abyssale de la faune moderne hydrothermale ?
2. Les espèces contemporaines montrent-elles réellement des signes forts d'adaptation à la température ? La stabilité des protéines est-elle une bonne approche pour évaluer cela ? En effet, les espèces froides montrent souvent un caractère eurytherme et peuvent survivre à des températures supérieures à leur température de confort sur des temps courts d'exposition.
3. La thermotolérance ou la perte de thermotolérance des espèces contemporaines est-elle apparue plusieurs fois au cours de l'évolution de la lignée ? Et si oui, les mécanismes moléculaires mis en œuvre sont-ils similaires ?
4. Si l'ancêtre des Alvinellidae était déjà une espèce thermophile, peut-on affirmer que l'espèce colonisait déjà le pôle chaud de l'habitat hydrothermal (murs des cheminées), ou y a-t-il d'autres hypothèses plausibles pour expliquer ce résultat ?

Pour répondre à cette problématique, nous avons choisi de reconstituer quelques protéines choisies de l'ancêtre de la lignée des Alvinellidae, et de comparer leur stabilité à la température avec celles d'espèces contemporaines de la lignée dont l'écologie est connue. Cette méthode est appelée Ancestral Sequence Reconstruction (ASR) et a déjà été proposée pour élucider les températures de vie d'organismes vieux de plusieurs centaines de millions d'années, aussi bien sur des séquences protéiques (Gaucher, Govindarajan et Ganesh, 2008) que nucléotidiques (Boussau et al., 2008).

I.3 Approche envisagée pour caractériser les phénotypes ancestraux des Alvinellidae

I.3.1 Mise en place d'une phylogénie complète et robuste des Alvinellidae

La première étape de la thèse, et qui constitue le premier chapitre de ce manuscrit, était d'établir une phylogénie des Alvinellidae qui intègre l'ensemble des données moléculaires disponibles actuellement pour la famille. Des extractions d'ARN ont été effectuées sur différentes espèces d'Alvinellidae, Ampharetidae et Terebellidae au sein de l'équipe DYDIV à la

suite de plusieurs campagnes dans l'océan Pacifique effectuées depuis 2004 et d'échantillons côtiers récoltés en Terre Adélie (station Dumont d'Urville, 2011) et sur la côte roscovite. A cela s'ajoute les données de séquençage du transcriptome obtenues pour l'espèce indienne *Paralvinella mira* par Han et al., 2021 et le génome assemblé au niveau chromosomique de l'espèce *A. pompejana* (El Hilali et al., 2024).

La famille des Alvinellidae, selon les descriptions morphologiques établies initialement par Desbruyères et Laubier, est divisée en deux genres, *Alvinella* et *Paralvinella* Desbruyères et Laubier, 1986. Le genre *Alvinella* regroupe seulement deux espèces *A. pompejana* et *A. caudata*, qui sont très similaires et vivent en syntopie. A l'origine, *A. caudata* avait été décrite comme une forme ontologique juvénile de *A. pompejana*, bien que les auteurs aient noté quelques singularités, notamment le fait que la forme juvénile (*A. caudata*) était plus grande (Desbruyères et Laubier, 1980). La subdivision en deux espèces distinctes a ensuite été rapidement effectuée à la suite de l'analyse génétique de leurs polymorphismes. Les espèces *Alvinella* ont toutes deux des filaments branchiaux secondaires plats, une paire de tentacules buccaux modifiés épais chez les mâles (qui intervient lors de l'accouplement) et un plastron ventral duquel les animaux sécrètent un tube résistant et minéralisé dans lequel ils vivent (Jollivet et Stéphane Hourdez, 2020). Ces critères les distinguent des espèces *Paralvinella*, qui possèdent des branchies sur lesquels les filaments secondaires, cylindriques, sont implémentés en peigne ou en buisson, des tentacules buccaux qui ont des formes variables au sein du genre et qui ne possèdent pas de plastron ventral, sécrétant un mucus qui peut être plus ou moins épais selon la température de vie de l'animal (Desbruyères et Laubier, 1991 ; Jollivet et Stéphane Hourdez, 2020). Ces critères peuvent cependant montrer une grande plasticité au sein de la famille. *Paralvinella* sp. nov., par exemple, présente un bouclier ventral moins marqué que les *Alvinella* mais remarquable, et sécrète un tube plus minéralisé bien que l'espèce présente tous les critères morphologiques des *Paralvinella* (Didier Jollivet, communication personnelle). En outre, les données moléculaires peuvent venir en contradiction avec la phylogénie proposée par Desbruyères et Laubier, notamment en ce qui concerne la monophylie du genre *Paralvinella*. La figure I.10 montre deux phylogénies établies par (Jollivet et Stéphane Hourdez, 2020). La phylogénie I.10A. est basée sur la concaténation de 278 transcrits orthologues, et supporte la monophilie du genre. Au contraire, la phylogénie I.10B. utilise le gène COI, et place les espèces *P. p. pandorae* et *P. p. irlandei* comme soeurs des autres espèces d'Alvinellidae. D'autres phylogénies moléculaires récentes, notamment Stiller, Tilic et al., 2020, supportent aussi l'hypothèse que les espèces *P. pandorae* sont soeurs des autres Alvinellidae.

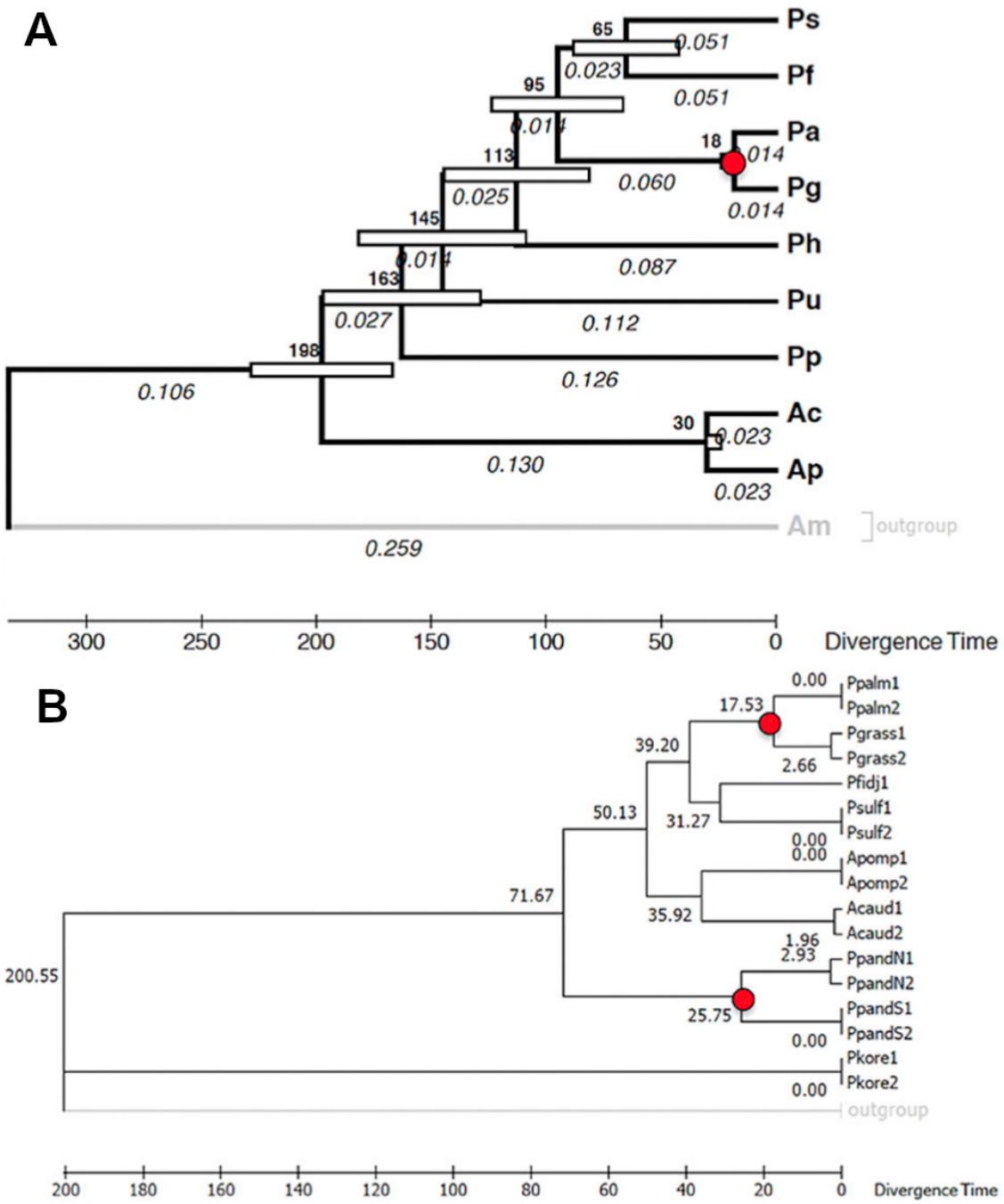


Figure I.10 – **(A) Phylogénie basée sur 278 transcrits orthologues obtenue par RaxML en utilisant le modèle GTR.** – Ps : *P. sulfincola*, Pf : *P. fijiensis*, Pa : *P. palmiformis*, Pg : *P. grasslei*, Ph : *P. hessleri*, Pu : *P. unidentata*, Pp : *P. p. irlandei*, Ac : *A. caudata*, Ap : *A. pompejana*, Am : *Amphitrides* spp. **(B) Phylogénie basée sur une séquence partielle du gène *Cox1* obtenue par PhyML en utilisant le modèle de substitution K2P.** – Ppalm : *P. palmiformis*, Pgrass : *P. grasslei*, Pfidj : *P. fijiensis*, Psulf : *P. sulfincola*, Apomp : *A. pompejana*, Ac : *A. caudata*, PpandN : *P. p. pandorae*, PpandS : *P. p. irlandei*, Pkore : *Pectinaria koreni*. Les points rouges indiquent la subduction de la plaque Farallon sous la plaque Américaine, 25 millions d'années. Tiré de Jollivet et Stéphane Hourdez, 2020.

Il est intéressant de remarquer que des données moléculaires pour l'espèce *P. unidentata* n'ont pas été intégrées dans la phylogénie de Stiller, Tilic et al., 2020. Or cette espèce a une morphologie particulière. En effet, l'absence de tentacules buccaux chez les mâles rapproche l'espèce de *Paralvinella pandorae*, ainsi que l'implantation des filaments branchiaux en peigne et l'absence de lobes notopodiaux digitiformes chez ces espèces. Ces critères avaient amené à regrouper les espèces *P. unidentata*, *P. p. irlandei* et *P. p. pandorae* au sein du sous-genre *Nautalvinella* (Desbruyères et Laubier, 1993). Pourtant, certains caractères comme la forme aplatie des filaments secondaires des branchies de *P. unidentata* sont partagés uniquement avec les deux espèces *Alvinella*. Cette espèce semble par conséquent emprunter des caractères aux deux genres, et sa place dans la phylogénie des Alvinellidae pourrait également permettre d'affirmer la position de *P. p. irlandei* dont elle pourrait être une espèce soeur.

Le genre *Paralvinella* est également divisé en deux autres sous-genres, *Miralvinella*, regroupant les espèces *P. hessleri*, *Paralvinella dela* et *Paralvinella bactericola*, ainsi que le sous-genre *Paralvinella*, regroupant *P. palmiformis*, *P. grasslei*, *P. fijiensis* et *P. sulfincola* (Desbruyères et Laubier, 1993). Ces deux sous-genres sont principalement définis par la forme des tentacules buccaux (respectivement pointus et torsadés ou tri-lobés), la taille des individus, comprenant moins de segments chez les *Miralvinella*, et la position du premier tore uncinigère, plus antérieur chez les *Miralvinella*. Han et al., 2021, sur la base des marqueurs COI, 16S et 18S, ainsi que de la forme des tentacules buccaux pointus et plus épais, classe *P. mira* parmi les *Miralvinella*, et notamment plus proche de *P. hessleri*. Cette espèce n'a pour l'instant pas été intégrée dans une phylogénie complète des Alvinellidae incluant de plus nombreux marqueurs moléculaires. A noter également que nous ne disposons pas de séquences pour les deux autres espèces soeurs de *Miralvinella*, *P. dela* et *P. bactericola*, qui sont plus rares et n'ont pas pu être échantillonnées pour un séquençage de transcriptomes.

En outre, la qualité des différents transcriptomes à notre disposition est variable d'une espèce à l'autre. Notamment, les données de l'espèce *P. p. irlandei* sont plus fragmentées et moins couvertes par l'effort de séquençage, or c'est l'une des espèces dont la place dans la phylogénie nous intéresse le plus. Notons qu'il s'agit des mêmes données moléculaires utilisées par Stiller, Tilic et al., 2020 dans une phylogénie plus large des Terebelloformia. Pour les autres espèces, des échantillonnages et séquençages plus récents des espèces ont permis d'obtenir des transcriptomes de très bonne qualité pour l'étude phylogénétique (développé en chapitre 1).

La divergence des premières lignées de la famille pourrait remonter au Crétacé (-70 millions d'années) voire au Jurassique (-200 millions d'année) selon Jollivet et Stéphane Hourdez, 2020, comme présenté en figure I.10. Cette évaluation s'appuie sur la date estimée de la subduction de la plaque Farallon sous la plaque Nord Américaine entre 23 et 34 millions d'années, qui aboutit à la séparation des champs hydrothermaux de Juan de Fuca et de la dorsale Est Pacifique (Tunnicliffe, 1988). Dans le cas des Alvinellidae, il s'agit d'un point pertinent car cet événement peut être associé à la divergence des espèces soeurs *P. p. irlandei* et *P. p. pandorae*, *P. grasslei* et *P. palmiformis* ainsi que *P. dela* et *P. bactericola*. La résolution de noeuds anciens pour lesquels le signal phylogénétique est moindre peut aboutir à des erreurs systématiques dans l'établissement des arbres optimaux, en particulier si des données sont manquantes pour une espèce importante dans la résolution du noeud,

comme cela peut-être le cas pour l'espèce *P. p. irlandei*. Outre les erreurs statistiques, de réels processus biologiques peuvent brouiller le signal de l'arbre des espèces. Ce point sera abordé en détail au chapitre 1, lors de la résolution de la phylogénie, mais quelques aspects théoriques peuvent être définis au préalable.

Lors de la divergence entre deux lignées, deux processus biologiques principaux peuvent aboutir à des discordances entre les histoires individuelles des gènes et les histoires des espèces, présentés en figure I.11. Le premier est le tri incomplet de lignées (ILS), qui suppose que le polymorphisme allélique précédent la divergence entre des lignées peut amener à la fixation d'allèles différents qui n'ont pas le même coalescent que celui prévu par l'arbre des espèces. Dans la figure I.11, le coalescent entre les allèles bleus précède la divergence des lignées A+B et C. Dans la lignée B, l'allèle bleu clair est fixé, tandis que l'allèle bleu foncé est fixé dans la lignée A, résultant dans un arbre du gène présentant une topologie différente de celle de l'arbre des espèces (Pease et al., 2018 ; L. Cai et al., 2021).

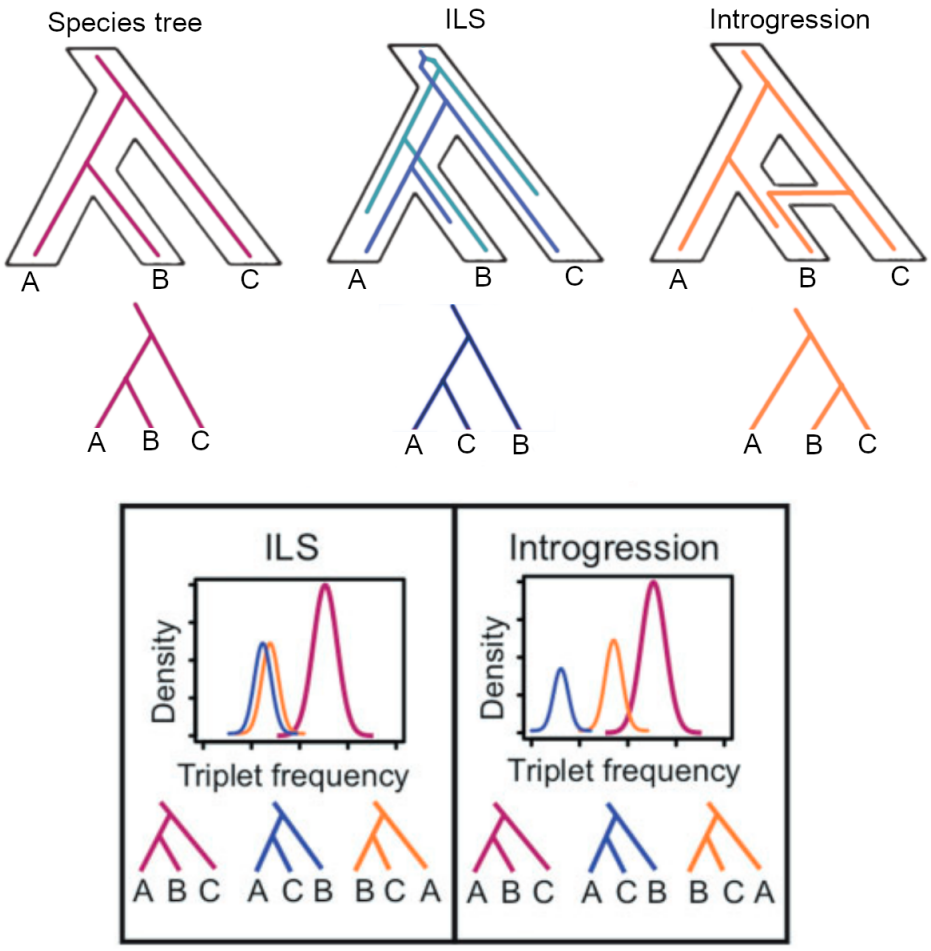


FIGURE I.11 – Tri de lignées incomplet et introgression interspécifique. Le tri de lignées incomplet (ILS), illustré par les allèles bleus, ainsi que l'introgression interspécifique illustrée par les allèles oranges aboutissent à des discordances entre arbres des gènes et arbre des espèces. Les fréquences des histoires de gènes doivent être biaisées dans le cas de l'introgression interspécifique, qui au contraire de l'ILS n'est pas aléatoire entre les lignées. Adapté de L. CAI et al., 2021.

Dans les méthodes de phylogénie classiques où les séquences de plusieurs gènes sont concaténées pour produire un arbre unique des espèces, ce problème est ignoré. Certains algorithmes permettent de partitionner le jeu de données pour permettre à différentes portions de l'alignement du supergène d'évoluer à des vitesses différentes (Chernomor, Von Haeseler et Minh, 2016), mais cela ne résout pas la question des incongruences dans les histoires de gènes. Hudson et Coyne, 2002 ont montré, dans l'hypothèse de deux populations séparées pour lesquelles les seules forces évolutives en jeu sont la mutation et la dérive génétique, que les temps nécessaires pour atteindre un tri complet des lignées est très long. Pour atteindre 50% de tri de lignées entre les deux branches descendantes de l'ancêtre (c'est à dire que 50% des loci de chaque branche sont fixés), il est nécessaire d'attendre 4 à 7 N générations, avec N la taille efficace de population. Pour atteindre 95% de tri de lignées, correspondant au cas d'un arbre unique pour les gènes, 9 à 12 N générations sont requises. Dans ce calcul, pour une espèce avec une taille efficace d'un million d'individu et des temps de génération de cinq ans, le tri de lignées n'est complet qu'après 50 millions d'années. Cet ordre de grandeur est vérifié dans des groupes comme les marsupiaux, dont la radiation date d'environ 60 millions d'années et pour lesquels 50% du génome de certains représentants montre du tri incomplet de lignées (Feng et al., 2022). Chez ces animaux, l'incongruence entre histoires des gènes et histoire des espèces a des conséquences directes sur certains traits morphologiques (formes des os et des dents) qui font se ressembler davantage sur ces caractères des espèces qui sont pourtant plus éloignées les unes des autres dans l'arbre des espèces.

Un second mécanisme qui peut intervenir pour expliquer la différence dans les histoires des gènes est l'introgression inter-spécifique. Dans ce cas, le transfert de gènes est spécifique entre deux lignées dont la divergence a commencé mais qui peuvent encore s'hybrider. Comme précédemment, cela aboutit à un arbre de gènes différent de l'arbre des espèces comme illustré en figure I.11. En revanche, comme ce processus n'est pas aléatoire entre les lignées, il aboutit à un déséquilibre dans les fréquences des histoires de gène observées (L. Cai et al., 2021 ; Pease et al., 2018). Ce déséquilibre peut être plus ou moins fort selon le coefficient de sélection sur le fragment de génome introgressé et le taux de migration entre les populations en cours de spéciation (Martin et Van Belleghem, 2017). Le déséquilibre entre les topologies de gènes peut varier entre les chromosomes et le long du chromosome. Le signal est donc plus simple à détecter dans le cas où l'on peut comparer des structures chromosomiques entières, sur lesquelles des fenêtres d'introgression peuvent apparaître. Dans le cadre de cette thèse, nous avons utilisé des données transcriptomiques qui peuvent provenir de l'intégralité du génome des espèces. Toutefois, les mêmes principes peuvent s'appliquer, y compris si l'on veut déceler l'effet de ces phénomènes dans des populations ancestrales.

Pour tenir compte des incongruences entre histoires de gènes et histoire des espèces, différentes méthodes ont été développées qui en général s'appuient sur l'établissement des arbres de gènes de façon indépendante, puis sur la résolution de l'histoire commune de ces gènes soit par modélisation empirique, soit par optimisation mathématique du super-arbre (Liu, Yu et Edwards, 2010 ; Zhang et al., 2018). Ce genre d'approche est possible avec des données de transcriptomes. Théoriquement, ces modèles surpassent l'approche de concaténation des gènes qui suppose une histoire évolutive unique. Toutefois, ils se heurtent aux erreurs d'estimation pour retracer la topologie de chaque arbre de gène (GTEE). Celles-ci

peuvent fausser les résultats. Molloy et Warnow, 2018 ont montré sur des simulations que la qualité des résultats de l'arbre des espèces dans un modèle de coalescent dépend très fortement de la qualité des arbres de gènes. Comme illustré en figure I.12, les modèles de coalescent ont de meilleures performances que la méthode de concaténation à partir d'un taux de tri de lignée incomplet élevé (40% de différence en moyenne entre les arbres de gènes et l'arbre des espèces) sous réserve que le taux d'erreur pour obtenir l'arbre soit relativement faible (moins de 50%). Pour de fort GTEE en revanche, la méthode de concaténation, qui permet de sommer l'information phylogénétique sur un plus grand nombre de sites, est systématiquement meilleure (bien que les performances soient dans tous les cas mauvaises pour un ILS fort couplé à un GTEE fort). Il s'agit d'une véritable limite à l'utilisation pratique des modèles de coalescent, puisqu'en général l'ILS va tendre à être important sur les branches les plus courtes, qui sont également celles captant le moins d'information phylogénétique et les plus propices à être mal estimées par des alignements de séquences courtes associées à la taille des gènes. Les méthodes phylogénétiques qui utilisent des données génomiques peuvent s'appuyer sur de plus longs alignements de séquences, mais ce n'est pas le cas dans les données de transcrits que nous utilisons dans ce projet. Les méthodes s'appuyant sur la concaténation des gènes restent donc pertinentes à explorer également.

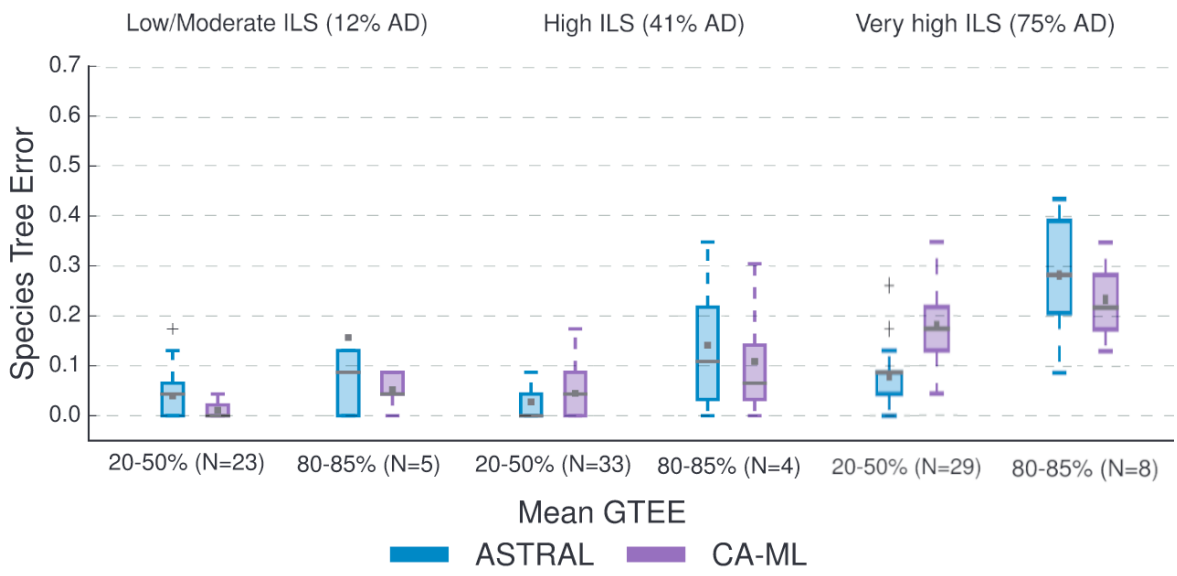


Figure I.12 – Effet de l'erreur d'estimation des arbres de gènes sur un arbre d'espèces dans la théorie du coalescent. CA-ML désigne la méthode de concaténation des gènes et résolution de la phylogénie selon une unique topologie. ASTRAL correspond à la résolution de l'arbre des espèces à partir des arbres de gènes par otpimisation du super-arbre, et est le logiciel utilisé dans le chapitre 1 de cette thèse. AD correspond à la distance de Robinson-Foulds normalisée moyenne entre l'arbre des espèces et l'arbre de gènes pour quantifier le niveau d'ILS, et entre l'arbre de gènes inféré et le véritable arbre de gènes pour quantifier le GTEE. Adapté de Molloy et Warnow, 2018.

I.3.2 Reconstruction de séquences probables de l'ancêtre des Alvinellidae

A partir de la phylogénie, la seconde étape du projet de thèse était de reconstruire certaines protéines ancestrales de la lignée des Alvinellidae et de caractériser leur stabilité thermique. Pour cela, nous avions dressé une liste de protéines candidates permettant d'apporter de potentielles informations supplémentaires sur le paléo-environnement de la lignée. Ces protéines et leur intérêt potentiel sont listées dans le tableau I.2.

Protéine	Longueur	NCBI	Fonction	Objectif	présence
thioredoxine	118	NP_199112.1	réduction de protéines et résidus oxydés (cystéine)	efficacité des systèmes d'oxydoréduction	trouvée chez toutes les espèces
thioredoxine reductase	516	NP_524216.1	réduction de la thioredoxine	efficacité des systèmes d'oxydoréduction	trouvée chez toutes les espèces
superoxyde dismutase Cu/Zn	157	APZ8838	dismutation des radicaux de l'oxygène	efficacité des systèmes d'oxydoréduction	trouvée dans toutes les espèces
prolyl hydroxylase	595	GAA37183.1	synthèse du collagène	activité en condition $< 10\%$ O_2 chez *A. pompejana* (Kaule et al., 1998)	non trouvée chez *P. p. irlandei*
hémoglobine intracellulaire	124	CAI56311	transport de l'oxygène	hypoxie du milieu, méthode ASR disponible (Isogai et al., 2018)	trouvée chez toutes les espèces
protéines inconnues 8 et 13	413 & 414		fonction inconnue	exprimées en condition hypoxiques (Mary et al., 2010)	trouvées et spécifiques chez tous les Alvinellidae
malate dehydrogenase cytosolique	403	NP_190336.1	oxydation du L-malate en oxaloacétate	métabolisme anaérobie, profondeur potentiellement (Dahlhoff et Somero, 1991)	présent dans toutes les espèces
métallothioneine	79	CAA06720.1	affinité pour les métaux Cu, Zn, Cd, Hg	concentration en métaux lourds	trouvée chez *P. grasslei*, *P. palmiformis*, *P. fijiensis*, *P. p. irlandei*, *P.* sp. nov.
nucleotide diphosphate kinase	168	BAU25892.1	hydrolyse de l'ATP	température de l'environnement chez procaryotes (Akanuma, 2017)	trouvée chez toutes les espèces
collagène fibrillaire	260	AAT85578.1	matrice extracellulaire	température de l'environnement Francoise Gaill et al., 1995 ; Sicot et al., 2000	trouvée dans toutes les espèces

Table I.2 – **Protéines candidates pour la reconstruction de séquences ancestrales** – La longueur est indiquée en nombre de résidus par rapport à la référence NCBI répertoriée.

Le choix des meilleures protéines pour conduire l'expérience de reconstruction de séquences ancestrales se fait d'abord en fonction de la présence des séquences contemporaines dans les transcriptomes à notre disposition, de la possibilité d'exprimer la protéine en système hétérologue, et de la pertinence des informations que ces protéines peuvent nous apporter sur le paléo-environnement des Alvinellidae.

I.3.3 Méthodes de reconstruction des protéines ancestrales

La méthode de reconstruction de séquences ancestrales (ASR) permet, à partir de l'alignement de séquences contemporaines orthologues et de leur lien phylogénétique, d'inférer aux différents nœuds de l'arbre dans un cadre probabiliste les séquences appartenant aux organismes ancestraux. Ces séquences peuvent ensuite être synthétisées pour étudier les propriétés des protéines correspondantes (Wheeler et al., 2016 ; Merkl et Sterner, 2016).

L'idée de l'ASR a d'abord été présentée par Linus Pauling et Emile Zuckerkandl (Pauling et al., 1963). Les auteurs présentaient un modèle où la comparaison de séquences trouvées dans des organismes contemporains permettait de reconstruire au moins partiellement la séquence des protéines d'un ancêtre, sur la base des relations phylogénétiques entre les organismes actuels. Ils anticipaient également que cette méthode, outre le fait de déterminer les séquences ancestrales d'organismes n'ayant pas laissé de traces fossilisables, permettrait de déduire dans quelles lignées certaines mutations fonctionnelles sont apparues dans les séquences codantes et d'établir un lien entre biologie évolutive et biologie fonctionnelle. L'ASR a par la suite été améliorée, notamment en passant d'un algorithme parcimonieux à un modèle probabiliste par Koshi et Goldstein, 1996 basé sur l'optimisation de la vraisemblance. Depuis, elle a été mise en oeuvre pour inférer des séquences d'organismes pouvant remonter jusqu'à LUCA il y a 3,8 milliards d'années, aussi bien sur des séquences protéiques que sur des séquences d'ADN ou d'ARNr (Boussau et al., 2008 ; Gaucher, Govindarajan et Ganesh, 2008).

La méthode repose sur un certain nombre d'hypothèses, notamment sur l'arbre du gène d'intérêt s'il est en conflit avec l'arbre des espèces. Dans ce cas, il est possible de construire les séquences ancestrales sous différentes hypothèses de phylogénie et de les comparer, ou bien d'intégrer l'incertitude directement dans le modèle de reconstruction (Hanson-Smith, Kolaczkowski et J. W. Thornton, 2010 ; Isogai et al., 2018). D'autres questions peuvent se poser, concernant par exemple l'estimation des longueurs de branches qui diffèrent entre celles estimées sur l'ensemble des gènes ayant permis d'établir la phylogénie et celles associées à un gène unique. La plupart des auteurs optimisent les longueurs de branches (et donc la dissimilarité potentielle entre séquences ancestrales) sur l'alignement de la protéine d'intérêt, mais la réduction de l'information phylogénétique à quelques sites variables engendre plus d'incertitudes dans l'estimation des divergences entre les séquences.

Dans le dernier chapitre de cette thèse, je me suis par conséquent intéressé aux modèles permettant d'effectuer les reconstructions ancestrales et aux facteurs qui permettraient d'améliorer les résultats. L'intégration vers plus d'informations concernant la protéine étudiée est une piste à privilégier, notamment la structure secondaire de la protéine (Si Quang Le et Olivier Gascuel, 2010 ; Moshe et Pupko, 2019) et l'épistasie entre mutations (Laine,

Karami et Carbone, 2019), qui pourrait être déterminée en observant les combinatoires de mutations au sein de la diversité des séquences homologues recensées dans les bases de données.

Le cas que nous souhaitons étudier est en outre particulier puisque nous nous intéressons à une famille pour laquelle on s'attend à ce que les ancêtres aient vécu à des températures contrastées, tout comme les espèces contemporaines d'Alvinellidae. Or nous avons établi que les températures de vie ont un effet important sur les compositions en acides aminés des protéines. Par conséquent, nous pouvons également envisager des modèles qui s'attaquent plus précisément à cette question en autorisant la variation de composition moyenne en acides aminés dans le temps, selon les biais connus pour corréler avec la température. Ces modèles seraient plus en accord avec l'observation que les espèces contemporaines ont des composition en acides aminés significativement différentes (Foster, 2004 ; Blanquart et Lartillot, 2008).

Un point essentiel est que les méthodes ASR sont fondamentalement probabilistes. La séquence ancestrale réelle ne peut jamais être connue exactement, en cela qu'elle n'est pas nécessairement la séquence la plus probable. Pour cette raison, plusieurs auteurs ont décidé d'exprimer la séquence la plus probable (la séquence dite de maximum de vraisemblance) ainsi que des séquences alternatives moins probables afin d'étudier la variabilité dans les propriétés inférées de la séquence ancestrale (Gaucher, Govindarajan et Ganesh, 2008 ; Hart et al., 2014). La figure I.13 illustre la reconstruction du site de liaison à l'ADN d'un récepteur à l'hormone stéroïde. La protéine de maximum de vraisemblance (ML), est similaire en terme de fonction aux protéines alternatives reconstruites, qui correspondent aux cinq variants de protéines ayant les plus fortes vraisemblances après la reconstruction ML. Toutes les protéines sont activées par la liaison avec les éléments de réponse aux œstrogènes (ERE), mais pas par les éléments de réponse aux stéroïdes (SRE). En revanche, la variation d'amplitude dans cette activation est importante, allant du simple au triple. Ainsi, si la fonction semble bien conservée (notamment parce qu'elle repose en général sur peu d'acides aminés conservés et pour lesquels le signal phylogénétique est fort), la quantification fonctionnelle de la protéine est sujette aux incertitudes de l'ASR (Eick et al., 2016).

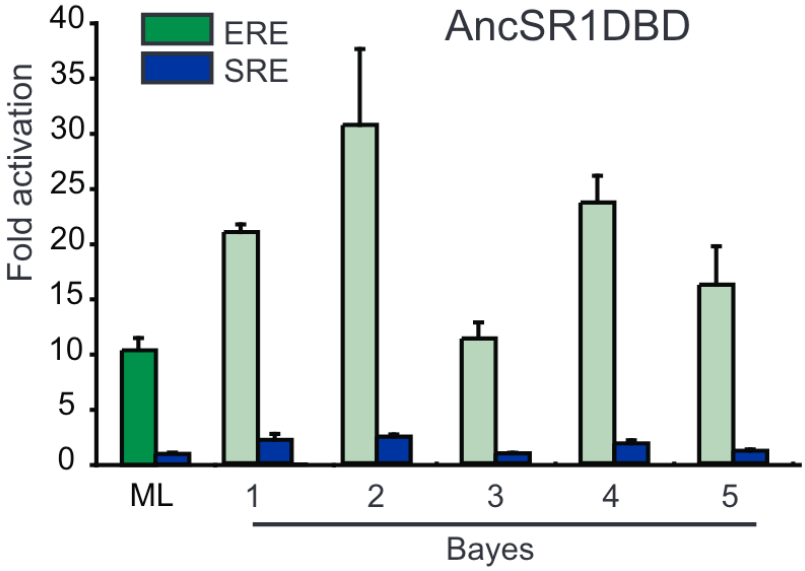


FIGURE I.13 – Variabilité phénotypique des variants ASR d'une protéine récepteur des hormones stéroïdes. L'activation par les éléments de réponse aux œstrogènes (ERE) et stéroides (SRE) du récepteur le plus vraisemblable (ML) est comparée avec cinq variants bayésiens ayant des vraisemblances élevées. Tiré de EICK et al., 2016.

Dans notre cadre d'étude, le caractère phénotypique qui nous intéresse est la stabilité thermique des protéines reconstruites. Or les méthodes ASR sont suspectées de biaiser les reconstructions en faveur de protéines plus stables qu'elles ne l'étaient réellement, observant que la plupart des protéines ainsi reconstruites étaient plus stables que les protéines contemporaines utilisées pour la reconstruction (Wheeler et al., 2016 ; Merkl et Sterner, 2016). Plusieurs raisons ont été invoquées pour expliquer cela (effet consensus, biais dans les matrices de mutations utilisées), mais l'explication la plus convaincante à mon sens est le choix systématique de la protéine de vraisemblance maximale (ML) dans le choix de la protéine ancestrale. Rationnellement, la protéine ML est effectivement la meilleure protéine que l'on puisse obtenir et certainement celle qui partage la plus grande identité avec la protéine ancestrale vraie. Cependant, le choix du résidu le plus probable à chaque position de la protéine indépendamment tendrait, en moyenne, à favoriser des protéines plus stables parce qu'on corrigerait les "erreurs" stochastiques de l'évolution qui parfois déstabilisent la protéine. Williams et al., 2006 ont ainsi montré, sur la base de la reconstruction de protéines hypothétiques dont on peut simuler l'énergie libre de dénaturation, que toutes les méthodes de reconstruction (parcimonie, maximum de vraisemblance ou bayésienne) aboutissent à des différences entre l'ancêtre vrai et l'ancêtre reconstruit, de l'ordre de 1 kcal/mol dans ces simulations en moyenne comme montré en figure I.14A. En revanche, dans le cas des reconstructions ML, cette différence est biaisée vers le choix de protéines plus stables (figure I.14B.). Ainsi, la reconstruction ML aboutit plus souvent, mais pas nécessairement, à une protéine plus stable de manière artéfactuelle. Étonnement, la reconstruction parcimonieuse qui n'intègre pas de modèle évolutif à proprement parler, et qui a plus d'erreurs dans la séquence ancestrale inférée en terme de pourcentage d'identité avec l'ancêtre vrai, a un biais orienté dans le même sens mais moins fort. Le fait que la méthode parcimonie incorpore plus d'erreurs dans la séquence tend à atténuer l'erreur systémique de la méthode ML. La meilleure méthode, dans ces simulations, est l'échantillonnage bayésien. Dans ce cas, plutôt que de choisir systématiquement le résidu le plus probable à chaque position de la protéine, les résidus sont sélectionnés aléatoirement selon leur probabilité postérieure. Cette méthode permet d'annuler l'orientation dans le biais de thermostabilité accrue de la méthode ML. En revanche, elle ne permet pas en soi de faire disparaître le problème que la séquence ancestrale réelle pouvait être plus ou moins stable que les séquences ancestrales reconstruites.

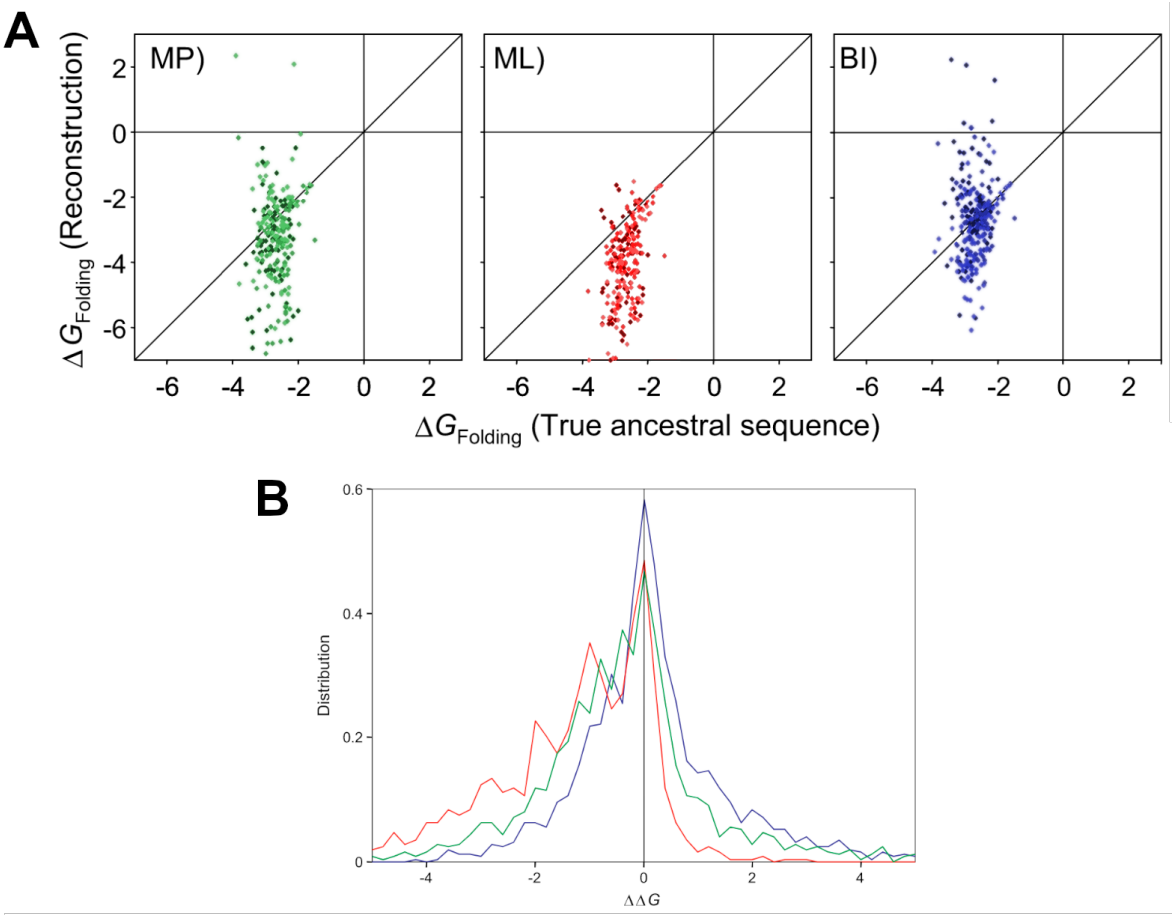


FIGURE I.14 – Effet du modèle de reconstruction des séquences sur la stabilité des protéines ancestrales. Les reconstructions de protéines simulées sont effectuées par maximum de parcimonie (MP, en vert), maximum de vraisemblance (ML, en rouge) ou échantillonnage bayésien (BI, en bleu). (A) Distribution de l'énergie libre de dénaturation ΔG^o en kcal/mol de l'ancêtre réel par rapport à celle de la protéine reconstruite pour un lot de protéines. (B) Déviation entre les énergies libre de dénaturation de la protéine vraie par rapport à l'ancêtre inféré en kcal/mol pour un lot de protéines. Adapté de WILLIAMS et al., 2006.

Récemment, Arenas et Bastolla, 2020 ont développé une méthode qui utilise directement la structure tri-dimensionnelle des protéines pour simuler la stabilité des protéines et effectuer des reconstructions de séquence qui tentent de maintenir la stabilité constante au cours de l'évolution. Cette approche est intéressante car elle tente directement de résoudre le problème de stabilité trop élevée des méthodes ML. Toutefois, cela suppose que la stabilité des protéines est relativement constante, or dans notre modèle d'étude nous cherchons justement à déterminer si c'est le cas. Ce genre de modèles est certainement prometteur mais peu d'études ont pour l'instant été réalisées pour confirmer la viabilité des résultats obtenus sur des protéines ayant des histoires de stabilité complexe.

L'étude de Williams et al., 2006 conclut que la meilleure stratégie est certainement de choisir un lot de protéines ancestrales échantillonnées depuis leurs probabilités bayésiennes, puis de les caractériser afin d'estimer la confiance que l'on a dans le phénotype mesuré des ancêtres. Cette approche est également défendue par Eick et al., 2016. Toutefois, ces derniers auteurs notent qu'un échantillonnage strictement bayésien aboutit à l'expression de protéines ayant de très faibles probabilités d'être vraies, et qui montrent parfois des phénotypes impossibles (par exemple des protéines non fonctionnelles). Cela renvoie à une question sous-jacente mais rarement mise en avant, qui est celle de la fiabilité en eux-mêmes des modèles phylogénétiques utilisés pour les reconstructions. La figure I.15 montre les résultats d'une reconstruction ancestrale effectuée sur une phylogénie expérimentale d'une protéine fluorescente rouge (RFP, 225 acides aminés) pour laquelle les séquences ancestrales réelles sont connues (Randall et al., 2016). Le point intéressant est que tous les modèles utilisés (maximum de vraisemblance par différents logiciels ou maximum de parcimonie) aboutissent à des séquences reconstruites globalement proches de la séquence réelle pour l'ancêtre 21 qui est le plus ancien de la phylogénie (plus de 95% d'identité, voir figure I.15). En revanche, on constate que les modèles probabilistes de phylogénie produisent des résultats qui ne sont pas beaucoup meilleurs que la méthode de parcimonie, qui n'est pourtant pas un modèle d'évolution. Dans cet exemple, les modèles d'évolution permettent essentiellement d'arriver au même résultat que la parcimonie sur les résidus faciles à résoudre (plus de 90% de la séquence), et ne permettent de mieux inférer que 10 à 20% de résidus ambigus de plus que la parcimonie, ce qui est assez faible.

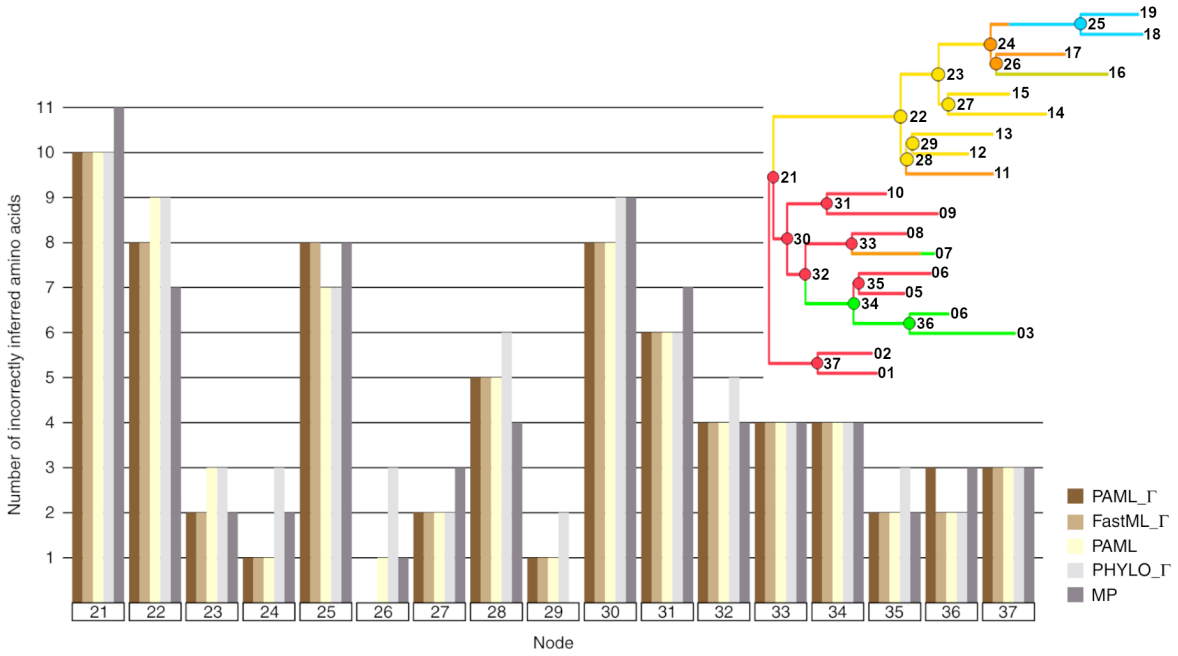


FIGURE I.15 – Erreurs dans les reconstructions d'ancêtres d'une phylogénie expérimentale d'une protéine RFP. La phylogénie est montrée et les couleurs des branches correspondent aux variations phénotypiques de la fluorescence de la RFP. Le graphique montre le nombre d'erreurs dans la séquence inférée par rapport à la séquence ancestrale réelle à différents noeuds de l'arbre, selon différentes implémentations de maximum de vraisemblance (PAML, FastML et PhyloBayes avec ou sans paramètres Γ) ou maximum de parcimonie (MP). Adapté de RANDALL et al., 2016.

Ce résultat est peut être dû au fait que les modèles de phylogénie sont d'abord pensés pour trouver les arbres phylogénétiques. Dans ce cas, la supériorité des approches probabilistes par rapport à la parcimonie n'est plus à démontrer, notamment pour limiter les cas d'attraction des longues branches. Toutefois, leur usage en ASR est quelque peu détourné et ces modèles ne sont pas spécifiquement optimisés pour bien évaluer les séquences ancestrales. Un échantillonnage bayésien, comme effectué par Williams et al., 2006 sur la base de séquences simulées se conformant exactement aux modèles évolutifs, n'a de sens que si les probabilités des résidus sont correctement estimées. Or le fait qu'une part importante de séquences échantillonnées dans l'espace bayésien selon Eick et al., 2016 s'avèrent non fonctionnelles (l'intégralité des 5 séquences bayésiennes de la protéine AncSR1LBD de l'article) interroge sur la fiabilité des probabilités accordées à chaque séquence. Par conséquent, l'incertitude probabiliste de la méthode n'est en soi pas un problème, mais l'évaluation de la qualité des probabilités postérieures est un point qui mérite d'être étudié.

Enfin, dans une dernière partie du chapitre, j'ai essayé de proposer une méthode pour reconstruire les événements d'insertion et délétions (indels) dans un modèle de vraisemblance. Quelques résultats ont pu être obtenus pour essayer de traiter les indels comme des caractères phylogénétiques à part entière (Simmons et Ochoterena, 2000 ; Rivas, 2005).

Chapitre II

Etablissement de la Phylogénie des Alvinellidae

L'objectif de cette première partie était de déterminer une phylogénie des Alvinellidae incluant le plus d'espèces possibles de la famille. Des extractions d'ARN ainsi que des séquençages avaient été préalablement effectués pour 11 des 14 espèces, incluant la nouvelle espèce *Paralvinella mira* échantillonnée dans l'océan Indien (Han et al., 2021) ainsi qu'une espèce encore non décrite du Bassin de Manus, *Paralvinella* sp. nov. Cette phylogénie doit servir de point de départ pour la reconstruction ultérieure des protéines de l'ancêtre des Alvinellidae, qui nécessite de connaître les relations évolutives entre les différentes séquences contemporaines utilisées.

Notre phylogénie est enracinée grâce à quelques transcriptomes d'Ampharetidae ainsi que des espèces plus distantes de Terebelloformia. Peu de données moléculaires sont disponibles parmi les Ampharetidae, aussi avons-nous utilisé les données de séquençage produites par Stiller, Tilic et al., 2020.

Nous avons résolu la phylogénie avec fiabilité au sein du genre *Paralvinella*. En revanche, les lignées les plus profondes de la famille montrent des signaux phylogénétiques discordant entre les arbres de gène. Nous attribuons cette discordance de signal à de forts taux de tri de lignées incomplet, ce qui est fréquent lors de la radiation rapide de plusieurs espèces, ainsi qu'à une forte introgression interspécifique entre les lignées ancestrales amenant aux actuelles espèces *Paralvinella*, *Alvinella* et *Nautalvinella* (*P. unidentata* et *Paralvinella pandorae*).

La phylogénie principale retenue est en désaccord avec la vision classique des Alvinellidae, séparés en deux genres *Alvinella* et *Paralvinella*. Dans cette nouvelle proposition, les espèces *Paralvinella pandorae* sont soeurs des autres espèces d'Alvinellidae, et le genre *Paralvinella* n'est plus monophylétique. L'espèce *P. unidentata* est groupée avec les espèces *Alvinella*, et par conséquent l'espère pourait appartenir au genre *Alvinella*.

Cette topologie, en revanche, n'est pas valable pour l'ensemble des gènes étudiés. Nous avons pris soin dans cette partie de disposer des meilleures données moléculaires possibles pour les lignées directement impliquées dans la topologie de la racine de la famille, bien

que la qualité du séquençage de *P. p. irlandei* soit de moins bonne qualité. Pour confirmer ces résultats, il serait nécessaire de disposer du génome complet de plusieurs espèces, notamment *A. pompejana*, *P. p. irlandei* et *P. unidentata*, afin de constater si la proposition d'introgression interspécifique entre les lignées ancestrales est confirmée au niveau de la structure des relations phylogénétique entre ces trois espèces le long du génome. Ce chapitre a fait l'objet d'un article soumis au journal Molecular Phylogenetics and Evolution et est actuellement en cours d'évaluation.

A step in the deep evolution of Alvinellidae (Annelida : Polychaeta) : a phylogenomic comparative approach based on transcriptomes

Pierre-Guillaume Brun[1], Stéphane Hourdez[2], Marion Ballenghien[1], Yadong Zhou[3], Jean Mary[1*] and Didier Jollivet[1*,]**

[1] Station Biologique de Roscoff, Sorbonne Université-CNRS, UMR 7144, Place G. Teissier, 29280 Roscoff, France
[2] Observatoire Océanologique de Banyuls, Sorbonne Université-CNRS, UMR 8222, 1 Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France
[3] Key Laboratory of Marine Ecosystem Dynamics, 310000 Hangzhou, China
[*] JM and DJ are joint senior authors

[**] jollivet@sb-roscoff.fr

II.0.0.0.0.1 Abstract The Alvinellidae are a family of worms that are endemic to deep-sea hydrothermal vents in the Pacific and Indian Oceans. These annelid worms, a sister group to the Ampharetidae, occupy a wide range of thermal habitats. The family includes the most thermotolerant marine animals described to date such as the Pompeii worm *Alvinella pompejana*, and other species living at much lower temperatures such as *Paralvinella grasslei* or *Paralvinella pandorae*. The phylogeny of this family has not been studied extensively. It is, however, a complex case where molecular phylogenies give conflicting results, especially concerning the monophyletic or polyphyletic character of the genus *Paralvinella*. We carried out a comprehensive study of the phylogeny of this family using the best molecular data currently available from RNAseq datasets. The study is based on the assembly of several hundred transcripts for 11 of the 14 species currently described or in description. The results obtained by the most popular phylogenetic inference models (gene concatenation with maximum likelihood, or coalescent-based methods from gene trees) are compared using a series of ampharetid and terebellid outgroups.

Our study shows that a high number of gene trees support the hypothesis of the monophyly of the *Paralvinella* genus, as initially proposed by Desbruyères and Laubier, in which the species *Paralvinella pandorae* and *Paralvinella unidentata* are more closely related within the subgenus *Nautalvinella*. However, the global phylogenetic signal favors the hypothesis of paraphyly for this genus, with *P. pandorae* being sister species of the other Alvinellidae. Gene trees separated equally between these two hypotheses, making it difficult to draw conclusions about the initial split of the MCRA as different genomic regions seem to have very different phylogenetic stories. According to molecular dating, the radiation of the Alvinellidae was rapid and took place in a short period of time between 70 and 80 million years ago. This is reflected at the genomic scale by high rates of incomplete lineage sorting between the first ancestral lineages with probable gene transfers between the ancestors of *Alvinella*, *Nautalvinella*, and the rest of the *Paralvinella* lineages. Alvinellidae, hydrothermal vents, phylogenomics, RNAseq, transcriptomics.

II.1 Introduction

After the discovery of hydrothermal vents and their associated communities on the Galapagos rift in the late 1970s, the Pompeii worm *A. pompejana* Desbruyères et Laubier, 1980, was one of the first emblematic species collected from the vent chimney environment following the sampling of a "black smoker" chimney at 21°N on the East Pacific Rise (EPR) (Monaco et Prouzet, 2015). Two morphological types living in close association (syntopy), initially viewed as ontogenic forms, were later found to represent two distinct species, *A. pompejana* and *A. caudata* Desbruyères et Laubier, 1986 (Jollivet et Stéphane Hourdez, 2020). While the two *Alvinella* species were initially described as aberrant forms in the family Ampharetidae, some obvious specificities, such as the lack of separation between the head and the rest of the body, a process of cephalization of the gills, uncini with a reduced number of teeth and notopodia with simple dorsal chaetae, led the authors to suggest the creation of the subfamily initially named Alvinellinae (Desbruyères et Laubier, 1980 ; Desbruyères et Laubier, 1982). Following the discovery of the species *P. grasslei*

Desbruyères et Laubier, 1982, at several sites in the EPR and the Guaymas basin (eastern Pacific), the Alvinellinae were split into two genera, *Alvinella* and *Paralvinella*. The distinction between the two genera was based in particular on the shapes of the mouth apparatus and of the secondary filaments of the gills, which are lamellar in *Alvinella* and filamentous in *Paralvinella* Desbruyères et Laubier, 1982.

Subsequently, new species of Alvinellinae were discovered, all endemic to hydrothermal vents in the Pacific Ocean. In the south-east Pacific, two additional species were found and described as *P. p. irlandei* Desbruyères et Laubier, 1986, (EPR) *P. p. irlandei* and *P. bactericola* Desbruyères et Laubier, 1991 (Guaymas basin). In the north Pacific, several new species were described from the Juan de Fuca Ridge with *P. p. pandorae* Desbruyères et Laubier, 1986, *P. palmiformis* Desbruyères et Laubier, 1986, *P. dela* Detinova, 1988, as well as *P. sulfincola* Desbruyères et Laubier, 1993 found on top of vent chimneys. Finally, three additional species were added to the family from expeditions in the western Pacific : *P. hessleri* Desbruyères et Laubier, 1989 is found in the Marianas back-arc basins and Okinawa Trough, while *P. fijiensis* Desbruyères et Laubier, 1993 and *P. unidentata* Desbruyères et Laubier, 1993 live in the south-west Pacific.

As the number of species in the subfamily increased, subsequent studies using both morphological characters and molecular data demonstrated that inside Terebelliformia, alvinellid worms form a monophyletic group, sister-group-to, but not nested within the Ampharetidae. The family was accordingly renamed Alvinellidae (Rousset, Rouse et al., 2003 ; Glasby, Hutchings et Hall, 2004 ; Stiller, Tilic et al., 2020). Recently, a new species, *P. mira* Han, Zhang, Wang & Zhou, 2021, genetically close to *P. hessleri*, has been described outside of the Pacific Ocean, from the hydrothermal vents of the Carlsberg Ridge in the Indian Ocean (Han et al., 2021). The Alvinellid worms' distribution has thus expanded beyond Pacific waters, supporting the purported observation of a small population of alvinellid worms in the Solitaire field on the Central Indian Ridge, though they have been neither sampled nor taxonomically resolved (Nakamura et al., 2012). Finally, another species, whose description is still in progress, was first discovered in 2011 from the Nifonea vent site along the volcanic arc of Vanuatu in the Manus Bassin. This species is referred to as *P.* sp. nov. in this study.

Alvinellid worms probably have a long evolutionary history of speciation, which could date back between 70 and 200 millions of years ago (Jollivet et Stéphane Hourdez, 2020). Considering the amino acid composition of reconstructed ancestral protein sequences and the very low evolutionary rates of proteins in the thermophilic lineages (Fontanillas et al., 2017) led Fontanillas et al. to hypothesise that the ancestor of the Alvinellidae was a thermophilic species, suggesting an old colonization of the hydrothermal environment. Yet the family has now diversified to cope with a wider variety of vent conditions. Alvinellidae notably include two of the most thermotolerant animals known to date, namely *A. pompejana* and its ecological homolog *P. sulfincola* thriving between 40 and 50°C (Girguis et Lee, 2006 ; Ravaux et al., 2013), while species such as *P. grasslei* are comfortable in temperatures around 15°C (Cottin et al., 2008).

The ecological diversity of these closely-related worms is particularly intriguing. The question of the evolutionary path taken by the alvinellids naturally arises, as well as the different stages that allowed its dispersal throughout the Pacific to the Indian Ocean, including

their putative initial link with the hydrothermal paleocoastal environment. A necessary step in understanding this evolutionary history is the establishment of a robust phylogeny of the family, including all currently known and described species. More specifically, the place of *P. p. irlandei* and *P. p. pandorae* in the phylogeny, which determines the monophyly of the genus *Paralvinella*, remains uncertain. Resolving this question would represent a important step in our understanding of the relationships between the early lineages of the family.

Several partial phylogenies of the Alvinellidae have been proposed. Based on detailed observations of morphological characters, Desbruyères and Laubier divided the family into two monophyletic genera, *Alvinella* and *Paralvinella* (Desbruyères et Laubier, 1993). Species of the genus *Alvinella* have gills with flattened-leaf secondary filaments, the anterior fourth and fifth setigers are modified with the insertion of two pairs of hooks, and a pair of thick tentacles is inserted on the buccal apparatus in males. Species of the genus *Paralvinella* have stalked bipennate gills with two rows of cylindrical filaments and no filaments at the tip of the stem, the seventh setiger is modified with a single pair of hooks, and the buccal tentacles of males are different (either three-lobed, coiled and tapered, or absent) (Jollivet et Stéphane Hourdez, 2020). Other non-morphological differences between the two genera have been noted : *Alvinella* species secrete thick parchment-like tubes including inorganic material and have epibiotic bacteria, whereas *Paralvinella* species secrete mucus tubes or cocoons (with the exception of *P.* sp. nov., personal communication) and are devoid of filamentous epibiotic bacteria (Desbruyères et Laubier, 1991).

Within the *Paralvinella* species, Desbruyères and Laubier further proposed to group the species *P. p. pandorae*, *P. p. irlandei* and *P. unidentata* within the subgenus *Nautalvinella*, considering the comb-like insertion of the secondary filaments of the gills as well as the lack of tentacles on the buccal apparatus of the males. However, some peculiarities are noted in the species *P. unidentata*, notably the leaf-like shape of the secondary filaments of the gills, similar to those of the *Alvinella* species (Desbruyères et Laubier, 1993 ; Jollivet et Stéphane Hourdez, 2020).

Recent molecular phylogenies, on the other hand, have produced conflicting results regarding the subdivision of the family into two monophyletic genera. Jollivet et al. supported this view, as well as the existence of the *Nautalvinella* subgenus, based on enzyme isoforms similarities (Jollivet, Desbruyères et al., 1995). Another study by Jollivet and Hourdez, based on a set of 278 orthologous transcripts, came to the same conclusion, with *P. p. irlandei* being a sister species to the other *Paralvinella* species (Jollivet et Stéphane Hourdez, 2020). Counter to this, a recent wide phylogeny for Terebelliformia by Stiller et al., based on several hundred orthologous genes derived from transciptomic datasets, proposed with high support that the *Paralvinella* genus is polyphyletic, with *P. p. irlandei* being sister to the other Alvinellidae (Stiller, Tilic et al., 2020). Interestingly, this later study also fitted a previous mitochondrial sequence analysis done by (Vrijenhoek, 2013) but neither of them include sequences from *P. unidentata*, in contrast to studies suggesting the monophyly of the genus, which then anchor the clades *P. p. irlandei*, *P. p. pandorae* and *P. unidentata* together in the phylogeny. It appears that the phylogenetic relationship between the three *Nautalvinella* species and the *Alvinella* species is therefore key to understanding the deep phylogeny of the Alvinellidae.

Here, we propose a new molecular phylogeny of the family Alvinellidae based on several hundred orthologous genes derived from RNAseq datasets, notably improved by a new sequencing of the species *P. unidentata* and other species from both the western Pacific and the Indian Ocean. The phylogeny therefore includes all described species of Alvinellidae, with the exception of *P. p. pandorae*, and the rare species *P. dela* and *P. bactericola* for which no exhaustive molecular data or samples exist to date. This phylogeny is rooted with species of the sister group Ampharetidae, as well as more distantly-related species of Terebelliformia and the outgroup coastal pectinariid worm *Pectinaria gouldii*, in accordance with the currently accepted phylogeny of the Terebelliformia worms (Rousset, Pleijel et al., 2007 ; Stiller, Tilic et al., 2020). Finally, the results obtained by different methods are compared, following a concatenation + Maximum Likelihood approach on the one hand, and a gene trees + coalescent approach on the other hand.

II.2 Materials and Methods

II.2.1 Animal Collection, Sequencing and Assembly

Transcripts of *A. pompejana* were predicted from the genome assembled at the chromosomal level by R. Copley (NCBI accession number PRJEB46503) with the Augustus webserver (Hoff et Stanke, 2013). Two individuals from 9°50N/EPR were used to obtain the genome, one given by C.G. Cary and the other one collected during the Mescal 2012 French cruise with ROV *Victor6000* and RV L'Atalante. The software Augustus was first trained on cDNA sequences provided by the MPI that contained previous *A. pompejana* EST assemblies obtained from a Sanger sequencing (Genoscope project, see Gagnière et al., 2010 for details). The prediction was performed on scaffolds of more than 60k bases (95.8% of the assembled genome). The resulting coding sequences were filtered with Kraken v.2.0.9 and then Transdecoder v.5.5.0 to ensure data homogeneity among all final transcriptomes.

RNAseq reads for the species *P. gouldii*, *Anobothrus* sp., *Amphicteis gunneri*, *Amphisamytha carldarei* and *Hypania invalida* were downloaded from the NCBI SRA database (accession numbers SRR2057036, SRR11434464, SRR11434467, SRR11434468, SRR5590961) (Kocot et al., 2016 ; Stiller, Tilic et al., 2020). Total RNA extractions from flash-frozen tissues were performed with Trizol. Libraries of *P. gouldii* were sequenced with Illumina HiSeq, libraries of *Anobothrus* sp., *A. gunneri* and *A. carldarei* were obtained with Illumina Hiseq 4000, and Illumina Hiseq 2500 was used for *H. invalida*.

An EST library was previously obtained for the species *P. sulfincola* in the framework of a Join Genome Initiative project (NCBI accession number PRJNA80027), led by P.R. Girguis and S. Hourdez (Girguis et Lee, 2006). The animals were collected on the Juan de Fuca Ridge during the jdFR oceanographic cruise in August and September 2008, with the submersible Alvin on board of the R/V Atlantis. Reads were obtained by 454 Roche technology, and 24,702 transcripts were assembled using Newbler (Margulies et al., 2005).

Other alvinellid species were collected during several oceanic cruises from 2004 to 2019 on board of the N/O L'Atalante and using either the ROV *Victor6000* or the manned submer-

sible Nautile with the exception of *P. palmiformis*, which was also collected with *P. sulfincola* during the same jdfR cruise on Juan de Fuca. Information about the sampling locations of the alvinellid worms collected along the East Pacific Rise are provided in Fontanillas et al., 2017. Other alvinellid species and the vent terebellid were sampled from different vent sites of the western Pacific back-arc basins during the Chubacarc 2019 cruise with the tele-manipulated arm of the ROV *Victor6000* and brought back to the surface in an insulated basket. *P. hessleri* was sampled from Fenway in the Manus basin, *P. fijiensis* from Big Papi (Manus Basin) and Tu'i Malila (Lau Basin), *P. unidentata* from Fenway (Manus Basin), *P.* sp. nov. from the volcano South Su (Manus Basin) and the vent terebellid Terebellidae gen. sp. (not yet described) at Snow Cap (Manus Basin). Finally, *P. mira* was sampled from the Wocan vent field in the northwest Indian Ocean on the Carlsberg Ridge by HOV Jialong during the DY38 cruise in March 2017 (Han et al., 2021). *Melinna palmata* and *Neoamphitrite edwardsi*, which are shallow-water species, were respectively sampled in the bay of Morlaix and Roscoff, France. Total RNA extraction from flash-frozen tissue were performed with Trizol after tissue grinding using a ball mill. RNAseq libraries were produced at Genome Québec following a polyA purification of mRNAs, and sequenced using the Novaseq 6000 technology.

With the exception of *A. pompejana* and *P. sulfincola* species, all transcriptomes were assembled using a common procedure. The reads were first cleaned of adapters and trimmed with Fastp v.0.20 to retain reads with a phred-score above 25 for each base (Chen et al., 2018). Kraken v.2.0.9 was then used to remove reads corresponding to prokaryotic contamination (Wood et Salzberg, 2014). The reads retained at this stage were assembled *de novo* using Trinity v.2.9.1 (Grabherr et al., 2011). Finally, the assembled transcripts with an identity greater than 99% on 50 consecutive bases were assembled with cap3 v.10.2011 (Huang et Madan, 1999). Results of the sequencing, filtration steps, and assembly metrics are detailed in table 1 of the Supplementary data, together with the associated command lines. Transcriptomes are available from the NCBI BioProject repository with the following accession numbers (*A. caudata* XXX, *P. unidentata* XXX, *P. p. irlandei* SRX1843887, *P. palmiformis* XXX, *P. grasslei* XXX, *P. mira* XXX, *P. hessleri* XXX, *P.* sp. nov. XXX, *P. fijiensis* XXX, Terebellidae gen. sp. XXX, *M. palmata* XXX, *N. edwardsi* XXX).

The Open Reading Frames (ORFs) were then identified with Transdecoder v.5.5.0, firstly trained with the 500 longest sequences accessible from each transcriptome. The predicted ORFs were retained, stripped of the 3' and 5' UTR regions. The metrics associated with these different steps are also presented in the table 1 of the Supplementary data.

In total, 25 transcriptomes were assembled and compared, from 19 different Terebelliformia species including 11 species of the family Alvinellidae.

II.2.2 Search for orthologous genes and Bioinformatic Processing

Orthologous sequences were determined from the assembled transcriptomes using Orthograph v.0.7.2 (M. Petersen et al., 2017). Following the recommendations of Orthograph, a reference database of 1997 orthogroups (OG) was first constructed from the ODB9 database (Zdobnov et al., 2021). These OGs correspond to single copy genes in all lophotrocho-

zoan species available when downloading ODB9 (*Lottia gigantea*, *Crassostrea gigas*, *Biomphalaria glabrata*, *Capitella teleta*, *Helobdella robusta*). Then, each Terebelliformia transcriptome was blasted against these OGs with Orthograph to identify orthologous genes. These steps attempted to identify all orthologous genes present in Terebelliformia that are expected to be single copies. The results of these ortholog identification steps are detailed in Supplementary data Table 1.

The main uncertainty in the phylogeny of the Alvinellidae lies at the root of the family (see Results section), and concerns the relationships between the branches leading to *P. p. irlandei*, *P. unidentata*, *Alvinella* species and other *Paralvinella* species. For this reason, we retained only orthologous genes groups that contained at least one gene for each of these species or clade.

The nucleotide sequences were aligned for each gene with MACSE v.2.05, which is a codon-gaps aligner for coding sequences (Ranwez et al., 2011). The translated amino acid sequences were aligned with Probcons v.1.12 (Do et al., 2005).

From these sequences alignments, we obtained the following datasets :

1. a set of 657 orthologous genes, in nucleotides, present in at least 20 transcriptomes and including a gene ortholog for the species *P. p. irlandei* and *P. unidentata*, and, at least, a gene ortholog for one of the two *Alvinella* species (*A. pompejana* or *A. caudata*) and one of the other *Paralvinella* species. Aligned fragments shorter than 60 nucleotides between two subsequent gaps were considered unreliable and discarded. The third nucleotide of each codon was also removed to reduce the saturation of the phylogenetic signal. Sequences shorter than 20% of the length of the total gene alignment were also discarded. Finally, each orthologous gene group was tested with IQ-TREE 2.0.3 (Minh et al., 2020) to guarantee that the phylogenetic assumptions about residue homogeneity and stationarity were not violated using the tests of symmetry at a 5% threshold (Naser-Khdour et al., 2019). If an orthologous group failed the test, biased sequences (excluding mandatory species) were eliminated until the alignment did reject the homogeneity and stationarity hypothesis. Uninformative sites at which at most one species had a residue were also removed ;

2. a set of 699 orthologous amino-acid translated genes with identical constraints to those described for the nucleotidic sequences, but with the removal of fragments shorter than 20 amino-acids when included between two subsequent gaps.

These sets were then concatenated into two supergenes. The first, in nucleotides (two first codon positions of the coding sequence), contained 657 genes and 499,036 sites, with *A. caudata* being the longest sequence with 461,042 sites, and *P. gouldii* the shortest with 245,846 sites. The second, in amino-acids, contained 699 genes and 277,900 sites, with *A. caudata* being the longest sequence (255,557 sites) and *P. gouldii* the shortest (139,067 sites).

II.2.3 Phylogenetic inference

Identifying well-resolved clades

A global unrooted phylogeny was evaluated from the supergenes with the maximum likelihood (ML) approach implemented in IQ-TREE 2.0.3. The supergenes were first partitioned according to each gene, and we used the merging strategy implemented in IQ-TREE to retrieve relevant merged partitions (Chernomor, Von Haeseler et Minh, 2016). For each merged partition, ModelFinder, was used to infer the best-fitting evolutionary model according to the minimized Bayesian Information Criteria (BIC) (Kalyaanamoorthy et al., 2017). For nucleotides and amino-acids sequences, heterogeneity of the evolutionary rate between sites was modelled by a Γ distribution with four discrete classes (Yang, 1994). For amino-acid sequences, empirical or ML equilibrium amino-acid frequencies were both tested, and mixture models were also added to the standard test of ModelFinder (Si Quang, Olivier Gascuel et Lartillot, 2008 ; Si Quang Le, Lartillot et Olivier Gascuel, 2008 ; S. Q. Le, Dang et O. Gascuel, 2012). Finally, the robustness of the tree was inferred *via* 1,000 ultra-fast bootstraps (Hoang et al., 2018).

Testing alternative topologies

It appeared that the nodes at the root of the Alvinellidae family were the most conflicting ones. Therefore, we compared 15 alternative topologies one to each other. In these topologies, both the *Paralvinella* clade (excluding *P. p. irlandei* and *P. unidentata*) and the set of outgroup species, which were well resolved in the first step of the tree reconstruction, were constrained to remain identical among the 15 topologies. The outgroup species arrangement obtained from our own results was also consistent with the Terebelliformia phylogeny proposed by Stiller, Tilic et al., 2020.

The 15 topologies were evaluated in a ML framework using IQ-TREE, using the GTR+Γ4+F model for the nucleotide-encoded supergene taken as a single partition, and the LG+Γ4+F model for the amino-acid encoded supergene also taken as one partition. The evolutionary model used for all supergenes corresponds to the most commonly chosen models obtained under a complete gene partition. The topologies were then tested against the best topology obtained for each supergene using the AU test at a 5% threshold with 10,000 RELL replicates implemented in IQ-TREE (Shimodaira, 2002). This test gives the probability that a topology scores better than the reference topology after a site-to-site resampling.

The topologies were also tested against each individual gene in the coalescent framework. To this end, gene trees, for either nucleotide or amino-acid encoded sequence alignments, were first optimized using IQ-TREE. Substitution models were tested with ModelFinder, including mixture models for amino-acid sequences. Species trees were then scored from the gene trees using ASTRAL v.5.7.8 (Zhang et al., 2018). Branches with low bootstrap support in the gene trees ($<$10%) were contracted beforehand, according to the recommendations of the authors. To test the relevance of the different topologies, the AU test used for ML species tree topologies obtained from supergenes was transposed to the coalescent framework. To this end, instead of testing ML scores of different topologies against a reference

one after site-to-site resampling, we tested the coalescent scores of the different species tree topologies, as expressed by the quartet scores, after gene-to-gene resampling. The multiscale boostrap sampling procedure described in Shimodaira, 2002 was recoded using the same scale parameters, carrying out 10 successive resamples of 10,000 bootstrap samples, each comprising from 0.5 to 1.4 times the total number of genes. The 10 multi-scale boostraps are used to score the species trees with an increasing number of genes involved, and the signed distance and curvature parameters are obtained with their maximum likelihood estimates, as given in Shimodaira, 2002. Scripts for this procedure are given in Supplementary material. Again, topologies were compared to the best-scoring species tree topology obtained for nucleotide or amino-acid-encoded gene-trees at a 5% threshold, to obtain the probability of a topology to score better than the reference one after gene resampling.

Tree reconstruction with constrained gene trees

It is known that coalescent methods, which aim at taking into account potential incomplete lineage sorting or introgression between lineages, are sensitive to gene tree estimation error (GTEE) (Roch et Warnow, 2015 ; Nute et al., 2018). Therefore, we tried to reduce the error in gene tree estimation by constraining the topology in clades that are well-resolved in the species tree. Indeed, branches in the *Paralvinella* clade (exluding *P. p. irlandei* and *P. unidentata*), as well as in the ougroup species, are longer than the branches at the root of the Alvinellidae family (see Supplementary data Figure 3 and 4). This could indicate that most of the gene tree discordance observed in these groups is attributable to GTEE, rather than ILS or introgression. Fixing species arrangement in these groups, as in the 15 candidate topologies, should therefore be closer to the true tree compared to free-running arrangements and improve the confidence on the probabilities for the nodes at the root of the family.

For each gene, the 15 candidate topologies were evaluated under ML using the best evolutionary model according to ModelFinder. Topology k was then given a Bayesian posterior probability from its likelihood L_k using the approximation $P_k = exp(L_k)/\sum_{i=1}^{15} exp(L_i)$. The species tree scores were obtained with ASTRAL, with confidence intervals obtained by 10,000 bootstraps on each gene tree topology sampled in accordance with their Bayesian posterior probability, to take into account the uncertainty of gene tree topologies.

This procedure was conducted on both nucleotide-encoded genes (657) and amino-acid encoded genes (699). To summarize the information coming from the two types of sequences, a global species tree score was calculated. Nucleotide normalized quartet scores, N between 0 and 1, were combined with amino-acid normalized quartet scores, A between 0 and 1, by considering the Euclidean distance d between pairs of sequence datasets (N, A) and an hypothetical best topology scoring $(1, 1) : d = \sqrt{(1-N)^2 + (1-A)^2}$. This topology would correspond to an ideal tree that would maximize the quartet scores both for nucleotide and amino-acid encoded genes. Thus, the lower the distance estimated, the better the topology.

The distance obtained for each candidate topology was then tested against the best-scoring topology with our transposed AU test at a 5% p-value threshold. To this end, we used 10 successive multi-scale resamples of 10,000 gene trees, drawn according to their

Bayesian posterior probability.

Finally, gene tree discordance was assessed in an attempt to distinguish between GTEE, ILS or interspecific gene flow. In the case of GTEE without ILS, one would expect the probabilities of the 15 candidate topologies to be more randomly distributed for genes with a lower number of phylogenetically informative sites, compared to highly informative genes that should give a stronger signal toward one or few preferred topologies. On the contrary, ILS implies that gene tree discordance remains high, independently of the load of the phylogenetic signal. Therefore, genes were split into four categories according to their number of parsimony-informative sites for the nodes at the root of the family Alvinellidae in order to observe whether gene tree discordance at the root of the family was mostly due to GTEE or if ILS was also involved.

To test whether interspecific introgression also occured, topologies' probabilities were mapped on a ternary plot, allowing us to evaluate the number of genes strongly associated with one particular topology ($Pr > 0.8$). In that case, the 15 topologies were reduced to only three possibilities corresponding to different quartets at the root of the Alvinellidae family : ((*P. p. irlandei*,*P. unidentata*),(*Alvinella*,*Paralvinella*)), ((*P. p. irlandei*,*Alvinella*),(*P. unidentata*,*Paralvinella*)), ((*Alvinella*,*P. unidentata*),(*P. p. irlandei*,*Paralvinella*)). In the case of high interspecific gene flow between ancestoral lineages in the early steps of the Alvinellidae radiation, one would expect to see one quartet supported by a majority of genes (corresponding to the species tree), one quartet supported by a high number of genes (corresponding to intraspecific gene flow) and a quartet supported by a lower number of genes, due to potential ILS or GTEE (L. Cai et al., 2021). In no introgression occured, only the quartet corresponding to the species tree should be highly supported, while the alternative quartet should have the same prevalence among gene trees.

Molecular dating of the family's radiation

The Alvinellid phylogeny was dated with the software Phylobayes v.4.1. (Lartillot et Philippe, 2004 ; Lartillot, Brinkmann et Philippe, 2007). The uncorrelated Gamma model (Drummond et al., 2006), Cox-Ingersoll-Ross (CIR) model (Lepage et al., 2007) and the log normal model (Thorne, Kishino et Painter, 1998) were tested on the concatenated amino acid-encoded genes for the two most relevant species tree topologies. Considering that the oldest known Terebelliformia fossils date from the Carboniferous/Devonian time (Sepkoski Jr., 2002), the age of the tree's root (divergence between *P. gouldii* and other terebellid, ampharetid and alvinellid species) is given by a Gamma distribution *prior* with a mean of 330 million years and a standard deviation of 200 million years. This places 95% of the age distribution between 300 and 360 million years ago. We also used two calibration points corresponding to the divergence between the species *P. palmiformis* and *P. grasslei*, estimated to be 23 to 34 millions of years ago (Ma) as a consequence of the subduction of the Farallon plate under the North American plate (Tunnicliffe, 1988)), as well as the divergence between the two geographic forms of the species *P. fijiensis* sampled in the Manus and Lau basins, estimated between 2 and 4 Ma (Desbruyères, Hashimoto et Fabri, 2006 ; Boulart et al., 2022).

II.3 Results

II.3.1 Global phylogenetic inference

The global phylogeny was performed under a partition gene model on both the nucleotidic and amino-acid supergenes, or using the coalescent method on the individual gene datasets, based on either nucleotide or amino-acid encoded sequence alignments. Both approaches identify three main clades inside terebellomorph species, which are the Alvinellidae family, the Ampharetidae family comprising *Anobothrus* sp., *H. invalida*, *A. carldarei* and *A. gunneri*, and a third group composed of the Terebellidae species (Terebellidae gen. sp. and *N. edwardsi*) as well as the melinnid species (*M. palmata*), as shown in Figure II.1. *M. palmata* is a sister species of other terebellidae species for all methods but the phylogeny obtained from the nucleotide-encoded supergene, where it is the sister species to the three polychaete families (Alvinellidae, Terebellidae and Ampharetidae excluding the outgroup *P. gouldii*). *Anobothrus* sp. also represents a sister species to other Ampharetidae species in phylogenies using the amino-acid encoded sequences (82% bootstrap for the concatenation approach or 0.6 posterior probability in the coalescent method), but is sister to other Ampharetidae and Alvinellidae species using nucleotide-encoded sequences with a greater robustness (100% bootstrap or 1.0 posterior probability). All phylogenies with their node supports are included in the supplementary data, Figure 1 to 4. In further analyses, however, we chose to only consider phylogenies that group *M. palmata* in the Terebellidae clade and separate Alvinellidae and Ampharetidae as two monophyletic families (Fig. II.1), according to current views on the Terebelliformia phylogeny (Stiller, Tilic et al., 2020).

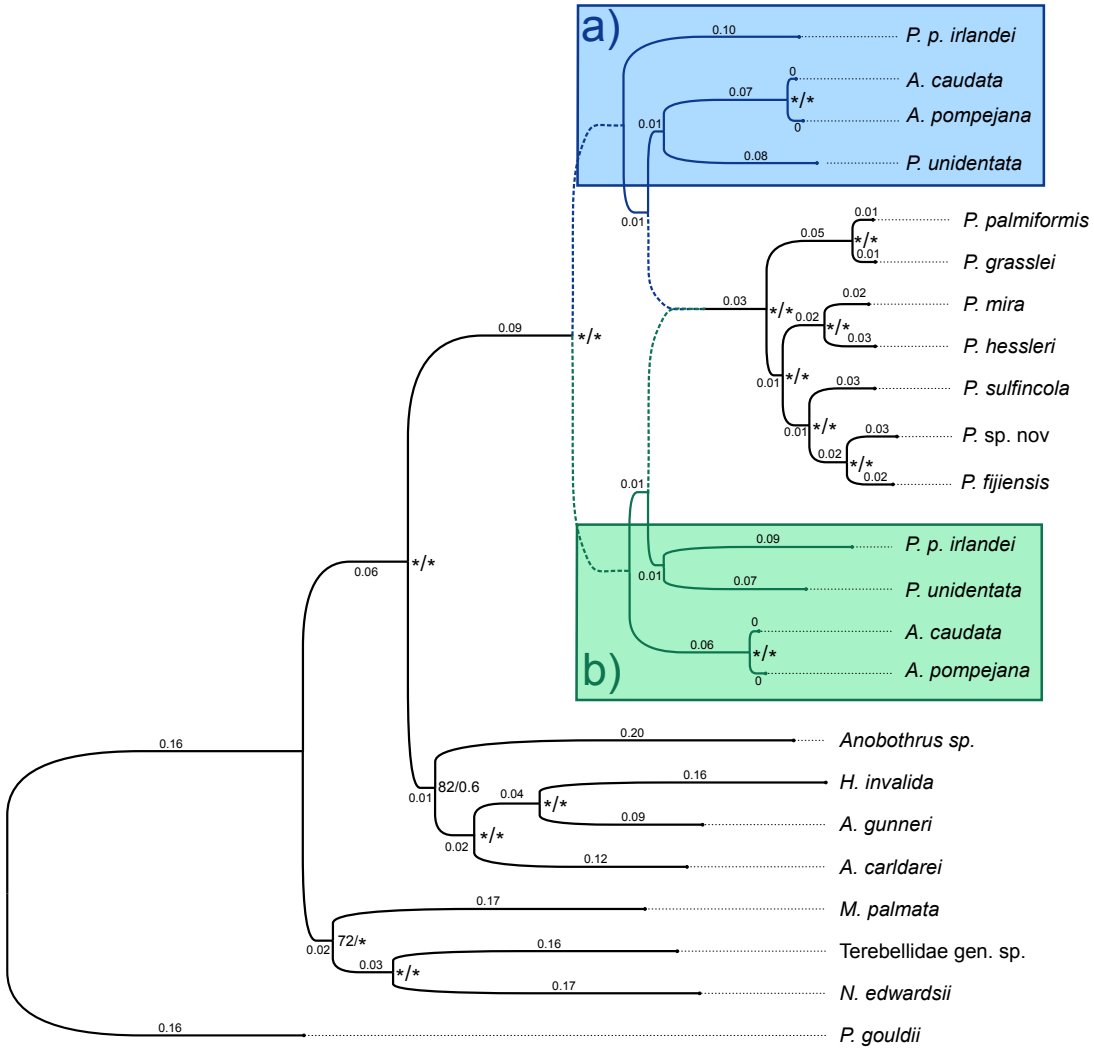


Figure II.1 – High-scoring alternative topologies according to different phylogenetic inference methods for amino-acid encoded genes. (a) Main topology, with *P. p. irlandei* basal to the family Alvinellidae and clustering of *P. unidentata* with the *Alvinella* species (b) Alternative topology with clustering of all *Paralvinella* species together. Branch lengths are optimized under a partitioned model from concatenated amino-acid encoded genes. Node supports are given in the following order : Ultrafast bootstrap (partitioned model)/ Posterior probability (coalescent model). (*) indicates a bootstrap of 100% or posterior probability of 1.

Within the Alvinellidae, all methods give a maximum bootstrap value to the grouping of the two *Alvinella* species. In the *Paralvinella* genus, with the exception of *P. unidentata* and *P. p. irlandei*, the phylogeny of species is also resolved without ambiguity, with a maximum of confidence on all nodes regardless of the phylogenetic method used. Conversely, the placement of the species *P. unidentata*, and more importantly *P. p. irlandei*, is not well supported. Depending on the number of genes considered, the evolutionary model used, or the types of sequences, the placement of these two species in relation to the well-defined groups of the other *Paralvinella* species, on one hand, and the two species of the genus *Alvinella,* on the other hand, is fluctuating at the root of the family Alvinellidae. Two high-scoring alternative trees are shown in Figure II.1a. and II.1b. For this reason, we chose to address more specifically the resolution of these deep nodes with a topology fixed for the outgroups and the other *Paralvinella* species, as they were well resolved.

II.3.2 Evaluating bifurcations at the root of Alvinellidae

In order to solve the alvinellid family topology, we exhaustively compared the phylogenetic results obtained for all possible root topologies. The outgroup subtree (Terebellidae+Ampharetidae rooted with *P. gouldii*) as well as the *Paralvinella* subtree (excluding *P. p. irlandei* and *P. unidentata*) are considered resolved and constrained. Figure II.2 lists all these topologies, numbered from 1 to 15.

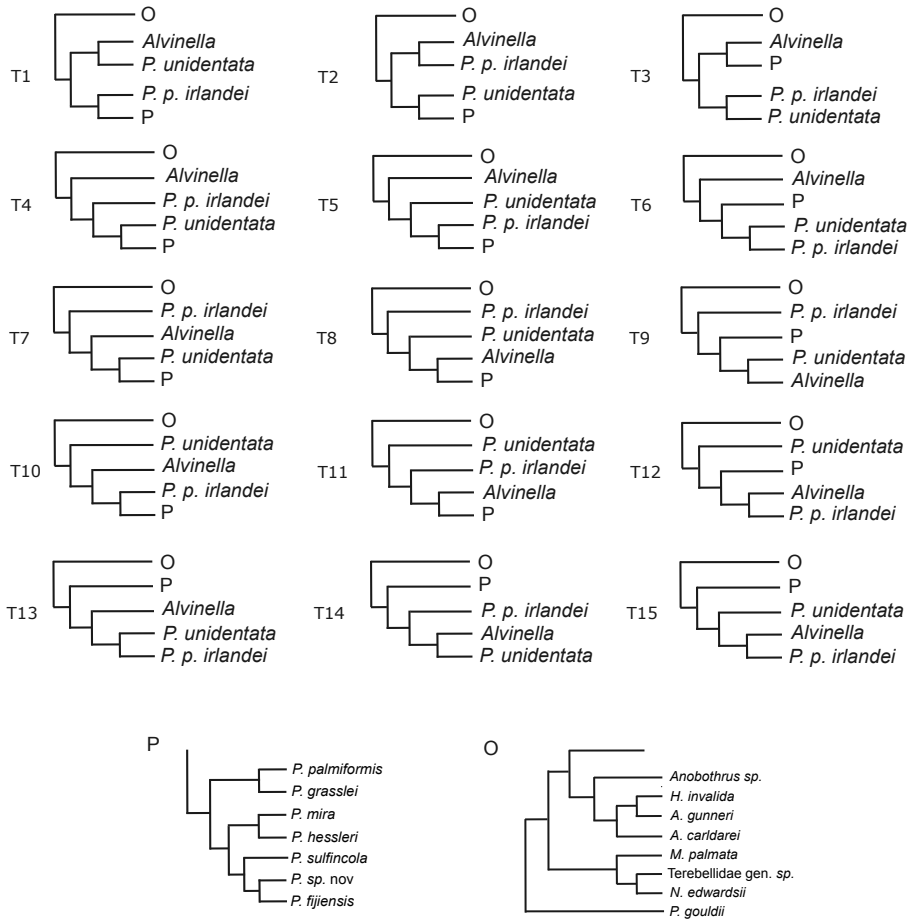


Figure II.2 – Constrained topologies. The topologies only differ in the arrangement of the species *P. unidentata*, *P. p. irlandei*, when compared with the genera *Paralvinella* and *Alvinella* at the root of Alvinellidae. The *Alvinella* group comprises the species *A. pompejana* and *A. caudata*, and the *Paralvinella* group (P) comprises all species *Paralvinella* except *P. p. irlandei* and *P. unidentata*. Trees are rooted with a set of outgroups (O) with the same species arrangement comprising species from other closely-related families of Terebelliformia (Terebellidae+Ampharetidae+*P. gouldii*).

The results of the inference methods on the two sequence datasets (CDS or translated proteins) are presented in Table II.1. From these analyses, four topologies stand out :

— Topology 7, 8 and 9, in which *P. p. irlandei* is sister to the all other Alvinellidae. Topology 9 brings the *Alvinella* and *P. unidentata* lineages closer together, *A. pompejana*, *A. caudata* and *P. unidentata* being sister species. Topology 7 brings the *Alvinella* lineage closer to *P. p. irlandei*, and topology 8 brings *Alvinella* species closer to other *Paralvinella*. According to the AU test, the best-ranking topology obtained in the different phylogenetic method (T9 or T8) was not significantly better than these topologies, except T7 which was significantly worse than T9 in the nucleotide+ML analysis (p-value=0.3%). Following all analyses, we finally retained T9 as the most appropriate topology over the three and this topology is detailed in Figure II.1a ;

— Topology 6, in which the *Alvinella* lineage is sister to *Paralvinella*, which forms a monophyletic genus within which *P. p. irlandei* and *P. unidentata* are sister species. According to the AU test, the best-ranking topology T9 (using the ML inference on nucleotide sequences), or the topologies T9 and T8 (using the coalescence approach with both nucleotide and amino acid-encoded genes) were not significantly better than T6. However, T6 was significantly worse than T9 in the amino acid+ML approach (p-value=2.1%). This T6 topology is detailed in Figure II.1b.

	Concatenation		Coalescence	
	499,036 sites nucleotide	277,900 sites amino acid	657 genes nucleotide	699 genes amino acid
T1	- 419	- 582	- 2,743	- 16,170
T2	- 600	- 619	- 13,197	- 14,886
T3	- 343	- 410	- 6,822	- 6,420
T4	- 312	- 428	- 2,915	- 6,627
T5	- 271	- 480	- 2,304	- 16,085
T6	- 119	- 208	- 665	- 9,202
T7	- 167	- 88	- 2,234	- 1,019
T8	- 77	- 93	- 6,330	0
T9	0	0	0	- 802
T10	- 482	- 632	- 10,956	- 19,554
T11	- 380	- 521	- 10,896	- 13,790
T12	- 548	- 645	- 22,097	- 27,472
T13	- 443	- 408	- 9,572	- 16,903
T14	- 421	- 492	- 8,951	- 17,059
T15	- 644	- 703	- 14,128	- 21,391

Table II.1 – Scores obtained for the 15 Alvinellidae topologies under different phylogenetic models. First and second columns : maximum likelihood on the nucleotide or amino-acid encoded supergenes, run as one gene partition under the GTR+Γ or LG+Γ model. Third and fourth columns : quartet score of species trees established from gene trees, either from nucleotide or amino-acid encoded genes. For each method, the topology with the highest score is set to 0, and the difference between the tree score and the best-scoring topology is shown. According to the AU test, species trees for which the top ranked topology is not better at a 5% threshold are in grey.

II.3.3 Phylogenetic discordance between genes

We then focused on the evaluation of the phylogenetic signal at the gene level. We estimated the likelihood of the 15 constrained topologies for each gene separately, either on their coding sequences or their translated protein counterparts. As shown in Figure II.3b. and II.3d., all topologies are supported by a high number of genes. Interestingly, genes containing more parsimony-informative sites, and thus having more power to distinguish between the contrained topologies (by increasing the phylogenetic signal at the root of the family), did not exclude some topologies, even though the topologies T6 and T9 are supported by a relatively higher proportion of genes as the number of informative sites increases. As a consequence, the discordance between gene histories is not solely attributable to gene tree estimation error and a loss of phylogenetic signal at some slow-evolving genes in which a higher GTEE is prone.

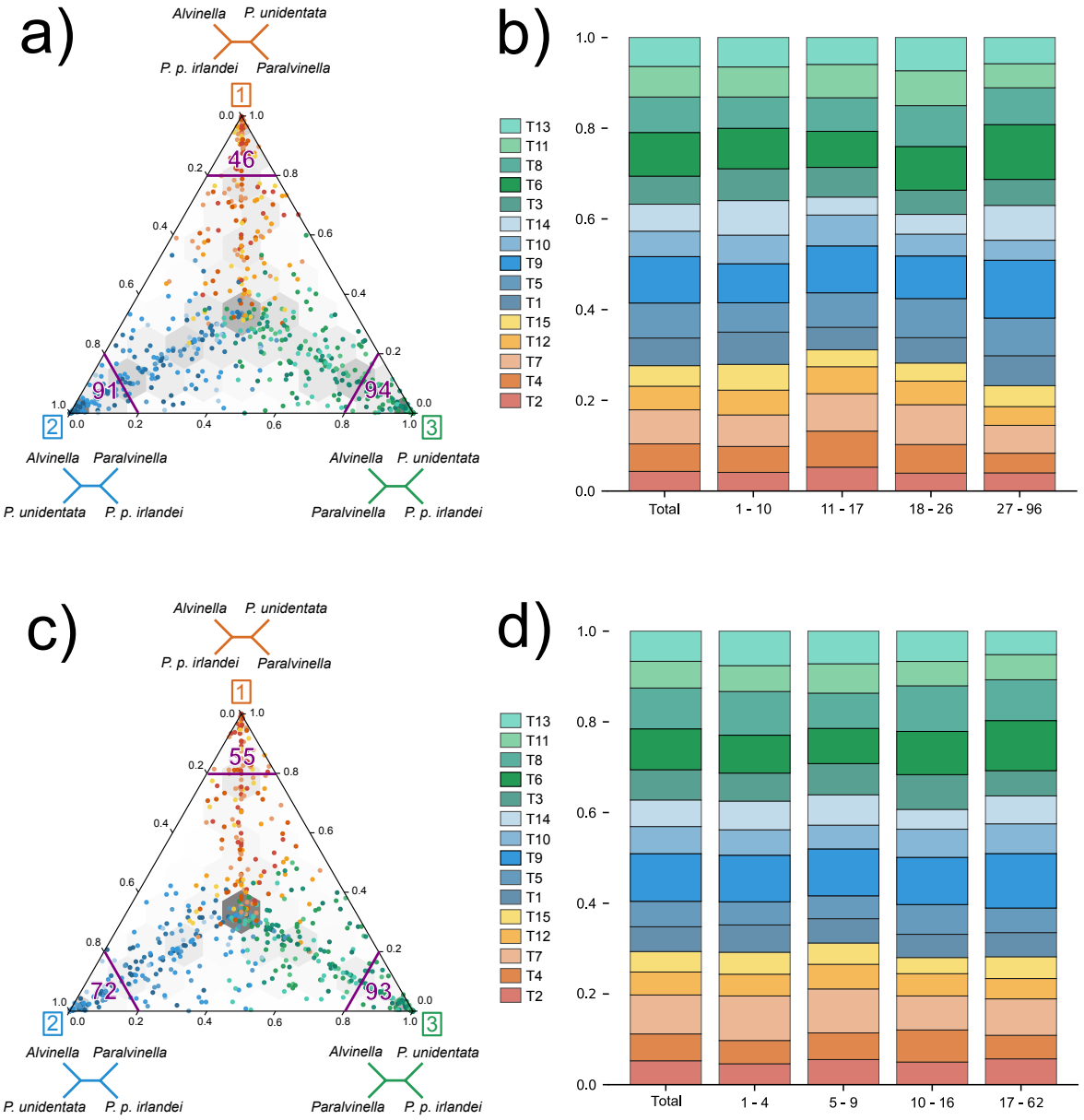


Figure II.3 – Topology weight across genes. Topologies are colored according to the maximum probability of their associated lineages' quartet. (a) Nucleotide-encoded gene associations with the three possible scenarios labelled 1, 2 and 3 at the root of the family. Number of genes strongly associated with one scenario ($Pr > 0.8$) are indicated in each pole. (b) Distribution of genes' probabilities according to the number of parsimony-informative sites for nucleotide sequences. (c) Amino acid-encoded gene associations with the three possible scenarios labelled 1, 2 and 3 at the root of the family. Number of genes strongly associated with one scenario ($Pr > 0.8$) are indicated in each pole. (b) Distribution of genes' probabilities according to the number of parsimony-informative sites for amino-acid sequences.

In Figure II.3a. and II.3c., we reassigned the 15 topologies to three simpler scenarios labelled 1, 2 and 3, to identify a potential bias of support in the shared gene histories between the *Alvinella*, *Paralvinella*, *P. unidentata* and *P. p. irlandei* lineages. Indeed, the three scenarios should account for the true species tree on one hand, and two alternative quartets on the other hand, resulting either from introgression, ILS or GTEE (L. Cai et al., 2021). Observing a bias in the sampling of the two alternative scenarios is the result of specific introgression between two lineages. We tested this bias by considering the null hypothesis that the two alternative quartets were equally sampled in gene trees, following a binomial distribution of parameters n, being the number of genes falling in the alternative quartets, and $p = 0.5$, the probability of choosing one quartet over the other (Pease et al., 2018). Considering nucleotide-encoded genes, 182 genes favor scenario 1 with a closer relationship between *Alvinella+P. p. irlandei* opposed to *P. unidentata +Paralvinella* (55 genes assigned with $Pr > 0.8$), 234 genes favor scenario 2 in which *Alvinella+P. unidentata* is opposed to *Paralvinella+P. p. irlandei* (91 with $Pr > 0.8$) and 241 genes are supporting scenario 3 where *Alvinella+Paralvinella* is opposed to *P. p. irlandei +P. unidentata* (94 with $Pr > 0.8$). For amino-acid encoded genes, these scenarios are respectively supported by 205 (55), 234 (72) and 260 (94) genes. For nucleotide-encoded genes, choosing scenario 2 or 3 as the true species tree results in scenario 1 being less supported than the alternative topology (scenario 1 against 2 : $X \sim B(416, 0.5)$, $Pr(X \leq 182) = 0.006$, scenario 1 against 3 : $X \sim B(423, 0.5)$, $Pr(X \leq 182) = 0.002$). On the contrary, scenario 2 and 3 are not distinguishable if scenario 1 in the true topology ($X \sim B(475, 0.5)$, $Pr(X \leq 234) = 0.39$). For amino-acid encoded genes, choosing scenario 2 or 3 as the true species tree results in scenario 1 being less supported than the alternative topology 3 ($X \sim B(465, 0.5)$, $Pr(X \leq 205) = 0.006$) or marginally less supported than scenario 2 ($X \sim B(439, 0.5)$, $Pr(X \leq 205) = 0.09$). Again scenario 2 and 3 are not distinguishable if scenario 1 in the true topology ($X \sim B(494, 0.5)$, $Pr(X \leq 234) = 0.13$). Consequently, one of the scenarios 2 or 3 is likely the true species tree while the second is the result of significant gene flow. In this case, gene trees supporting scenario 3, which brings closer the *Alvinella* species and *P. p. irlandei*, are the result of ILS and potential GTEE.

II.3.4 Species tree reconstruction with constrained gene trees and the coalescent approach

In our final approach, we attempted to reconstruct the species tree using the coalescent method from gene trees constrained to one of the 15 topologies. The goal was to reduce the overall gene tree estimation error (GTEE), assuming that incomplete lineage sorting (ILS) and inter-species gene flow were low at all nodes of the tree with the exception of the two nodes at the root of the family Alvinellidae. When evaluating the 15 topologies with this approach (see : Figure II.4a), T7 had on average the highest quartet scores both for nucleotide-encoded genes and their translated amino-acid counterparts, although this result was sensitive to the resampling of genes and gene trees uncertainty. Figure II.4b. displays the combined quartet scores obtained by each topology as the Euclidean distance to a hypothetical ideal topology at coordinates (1,1) that would maximize the quartet scores for both types of sequences. In this representation, T7 also obtains the best combined score, although not statistically better than T8 and T9 ($p - value = 0.23, 0.34$ respectively).

In all these topologies, *P. p. irlandei* is paraphyletic with the other *Paralvinella* species. The relationship of *P. unidentata* with other alvinellid clades is however different in these three topologies, being sister species to *Paralvinella* in T7, to *Alvinella* in T9 or to *Alvinella+Paralvinella* in T8. Compared to previous analyses for which T6 and T9 were preferred, these topologies are less supported here, even if the score of T7 is only significantly better than T6 ($p - value = 0.009$) but not T9. This may be explained by the fact that T7 represents an intermediate topology between T6 and T9, which are two most frequent co-occurring topologies in the gene trees sets (see Discussion, Monophyly of the genus *Paralvinella*).

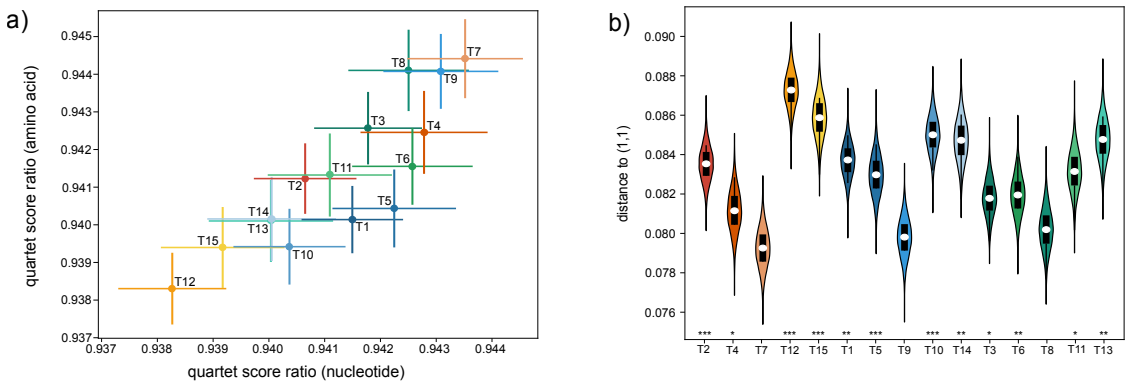


FIGURE II.4 – Fitting scores of the species tree topology by ASTRAL from the 15 constrained topologies, with standard deviations. Each evaluation is the result of 10,000 resamplings of gene trees. (a) biplots of quartet scores (in %) obtained for each topology from the amino acid and nucleotide sequence datasets. A score closer to 1 indicates that the species-tree topology is more in agreement with the set of gene trees. (b) Euclidean distance d estimated for each topology to the point (1,1), which represents an ideal case where all gene trees agree with the proposed species tree topology. For each topology TX, the hypothesis $d_7 < d_X$ is tested with a transposed AU test based on gene re-sampling. If the test fails, then T7 is not better than TX : $^{*} : p - value < 5\%$, $^{**} : p - value < 1\%$, $^{***} : p - value < 0.1\%$.

II.3.5 Molecular dating of the alvinellid radiation

We estimated the age of the alvinellid worm's radiation under the three main topology hypotheses. The three models used for molecular dating (Log-normal, CIR, and uncorrected Gamma multipliers) gave different results about the age of the radiation, ranging from 45 to 100 millions of years ago (Ma). However, the log-normal model appears to significantly underestimate ages, since the root of the tree is estimated to be between 79 and 89 millions of years old (My) depending on the tree topology, which is far from our expectations based on the age of the terebellomorph fossils from the Devonian-Carboniferous period (Sepkoski Jr., 2002). On the other hand, the uncorrelated Gamma multipliers model, which assumes that the evolutionary rates of branches are uncorrelated with time, places the age of the tree's root between 291 and 335 My (confidence intervals : 132-599 Ma), which is very close to the expected age. The radiation of the Alvinellidae is estimated between 92 and 99 Ma (47-180 Ma). Yet the 95% confidence intervals are still very large with this model, and the estimated ages are far from the ages with the model run under the *prior*, which dates the MCRA around 70-80 Ma (between 36 and 214 Ma). Finally, the CIR model proposed the most consistent estimates (see : Figure II.5). The age of the terebelliform MCRA is 120-160 My (98-198 My). The age of the Alvinellidae radiation stands between 67 and 91 Ma (55-112 Ma). These ages also better agree with the model run under the *prior* (age of the root ranged between 34 and 210 My, with a mean between 75 and 81 Ma depending on the topology). In these two last models, the radiation of the Alvinellidae was a rapid event occurring in the Late Cretaceous. The estimates obtained with the CIR model for alternative topologies are given in the supplementary data, Figure 5.

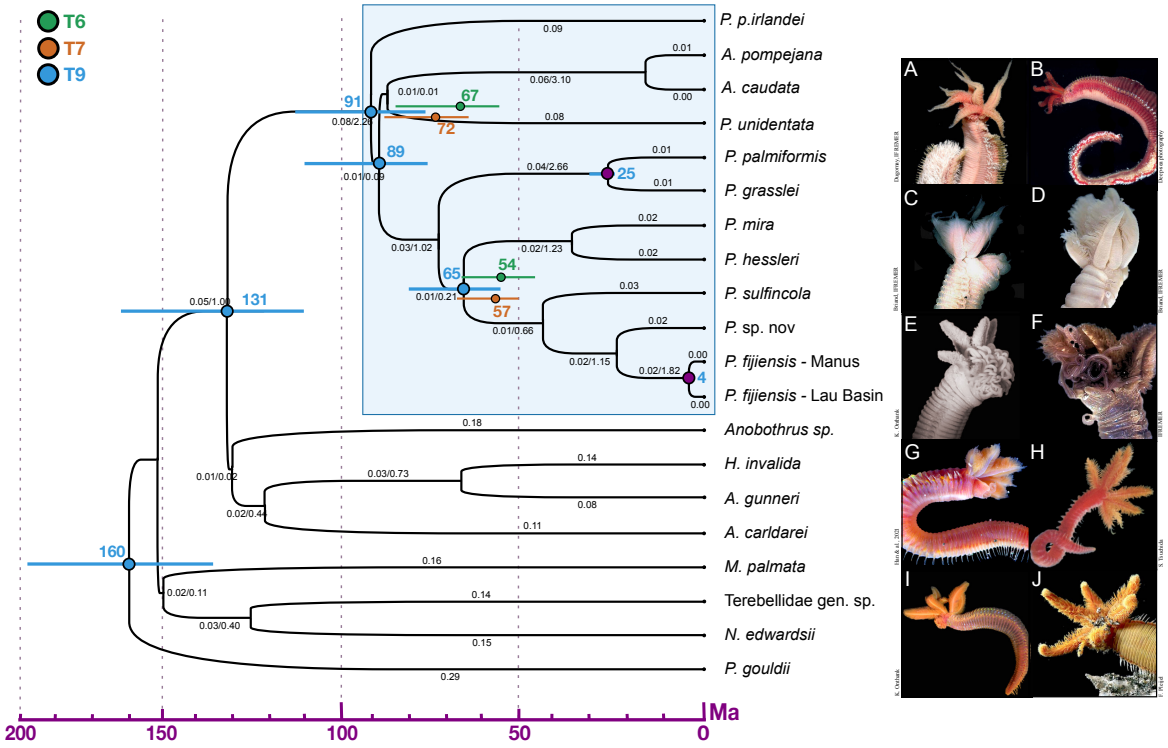


FIGURE II.5 – Chronogram of the family Alvinellidae under the CIR model with two calibration points (in purple), assuming T9 topology. Branch lengths are reported in expected numbers of amino-acid substitutions (LG+Γ) and coalescent time unit (obtained from unconstrained gene trees). The ages of some nodes of interest with 95% confidence intervals are reported in millions of years ago (Ma). The ages of birth of the Alvinellidae family and its spread in the eastern Pacific are also reported assuming the T6 and T7 topologies with 95% confidence interval. Illustrations : (A) *A. pompejana*, (B) *A. caudata*, (C) *P. unidentata*, (D) *P. p. irlandei*, (E) *P. palmiformis*, (F) *P. grasslei*, (G) *P. mira*, (H) *P. hessleri*, (I) *P. sulfincola*, (J) *P. fijiensis*.

II.4 Discussion

II.4.1 Monophyly of the genus *Paralvinella*

The initial description of the family by Desbruyères and Laubier subdivided the family into two genera, mainly based on the shape of the gills (four pairs of lamellar gills in *Alvinella*, *versus* simple cylindrical gills with no filaments at the tip in *Paralvinella*), the position and shape of the ventral uncini (on segment 9 in *Paralvinella* or 9 and 10 in *Alvinella*) and the position of the first setiger (segment 3 in *Paralvinella* and 6 in *Alvinella*) (Desbruyères et Laubier, 1986 ; Jouin et Françoise Gaill, 1990 ; Jollivet et Stéphane Hourdez, 2020). Based on these morphological criteria, these authors therefore proposed a taxonomic arrangement in which *P. unidentata* was closer to the sibling pair of species *P. p. irlandei* and *P. p. pandorae* within the *Paralvinella* genus, and subsequently grouped into the Nautalvinella subgenus (Desbruyères et Laubier, 1993). This grouping proposal is equivalent to the T6 topology, which is one of the best topologies identified by the molecular phylogeny on supergenes of this study, and which is also well supported by individual gene trees when constrained among the 15 most relevant tree topologies. The dichotomy between the genera *Alvinella* and *Paralvinella* was also suggested by Jollivet and colleagues who proposed the monophyly of *Paralvinella* on the basis of 14 allozymes or a set of 278 concatenated orthologous sequences (Jollivet, Desbruyères et al., 1995 ; Jollivet et Stéphane Hourdez, 2020). Interestingly, these two articles, like this study, include sequences of the species *P. unidentata*, which are mandatory to resolve the initial quartet at the root of the family clade.

However, although the alvinellid family appears to be monophyletic, the split of the first alvinellid ancestors is poorly resolved at the root of this clade. Other topologies, notably T7 and T9 are also well supported by numerous genes suggesting that the genus *Paralvinella* could be paraphyletic. Topologies T7, T8 and T9, within which *P. p. irlandei* is a sister species to the other Alvinellidae, got very good scores and agree with previous studies proposing phylogenies based on either mitochondrial Cox1 sequences or sets of orthologous sequences derived from transcriptomes and combined with some morphological characters (Vrijenhoek, 2013 ; Stiller, Tilic et al., 2020 ; Jollivet et Stéphane Hourdez, 2020). In our analysis, the T9 topology - but also the T7 topology to a lesser extent - is well supported by the gene-to-gene approach (using a coalescent method similar to the one employed by Stiller, Tilic et al., 2020) or partitioned phylogeny. In the present study, we ensured that each orthologous gene cluster contained a sequence for *P. p. irlandei*, but additional information from its sister species *P. p. pandorae* from the Juan de Fuca Ridge would have certainly helped to improve the confidence at the first nodes of the family. These three topologies differ in the positioning of *P. unidentata* in relation to the rest of the Alvinellidae. In T7, *P. unidentata* is closer to other *Paralvinella* species, while it is closer to *P. p. irlandei* in T8, and closer to *Alvinella* in T9. It is worth noting that the grouping of *P. unidentata* and *Alvinella* (T9) is in agreement with some traits concerning the gills' morphology : in particular, the filaments of *P. unidentata* are flattened, which is unique compared to other *Paralvinella* species, and similar to the lamellar shape of the gills of the two *Alvinella* species. This topology thus raises the question of the monophyly of *Alvinella* in the traditional view that separates this genus from all other species of Alvinellidae. However, the overall

structure of the gills of *P. unidentata*, with a tip devoid of filaments and funnel-like inserted secondary filaments, brings the species closer to other *Paralvinella* and especially to *P. p. irlandei* (Desbruyères et Laubier, 1993). The study of the gills in Alvinellidae is a good example of the importance of morphological variations that can be observed within the Terebelliformia at the level of palps, chaetae and gills, as well as the shape and position of the uncini (Day, 1964 ; Stiller, Tilic et al., 2020), which can be highly diverged even in closely-related species because of difference in their adaptive trajectories. To this extent, the shape of gills and their size are likely to be sensitive to natural selection as they are crucial to enduring long periods of hypoxic conditions, while being able to rapidly capture massive amounts of oxygen in well-oxygenated water when the worm is no longer exposed to the fluid mixing (Jouin et Françoise Gaill, 1990). Depending on the species' ecology, these periods of oxygenation may greatly vary from very brief periods of times (*Alvinella*) to longer ones (*P. p. irlandei*).

From our results, the paraphyly of *Paralvinella* seems to be the preferred solution. It is noteworthy that this conclusion is more supported by amino-acid encoded genes when compared to nucleotide encoded genes. This could be partially due to long-branch attraction artifacts, given that the average amino acid composition of *P. p. irlandei* sequences deeply contrasts with the average composition of other alvinellid species (see Methodological control - Data quality). This bias in the amino-acid composition of *P. p. irlandei* is tricky but could be explained by the ecology of the worm, which lives under colder conditions and seems to be less adapted to hypoxia (i.e. comb-like gills). Nevertheless, we chose to retain T9, as shown in Figure II.5, as the most reliable species tree topology for the Alvinellidae. This contrasts with the initial view of a family split into two genera with *Alvinella* species sister to *Paralvinella* species. However, most gene trees are inconsistent with this topology, even for the most phylogenetically-informative genes. The radiation of the Alvinellidae is undoubtedly a difficult case to resolve where multiple short branches (divergence $< 1\%$) are buried deep in the tree topology. The radiation leading to the separation of the *Alvinella*, *P. unidentata*, *P. p. irlandei*, and other *Paralvinella* lineages is likely a fast event in the timescale of the whole family, and possibly linked to some adaptive strategies to cope with different thermal habitats. This led to high incomplete lineage sorting and cross-species gene flow, possibly adaptive, which appears with branches shorter than 0.1 coalescent time units and a mixing of genes pointing to either the T6 or the T9 topologies. This also agrees well with the observation that the evolutionary rate may be slow in alvinellid worms, consequently maintaining ancestral polymorphisms for a very long time (Fontanillas et al., 2017). For example, in *P. unidentata* and *P. fijiensis*, present-day populations that have been separated for 3 to 4 million years also display short coalescent times, between 0.05 and 0.35 (see supplementary data Fig. 1), and Jang et al. estimated the split of the the northern East Pacific Rise and the northeastern Pacific Antarctic Ridge populations to about 4.2 million years ago for *A. pompejana* (Jang et al., 2016), despite the maintenance of shared polymorphisms between the two metapopulations.

Finally, the distribution of gene tree topologies is not random, but biased toward two alternative scenarios where *P. unidentata* is either closer to the *Alvinella* species, or closer to *P. p. irlandei*. Such a distribution is expected if one of these topologies is reflecting the true species tree, while the other is the consequence of specific gene flow between the two incriminated lineages. Yet these two scenarios are not directly compatible with one ano-

ther if we consider than T9 is the true species tree. On the contrary, T7 is an intermediate topology where high asymmetric gene flow from the ancestor to the *P. unidentata* lineage to the ancestor of *P. p. irlandei* would result in T6, while cross-species gene flow between ancestors of *P. unidentata* and the *Alvinella* lineage would result in T9. Thus, even though it is less well supported by the phylogenetic inferences from concatenated supergenes or unconstrained gene trees, T7 is an interesting topology that could reconcile the main phylogenetic relationships that appear in the multi-gene analysis. Lastly, while T8 is closer to the T9 topology, it is hardly compatible with T6 as the transition from T8 to T6 requires a higher number of tree rearrangements. This make T8 less likely to represent the true species tree.

Consequently, in the case of the Alvinellidae, the rapid radiation of the first ancestral lineages led to some probable introgressive gene flow between the ancestor of *P. unidentata* and both the ancestors of *P. p. irlandei* and the two *Alvinella* species which may have been coupled with the maintenance of ancestral polymorphisms due to a very high level of ILS. Following this view, the topologies T6, T7, and T9 appear to be equally likely to explain the evolutionary history of this peculiar and highly adapted family of worms. This casts doubt on the relevance of a binary tree of species at the root of the family, as different genomic regions more likely have different phylogenetic relationships.

II.4.2 Division of the genus *Paralvinella*

Desbruyères and Laubier proposed to divide the genus *Paralvinella* into three subgenera on the basis of morphological characters (Desbruyères et Laubier, 1993) :

— *P. Paralvinella*, which groups the species *P. palmiformis*, *P. grasslei*, *P. fijiensis* and *P. sulfincola* ;

— *P. Miralvinella*, which groups the species *P. hessleri*, *P. dela* and *P. bactericola* ;

— *P. Nautalvinella*, which includes the species *P. unidentata*, *P. p. irlandei* and *P. p. pandorae*.

Several criteria were used to differentiate these subgenera, notably the presence of specialized sexual tentacles (used in the male-female pairing during reproduction) on the mouth apparatus of males in *P. Paralvinella* (three-lobed) and *P. Miralvinella* (pointed and coiled) which are absent in *P. Nautalvinella*, the shape of the gills (filaments on the opposite sides in *P. Paralvinella* and *P. Miralvinella*, and comb-like in *P. Nautalvinella*), as well as the presence of digitiform notopodial lobes on the anterior part of *P. Paralvinella* and *P. Miralvinella* which are absent in *P. Nautalvinella*. Other morphological criteria were used to differentiate these subgenera despite being more variable, such as the number of segments of the animals (55 to 180 setigers in *P. Paralvinella versus* 50 to 75 in *P. Miralvinella* and *P. Nautalvinella*), or the position of the first uncini (setigers 12 to 26 in *P. Paralvinella*, 15 to 26 in *P. Miralvinella* and 5 to 32 in *P. Nautalvinella*) (Desbruyères et Laubier, 1993 ; Jollivet et Stéphane Hourdez, 2020 ; Han et al., 2021).

Han et al. also argued that the new species *P. mira* is related to the subgenus *Miralvinella* by considering the pointed shape of the sexual tentacles, the shape of gills and the presence of digitiform lobes on the notopodia. The authors still noted peculiarities, such as the first

three setigerous segments, which are not fused in *P. mira*, and the insertion of numerous slender oral tentacles on the buccal apparatus (Han et al., 2021). In our analyses, the phylogeny of the species which belong to the subgenera *Miralvinella* and *Paralvinella* is well resolved with a high level of confidence. The species *P. mira*, as suggested by Han et al., 2021, effectively groups with *P. hessleri*, which is the only other *Miralvinella* species included in this phylogeny. According to Han et al., *P. mira* is morphologically closer to *P. hessleri* than to *P. dela* or *P. bactericola* when considering the shape of the buccal apparatus (stronger tentacles in *P. mira* and *P. hessleri*) and the position of the first uncinigerous neuropodial tori on chaetiger 16 or 18 in *P. mira* and *P. hessleri vs.* 32 for *P. dela* and *P. bactericola* (Han et al., 2021 ; Jollivet et Stéphane Hourdez, 2020). The geographical distribution of the species also agrees with this view, as *P. mira* and *P. hessleri* are found in the Indian Ocean and south-west Pacific Ocean, while *P. dela* and *P. bactericola* are sister species inhabiting the Juan de Fuca Ridge and the Guaymas basin in the eastern Pacific Ocean (Desbruyères et Laubier, 1993 ; Jollivet et Stéphane Hourdez, 2020). Unfortunately, molecular data are not available for the rarer species *P. dela* and *P. bactericola* to confirm the monophyly of the *Miralvinella* subgenus. The subgenus *Paralvinella* is however no longer monophyletic since the *Miralvinella* species are included in this group. The three-lobed tentacles of the mouth apparatus would then represent a symplesiomorphy of the genus *Paralvinella*, derived later as pointed tentacles in *Miralvinella*.

Finally, the subgenus *Nautalvinella* as described by Desbruyères & Laubier is also no longer monophyletic. If T9 or T7 are kept as the most probable species trees, the species *P. p. irlandei* and *P. p. pandorae* are sister to the *Alvinella* and *Paralvinella* clades, and can be grouped under a reduced *Nautalvinella* genus as proposed by Stiller, Tilic et al., 2020. It is noteworthy that in T9, *P. unidentata* is also a sister species to the two *Alvinella* species. This casts doubts on the definition of the *Alvinella* genus, which groups together two species with very little divergence ($< 1\%$) but not *P. unidentata* from which they greatly diverge. Yet single gene phylogenies are equally distributed between topologies that group *Alvinella* and *P. unidentata* (mainly T9) and those which group *P. p. irlandei* and *P. unidentata* together (mainly T6, see Fig. II.3). Thus, the acquisition of comb-like gills, the loss of the pair of sexual tentacles on the buccal apparatus of males and the loss of the digitiform notopodial lobes, which are the main elements in Desbruyères & Laubier's diagnosis of the subgenus *Nautalvinella* (Desbruyères et Laubier, 1993), may have arisen as a result of gene transfer between the *P. unidentata* and *P. p. irlandei* lineages.

II.4.3 The spread of alvinellid worms in and outside of the Pacific Ocean

Our molecular dating estimates rely on two calibration points. The first point corresponds to the recent opening of the Manus (3-4 Ma) and Lau (1-2 Ma) basins, which are assumed to form two disjoint ridge systems (Boulart et al., 2022). The second refers to the subduction of the Farallon Plate beneath the North American Plate, which caused a vicariant event between hydrothermal vent communities of the Eastern Pacific between 34 and 23 My (Tunnicliffe, 1988). This vicariant event has already been used to date other phylogenies of vent species distributed between the EPR and JdF, if we consider that these

species disperse mostly along the axial valley of oceanic ridges (Stiller, Rousset et al., 2013 ; Jollivet et Stéphane Hourdez, 2020). This last calibration point is particularly interesting in the case of the Alvinellidae, since it separates several pairs of alvinellid sibling species distributed throughout the phylogeny : *P. palmiformis* and *P. grasslei*, *P. p. irlandei* and *P. p. pandorae*, as well as *P. dela* and *P. bactericola* (Tunnicliffe, 1988 ; Desbruyères et Laubier, 1993 ; Jollivet et Stéphane Hourdez, 2020). Our estimates would certainly become more accurate through a more comprehensive sampling of the different species of the family, as there are no molecular data for the rarer species *P. dela* and *P. bactericola*. Similarly, only the *Cox1* gene of *P. p. pandorae* is available (Jollivet et Stéphane Hourdez, 2020). Based on this sequence, the level of divergence estimated between *P. p. irlandei* and *P. p. pandorae* was nearly identical to the divergence estimated by Chevaldonné et al. for the species *P. palmiformis* and *P. grasslei*, and supports the *Cox1* divergence rate of 0.13%/My established by these authors for the Alvinellidae (Chevaldonné, Jollivet et al., 2002). Following this rate, the estimates for dating the age of alvinellid radiation using our genomic datasets were not very different from the mitochondrial estimate with the same set of species (Jollivet et Stéphane Hourdez, 2020).

Most of the species in the family, namely *A. pompejana*, *A. caudata*, *P. p. irlandei*, *P. grasslei* and *P. palmiformis*, are endemic to the East Pacific Rise or the Juan de Fuca Ridge (Desbruyères et Laubier, 1986). The most parsimonious explanation for this distribution regarding the phylogeny is an origin of the family in the Eastern Pacific Ocean. This is consistent with the distribution of the sister family Ampharetidae, for which species diversity was the greatest in the eastern Pacific, which may suggest a common origin for both families in the now-extinct or subducted Pacific-Farallon and Mathematician ridges. These ridges have been active since the Early Jurassic (200 Ma) and have played a central role in the initial dispersal of the hydrothermal vent fauna in the Pacific (Mammerickx, Herron et Dorman, 1980 ; Bachraty, Legendre et Desbruyères, 2009). The split between the Alvinellidae and Ampharetidae occurred between 100 and 130 Ma according to the CIR model, at a time when high spreading rates and off-ridge volcanism formed sub-aerial volcanic environments, which are now submerged in the Pacific (Röhl et Ogg, 1996). These emerged masses may have served as refuges for vent faunas during episodes of deep-sea anoxia from the Aptian to the Cenomanian ages (125 to 94 Ma) (Jacobs et Lindberg, 1998). In contrast to the Ampharetidae, for which colonization of the deep hydrothermal environment seem to have happened several times independently, all Alvinellidae are endemic to the hydrothermal environment and it is likely that the ancestor of the family was already associated with vent sites.

Depending on the model used for the molecular dating, the subsequent radiation of the family Alvinellidae occurred between 92-99 Ma (uncorrelated gamma model, 47-180 Ma) or between 67-91 millions years ago (CIR model, 95% confidence interval 55-112 Ma) . These estimates suggest that alvinellid worms have a long history of speciation, dating back to the Cretaceous period (Haymon, Koski et Sinclair, 1984 ; Vrijenhoek, 2013). Using 278 orthologous genes or the *Cox1* mitochondrial gene, Jollivet and Hourdez proposed two dates for the radiation of the Alvinellidae, one at 198 Ma and the other at 72 Ma (Jollivet et Stéphane Hourdez, 2020). The first dating, obtained from only one calibration point, is likely too old as molecular phylogeny analyses show that most of the invertebrate taxa underwent a radiation in the Cenozoic (about 66 Ma), and the fossil record exhibits important transitions

of the chemosynthetic faunas during the late Mesozoic (145-66 Ma) (Vrijenhoek, 2013). Moreover, the molecular dating did not take into account the upper age limit associated with the first occurrence of Terebelliformia fossils. Our estimates, under the hypotheses of the three tested topologies (T6, T7 and T9), are in line with a radiation dating back to 70-80 Ma. This is in agreement with the second dating of Jollivet and Hourdez at 72 Ma, established from *Cox1* with one calibration point for two pairs of sibling species. The radiation is therefore likely to follow the Cenomania/Turonian extinction phase, about 94 Ma. This major anoxic/dysoxic event has been argued to have created opportunities for modern taxa to invade deep environments (Thomas, 2007 ; Vrijenhoek, 2013).

Finally, considering the alternation in the phylogeny between species associated with the eastern Pacific (*A. pompejana*, *A. caudata*, *P. p. irlandei*, *P. grasslei*, *P. palmiformis*, *P. sulfincola*) or with the western Pacific (*P. unidentata*, *P. hessleri*, *P.* sp. nov., *P. fijiensis*), the colonization of the Pacific Ocean must have occurred several times (Desbruyères et Laubier, 1993). The divergence of the *Miralvinella* species would range between 35 and 65 My (95% confidence interval : 27-80 according to the CIR model), which corresponds to the spread and subduction of the Kula ridge (between 60 and 43 My, Smith, 2003). This ridge has been indeed proposed to represent a bridge between the western and eastern Pacific hydrothermal basins, which are now isolated from each other (Hessler et P. F. Lonsdale, 1991 ; Bachraty, Legendre et Desbruyères, 2009). The colonization of the Southwest Pacific region would be a second event between 45 and 22 My (55-18), potentially concurrent with the subduction of the Pacific plate under the Ontong-Java plateau between 45 and 30 Ma and the subsequent opening of modern-day basins (Schellart, Lister et Toy, 2006). The recent description of *P. mira* in the Indian Ocean, at Wocan and Daxi vents, shows that the alvinellid worms have also spread outside the Pacific Ocean. The divergence between *P. mira* and *P. hessleri* is dated back to 35-37 My (27-48) and could be an argument for the recent colonization of the Indian Ocean. The connection between the western Pacific and the Indian Ocean through Northern Australia was closed between 60 and 40 Ma, with the advance of the Philippines and the opening of the China Sea (Parker et Gealey, 1985 ; Moalic et al., 2012). Although the tectonic history of this region makes it difficult to reconstruct ridge connectivity between oceanic basins, the colonization of the Indian Ocean by Alvinellidae was made possible through the Antarctic Ridge, whose spread started about 50 Ma and connects the Indian Ocean with the Pacific Ocean (Parker et Gealey, 1985). This ridge is indeed a link to most of the vent sites, from the EPR to the mid-Atlantic Ridge through the Indian Ocean (Moalic et al., 2012). Although Alvinellid worms are rare in the Indian Ocean, an undescribed species of *Paralvinella* was found at the Solitaire vent field on the Central Indian Ridge (Nakamura et al., 2012). The sampling of this species may be of particular interest to investigate the possibility that the Central Indian Ridge acted as a stepping stone to the Wocan and Daxi sites in the dispersal of Alvinellidae.

II.4.4 Methodological control of the phylogenetic reconstructions

Constrained topologies

To identify the most likely tree topology for the Alvinellidae, we chose to exhaustively evaluate the most relevant topologies *via* different phylogenetic inference methods. These analyses provided robust inferences for some specific groupings. Species associated with the subgenera *Paralvinella* and *Miralvinella* (without both *P. unidentata* and *P. p. irlandei*) were thus constrained, taking into consideration that this group was correctly resolved with reliable and consistent positioning of all species regardless of the method used (concatenation or coalescent-based approaches, on either coding sequences or translated proteins). The outgroup species (Ampharetidae, *M. palmata*, *N. edwardsi*, Terebellidae gen. sp. and *P. gouldii*) were also constrained as an outgroup subtree. The grouping of *M. palmata* with Terebellidae gen. sp. and *N. edwardsi* was consistent with the phylogeny proposed by Stiller et al. and clusters the Melinnidae and the Terebellidae together with strong bootstrap support (Stiller, Tilic et al., 2020). The species positioning in the family Ampharetidae is more questionable. *Anobothrus* sp. was indeed sister to the family Ampharetidae from analyses based on amino-acid encoded sequences (bootstrap = 82% and posterior probability = 0.6), but the nucleotide sequence data suggested that *Anobothrus* sp. could be sister to both Alvinellidae and Ampharetidae. Here we chose to consider only the topology in which Ampharetidae species form a monophyletic group to reduce the variance of our tree likelihood estimates with another flexible species position, but this constraint is likely a simplification of the true gene topologies.

The question of whether the Alvinellidae form a sister family to the Ampharetidae or whether the family should be included within the Ampharetidae cannot be fully resolved without a more exhaustive sampling of species within the latter family. Both families have retractable buccal filaments attached to the dorsal curtain, which is an important trait in establishing the monophyly of Ampharetidae. While several morphological and molecular studies have concluded a clear separation of the two families (Rousset, Rouse et al., 2003 ; Stiller, Tilic et al., 2020), Eilertsen et al. suggested on a broader ampharetid phylogeny based on classical phylogenetic markers (mtCOI, mitochondrial 16S rDNA, nuclear 18S and 28S rDNA) that Alvinellidae should be included within Ampharetidae (Eilertsen et al., 2017). The relationship between the two families is presented as a trifurcation between the Alvinellidae, *Amphysamytha* and [*Anobothrus*+*Amphicteis*] lineages. In contrast to this last result, our study concludes that the *Amphisamytha* species (*A. carldarei*) is sister to the *Amphicteis* species (*A. gunneri*). The unresolved trifurcation proposed by Eilertsen et al. is resolved in our analysis, with Alvinellidae being a monophyletic family, sister to the clade and *(Anobothrus*,(*Amphicteis*,*Amphysamytha*). As the question of polyphyly of the Ampharetidae group was not the purpose of the present study, we therefore decided to follow the more consensual statement considering Ampharetidae and Alvinellidae as two separate monophyletic groups, in accordance with phylogenies including a broader range of Terebelliformia species. Nevertheless, the monophyly of Ampharetidae with respect to Alvinellidae remains under question.

Data quality

Phylogenetic reconstruction is sensitive to differences in species sequence composition. Indeed, compositional biases affecting divergent species in the same way may be explained by evolutionary convergence events that can be confused with synapomorphies. Conversely, the absence of a compositional bias in some species compared to others can have the same effect (the symplesiomorphy trap), which is a special case of long branch attraction. This phenomenon has been observed in mitochondrial data of Ampharetidae and Alvinellidae, which exhibit strong amino-acid and base composition biases (Zhong et al., 2011). In addition, Alvinellidae colonize contrasting thermal environments, which is known to have an effect on the amino-acid composition of proteins between mesophilic and thermophilic eukaryotes (G.-Z. Wang et Lercher, 2010 ; Noort et al., 2013 ; Bock et al., 2014).

The evaluation of the amino-acid composition bias in the alvinellid orthologous gene clusters shows strong disparities between species (see supplementary data Fig. 6). In particular, *P. p. irlandei* shows very different values of amino-acid composition when compared to other species, with great differences in indices such as CvP bias, Serine bias and purine load : criteria which have been suggested to discriminate between warm and cold species (Hickey et Singer, 2004 ; Fontanillas et al., 2017). Therefore, we carefully filtered out genes that exhibited a high compositional heterogeneity in our phylogenetic analyses. We also ensured each gene alignment contained a sequence of *P. p. irlandei* in the analysis, although most of the *P. p. irlandei* sequences were on average shorter and more fragmented (158,783 total nucleotide sites and 292,610 amino acid sites). This limits the available phylogenetic information for this species compared to other alvinellid species of the dataset, and may have played a role in the placement of this species in the alvinellid tree. One might indeed expect that the strong compositional bias found in *P. p. irlandei* would tend to make this species diverge more rapidly among the Alvinellidae in the phylogenetic reconstruction. Moreover, the average composition of *P. p. irlandei* is not only different from the other species of Alvinellidae, but also differs from the species chosen as outgroup (Ampharetidae + Terebellidae + *P. gouldii*). The positioning of *P. p. irlandei* in the alvinellid tree could be therefore highly gene-dependent when estimating tree topologies (and to a lesser extent when fitting constrained gene trees). One must therefore be cautious with the good scores of the T7, T8 and T9 topologies, which place *P. p. irlandei* sister to other Alvinellidae species, as it may be a typical consequence of long-branch attraction.

The rapid radiation of the family probably occurred at the end of the Cretaceous period over a few million years, as shown by the short internal branches found at the root of the family. This is another difficulty for resolving tree bifurcations in this phylogeny. Rapid divergence can lead to a high rate of incomplete lineage sorting (ILS) and result in larger gene tree estimation errors (GTEE) (Molloy et Warnow, 2018). We must therefore be especially vigilant of the main biases that can lead to a difference between the gene trees and the species tree, namely ILS, inter-species gene flow, gene duplications and GTEE (L. Cai et al., 2021).

To limit the effects of gene duplications, we focused on orthologous gene clusters considered single-copy in lophotrochozoans based on ODB9 (see Material & Methods), and paralogy should therefore not be a concern in our analysis. In the case of very high ILS,

coalescent-based methods should give better results than concatenation methods for estimating the true species tree. Indeed, concatenation methods impose the same evolutionary history on all genes, and do not take into account the possibility of gene introgression or ILS. However, the superiority of coalescent methods depends on the confidence level put in gene tree estimation. For low or medium ILS or inter-species gene flow, the concatenation method, which benefits from longer alignments, allows the reduction of GTEE, and is still more reliably in practice (Roch et Warnow, 2015 ; Molloy et Warnow, 2018). Generally speaking, in the case of high ILS or large amounts of inter-species gene flows, coalescent methods are preferable provided that GTEE remains low, particularly with the longest possible gene alignments (Nute et al., 2018). A simulation by Molloy and Warnow on 26 species and 1000 genes - conditions relatively similar to our study - further shows that if the ILS is high, increasing the number of genes is preferable over filtering out gene trees on the basis of GTEE. Nevertheless large GTEE severely limits the ability of both concatenation or coalescent methods to find true species trees (Molloy et Warnow, 2018).

To distinguish between GTEE and true gene tree variability at the root of the Alvinellidae family, we considered a second step in which ILS and gene flow were negligible at all nodes of the tree. Such an assumption was globally well supported by all methods, except for the basal nodes of the Alvinellidae clade. We recomputed gene trees by reducing the topology space to trees T1 to T15, with the objective of limiting GTEE at these nodes by imposing a strong *prior* on these probable topologies. Surprisingly, we find that all 15 tested topologies are supported by a significant number of genes, regardless of the number of informative sites at the gene level (see Fig. II.3). If choosing between gene topologies was guided only by GTEE, one would expect to see certain topologies to prevail over others as the phylogenetic signal strengthens. This was not the case here, suggesting that other processes such as ILS and inter-species gene exchanges are causing the discordance between gene trees. Furthermore, unlike ILS which randomly segregates alleles between lineages at a given node, gene introgression occurs specifically between two lineages in some specific regions of their genome. Therefore, in the case of high gene flow from one lineage to another, one would expect to see some secondary topologies emerging to become more represented than others. In our case, *P. unidentata* was more frequently closely associated with *P. p. irlandei* and *Alvinella* than with other *Paralvinella* species, suggesting that secondary gene exchanges occurred after the first speciation events with at least one of these two sister lineages.

After assessing the effect of GTEE on gene topologies, we also evaluated the species tree topology from the 15 constrained gene trees. This approach aimed at reducing the GTEE, which is an essential assumption of the coalescent-based methods. In this case, topologies T7, T8, and T9 were the three preferred solutions with very similar scores. These three topologies also obtained good scores from the phylogenies on supergenes (Table II.1). In general, T9 was the most supported topology in all approaches. Yet, T7 draws our attention as it is the only species tree topology that can easily explain the high prevalence of both T6 and T9 at the gene level, with subsequent gene transfers between the *P. unidentata*, *P. p. irlandei* and *Alvinella* branches. The sequencing effort, which includes several hundred genes, is already very substantial and it is unlikely that increasing the number of sequences will improve this result significantly. In the future, mapping the genes carrying differences in the phylogenetic signal onto the genome of one of these three species (and

potentially enhancing the sequencing to non coding regions to improve genetic resolution) could enable us to see whether the genes associated with some specific topologies belong to specific regions of the genome. This would help to conclude whether our gene topology discrepancies were the result of ancestral gene transfers between these lineages and better quantify their relative impacts.

II.5 Conclusion

To conclude, our phylogenetic analysis of the alvinellid worms led to some uncertainty in choosing the right tree species topology between T6 (monophyly of the *Paralvinella* genus) and T9 (polyphyly of the *Paralvinella* genus with *P. p. irlandei* being sister to other alvinellid species). Different genes can have different phylogenetic histories, which is probably a consequence of the rapid radiation of several lineages from an ancestral worm population, resulting in high incomplete lineage sorting and intraspecific gene introgression. Moreover, these two topologies can be equally supported by some morphological traits depending on whether some of them represent synapomorphies or symplesiomorphies. Some traits such as the gill shape or the amino-acid residue biases in proteins are likely to be the result of natural selection in the face of the contrasting thermal regime encountered by the worms.

CRediT authorship contribution statement

Pierre-Guillaume Brun - Data Curation, Methodology, Formal Analysis, Writing original draft and review. Stéphane Hourdez - Funding acquisition, Resources, Reviewing original draft. Marion Ballenghien - Methodology. Yadong Zhou - Resources, Reviewing original draft. Jean Mary - Supervision, Funding acquisition, Reviewing original draft. Didier Jollivet - Supervision, Project administration, Funding acquisition, Resources, Writing original draft and review.

Supplementary Material

Data available from the Dryad Digital Repository : `https://doi.org/10.5061/dryad.dbrv15f6f`.

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was supported by the "Projet Emergence" grant (Sorbonne Université).

II.6 Annexes

A step in the deep evolution of Alvinellidae (Annelida: Polychaeta): a phylogenomic comparative approach based on transcriptomes

Supplementary Data

	Raw sequencing reads ($x10^6$) and bases ($x10^9$)	Sequencing reads ($x10^6$) and bases ($x10^9$) after filtration	Assembled transcripts ($x10^3$)	% prokaryotic contamination	ORF ($x10^3$)	Busco % complete or % fragmented transcripts	Orthologous genes
Pectinaria gouldii	/ (1)	/ (1)	55	0.31	40	53.5/21.6	1235
Melina palmata	334.5/50.51	223.8/29.4	872	0.07	515	99.0/0.4	1939
Terebellidae. sp. (2 individuals)	2014.3/304.2	1545.1/195.4	2340	0.08	1308	95.0/4.0	1879
Neoamphitrite edwardsii	280.9/42.4	186.6/24.4	366	0.08	183	98.1/0.3	1908
Anobothrus spp.	76.8/7.7	56.6/5.4	432	0.08	193	90.3/6.3	1753
Amphisamytha carldarei	78.5/7.9	43.8/4.2	280	0.07	126	76.4/19.7	1670
Amphicteis gunneri	95.9/9.6	71.3/6.9	408	0.07	182	86.1/9.9	1688
Hypania invalida	49.5/4.7	27.1/2.5	195	0.06	124	97.4/1.8	1905
Alvinella caudata	230.1/34.8	154.9/20.4	394	0.10	299	99.2/0.3	1925
Alvinella pompejana	/ (2)	/ (2)	45	6.52	39	94.9/2.8	1871
Paralvinella pandorae irlandei	13.8/1.0	1.0/0.7	195	0.27	80	35.4/31.1	1009
Paralvinella unidentata 1	311.1/31.1	244.3/23.6	488	0.50	369	98.5/0.6	1942
Paralvinella unidentata 2	407.2/61.5	321.5/41.3	345	0.22	227	96.1/2.6	1907
Paralvinella unidentata 3	316.9/47.9	248.8/31.8	374	0.27	488	98.3/0.8	1925
Paralvinella palmiformis	270.6/27.1	221.7/21.3	372	0.30	194	98.7/0.4	1916
Paralvinella grasslei (2 individuals)	56.8/5.5	46.1/4.4	223	0.23	132	76.2/19.1	1595
Paralvinella mira	13.4/2.0	11.0/1.5	127	0.16	86	91.0/4.5	1849
Paralvinella hessleri 1	418.6/41.9	330.5/31.9	467	0.29	268	97.8/1.3	1925
Paralvinella hessleri 2	685.2/103.5	531.0/68.3	331	0.22	179	96.5/2.4	1890
Paralvinella sulfincola	/ (1)	/ (1)	21	0.41	32	68.1/14.6	1126
Paralvinella sp. nov. (2 individuals)	894.7/135.1	669.4/85.4	560	0.35	276	83.4/13.3	1460
Paralvinella fijiensis Manus Basin	1092.2/164.9	858.9/109.8	432	0.34	269	98.5/0.6	1921
Paralvinella fijiensis 1 Lau Basin	326.5/32.6	252.8/24.4	496	0.52	359	98.6/0.5	1938

	Raw sequencing reads ($x10^6$) and bases ($x10^9$)	Sequencing reads ($x10^6$) and bases ($x10^9$) after filtration	Assembled transcripts ($x10^3$)	% prokaryotic contamination	ORF ($x10^3$)	Busco % complete or % fragmented transcripts	Orthologous genes
Paralvinella fijiensis 2 Lau Basin	493.0/74.4	384.1/49.4	347	0.24	223	97.3/1.9	1866
Paralvinella fijiensis 3 Lau Basin	430.3/65.0	336.1/43.3	325	0.24	198	97.4/1.5	1892

Table 1. ***Metrics of the different assembly steps and identification of orthologous genes.*** Reads are filtered with Fastp. Transcript assembly is performed with Trinity. Prokaryotic transcript cleaning is performed with Kraken. ORF delineation and trimming of 5' and 3' UTRs is performed with Transdecoder. Transcriptome completeness assessment is performed with Busco in comparison with the ODB9 Metazoa database (978 Buscos). Orthologous gene clusters are retrieved with Orthograph, based on a database of 1997 orthogroup genes identified in lophotrocozoa as single-copy genes. (1) Transcriptomes provided by Didier Jollivet. (2) Transcripts retrieved with Augustus from the *A. pompejana* genome.

Command lines

FASTP

```
fastp -i $1 -o reads_fastp_R1.fastq.gz -I $2 -O reads_fastp_R2.fastq.gz $phred
--compression=9 --detect_adapter_for_pe --adapter_fasta adapters-primers.txt -n
10 -q 25 -u 50 -l 50 -y 10 --cut_right --cut_right_window_size=5 --
cut_right_mean_quality 30 --correction --overlap_len_require 20 --
overlap_diff_limit 2 --overlap_diff_percent_limit 5 --cut_front --
cut_front_window_size 1 --cut_front_mean_quality 25 --trim_poly_g --
poly_g_min_len 20 --trim_poly_x --poly_x_min_len 20
```

TRINITY ASSEMBLY

```
Trinity --seqType fq --max_memory 130G --left reads_fastp_kraken_R1.fq.gz --
right reads_fastp_kraken_R2.fq.gz --CPU 8 --min_contig_length 50 --output
transcriptome_trinity --full_cleanup
```

CAP3

```
cap3 transcriptome.fasta -o 50 -p 99 > transcriptome.cap3.fasta
```

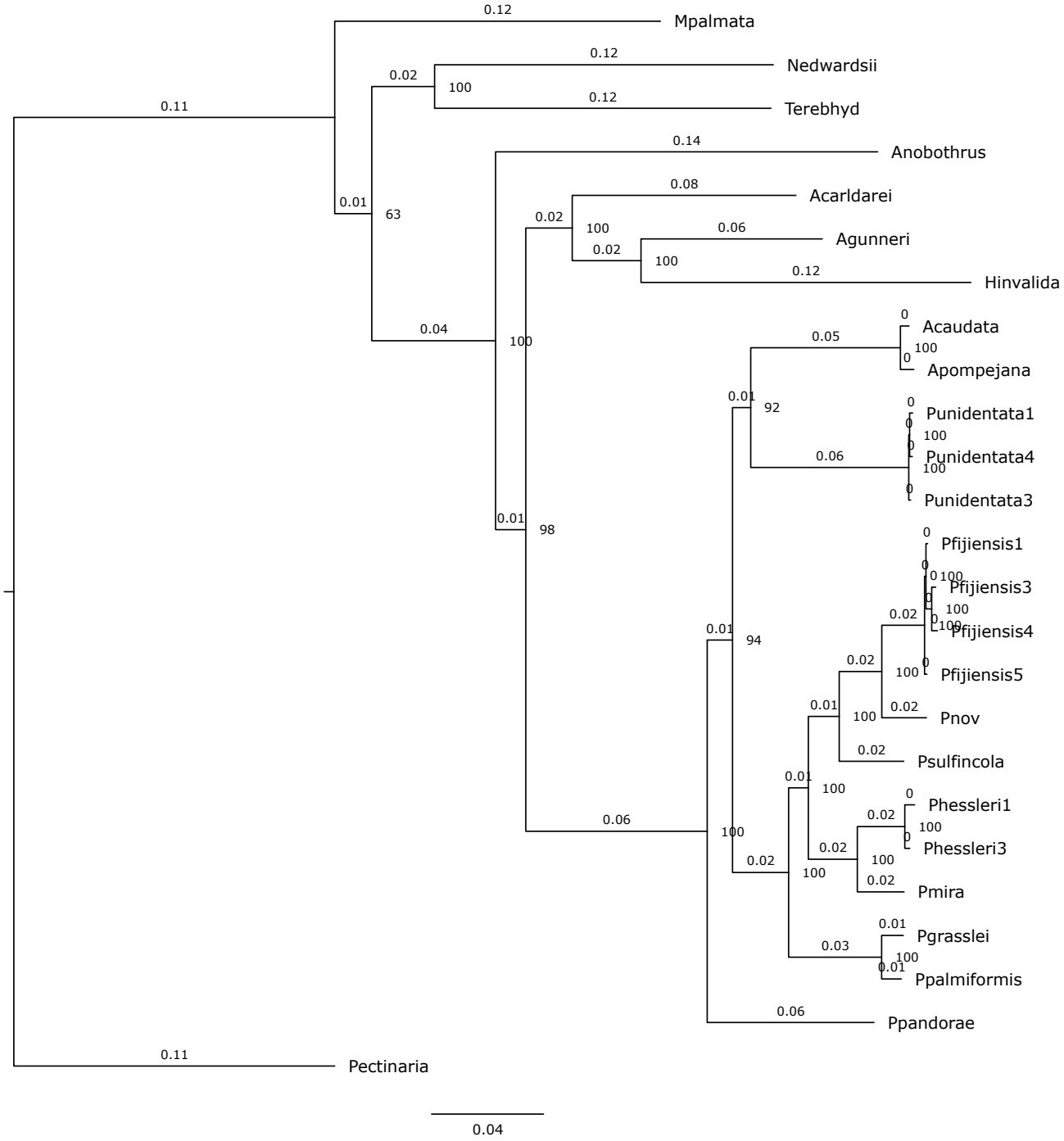


Figure 1. ***Phylogeny obtained from 657 concatenated nucleotide genes*** (499,036 sites) shared by at least 20 transcriptomes. IQ-TREE 2.0.3, partitioned model. 1000 ultrafast bootstraps.

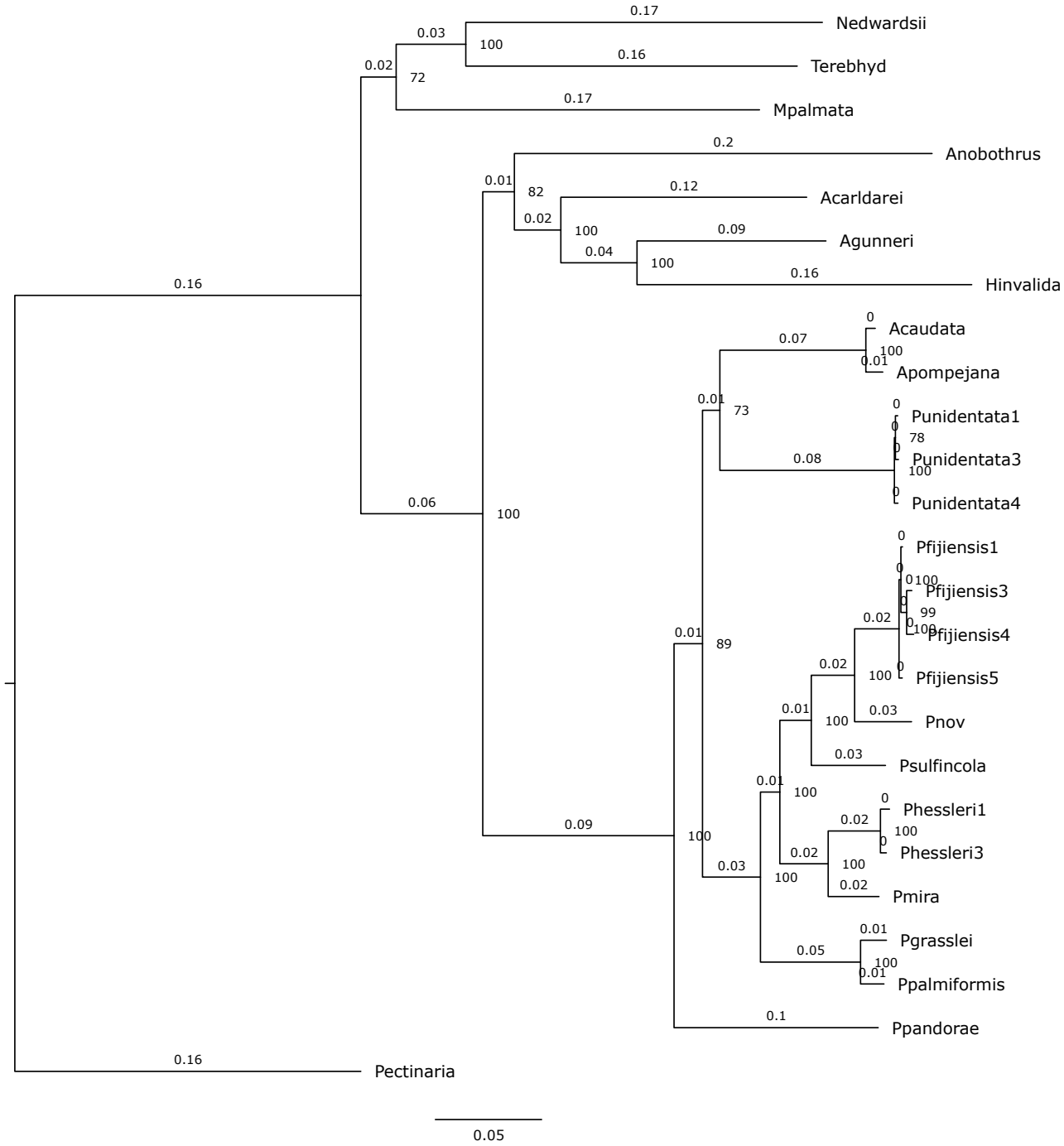


Figure 2. ***Phylogeny obtained from 699 concatenated amino acid genes*** (277,900 sites) shared by at least 20 transcriptomes. IQ-TREE 2.0.3, partitioned model. 1000 ultrafast bootstraps.

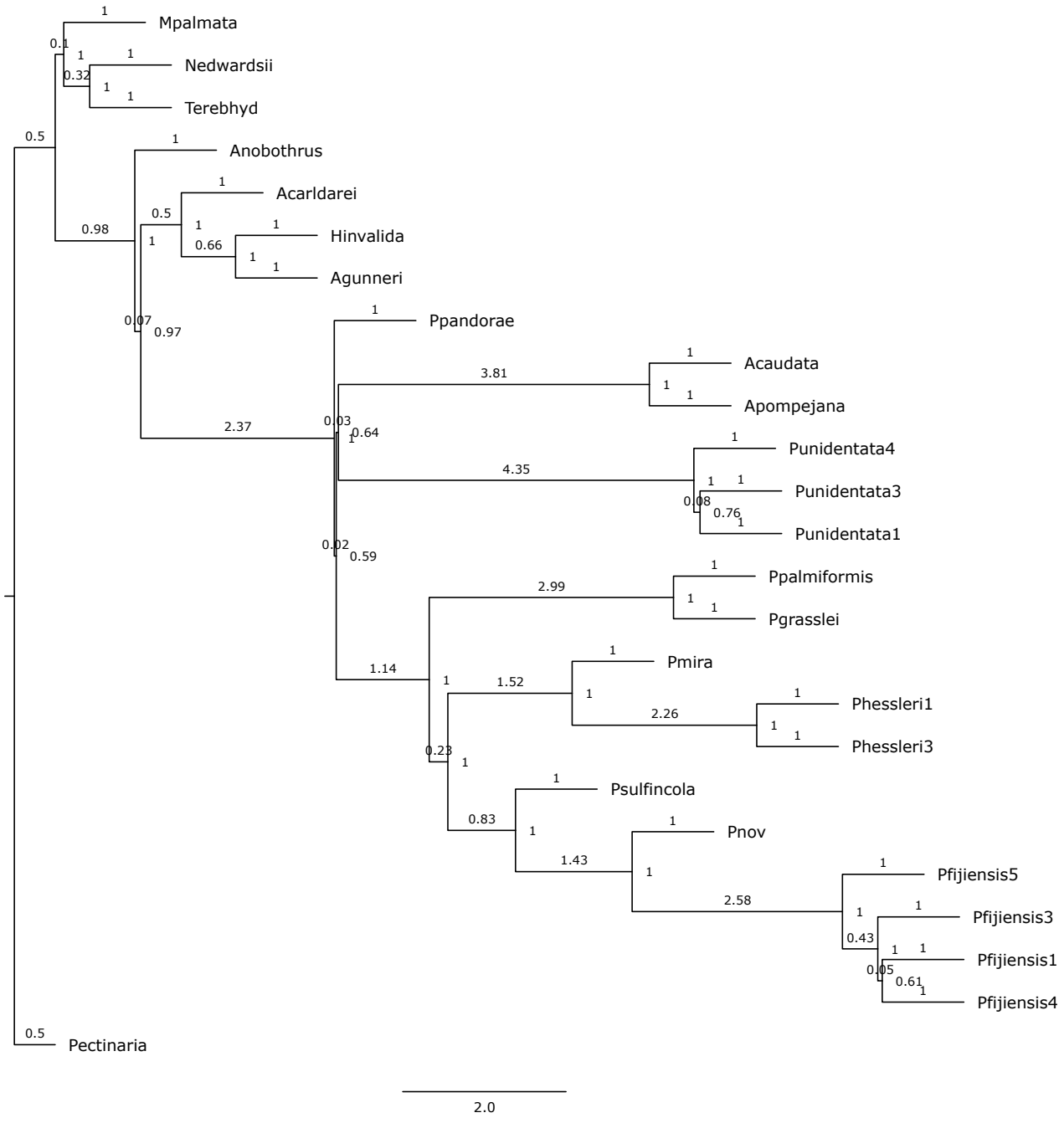


Figure 3. ***Phylogeny obtained from 657 nucleotide gene trees*** shared by at least 20 transcriptomes. Gene trees are obtained with IQ-TREE 2.0.3. The species tree is obtained with ASTRAL 5.7.8. Posterior probabilities are indicated at each node.

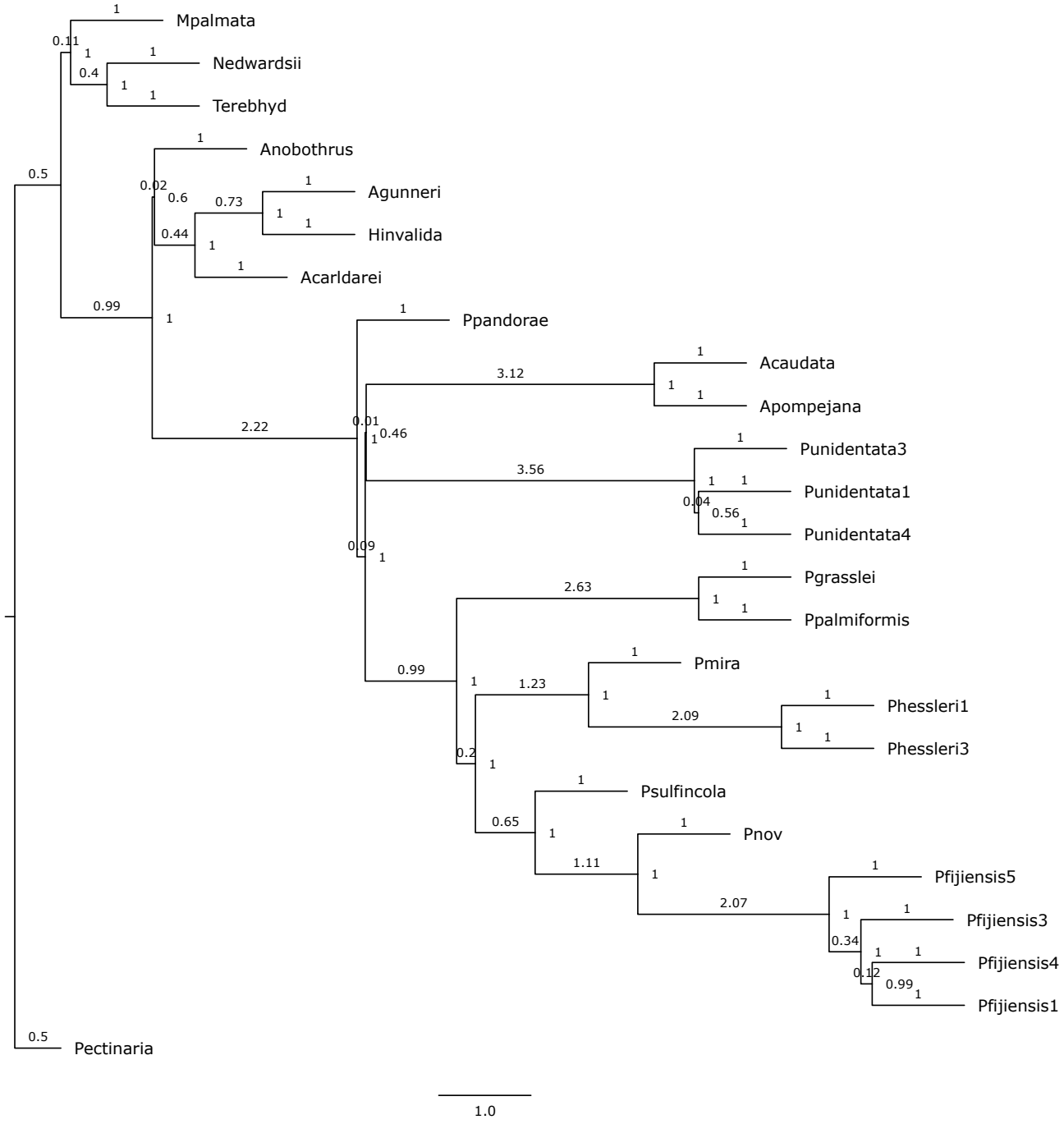


Figure 4. ***Phylogeny obtained from 699 amino acid coded gene trees*** shared by at least 20 transcriptomes. Gene trees are obtained with IQ-TREE 2.0.3. The species tree is obtained with ASTRAL 5.7.8. Posterior probabilities are indicated at each node.

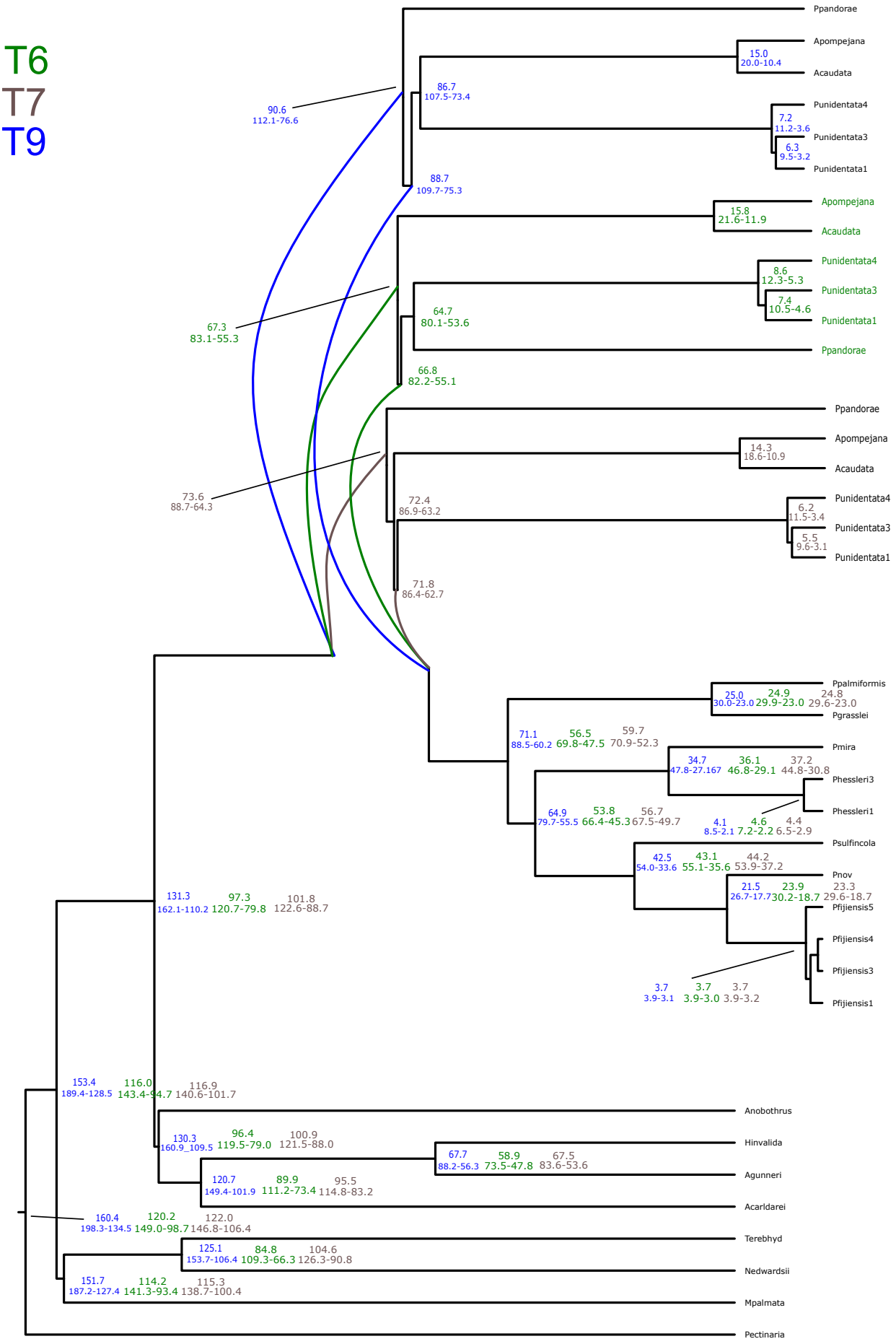


Figure 5. ***Age estimates according to the CIR model***, obtained with Phylobayes 4.1 under the three topologies T6, T7 and T9 with 95% confidence intervals. Date estimates for other molecular clock models (log Normal and Uncorrelated gamma) are given in supplementary material as chronograms.

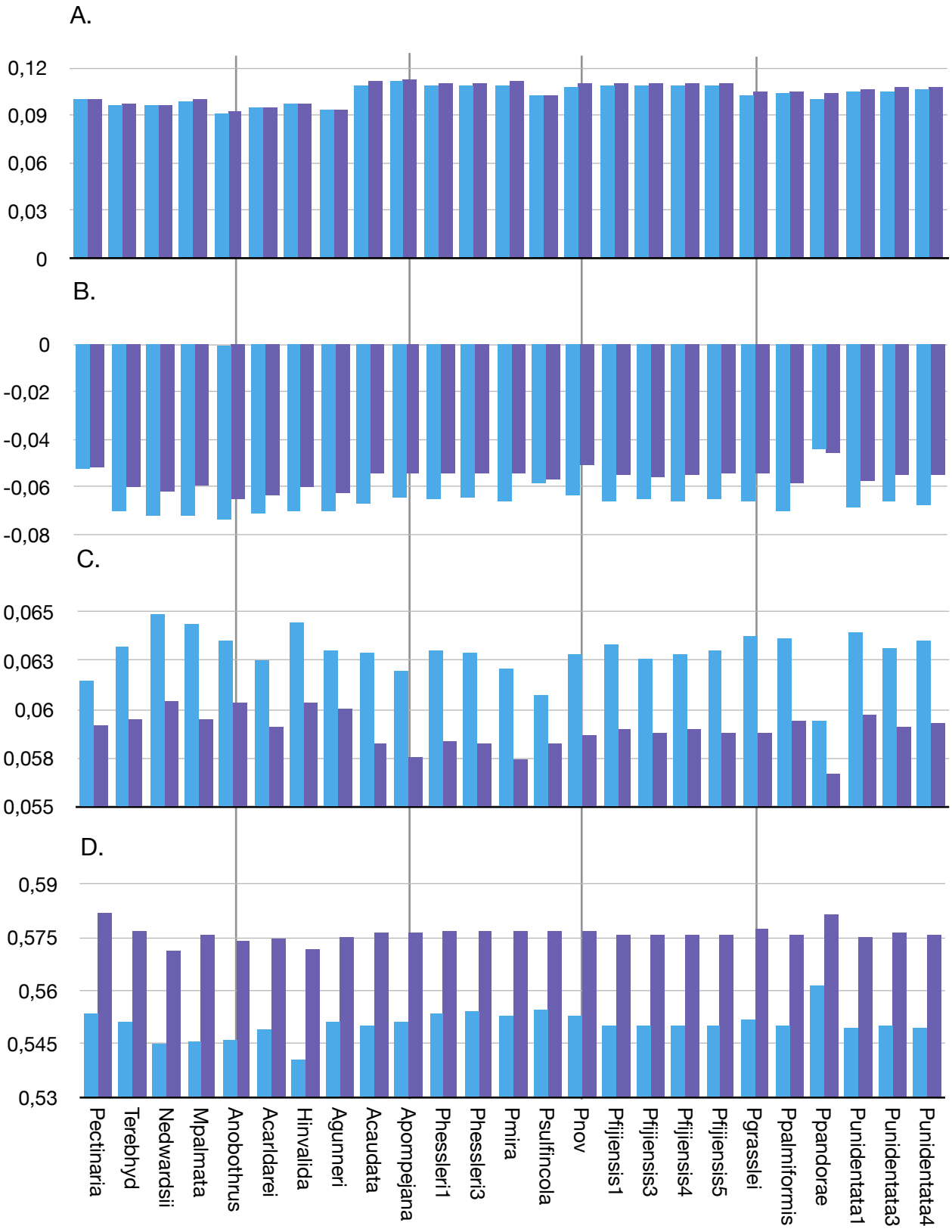


Figure 6. ***Composition of the transcriptomes.*** The blue bars correspond to the raw compositions of the total assembled transcripts, trimmed of the 5' and 3' UTRs and assigned to an ortholog group by Orthograph. The purple bars correspond to the compositions after filtration (composition/stationarity test as well as 3rd codon nucleotide removal for nucleotide-encoded genes). The sequences used for phylogenetic inference correspond to the filtered genes. **A.** PAYLE vs. DGMS bias. **B.** CvP bias (EDKR vs. GHNPQST). **C.** Serine bias. **D.** Purine load (AG vs. TC).

Chapitre III

Thermophilie de l'ancêtre des Alvinellidae

Cette seconde partie de la thèse a visé à reconstruire et exprimer certaines protéines ancestrales des Alvinellidae, ainsi que de caractériser leur stabilité. Nous formulons l'hypothèse que chez les Alvinellidae, les protéines d'espèces thermophiles sont en moyenne plus stables que les protéines d'espèces froides, comme suggéré par plusieurs études antérieures (Jollivet, Desbruyères et al., 1995 ; Rinke et Lee, 2009). Nous devons cependant déterminer si cela reste vrai pour les protéines sélectionnées dans ce chapitre.

La reconstruction des protéines ancestrales s'appuie notamment sur la phylogénie des espèces établie dans le chapitre 1. Pour les espèces du genre *Paralvinella*, nous considérons que les relations sont bien résolues. En revanche, une incertitude persiste quand aux relations évolutives entre les lignées profondes de la famille. En particulier, nous avons conclu que différents gènes puissent ne pas avoir la même histoire évolutive du fait de tri incomplet de lignées et d'introgression interspécifique ancestrale. Pour cette raison, nous choisissons de reconstruire les séquences ancestrales sous deux hypothèses phylogénétiques différentes, correspondant aux propositions les plus vraisemblables.

Nous avons choisi de reconstruire les malate déshydrogénase cytosolique (MDHc), superoxyde dismutase Cu/Zn (SOD) ainsi qu'une hémoglobine intracellulaire. L'objectif étant, potentiellement, d'obtenir d'autres informations concernant le paléo-environnement comme l'hypoxie, les conditions oxydo-réductrices et le contenu en métaux, qui sont des paramètres importants pour caractériser le milieu hydrothermal. Les ancêtres sélectionnés correspondent au dernier ancêtre commun de la famille, ainsi qu'à cinq autres ancêtres importants pour décrire l'évolution de la thermophilie au sein de la lignée. A cela s'ajoutent les protéines de 8 espèces contemporaines, quatre espèces considérées chaudes (*A. pompejana*, *P. sulfincola*, *P. mira*, *P. fijiensis*) et quatre espèces froides (*P. unidentata*, *P. p. irlandei*, *P. grasslei*, *P. palmiformis*).

Au total, 44 protéines ont été exprimées et purifiées. La stabilité de ces protéines a été suivie expérimentalement par nanoDSF ou nanoDSC grâce à des collaborations avec la plateforme de biophysique moléculaire de l'Institut Pasteur et la plateforme de Biologie struc-

turale de l'Institut intégrative de la cellule (I2BC) de l'Université Paris-Saclay. L'incertitude sur les reconstructions ancestrales, due à des incertitudes concernant la topologie des arbres de gène ou bien à l'inférence de la séquence en elle-même, a également été évaluée via une approche *in silico* de simulation des stabilités de protéines alternatives.

Ce chapitre présente les résultats obtenus pour la MDHc ainsi que la SOD Cu/Zn. La mesure de la dénaturation de l'hémoglobine intracellulaire ayant été plus difficile à obtenir, des résultats partiels pour cette protéine sont présentés en annexe. Néanmoins, considérant les mesures obtenues pour le dernier ancêtre commun des Alvinellidae pour ces trois familles de protéines, nous pouvons conclure avec une bonne confiance que l'ancêtre de la famille était déjà une espèce thermophile, y compris en prenant en compte les différentes sources d'incertitudes liées à la reconstruction des séquences.

Protein resurrection shed light into the thermophilic trait of the Last common ancestor of the Alvinellidae family worms

Pierre G. Brun, Anne-Sophie Le Port, Sébastien Brûlé, Magali Aumont-Niçaise, Lionel Cladière and Jean Mary*

*jmary@sb-roscoff.fr

III.0.0.0.0.1 Abstract Alvinellid worms are endemic worm species from the hydrothermal vents of the Pacific and Indian oceans. They are believed to have a long history of speciation in these harsh environments, dating back to the Cretaceous era. Contemporary species face contrasted thermal regimes, as some species are associated with the walls of hydrothermal chimneys, experiencing higher temperatures than closely-related species which are not on the chimneys. Using ancestral protein reconstruction on both the enzymes Cu/zn superoxyde dismutase (SOD) and the cytosolic malate dehydrogenase (cMDH), we show that these two proteins have a last common ancestor for the Alvinellidae as stable as the proteins of nowadays warm alvinellid species. Combining experimental thermodynamic measures of protein unfolding (micro-calorimetry and differential scanning fluorimetry) on ancestral recombinant proteins and computer simulations, we demonstrate that this result is reliable in the face of reconstructions' uncertainties. We thus conclude that the ancestor of the Alvinellidae was a thermotolerant species. The adaptation to colder environments in some current species is a trait that was gained recently, independently in several lineages.

III.1 Introduction

Alvinellide are a family of marine worms endemic to deep-sea hydrothermal vents from the Pacific and Indian Ocean (Desbruyères et Laubier, 1986 ; Han et al., 2021). Their name come from the Alvin submersible which was used to discover for the first time deep-sea hydrothermal vents in 1977 on the Galapagos rift at a depth of 2,500 m (P. Lonsdale, 1977). Hydothermal vents are acknowledged as extreme environments, where highly specialized chimiosynthetic-based communities can thrive using the reduced chemicals present in the hydrothermal fluid (Hourdez et Weber, 2005). This fluid indeed results from hot end-members' emissions of sea water percolating through the oceanic crust to the magmatic chamber that subsequently mix with cold deep-sea water at the vent chimneys. This produces steep but unstable environmental gradients, both physical (contrasting temperatures) and chemical (pH, oxygen, hydrogen, sulfide) (Lelièvre et al., 2018). While the *in situ* measurement of the parameters of the fluid mixing remains challenging, notably due to its high heterogeneity and complex non-conservative processes occurring during the mixing, the end-member fluid is hot (temperatures above 300 °C), sulfide-rich (hundreds of micromolar), enriched in metals (Fe^{2+}, Zn^{2+}, Cd^{2+}, Cu^{2+}, Pb^{2+}), acidic pH (reaching 2-4), CO_2-rich (3.5-6 mM) and O_2-depleted (Le Bris et Françoise Gaill, 2007). This implies that some environmental parameters tend to correlate to form an abrupt chemo-physical gradient from the source to the abyssal vicinity . Thus, higher temperatures of the fluid are generally sulfide-rich, and more depleted in O_2 as a consequence of the spontaneous reaction with sulfide : $2\,H_2S + 3\,O_2 \longrightarrow 2\,H_2O + 2\,SO_2$ (Johnson et al., 1986 ; Stéphane Hourdez et Lallier, 2007). As a consequence, hydrothermal species closest to the end-member fluid are expected to experience regular bursts of temperatures, chronic hypoxia and high oxidative stress.

In this view, Alvinellidae, which have adapted to a wide range of environmental conditions within hydrothermal vents, represent an outstanding family for the study of adaptations to extreme environments. The most striking adaptive features displayed by Alvinellidae are arguably linked to the apparent constrasting thermal regimes they experience. *A. pompejana* is indeed one of the most thermotolerant metazoan described to date, with a

thermal optimum in pressurized aquaria between 42 °C and 50 °C (Ravaux et al., 2013). Other species of the family, such has the ecological homologous*P. sulfincola*, also appear to have a thermal preference above 40 °C (Girguis et Lee, 2006). On the contrary, other Alvinellidae species such as the sister-species *P. palmiformis* and *P. grasslei* are comfortable below 35°C and 20°C respectively (Girguis et Lee, 2006 ; Cottin et al., 2008). The first adaptation reported *in situ* allowing *Alvinella* species to cope with high temperatures is the secretion of a thick mineralized tube with higher thermal and chemical stability than other annelid tubes. The medium inside these tubes contains 72-91 % seawater, with a near-neutral pH and lower temperatures (14-59 °C) compared to the end-member fluid (Le Bris, Zbinden et Françoise Gaill, 2005 ; Le Bris et Françoise Gaill, 2007). *A. pompejana* spends most of the time in its tube, which probably acts as a protection against the extreme and fluctuating external environment Chevaldonné et Jollivet, 1993 ; Desbruyères, Chevaldonné et al., 1998. Other "hot" alvinellid species, such as *P. sulfincola*, *P. hessleri*, *P. mira* or *P.* sp. nov. also produce thick tubes, while colder species such as *P. grasslei*, *P. palmiformis*, *P. p. irlandei* produce a thinner mucus cocoon Desbruyères et Laubier, 1991 ; Le Bris et Françoise Gaill, 2007. *P. grasslei* is also reported to show escape behaviors in the face of increasing temperatures, contrary to *A. pompejana* (Chevaldonné et Jollivet, 1993). Physiological and molecular studies also reveal different physiological performances linked to differential temperature tolerance. Isolated mitochondria of *A. pompejana* and *A. caudata*, both thermotolerant, are more resistant to high temperatures with a break in the Arrhenius plot (consumption of O_2/min-mg protein at different temperatures) at 48-49 °C against 21-33 °C for deep-sea cold species or cool shallow waters species (Dahlhoff, O'Brien et al., 1991). At a molecular scale, Jollivet et Stéphane Hourdez, 2020 reported that several proteins' activities (aspartate aminotransferase, glucose-6-phosphate isomerase, phosphoglucomutase) of warm species (*A. pompejana*, *A. caudata*, *P. sulfincola* and *P. hessleri*) were less sensitive to temperature than their homologous proteins from cold-adapted species (*P. palmiformis*, *P. grasslei*, *P. p. pandorae* and *P. p. irlandei*). Thus, there is a body of evidence that all alvinellid species are adapted to a wide range of environmental temperatures, even though the precise *in situ* measure of their optimal temperatures remains challenging (Chevaldonné, Fisher et al., 2000 ; Le Bris et Françoise Gaill, 2007).

Alvinellidae are thought to have a long evolutionary history of speciation from a common ancestor already adapted to hydrothermal vents at the end of the Cretaceous period (Jollivet et Stéphane Hourdez, 2020 ; Brun et al., 2023). The characterization of this ancestor would bring valuable information in understanding how contemporary lineages have gained or lost traits to adapt to different thermal regimes. Based on mean amino-acid composition of *in silico* reconstructed ancestral protein sequences, as well as the inference that lineages leading to cold species evolved more rapidly than lineages leading to warm species, Fontanillas et al., 2017 proposed that the ancestor of the family was adapted to warmer environments. This hypothesis of a thermotolerant ancestor is in good agreement with the higher temperature increase of 10 to 15 °C of the Pacific deep waters during the late Mesozoic between 100 and 60 Myr, with a peak reached between 55 and 60 Myr (Savin, 1977 ; Vrijenhoek, 2013). In this study, our objective is to test this hypothesis and refine the evolutionary history that led to the nowadays species diversity in the family *via* an experimental approach.

To this end, we considered ancestral protein sequence reconstruction (ASR) as a manner

to infer the phenotype of the ancestor of the Alvinellidae and test its performance toward temperature. This method aims at reconstructing ancestral amino acids at each site of homologous contemporary sequences, taking the phylogenetic relationships and estimated divergences between modern sequences as a probabilistic model (Koshi et Goldstein, 1996). This strategy has already proven effective to estimate paleoenvironmental living conditions of extinct organisms such as salinity conditions or temperatures as old as 3.5 billion years ago (Gaucher, Govindarajan et Ganesh, 2008 ; Blanquart, Groussin et al., 2021). Indeed, concerning temperature, proteins evolve in such a way as to remain marginally stable at the temperature at which they are supposed to function (Taverna et Goldstein, 2002 ; Goldstein, 2011). For ectothermic organisms, such as alvinellid worms, this should translate into melting temperatures (T_m, temperatures at which 50 % of the protein population is in its folded/native state) that are linearly positively correlated with environmental temperature (Wheeler et al., 2016). Moreover, although it remains difficult to describe obligatory mechanisms underlying protein stability, it is known that stability is well assessed from the primary structure and composition of proteins (Vihinen, 1987 ; Jaenicke et Böhm, 1998 ; Farias et Bonato, 2003 ; Pack et Yoo, 2004 ; Fraser et al., 2016). Thus, the *in vitro* expression of ancestral alvinellid proteins and the characterization of their thermostability should be a reliable proxy to to infer the life temperature of the alvinellid ancestors.

Our take in this study was to fully characterize the thermodynamic stability of several reconstructed ancestral proteins at key nodes for the most probable gene trees expected for alvinellid sequences (Brun et al., 2023). We chose to reconstruct the ancestors of the cytosolic malate dehydrogenase (cMDH) and Cu/Zn superoxyde dismutase (SOD) in order to compare the coherence in the signal given by these two proteins. The cMDH is a dimeric protein responsible for the conversion of malate into oxaloacetate (OAA) with the reduction of the coenzyme nicotinamide adenine dinucleotide, NAD^+ (Minárik et al., 2002). The K_m of the NADH for this enzyme in *A. caudata* and *A. pompejana*, two warm alvinellid worms, has been shown to be weakly affected by temperature and to remain pressure-insensitive, in contrary to cold shallow-water invertebrates (Dahlhoff et Somero, 1991), which suggests adaptations of the catalytic properties of this enzyme to hydrothermal conditions. The cMDH is also involved in the redox balance of the cytosol and in anaerobic metabolism in O_2-depleted environments (Hand et Somero, 1983 ; Dahlhoff et Somero, 1991). The second enzyme Cu/Zn SOD is one of the most important enzyme involved in redox regulation. Reactive oxygen species (ROS, such as the superoxide radical anion $O_2^{\bullet -}$), which are produced during aerobic respiration or caused by exposure to pollutants, can damage macromolecules and microstructures in the cell (Bordo, Djinovic et Bolognesi, 1994 ; Zeinali, Homaei et Kamrani, 2015). SOD catalyses the reaction $2\,O_2^{\bullet -} + 2\,H^+ \longrightarrow H_2O_2 + O^2$, while hydrogen peroxide is later converter into water by the catalase, glutathione peroxidase or other peroxidases Zeinali, Homaei et Kamrani, 2015. The *in vivo* expression and activity of SOD is known to be up-regulated by metal tissue-content, also in polychaetes (Rhee et al., 2011 ; Zeinali, Homaei et Kamrani, 2015), or even by industrial pollutants exposure (Xia et al., 2016), indicating the central role of SOD in ROS regulation. In Alvinellidae, the structure of this dimeric protein has been elucidated in *A. pompejana* (Shin, DiDonato, Barondeau, Hura, Berglund et al., 2010). As expected for these animals which are exposed to high oxidative threat, the SOD activity is reported to be high in *P. grasslei* (Marie et al., 2006) and even higher in *A. pompejana* (Genard et al., 2013). Interestingly, these activities were particularly high in gut tissues, potentially due to respiration-independent ROS

generation caused by metal exposure (Marie et al., 2006 ; Genard et al., 2013). The comparative study of these two central enzymes between present-days species an their direct ancestors could bring valuable information about the paleoenvironmental conditions of the Alvinellidae (in terms of temperature, but also oxygen, metal and sulfide content), as well as the current living conditions of modern species which can be challenging to assess and not always known with certainty.

III.2 Material & Methods

III.2.1 Obtaining the genes for contemporary species

We were interested in reconstructing three families of ancestral proteins, namely the cytosolic malate dehydrogenase (cMDH) and the Cu/Zn superoxyde dismutase (SOD).

For this purpose, we collected the coding sequences for these three proteins in all transcriptomes available for alvinellid and several outgroup species belonging to the Terebelliformia sub-order, namely *H. invalida* (SRR5590961), *A. gunneri* (SRR11434467), *A. carldarei* (SRR11434468), *Anobothrus* sp. (SRR11434464), *P. gouldii* (SRR2057036), Terebellidae gen. sp. (XXX), *N. edwardsi* (XXX), *M. palmata* (XXX) previously obtained from a series of RNAseq data used to establish their phylogeny (Bioproject number XXX), Methods regarding animal collection, RNA extractions, transcriptome assembly and open reading frame predictions are provided in Girguis et Lee, 2006 ; Fontanillas et al., 2017 ; Stiller, Tilic et al., 2020 and Brun et al., 2023.

Genes of *A. pompejana* were predicted from the 400 Mb genome assembled by R. Copley (NCBI accession number PRJEB46503, El Hilali et al., 2024) with the Augustus webserver (Hoff et Stanke, 2013), trained on cDNA sequences from previous *A. pompejana* EST assemblies (Genoscope project, see Gagnière et al., 2010). The three proteins were retrieved with BLASTP against the predicted genes. The best-hit genes were manually filtered to remove diverging sequences and identify potential paralogy. The *A. pompejana* coding sequences were then used to identify the genes in all other species' unfiltered transcriptomes *via* BLASTP. The best-hit sequences were filtered to remove paralogy, both manually and with the help of PhyML to regroup our sequences of interest (Guindon et al., 2010). Final coding sequences were aligned with ProbCons v.1.12. (Do et al., 2005) and manually refined when needed.

III.2.2 Obtaining the genes for ancestral species

Amino-acid encoded genes from contemporary species were used to infer the coding sequences of all alvinellid ancestors. Ancestral protein reconstruction was performed with FastML v.3.11, using branch length optimization, Γ parameter divided in eight discrete classes and parsimony reconstruction for indels (Ashkenazy, Penn et al., 2012). The JTT matrix was used for the reconstruction of the MDHc and the Hb, while the WAG matrix

was used for the reconstruction of the Cu/Zn SOD as the model had a better likelihood with this later matrix for the reconstruction of ancestors.

The species tree topology for the Alvinellidae is not fully resolved at its deepest nodes, as it is likely that the family experienced a rapid radiation leading to high incomplete lineage sorting at the root of the family (Brun et al., 2023). We therefore retained two main alternative hypotheses for the tree topology to use in the reconstruction of experimentally expressed ancestral proteins. The first one corresponds to the case where the *Alvinella* species are sister to the *Paralvinella* species, with *P. p. irlandei* and *P. unidentata* grouped together in the *Nautalvinella* sub-genus. This topology H1 correspond to the proposal of Desbruyères et Laubier, 1986, based on the study of morphological characters. This is also a topology supported by a high number of genes, according to Brun et al., 2023. The second topology H2 corresponds to a topology where *P. p. irlandei* is sister to all other Alvinellid worms, while *P. unidentata* is grouped with the *Alvinella* species. This topology is also supported by a high number of genes, and the most supported species tree according to a coalescent tree based on individual gene trees (Brun et al., 2023). These two topologies were considered for the reconstruction of ancestral sequences of the family.

To explore the potential effect of tree uncertainty on the ancestral protein reconstructions, we also compared *in silico* the ancestral protein reconstructions obtained with the 13 other topologies hypotheses proposed in Brun et al., 2023. The aim of this step is to quantify the extent to which the two experimentally expressed proteins under H1 and H2 hypothesis are sensitive to tree uncertainty. The tree topology uncertainty was explored for the two protein families using IQTREE v.2.0.3 (Minh et al., 2020) and the implementation of ModelFinder (Kalyaanamoorthy et al., 2017) in IQTREE. The posterior topology probabilities for each protein families were approximated from their tree ML estimates with $Pr(n) = \frac{L_n}{\sum_{i=1}^{15} L_i}$, with L_n the likelihood for the n-th topology. Ancestral proteins were inferred using FastML for the most relevant alternative tree topologies. The expected number of rightly-inferred residues, E, for the LCA sequences was calculated from the marginal probabilities of the ML residues, $E = \sum_{i=1}^{n} Pr(i)$ with n the number of amino acids in the sequence and $Pr(i)$ the marginal probability of the i-th residue in the ML sequence. We also compared these potential LCA sequences with the ancestral proteins we actually expressed and measured. The similarity between these sequences is simply taken as the identity between the ML reconstruction and the different expressed proteins.

Finally, the ML protein reconstructions for different ancestors under the H1 and H2 hypotheses were retro-translated according to recommendations of Boël et al., 2016 to get a full CDS sequence ready to be inserted in an overexpression plasmid. Scripts for the retro-translation are available as Supplementary material.

III.2.3 Production of recombinant ancestral proteins

Proteins corresponding to a few selected ancestors and contemporary species have been produced as recombinants in a heterologous system. The expressed contemporary species were *A. pompejana*, *P. sulfincola*, *P. fijiensis*, *P. mira*, *P. grasslei*, *P. palmiformis*, *P. p. irlandei* and *P. unidentata*, the first four species being considered as warmer species than the last

four ones. The expressed ancestors correspond to all ancestors of the Alvinellidae, excluding the last common ancestor (LCA) of *A. pompejana* and *A. caudata*, the LCA of *P. palmiformis* and *P. grasslei*, the LCA of *P. hessleri* and *P. mira*, and the LCA of *P.* sp. nov. and *P. fijiensis*, as these pairs of species are very close to one another and share similar thermal habitat. Alternative ancestral sequences either under the hypothesis H1 or H2 were considered for the protein expression.

For each protein construction, T7 Express lysY/Iq Competent *E. coli* (NEB C3013) were transformed with pET100/D-TOPO expression vectors (ThermoFisher) containing the protein coding sequence fused with a $(His)^6$ tag. The bacteria were grown in LB medium containing 100 μg/mL Ampicillin and the expression was induced by 1 mM IPTG for four hours at 37°C when the OD600 reached between 0.4 and 0.6. For Hb expression, 5-Aminolevulinic acid hydrochloride 1 mM was added to the medium. Cultures were then harvested, and pelleted cells were re-suspended into His-Bind Buffer Kit (Novagen) for the purification steps. Bacteria were sonicated and DNAse 1‰was added for 30 min on ice. The lysate was centrifuged and the proteins were then purified from the supernatant by Ni^{2+} chelation chromatography in 2 mL columns with the His-Bind Kit of Novagen and eluted into a 20 mM Tris-HCl, 500 mM NaCl, 500 mM Imidazole pH 7.4 buffer. A second purification step was performed by a size-exclusion chromatography with a HiLoad 16/60 Superdex 75 (Cytiva) column connected to an ÄKTA Avant system in 20 mM Tris-HCl, 200 mM NaCl pH 7.4 buffer. The elution was monitored at 280 nm for cMDH, 205 nm for the SOD, and 280 and 414 nm for Hb. Proteins were then concentrated between 0.6 and 1.4 mg/mL, depending on the total harvested quantity.

III.2.4 Thermal denaturation measurements

Half-denaturation temperatures of the cMDH were measured by nano-format of Differential Scanning Fluorimetry (nanoDSF) using the Prometheus NT48 (Nanotemper). This technique requires low amount of proteins (30 μg for one experiment in triplicate), and is well suited for the cMDH which possess 6 tryptophanyl residues. The 350/330 nm fluorescence ratio, $r(T)$, was measured from 20°C to 95°C with a temperature increase of 1°C/min. $r(T)$ was fitted to the fraction of unfolded protein $f_u(T)$, against the temperature T in Kelvin, with a linear correction (Yadav et Ahmad, 2000) according to :

$$r(T) = f_u(T) \times (a_u \times T + b_u) + (1 - f_u(T)) \times (a_n \times T + b_n) \tag{III.1}$$

with a_u, b_u, a_n, b_n scalars adjusting for the linear correction of the signal. The fraction of unfolded proteins is linked to the equilibrium constant $K_{eq}(T)$, according to :

$$f_u(T) = \frac{K_{eq}(T)}{1 + K_{eq}(T)} \tag{III.2}$$

and the equilibrium constant is linked to the standard free energy variation during unfolding $\Delta G^o(T)$ by :

$$K_{eq}(T) = exp(-\frac{\Delta G^o(T)}{RT}) \tag{III.3}$$

with R the molar gas constant. Finally, $\Delta G^o(T)$ is obtained with the Gibbs-Helmholtz relation (Razvi et Scholtz, 2006) :

$$\Delta G^o(T) = \Delta H_m \times (1 - \frac{T}{T_m}) - \Delta C_p \times (T_m - T + T \times ln(\frac{T}{T_m})) \tag{III.4}$$

with T_m the melting temperature at which half of the proteins are denatured, ΔH_m the enthalpy variation at T_m, ΔC_p the heat capacity change of the protein during unfolding. These parameters are estimated from the measures of the nanoDSF.

Thermodynamic parameters for the Cu/Zn SOD were measured by differential scanning micro-calorimetry using the Microcal PEAQ-DSC (Malvern), as these proteins do not contain any tryptophanyl residue to monitor its fluorescence. The excess heat capacity (C_p^{exp}) was measured between 20-110°C, with a temperature increase of 1°C/min. $C_p^{exp}(T)$ is related to the fraction of unfolded protein by (Lepock, Frey et Hallewell, 1990)

$$C_p^{exp}(T) = \Delta H_c \frac{df_u}{dT} \tag{III.5}$$

with ΔH_c the calorimetric enthalpy variation. Using equation IV.2,

$$\frac{df_u}{dT} = \frac{1}{(1 + K_{eq}(T))^2} \times \frac{dK_{eq}}{dT} \tag{III.6}$$

and combining with equation IV.3

$$\frac{dK_{eq}}{dT} = K_{eq}(T) \times (\frac{\Delta G^o(T)}{RT^2} - \frac{1}{RT} \times \frac{d\Delta G^o}{dT}) \tag{III.7}$$

$\Delta G^o(T)$ is obtained from the Gibbs-Helmholtz relation (equation IV.4), and consequently

$$\frac{d\Delta G^o}{dT} = \frac{\Delta H_m}{T_m} - \Delta C_p \times ln(\frac{T}{T_m}) \tag{III.8}$$

Equations IV.2 to IV.7 can be used to express $C_p^{exp}(T)$ against T, depending on ΔH_c, T_m, ΔH_m and ΔC_p, which can be determined by fitting the curve to the experimental points. Considering the interpretation of SOD unfolding transitions provided by Rodriguez et al., 2002 and our own results, we considered three potential successive denaturation events. The measured signal is consequently assumed to be the sum of three distinct C_p^{exp}, with their own thermodynamic parameters to estimate.

From the thermodynamic parameters associated with each protein, we obtained the melting temperature T_m, corresponding to the temperature at which half of the proteins are in a denatured state. This also allowed us to calculate the temperature of the beginning of the denaturation, that we name T_d and define as the temperature at which 5% of the proteins are in a denatured state, using the Gibbs-Helmholtz relation (equation IV.4).

III.2.5 Assessing ancestral reconstruction uncertainty

In order to assess the ancestral protein's stability taking into account ancestral residues' uncertainty, we simulated the protein stabilities with replacements of the ambiguous residues under the H1 or H2 hypothesis.

The 3D structure of the most ancestral proteins, corresponding to the last common ancestor (LCA) of the Alvinellidae family under the H1 or H2 hypothesis, was modelled with the SWISS-MODEL webserver. For the MDHc, we used the Human Malate Dehydrogenase I as a template (PDB 7RM9, 1.65Å, 71% residue identity), and for the SOD, we used the *A. pompejana* Cu,Zn Superoxide Dismutase as a template (PDB 3F7K, 1.35Å, 96% residue identity) for the SOD.

Stabilities simulations were performed with FoldX 4. First at all, 3D structures were optimized using the RepairPDB function (Schymkowitz et al., 2005). Then, we checked if FoldX was able to achieve good predictions of proteins' stabilities. We first used the FoldX function BuildModel to simulate the experimentally measured proteins (14 cMDH, 11 SOD), taking as a reference the stability modeled for the LCA's protein. The difference between the simulated energies obtained with FoldX were plotted against the measures for these proteins. We calculated the linear correlation between the two sets, according to

$$\Delta G_{exp}(protein) - \Delta G_{exp}(Ancestor) = k \times [\Delta G_{foldx}(protein) - \Delta G_{foldx}(Ancestor)] \tag{III.9}$$

with T the temperature, ΔG_{foldx} the free energy simulated by FoldX, ΔG_{exp} the free energy determined experimentally, according to equation IV.4. $protein$ is the tested protein while $Ancestor$ is the reference protein of the LCA. k is the parameter being optimized in the correlation. The temperature used to calculate ΔG_{exp} was chosen to obtain the best correlation between the simulation and the measures in the linear model. Proteins that were poorly predicted by FoldX were removed from the calculation of the linear coefficient k.

Finally, 500 alternative proteins to the maximum likelihood estimate were drawn from the posterior sequence distribution of the LCA's proteins. Residues with low marginal probabilities below 0.01 are rounded to 0 to avoid the introduction of very unlikely mutations that are not biologically relevant. These alternative sequences' stabilities, which contained a small number of mutations compared to the ML estimate, were simulated using FoldX and equation IV.9. This allowed us to obtain a stability distribution uncertainty to the ancestor's protein around the ML estimate, experimentally characterized.

III.3 Results

III.3.1 Contemporary proteins and ancestral sequence reconstructions

For the cMDH and Cu/Zn SOD, we retrieved one sequence for each species in the dataset, including outgroup species *Anobothrus* sp. and *P. gouldii*. When several transcripts were found for a species with only one or two mutations, we kept the most likely sequence according to the Alvinellid phylogeny supported by Brun et al., 2023.

For the ancestral sequence reconstruction, we first assessed which gene tree topology was the most suitable to our sequences. The different gene tree candidates were taken from Brun et al., 2023, with T6 corresponding to H1 in the present article, and T9 corresponding

to H2. The likelihoods obtained for each are presented in table III.1. The difference of tree likelihoods between the most and least probable topologies are small, between 9.49 for the cMDH and 7.07 for the SOD, which is less than the mean phylogenetic signal per residue in the alignment (9.62 and 12.63 respectively). From such short alignments, it is not possible to distinguish the candidate topologies. The consequence is that the ML sequences corresponding to the LCA of the Alvinellidae are reconstructed with high confidence and very close to one another under the different gene tree hypotheses, with more than 99% identity to the ancestors corresponding to the topologies H1 and H2. The only exception was the reconstructed cMDH under the T3, T11 and T12 hypothesis, which are closer but more different from the expressed protein Anc6-H1 and *P. sulfincola* with three, four and five mutations between the sequences.

Finally, the sequence for the Cu/Zn SOD LCA under the H2 hypothesis (T9 topology) in this reconstruction is also closer to the SOD expressed under the H1 hypothesis (T6 topology) when it should be similar to Anc1-H2. Compared to the reconstructed and experimentally expressed proteins, rerunning the reconstruction led to a slightly different solutions regarding the estimated branch lengths of the gene tree and the resulting ancestral sequence for H2. The two sequences of SOD for Anc1-H1 and Anc1-H2 are only different for one mutation at position 101 (lysine for H1 and serine for H2). The residue is ambiguous in the two reconstructions (under H2 in this second reconstruction, K has a probability 0.56 whereas S has a probability 0.44). Thus the difference in these reconstructions falls under the uncertainty of the inference and is addressed later in the article.

	Tree likelihood		Sequence confidence	
	MDHc LG+G4	SOD WAG+G4	MDHc	SOD
T1	-3203.218 0	-1939.418 0	–	–
T2	-3199.030 0	-1939.574 0	–	–
T3	-3196.139 7	-1935.004 19	330 (99%) Anc6-H1 (329, 99%)	151 (99%) Anc1-H1 (153, 100%)
T4	-3200.333 0	-1940.827 0	–	–
T5	-3200.333 0	-1939.425 0	–	–
T6	-3198.737 0	-1935.004 19	330 (100%) Anc1-H1 (332, 100%)	151 (99%) Anc1-H1 (153, 100%)
T7	-3202.180 0	-1940.925 0	330 (99%) Anc1-H2 (330, 99%)	151 (99%) Anc1-H1 (153, 100%)
T8	-3195.904 8	-1940.925 0	329 (99%) Anc1-H2 (331, 100%)	–
T9	-3202.179 0	-1940.352 0	330 (99%) Anc1-H2 (330, 99%)	151 (99%) Anc1-H1 (153, 100%)
T10	-3199.093 0	-1939.463 0	–	–
T11	-3193.728 75	-1940.892 0	330 (100%) Anc6-H1 (328, 99%)	–
T12	-3195.990 8	-1939.574 0	331 (100%) *P. sulfincola* (327, 98%)	–
T13	-3200.386 0	-1933.852 60	–	150 (98%) Anc2-H1 (150, 98%)
T14	-3202.298 0	-1939.016 0	–	–
T15	-3198.499 1	-1938.794 0	–	–

Table III.1 – Hypothesis testing for the ancestral reconstructions. Left table : potential tree topologies likelihood for each multiple sequence alignment of either MDHc or SOD under the specified evolutionary model. The approximate posterior probability of each topologies is reported below the likelihood. Alternative candidate tree topologies to the most probable H1 and H2 trees are taken from Brun et al., 2023. Right table : confidence in the reconstruction of the ancestral sequence corresponding to the LCA of the family for the most probable topologies, as well as T6, T7 and T9 which are likely gene topologies according to Brun et al., 2023. Number of expected correct residues and sequence percentage identity are indicated, as well as the closest experimentally expressed sequence (number of identical residues and percentage of identity with the ML sequence).

III.3.2 Experimental characterisation of the ancestral proteins

We chose to express and characterize the thermostability of the cMDH and Cu/Zn SOD of *A. pompejana*, *P. sulfincola*, *P. fijiensis*, *P. mira*, *P. grasslei*, *P. palmiformis*, *P. p. irlandei* and *P. unidentata*, of the LCA of the Alvinellidae under the H1 and H2 hypotheses, as well as a few other ancestors of the phylogeny, key point in the reconstruction of the evolutionary history of thermostability within the Alvinellidae lineage. The first four contemporary species are considered to live in warm environments on chimney walls, whereas the last four are described to live in rather cold ones (Jollivet et Stéphane Hourdez, 2020 ; Han et al., 2021). For ectothermic species, the thermal stability of the proteins should reflect the temperature at which they thrive, at least for species living in a warm environment. This must hold true in order to conclude if the phenotype of the ancestral proteins correspond to warm or cold ancestors.

Example of denaturation curves for contemporary Cu/Zn SOD and cMDH are shown in figure III.1.

For the cMDH, thermal denaturation was monitored by nanoDSF. All the cMDH (Fig. III.1) gave a single transition corresponding to a single unfolding event.

As the SOD do not contain tryptophanyl residue, the thermal denaturation was monitored by nanoDSC. The overall signal was decomposed using a 3-peak model, corresponding to three endothermic unfolding transitions with distinct T_m (Fig III.1). According to Rodriguez et al., 2002, these three peaks should correspond to the denaturation of the partially and fully-metallated Cu/Zn SOD. The peak for the first T_m was only absent for the Anc6-H2 protein. Despite the third peak being generally lower than the two others, it corresponds to the unfolding of the fully-metallated SOD. We therefore considered this unfolding event only in further analysis.

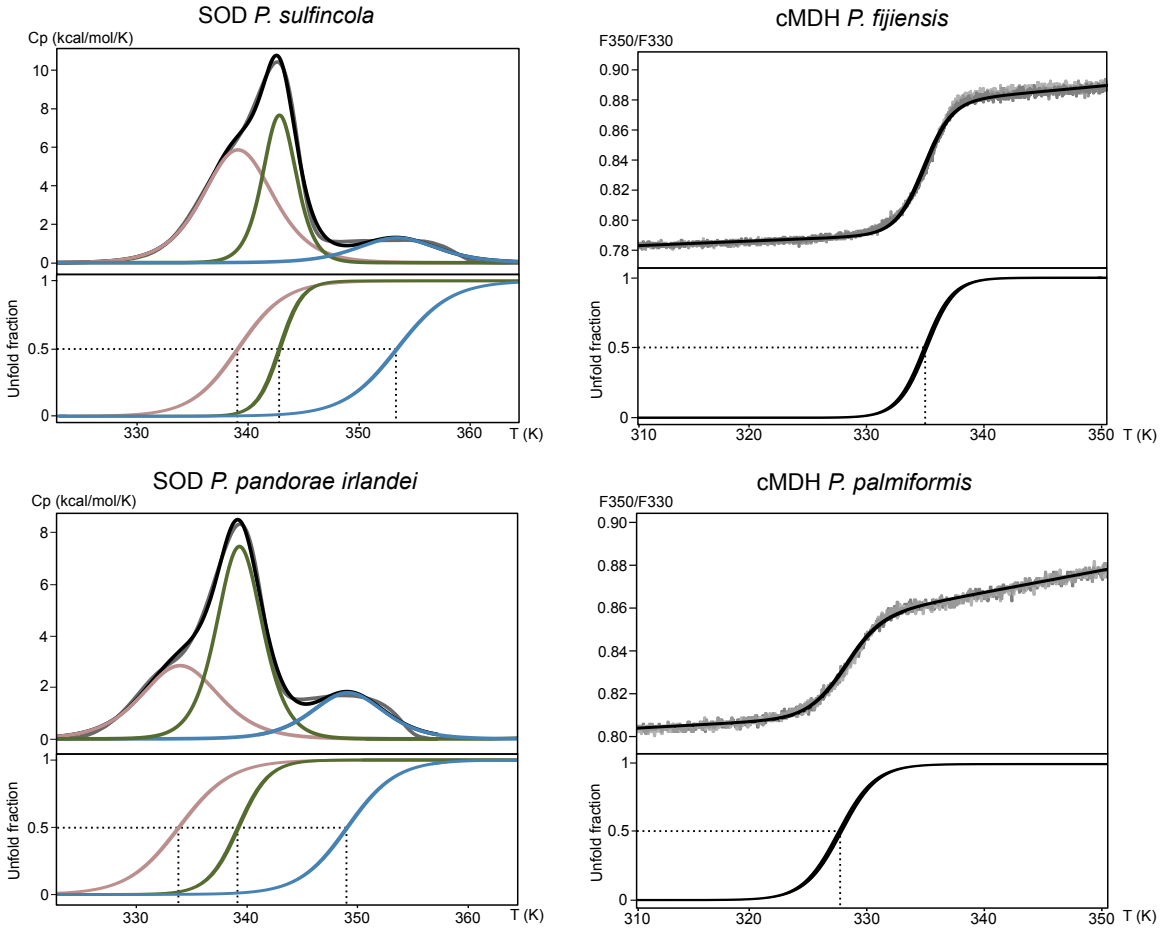


FIGURE III.1 – Experimental measures for the unfolding of proteins. Left panels show the denaturation curves obtained by nanoDSC for two SOD proteins of *P. sulfincola* and *P. p. irlandei* (heat capacity against temperature). Grey line : experimental curve. Black line : theoretical fit. Red, green and blue lines : decomposition of the theoretical fit into three denaturation events. Bottom panels show the integral of the theoretical denaturation curves, with the T_m indicated by dotted lines. Right panels show the denaturation curves obtained by nanoDSF for two cMDH proteins of *P. fijiensis* and *P. palmiformis* (350/330 nm fluorescence ratio against temperature). Grey lines : experimental curves. Black line : theoretical fit. Bottom panels show the integral of the theoretical denaturation curve, with the T_m indicated by dotted lines.

All thermodynamic parameters deduced from experimental measurements (T_m, ΔH_m and ΔC_p) are presented in table III.2. The cMDH had T_m, ranging between 55.8°C (*P. grasslei*) and 65.7 °C (*A. pompejana*), whereas the fully-metallated SOD had T_m ranged between 71.9°C (*P. palmiformis*) and 82.1°C (*P. fijiensis*). Several proteins, especially SOD, had unreliable measured thermodynamic parameters due to the apparently low-content of metallated SOD in some samples such as *P. palmiformis*. However, the T_m was easier to measure and can be trusted.

cMDH

	ΔHm (kJ/mol)	ΔCp (kJ/mol/K)	Tm (°C)
P. unidentata	704 ± 4	32.7 ± 0.4	58.2 ± 0.0
P. p. irlandei	634 ± 14	27.8 ± 0.8	60.3 ± 0.0
P. grasslei	668 ± 8	30.6 ± 0.1	57.0 ± 0.0
P. palmiformis	598 ± 3	29.7 ± 0.2	55.8 ± 0.0
A. pompejana	751 ± 12	28.9 ± 0.7	65.7 ± 0.4
P. fijiensis	773 ± 6	31.9 ± 0.1	61.5 ± 0.0
P. sulfincola	624 ± 19	27.1 ± 0.9	61.4 ± 0.1
P. mira	729 ± 7	31.7 ± 0.7	60.4 ± 0.6
Anc1-H1	737 ± 6	31.4 ± 0.3	61.4 ± 0.3
Anc2-H1	710 ± 17	31.4 ± 0.6	60.1 ± 0.4
Anc3-H1	633 ± 8	27.1 ± 0.3	60.9 ± 0.3
Anc4-H1	733 ± 2	30.6 ± 0.1	62.0 ± 0.0
Anc5-H1	624 ± 19	27.1 ± 0.9	61.4 ± 0.1
Anc6-H1	639 ± 12	28.1 ± 0.3	59.5 ± 0.1
Anc1-H2	752 ± 12	33.3 ± 0.7	60.1 ± 0.1
Anc2-H2	710 ± 17	31.4 ± 0.6	60.1 ± 0.4
Anc3-H2	633 ± 8	27.1 ± 0.3	60.9 ± 0.3
Anc4-H2	733 ± 2	30.6 ± 0.1	62.0 ± 0.0
Anc5-H2	624 ± 19	27.1 ± 0.9	61.4 ± 0.1
Anc6-H2	710 ± 17	31.4 ± 0.6	60.1 ± 0.4

SOD

	ΔHm_1 (kJ/mol)	ΔCp_1 (kJ/mol/K)	Tm_1 (°C)	ΔHm_2 (kJ/mol)	ΔCp_2 (kJ/mol/K)	Tm_2 (°C)	ΔHm_3 (kJ/mol)	ΔCp_3 (kJ/mol/K)	Tm_3 (°C)
P. unidentata	–	–	57.4	–	–	62,8	–	17.0	76.3
P. p. irlandei	413	0	60.2	660	38.5	65.8	423	17.0	75.4
P. grasslei	–	–	61.6	–	–	64.2	–	0	75.3
P. palmiformis	–	–	56.3	–	–	60.2	–	–	71.9
A. pompejana	495	0	68.7	953	50,8	70.5	481	17.4	80.4
P. fijiensis	490	0	68.7	987	67.3	70.6	442	19.1	82.1
P. sulfincola	459	7.9	65.7	924	53,6	69.6	386	14.8	79.7
P. mira	–	–	–	–	–	–	–	–	–
Anc1-H1	574	0	70.0	1009	67.1	70.8	488	21.3	79.4
Anc2-H1	574	0	70.0	1009	67.1	70.8	488	21.3	79.4
Anc3-H1	421	25.9	65.8	780	48.1	69.0	464	22.1	79.4
Anc4-H1	421	25.9	65.8	780	48.1	69.0	464	22.1	79.4
Anc5-H1	629	0	67.9	613	0	72.6	295	10.1	83.0
Anc6-H1	688	26.7	66.3	966	66.4	68.5	609	31.9	79.1
Anc1-H2	544	0	68.0	1168	75.5	69.3	528	26.0	78.5
Anc2-H2	544	0	68.0	1168	75.5	69.3	528	26.0	78.5
Anc3-H2	421	25.9	65.8	780	48.1	69.0	464	22.1	79.4
Anc4-H2	421	25.9	65.8	780	48.1	69.0	464	22.1	79.4
Anc5-H2	629	0	67.9	613	0	72.6	295	10.1	83.0
Anc6-H2	–	–	–	601	18.9	70.3	387	17.4	79.9

TABLE III.2 – Thermodynamic parameters measures for the unfolding of different proteins. ΔH, ΔC_p and T_m are the enthalpy variation, heat capacity variation and half-denaturation temperature.

III.3.3 Assessing the thermal phenotype of the alvinellid ancestors

The denaturation temperatures of the different expressed proteins are reported on the phylogeny backbone in figure III.2. Considering the T_m, the measured proteins for contemporary species are in good agreement with the hypothesis that thermotolerant species should have more stable proteins than colder species. The three species well characterised as warm, *A. pompejana*, *P. fijiensis* and *P. sulfincola* have a T_m greater than 61°C for the cMDH, whereas the three expected cold species *P. unidentata*, *P. grasslei* and *P. palmiformis* have T_m lower than 59°C. Two species, *P. mira* and *P. p. irlandei* which are considered warm and cold respectively have similar T_m for MDH, close to 60°C. For the SOD, the warm species lay above 80°C while cold species are under 76°C. Proteins of Anc1, the last common ancestor of the Alvinellidae, displayed T_m in line with the bottom spectrum of current thermotolerant species, 61°C for cMDH and 79°C for SOD, under H1. Under H2, the T_m were slightly lower (60°C and 78°C respectively), which would lead to the same conclusions. Thus, ancestors of the family appear to be rather on the warm side of the spectrum.

T_m is generally taken to experimentally compare protein's stability as it is a reliable parameter to estimate. Yet, T_m, temperature at which half of the proteins are denatured, lies well above the physiological temperature of the animals and do not correspond to the functional temperature range of enzymes. Therefore, we estimated the temperature of the beginning of the denaturation of proteins, T_d, potentially a better reflection of the upper thermal limit of species. Indeed, T_m generally correlates well with T_d, but not necessarily. We set T_d at 5% of the denatured protein, presented in figure III.2. For the cMDH, T_d is about 4°C lower than T_m, with a good correlation between the two. For the SOD, T_d is 6 to 10°C lower than T_m. Interestingly, even if the SOD has higher T_m than the cMDH, its sensitivity to temperature is greater by comparison with greater difference between T_m and T_d. The behavior of proteins at lower temperature can be different than at T_m : for the SOD of *A. pompejana* and *P. sulfincola*, the two T_m are the same, but *A. pompejana*'s SOD starts unfolding three degrees higher than *P. sulfincola*'s, showing that the stability of *A. pompejana*'s SOD at environmental temperatures (probably around 40°C) is likely greater. However, the conclusions of the species' temperatures derived from the T_d are essentially the same as derived from the T_m.

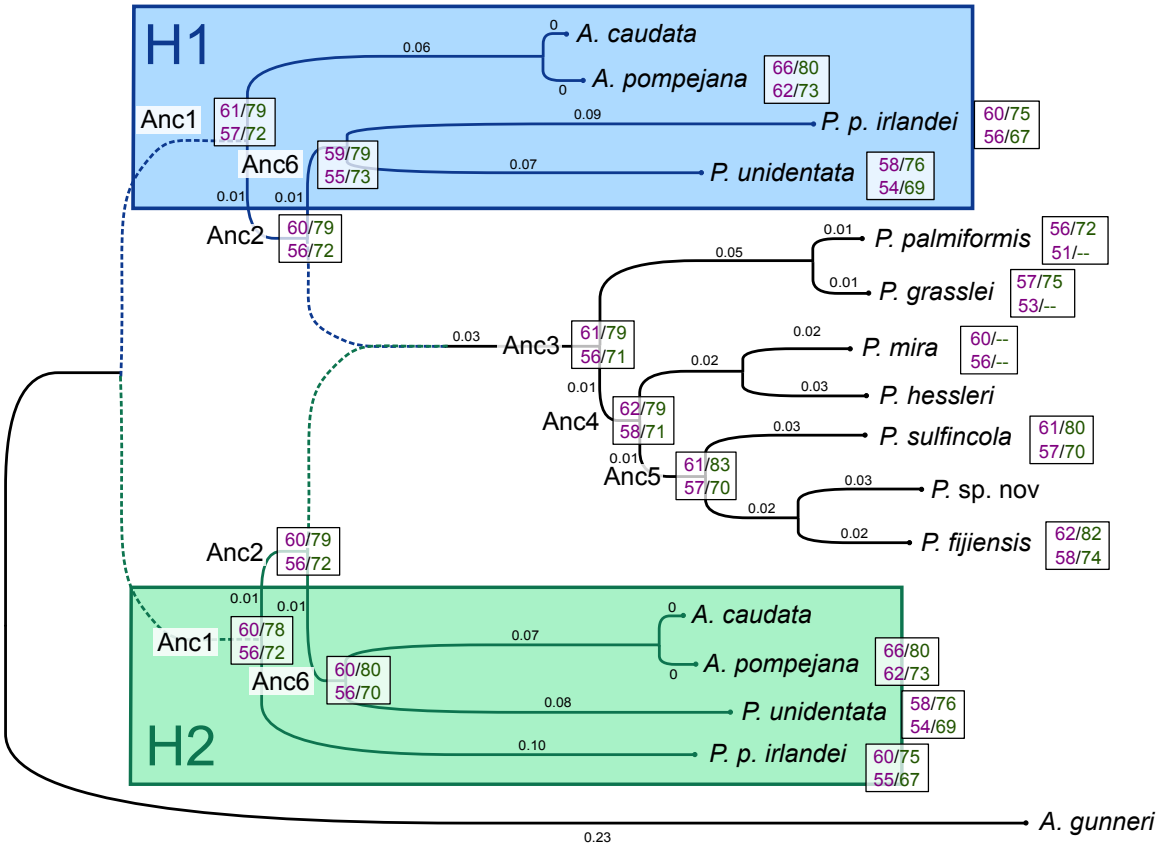


FIGURE III.2 – Denaturation temperature for contemporary and ancestral proteins in the two phylogenetic hypotheses. Each box indicate on the first line the half-denaturation temperature for the cMDH/SOD. On the second line, the temperature corresponding to beginning of the denaturation of the cMDH/SOD is reported. Branches are also labelled with the mean number of mutations per residue on whole species' transcriptomes according to BRUN et al., 2023. The tree is rooted with the outgroup ampharetid species *A. gunneri*.

Finally, ancestral proteins are ML reconstructions which bear uncertainty over several residues. Thus we tried to estimate the range of stability of the different potential ancestral reconstruction that were not experimentally tested. Alternative sequences for the proteins of the LCA of the Alvinellidae (Anc1-H1 and Anc1-H2) were drawn from the posterior distribution of the sequences, and their ΔG were simulated with FoldX. To ensure that FoldX was able to correctly simulate ΔG, we compared FoldX's results with experimental measures obtained with the expressed proteins. The temperature taken to calculate ΔG from the measured parameters was chosen in order to obtain the best prediction of true stability from FoldX's simulations. Figure III.3 panel MDH-H1 shows the difference in free energy of the different MDHs relative to Anc1-H1, experimental data ($\Delta\Delta G_{exp}$) *versus* that predicted by FoldX ($\Delta\Delta G_{foldx}$). We obtained a good correlation between the experimental results and the FoldX predictions, with the exception of the MDH of *P. fijiensis*. Without this last MDH measure, we obtained a $R^2 = 0.85$ and the correlation $\Delta G_{exp} = 0.21 \times \Delta G_{foldx}$ at 56 °C. Using Anc1-H2 as reference (Fig III.3 panel MDH-H2) and without considering the MDHs of *P. fijiensis* and *P. unidentata* we obtained a $R^2 = 0.96$ and the correlation $\Delta G_{exp} = 0.19 \times \Delta G_{foldx}$ at 56 °C. Then, as described in Materials and Methods we used these two correlations to predict the stability of 500 different alternative MDH sequences for Anc1-H1 and Anc1-H2, sampled from the ancestral sequence's probability distributions. The distribution of these stabilities for the cMDH of the two potential LCA of the Alvinellidae is shown Fig III.3 panel MDH. In the H1 hypothesis these stabilities were mostly comparable to warm species such as *P. fijiensis* and *P. mira*. It should be noted that some variants obtained under this hypothesis are much more stable, reaching the stability of *A. pompejana*. In the H2 hypothesis, most values fell within the values obtained from warm species but the tail of the distribution covers the stability of the colder species *P. p. irlandei*.

The same approach was done for SOD (Figure III.3 panel SOD-H1 and SOD-H2). Using Anc1-H1 as reference and excluding the protein of Anc3-H1 and Anc6-H1 whose results with FoldX do not correlate with the experimental ones, we obtained a $R^2 = 0.85$ and the correlation $\Delta G_{exp} = 0.12 \times \Delta G_{foldx}$ at 75 °C. Using Anc1-H2 as reference and excluding the protein of Anc3-H1 and Anc6-H1, we obtained a $R^2 = 0.77$ and the correlation $\Delta G_{exp} = 0.12 \times \Delta G_{foldx}$ at 75 °C. The distribution of stabilities for the SOD of Anc1 under both hypotheses is also comparable to warm species such as *A. pompejana* or *P. sulfincola*. In conclusion, the range distribution of stabilities for marginal reconstructions, for SOD, are overall comprised between -0.5 and +0.5 kcal/mol. For cMDH, this distribution is slightly broader with alternatives displaying a stability close to that of *A. pompejana* or *A. caudata* cMDH. These results confirm our main ecological hypothesis, namely that the SOD and cMDH proteins of the LCA of Alvinellidae have a thermostability closer to that of the contemporary proteins of the warm species. Overall, this leads us to propose that the ancestor of the Alvinellidae was a rather thermophilic organism.

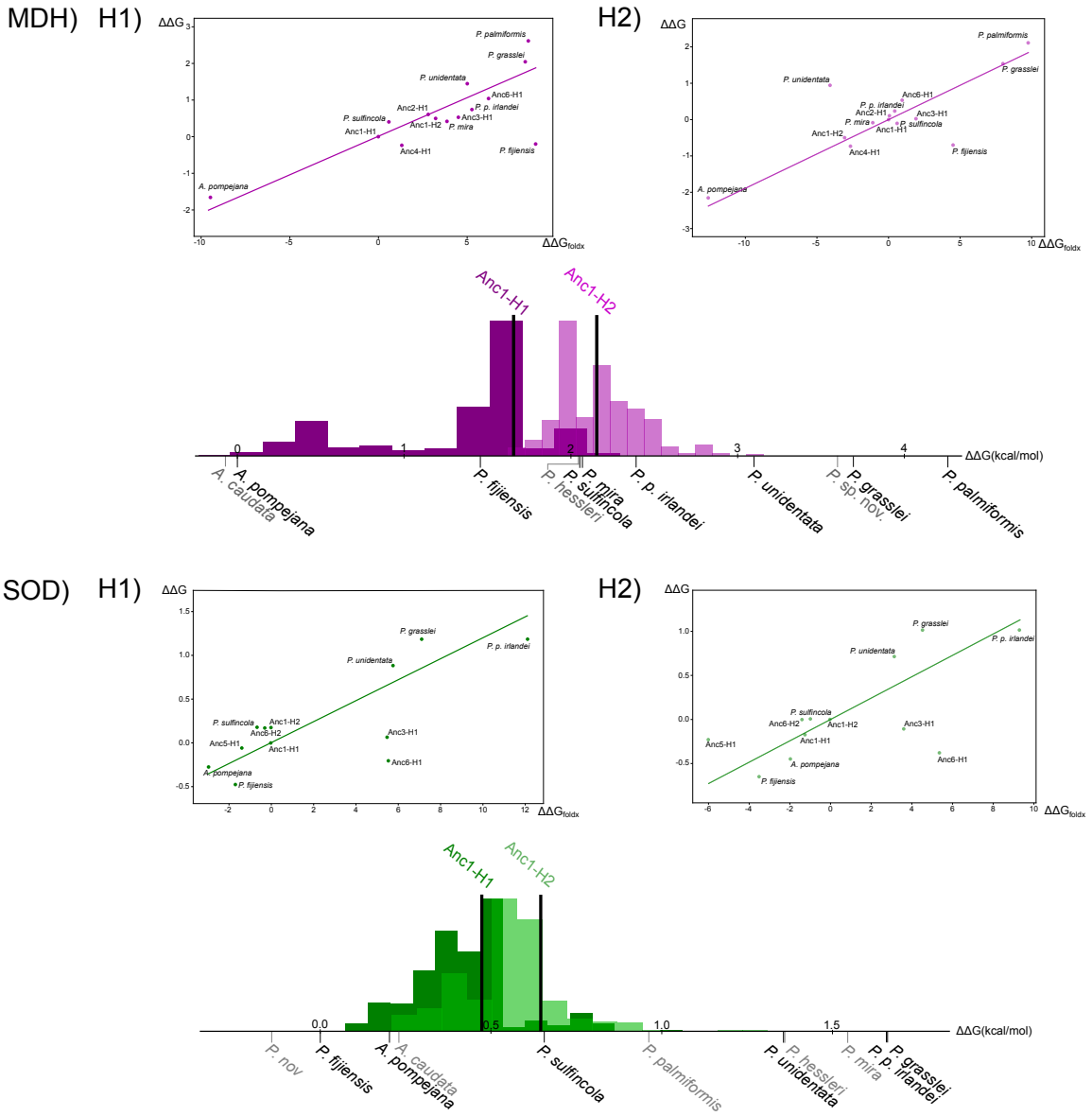


FIGURE III.3 – Thermostability range for the potential ancestral proteins. 500 potential ancestral sequences (1000 for Hb-H2) corresponding to the LCA of Alvinellidae are drawn from the posterior sequence distribution. The stability of the sequences are inferred with FoldX and compared with contemporary proteins of known species. The conversion between the FoldX prediction and the actual protein stability variation is obtained with a linear correlation calibrated with experimentally measured proteins. For contemporary species reported on the $\Delta\Delta G$ axis, measured proteins are in black while simulated ones are in grey. Measured ML sequences for the different ancestors are also reported on the main axis.

III.4 Discussion

III.4.1 Thermotolerance is an ancestral character of the Alvinellidae

III.4.1.0.0.1 Modern species phenotype variability Current Alvinellid worms exhibit a wide range of growing temperatures both within and between species. Species such as *A. pompejana*, *A. caudata*, *P. mira*, *P. hessleri*, *P. sulfincola*, *P.* sp. nov. and *P. fijiensis* are associated with chimney walls and are considered warm species with the building of tubes, whereas species such as *P. p. irlandei*, *P. p. pandorae*, *P. palmiformis*, *P. grasslei*, *P. dela* and *P. bactericola* live further from the fluid emissions in mucus shreads (cocoons) and are considered colder (Desbruyères et Laubier, 1986 ; Jollivet et Stéphane Hourdez, 2020 ; Han et al., 2021). However, most of them are subjected to high variations of temperatures during their life from minute to more longer times. We chose to express the proteins from eight contemporary species to establish whether the difference in environmental temperatures were reflected in the protein stabilities. Considering the melting temperatures of two enzyme families, experimental measures were in good agreement with our expectations, with proteins from *A. pompejana*, *P. sulfincola*, *P. mira* and *P. fijiensis* reaching higher T_m than *P. palmiformis*, *P. grasslei*, *P. p. irlandei* and *P. unidentata*. The difference between the least stable proteins from warm species and the most stable proteins from cold species could however be narrow, especially for the cMDH which is the slower evolving protein between the two studied families (maximum of 23 mutations between *P. p. irlandei* and *P. fijiensis*). This is probably explained by the eurythermal nature of many alvinnelid worms, which can face high burst of temperature in their environment, also for colder species. *P. palmiformis* or *P. grasslei*, despite being the coldest species in our sample according to the stability of their proteins, have been measured to withstand upper thermal limits (UTL) between 30°C and 35°C, way above their comfort temperatures around 10-15°C (Girguis et Lee, 2006 ; Cottin et al., 2008). This is lower than the accepted UTL of warmer species such as *P. sulfincola* or *A. pompejana*, reaching 50°C, and in the case of *A. pompejana* that show temperature-induced stress under 20°C (Girguis et Lee, 2006 ; Ravaux et al., 2013). Thus, the true temperature at which these animals can thrive is difficult to evaluate *in situ* with some probable overlap between the extreme living conditions of warm and cold species. Nevertheless, we were able to distinguish two groups of thermal stabilities from enzyme denaturation measures. Our result is in agreement with Jollivet, Desbruyères et al., 1995, who showed that the residual activities of several proteins (aspartate aminotransferase, glucose-6-phosphate isomerase, and phosphoglucomutase) of species living under warm conditions (*A. pompejana*, *A. caudata*, *P. sulfincola*, *P. hessleri*) were functionnally less sensitive to temperature than proteins from colder species (*P. palmiformis*, *P. p. irlandei* and *P. grasslei*). In another study, Rinke et Lee, 2009 showed that the activity of lactate dehydrogenase, alanopine dehydrogenase, strombine dehydrogenase and citrate synthase is maintained up to 50 to 55°C in *P. sulfincola*, whereas it decreased earlier in *P. palmiformis*, between 35 and 50°C. These studies and our results prove that inference of environmental optimum temperature is achievable from protein denaturation or activity assays in Alvinellidae. In particular, the proteins chosen for this study, taken together, appear as good proxys to distinguish between warmer and colder species even if the T_m differential is narrow.

III.4.1.0.0.2 Evolution of thermotolerance in the Alvinellidae Reconstruction of ancestral cMDH and Cu/Zn SOD suggest that the last common ancestor of the Alvinellidae was already a thermotolerant species, comparable to current species such as *P. mira* or *P. sulfincola* (see fig. III.3). This holds true under H1 or H2 hypotheses of phylogenetic reconstruction of the worm family, and the thermotolerance appears as a shared trait between all alvinellid ancestors, except potentially Anc6 under the H1 hypothesis (cMDH T_m at 59°C). The adaptation toward colder temperatures was gained independently in the branch leading to *P. grasslei* and *P. palmiformis*, as well as the branches leading to *P. p. irlandei* and *P. unidentata*. This could also explain why some proteins from colder species are still fairly stable, if we consider that they may still retain some heritage of past thermostability, and may have not reached their thermodynamic equilibrium yet, which would be mostly driven by evolutionary drift. This is in agreement with results from Fontanillas et al., 2017, who proposed, based on the *in silico* study of six alvinellid transcriptomes, that the ancestor of the family was a thermotolerant species, considering the amino-acid composition of 423 ancestor's loci as well as the higher mutation rate in "cold" branches, interpreted as a slight selection relaxation toward colder species. It should also be noted that according to Brun et al., 2023, the radiation of Alvinellidae was a fast event during which the diverging worm populations, distinguished in the phylogeny by Anc1, Anc2 and Anc6, coexisted and likely experienced high level of intraspecific introgression. As such, the alleles of the cMDH and Cu/Zn SOD must fall within one of the tested tree topologies, but the phenotypes of the corresponding ancestor were close to one another, as they must have shared similar environments.

At last, the results obtained for Anc3 and Anc4 may be affected by the addition of *P. dela* and *P. bactericola*. Indeed, these two species are missing in our study due to the fact that these two rare species have not been yet sampled for sequencing. They are considered living in cold environments, and are expected to be sister-species to *P. mira* and *P. hessleri* based on morphological features (Jollivet et Stéphane Hourdez, 2020 ; Han et al., 2021). The addition of the sequences of these species in the model could therefore potentially influence the result obtained for Anc4, and Anc3 to a lower extent. However, a strong influence of these species on the result of Anc1 is very unlikely, as this ancestor is much further from *P. dela* and *P. bactericola* in terms of numbers of nodes in the tree.

III.4.1.0.0.3 Thermal denaturation kinetics of contemporary and ancestral proteins One important point to consider is that T_m is generally taken as a proxy to describe the protein's stability, as it is a convenient and generally good approximation of the protein unfolding, but this relationship does not necessarily holds if T_m is not strongly correlated with the true free energy difference, ΔG^o, between the native and denatured states at lower temperatures Razban, 2019. Indeed, because T_m is the temperature where half of the proteins are denatured and non functional, it is much higher than the maximal environmental temperature of the animals and does not contain by itself all of the thermodynamic information associated with the denaturation process. For this reason, the temperatures T_d where the proteins start unfolding are presented in figure III.2. T_m is overall well correlated with T_d, but it is not the case for a few alvinellid proteins. The SOD of *A. pompejana* has a higher T_d than the SOD of *P. sulfincola* or Anc5, despite similar or lower T_m, meaning that its transition between the native and unfolded state is steeper due to its higher enthalpy

(see III.2). This implies that the Cu/Zn SOD of *A. pompejana* is actually stable on a wider range of temperatures when compared with other species and ancestors. This difference of effective stability should be even greater at the lower temperatures at which the animal thrives, between 40 and 50°C (Ravaux et al., 2013). In theory, all the proteins of the same organism should be stable for the same range of temperature and we should be able to correlate the stabilities coming from different protein families, assuming that the protein buffers affects the stabilities in an analogous way. This would however require more proteins families to be studied. This approach could be useful to assess the environmental temperatures of current species, and to control how well the reconstructed proteins of the same ancestors, which are independently statistically assessed, agree with each other.

III.4.2 Confidence in the stability of the ancestral proteins

III.4.2.0.0.1 Hypotheses regarding the stabilities of proteins Temperature is probably the environmental factor that has the strongest influence on the physiological adaptations of the alvinellid species. The main assumption in this article is that environmental temperatures are directly mirrored by the relative stabilities of contemporary and ancestral proteins of the Alvinellidae. Yet, aside temperature, protein marginal stability is thought to mostly depend on the protein's abundance inside the cell, species population size (Goldstein, 2011 ; Serohijos et Eugene I Shakhnovich, 2014) and function (Matsumura, Yasumura et Aiba, 1986 ; Tokuriki et al., 2008). Thus this correlation only holds if these parameters remained fairly constant during a considered evolutionary period of time. These confounding effects are well illustrated by the increase of hemoglobin stabilities and solubilities in diving mammals, likely to prevent pigment aggregation as the concentration of intracellular hemoglobins increased along with diving capacity (Mirceta et al., 2013 ; Holm, Dasmeh et Kepp, 2016). In this study, special care was taken to select proteins which function and abundance are not expected to have varied over the last 100 millions years in Alvinellidae. As temperature is supposed to affect the whole proteome in the same way, we therefore preferred to express several ancestral proteins in different enzyme families, to observe if the conclusions drawn from at least two families were concordant, in contrary to several studies relying on ASR that express one protein of interest with potential sequence alternatives. For the two families, the proteins of the LCA were in the high range of observed stabilities, which gives good confidence that the ancestor of Alvinellidae was a thermotolerant species.

III.4.2.0.0.2 Systematic biases in sequence reconstructions Another limit to consider is that ASR remains a probabilistic method which relies on many uncertainties, regarding the protein evolutionary history (which can be different from the species phylogenetic tree) or the inference of some ancestral residues which have lost the phylogenetic signal because of their either high mutability or sensitivity to selection (Koshi et Goldstein, 1996 ; Hanson-Smith, Kolaczkowski et J. W. Thornton, 2010 ; Eick et al., 2016). Quoting Williams et al., 2006, "any conclusion drawn from such studies are only as good as the accuracy of the reconstruction method", and special care must be taken in order to quantify the uncertainty of the results drawn from ASR. Several authors have addressed this issue by

considering the properties of reconstructed ancestral protein sequences obtained from several but equally probable trees (Isogai et al., 2018 ; Blanquart, Groussin et al., 2021) and bearing alternative amino acids for ambiguously reconstructed residues (Gaucher, Govindarajan et Ganesh, 2008 ; Hart et al., 2014). ASR was also suspected to produce ancestral proteins that are generally more stable than the true ancestral ones, especially for billion years old proteins due to consensus effects that tend to favor stabilizing residues in ancestral sequences (Trudeau, Kaltenbach et Tawfik, 2016 ; Wheeler et al., 2016). The consensus effect still is an odd explanation to this phenomenon, as ASR relies on a probabilistic model imposed by the gene tree topology and is devoid of any majority-rule. On the contrary, some type of residues can be favored on longer branches due to their expectations at equilibrium in mutation matrices used for the reconstructions. In our study, the reconstructed proteins are relatively recent with a dense sampling of contemporary sequences (Brun et al., 2023), and the phylogenetic signal associated with each residue position of the sequence remained strong in the deepest nodes, as hinted by the high marginal probabilities obtained for the sequences' ancestral residues whatever the realistic hypotheses of species topology used.

A more convincing concern is the use of the ML method to estimate ancestral sequences. Indeed, while ML sequences are in fact the best-point estimate for ancestral sequences, systematically choosing residues with the highest *posterior* probabilities may select residues which are beneficial for the protein and increase its stability. (Williams et al., 2006) showed from simulations that all reconstruction methods (parsimonious, ML or bayesian) led to ancestral sequences that were different from the true ancestral ones, but in the case of a ML inference, $\Delta\Delta G$ between the predicted sequence and the true ancestor was more often biased in favor of more stability. This bias was on average weaker in parsimonious reconstruction, and non-existent for bayesian reconstruction despite these reconstructions bore more falsely-inferred residues. In our study, the reconstructed ML sequences show a lower number of mutations compared to contemporary ones, which could be the consequence of conservative ML reconstruction. To overcome this issue, we decided to simulate batches of alternative sequences drawn from the *posterior* distribution of the sequences in a bayesian way. The stability of the sequences were simulated with FoldX with the 3D models of the ML sequence. Eick et al., 2016 showed that the sampling of sequences from the *posterior* distribution led to a majority of very unlikely sequences with very low probabilities. When the corresponding proteins were experimentally expressed, they could be nonfunctional especially for the deepest nodes, which sounds like an artifact of a naive bayesian sampling. Therefore, the authors suggested an "AltAll" strategy where non-ML states are excluded if their probabilities are under a defined cutoff, set at 0.1 in their study. Although we rely on simulations dealing with protein stability and not function that could cope with nonfunctional proteins, we chose to replicate this strategy by excluding biologically non-relevant alternative proteins, with a lower cutoff set at 0.01.

In our study, FoldX was able to infer with good accuracy the stability of 88% of experimentally characterized proteins. The correlation between $\Delta\Delta G^o$ obtained with FoldX and the measured $\Delta\Delta G^o$ is the best at a temperature close to the mean T_m of the proteins' families, whereas FoldX is supposed to simulate the stability of proteins at 25°C (Schymkowitz et al., 2005). This could be the consequences of uncertainties in the measured thermodynamic parameters that deviate the measured ΔG^o from the true one at temperatures far from

T_m. Notably, the measured ΔC_p for all proteins, although consistent in each family, was surprisingly high (between 3.5 and 7 kcal/mol, when ΔC_p is generally lower than 3 kcal/mol for small proteins Razvi et Scholtz, 2006 ; Pucci, Kwasigroch et Rooman, 2017). Nevertheless, applying a conversion coefficient onto the inferred FoldX stabilities allowed us to to obtain good predictions of the relative stabilities of expressed proteins at temperatures around T_m. Based on this correction, FoldX was likely able to accurately simulate the relative stability of alternative ancestral sequences, given that the number of mutations between these alternative sequences and the experimented ancestral ones was generally low (less than 5 mutations). The stabilities of all alternative proteins for the alvinellid LCA lay in the highest range of the protein thermostability obtained from contemporary species, between -0.5 and +0.5 kcal/mol around the ML estimate. Moreover, the stabilities obtained with the H1 and H2 alternative topologies had a good overlap, as ambiguous residues were generally the same in both cases. Contrarily to the conclusions of Williams et al., 2006, our ML estimate of the LCA proteins did not appear more stable than alternative sequences. This could be due to the influence of contemporary sequences obtained from a series of species with different living temperature that overweight the systematic bias associated with the ML ASR reconstructions. For instance, in III.3 panel MDH, alternative sequences for the MDH under the H1 hypothesis show very high stabilities. These sequences, indeed, borrow more residues similar to the *A. pompejana* sequences than the ML estimate, and are expected to be very stable. Finally, one should keep in mind that the marginal probabilities obtained in ASR are a reflection of the model of construction. No current phylogenetic model can tell what the exact probabilities are. As a consequence, the shape of the distribution of alternative sequences, as shown in III.3, must be taken with caution. On the upside, we can expect these models to accurately distinguish between the almost-certain residues, and the other ones. In our simulations, the range of the distribution covers nowadays thermotolerant species. As such, the thermotolerance of the LCA is assessed with good confidence.

III.4.3 Future Directions

The two protein families used in the ancestral protein reconstructions were chosen as potential proxies for several key environmental parameters : temperature, which is the focus of this article, and the oxidative stress. Hydrothermal vents are indeed reputed as highly oxidative due to the high content of sulfides and metals (Monaco et Prouzet, 2015). The SOD expression and activity are up-regulated by metal tissue-content, such as Cu, in contemporary polychaetes (Rhee et al., 2011 ; Zeinali, Homaei et Kamrani, 2015), and has been reported to be particularly high in *P. grasslei* and *A. pompejana*, especially in gut tissues (Marie et al., 2006 ; Genard et al., 2013). It is however, not known if the high total tissue-activity of the SOD observed for these worms also translates into specific enzymatic properties at the protein-level for hydrothermal species. In alvinellid Cu/Zn SOD sequences, the residues at position 130, 131 and 135 help to distinguish the species *A. pompejana*, *A. caudata*, *P. p. irlandei*, *P. unidentata* from the other *Paralvinella* species. These residues are known to be part of the lower rim of the active site of the enzyme, and shape the strength of the electric field around the catalytic Cu^{2+} (Bordo, Djinovic et Bolognesi, 1994 ; Rhee et al., 2011). However, the study of the enzymatic kinetics of SOD relies on indirect measures, such as the inhibition of Nitroblue Tetrazolium reduction by photoche-

mically produced $O_2^{\cdot -}$ (Beauchamp et Fridovich, 1971). This protocol may be difficult to implement if we want to measure the sensitivity of the catalysis at several temperatures between hydrothermal and non-hydrothermal homologs of the Cu/Zn SOD.

On the contrary, the activity of cMDH can be directly assessed by following the absorbance at 340 nm of NADH, with reduction of oxaloacetate to malate by cMDH. Dahlhoff et Somero, 1991 showed that the Km of cMDH for NADH is inversely related to the upper thermal limit of several marine invertebrates, including *A. pompejana* and *A. caudata*. For this protein, it would be particularly interesting to compare the results of the thermodynamic measurements presented in this article with their enzymatic properties. Indeed, while maximum likelihood ASR methods present a potential bias in reconstructing ancestral proteins that are too much thermostable, the uncertainties associated with the functional characterisation of ancestral proteins do not seem to be systematically biased (Eick et al., 2016 ; Trudeau, Kaltenbach et Tawfik, 2016). Moreover, the cMDH of shallow-water invertebrates were shown to be pressure-sensitive, as it is also the case in fishes, whereas cMDH from vent invertebrates such as *A. pompejana*, *A. caudata* or *Riftia pachyptila* worms, are not (Siebenaller et Somero, 1978 ; Dahlhoff et Somero, 1991). In the sister family Ampharetidae, colonization of the deep-sea hydrothermal vents happened several times independently (Eilertsen et al., 2017), and the common ancestor of Alvinellidae and Ampharetidae, which emerged between 100 and 130 million years ago (Brun et al., 2023) is not expected to be an hydrothermal species. Characterizing the sensitivity of the ancestral alvinellid cMDH to hydrostatic pressure could hint if the ancestor of the Alvinellidae was already a deep-sea species, or inhabited shallow hydrothermal sites that were proposed as potential refuges for vent faunas between 125 and 94 million years ago, during episodes of deep-sea anoxia (Röhl et Ogg, 1996 ; Jacobs et Lindberg, 1998 ; Vrijenhoek, 2013).

Conflicts of interrest

None declared.

III.5 Annexes

Annexe - Alvinellid intracellular hemoglobin

1 introduction

Alvinellid worms exhibit a wide range of physiological and functional adaptations to deal with the harsh vents' conditions, such as four pairs of enlarged gills crossed by intraepidermal vessels in which the coelomic fluid circulates with a reduced diffusion distance to the sea water to cope with the hypoxic conditions (Jouin and Gaill, 1990; Hourdez and Lallier, 2007). They contain two types of hemoglobins (Hb), a circulating extracellular giant one in the vascular system and a non-circulating cytoplasmic one, both having high affinities for oxygen, not affected by sulfide content and weakly affected by temperature Somero et al. (1989); Martineu et al. (1997); Hourdez and Lallier (2007). They also possess a protein arsenal allowing them to thrive under low O_2 condition Mary et al. (2010); Dilly et al. (2012) and high H_2S concentration in body fluids, maintaining aerobic activity similar to related shallow-living species (Hand and Somero, 1983; Martineu et al., 1997). In a large transcriptomic analysis, Gagnière et al. (2010) reported that ta large part of transcripts in *Alvinella pompajena* were related to oxygen homeostasis (respiratory chain proteins and hemoglobins). The intracellular Hb is a small monomeric protein (≈ 14.8 kDa) contained in coelomocytes. These coelomocytes are found in high concentration in a periesophagal pouch, and can circulate between the pouch and the peribuccal tentacles. This Hb in Alvinellida is thought to act as a O_2 reservoir when little oxygen uptake is possible from the environment to supply the head region (Hourdez et al., 2000). Hemoglobins are interesting evolutionary indicators of physiological processes. They were shown to have gained higher intrinsic affinities for oxygen in several independent lineages of species confronted to hypoxic conditions (Bunn, 1981) and structural stability to avoid aggregation in species where they are found in very high concentrations, such as diving mammals (Isogai et al., 2018). In *A. pompejana*, Hourdez et al. (2000) reported that the intracellular hemoglobin has a higher oxygen affinity ($P_{50} = 0.66$ Torr at 20°C, pH 7.5) compared to non-vent species (Hourdez and Jollivet, 2020)). The hemoglobin however is still strongly influenced by temperature and loses its oxygen affinity above 50 °C (Hourdez and Weber, 2005). Others functional characteristics linked to hydrothermal vents have also been suggested for this hemoglobin. In *Paralvinella palmiformis* (but apparently not for *A. pompejana*), the intracellular Hb may have sulfide-binding properties, buffering the harmful redox effects of sulfide exposure (Martineu et al., 1997). Therefore, this hemoglobin could be involved in oxidative stress regulation as well.

2 Material and Methods

2.1 Hemoglobin expressions

Transcripts were retrieved as described in chapter 2. Three intracellular hemoglobin families were identified. These families are named HB1, HB2 and HB3 below. As we could not precisely infer the phylogenetic relationship between the three of them, only two families HB1 and HB2 were kept for ancestral sequence reconstructions. The ancestral proteins of HB1 were expressed and characterised, as described in chapter 3. HB1 contains one tryptophanyl residue. Therefore, denaturation assays by temperature were monitored both by nanoDSF and microDSC.

2.2 Hemoglobin simulations

For structural predictions of HB1 with Swiss Model, we used the Alphafold prediction of the *Arichlidon gathofi* hemoglobin (A0A6M8AUJ7.1.A) as a template. This hemoglobin has a 85-87% identity with HB1. Predictions of HB1 stabilities were performed with FoldX, as described in chapter 3. For HB1 under the H1 tree hypothesis, 500 alternative sequences were drawn from the posterior distribution and simulated. For HB1 under the H2 tree hypothesis, 1000 alternative sequence were drawn, as the number of ambiguous residues in this reconstruction was higher.

3 Results

3.1 Contemporary proteins and ancestral sequence reconstructions

Three families of closely-related transcripts were identified, named HB1 to HB3 below. HB1 and HB2 had transcripts for every alvinellid species in the dataset, while the transcripts for *Paralvinella sulfincola* and *Paralvinella unidentata* were missing in HB3. As we could not resolve with good confidence the relationships between these three families, we chose to discard HB3 for further analysis. Moreover, we could not find homologous sequence for each of the individual families in any closely-related species of the dataset (ampharetid species *Hypania invalida*, *Amphisamytha carldarei*, *Amphicteis gunneri*, mellinid species *Melinna palmata* and terebellid species *Neoamphitrite edwardsi* and Terebellidae gen. sp.). We concluded that HB1, HB2 and HB3 arose from gene duplication specific to the Alvinellidae.

The difference of likelihoods between the most and least probable hemoglobin HB1 tree is 15.62, as shown in figure 1. The difference is small compared to the mean phylogenetic signal per site in the alignment (12.63). HB1 corresponding to the H1 and H2 tree hypothesis were expressed, while uncertainty about tree topology is addressed with simulations.

	Tree likelihood		Sequence confidence
	HB1 Dayhoff+G4	HB2 Dayhoff+G4	HB1
T1	-2826.736 0	-2825.537 0	–
T2	-2820.185 0	-2821.645 0	–
T3	-1827.027 0	-2826.805 0	–
T4	-1820.131 0	-2813.926 0	–
T5	-2811.413 100	-2818.254 0	136 (99%) Anc1-H1 (137, 100%)
T6	-2825.330 0	-2818.235 0	136 (99%) Anc1-H1 (137, 100%)
T7	-2819.500 0	-2821.645 0	133 (97%) Anc1-H2 (136, 99%)
T8	-2825.811 0	-2805.787 100	–
T9	-2825.602 0	-2826.676 0	132 (97%) Anc1-H2 (137, 100%)
T10	-2827.027 0	-2827.621 0	–
T11	-2827.027 0	-2828.177 0	–
T12	-2826.810 0	-2828.638 0	–
T13	-2826.965 0	-2827.316 0	–
T14	-2826.730 0	-2826.676 0	–
T15	-2826.732 0	-2828.638 0	–

Table 1 – Hypothesis testing for the ancestral reconstructions. HB1 and HB2 correspond to the two hemoglobin families used for ancestral reconstructions. Left table : potential tree topologies likelihood for each multiple sequence alignment under the specified evolutionary model. The approximate posterior probability of each topologies is reported below the likelihood. Candidate topologies are taken from Brun and others (2023). Right table : confidence in the reconstruction of the ancestral sequence corresponding to the LCA of the family for the most probable topologies, as well as T6, T7 and T9 which are likely gene topologies according to Brun and others (2023). Number of expected correct residues and sequence percentage are indicated, as well as the closest experimentally expressed sequence (number of identical residues and percentage of identity with the ML sequence).

3.2 Experimental characterisation of the ancestral proteins

In the case of HB1, a clear denaturation transition could not be observed for some proteins with nanoDSF monitoring. The signal in microDSC was also low, likely due to protein degradation. A few T_m could still be measured, ranging between 51.3°C (*P. palmiformis*) and 67.0°C (Anc5-H1 and H2). For contemporary species, the only warm species measured was *Paralvinella mira*, with a T_m at 64°C. Cold species (*Paralvinella pandorae*

irlandei, *P. unidentata*, *Paralvinella grasslei*, *P. palmiformis*) had T_m between 51 and 60°C. The ancestor of the family had a T_m of 65°C under H1 hypothesis and 63°C under H2 hypothesis. The T_m of measured proteins are reported on figure 1, and associated thermodynamic parameters are given in table 2.

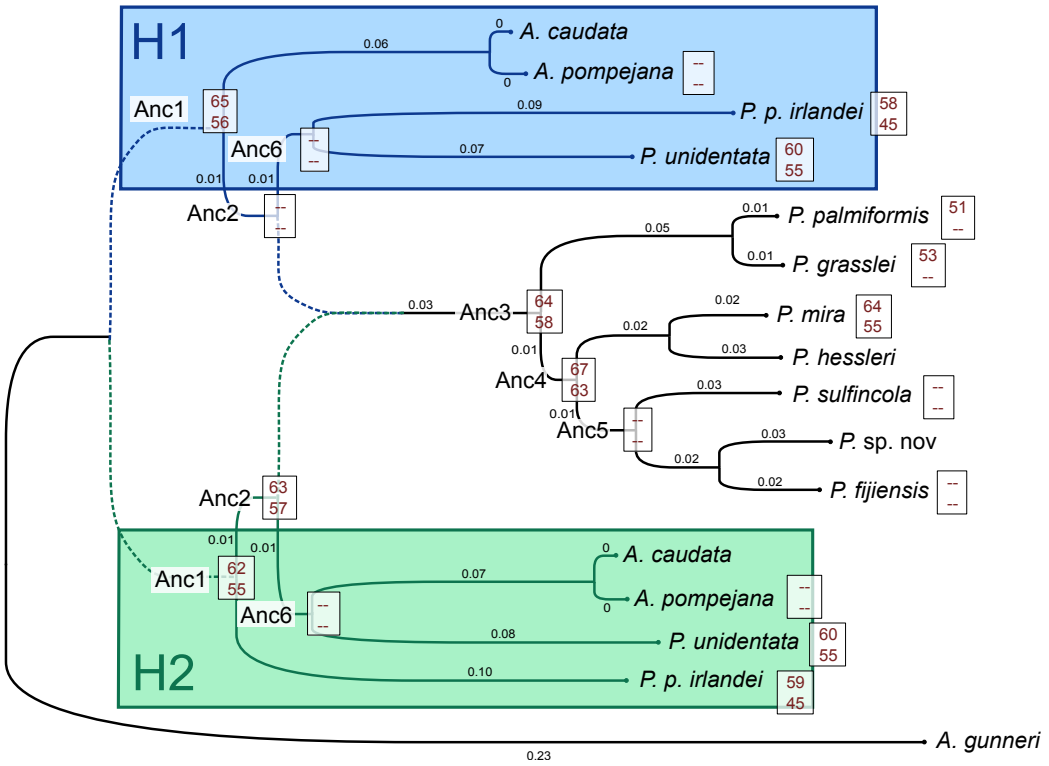


FIGURE 1 – Denaturation temperature for contemporary and ancestral proteins in the two phylogenetic hypothesis. Each box indicate on the first line the half-denaturation temperature for HB1. On the second line, the temperature corresponding to beginning of the denaturation (5% protein unfolded) for HB1 is reported. Branches are also labelled with the mean number of mutation per amino-acid on whole species' transcriptomes according to Brun and others (2023). The tree is rooted with the outgroup ampharetid species *A. gunneri*.

3.3 Assessing the thermal phenotype of the Alvinellid ancestors

FoldX failed to simulated the protein of Anc4-H1 from the Anc1-H1 3D model. For other proteins under H1 hypothesis, we obtained a good correlation between experimental measures and FoldX predictions ($R^2 = 0.86$, $\Delta G_{exp} = 0.24 \times \Delta G_{foldx}$ at 64 °C). FoldX failed to simulated the protein of Anc4-H1 from the Anc1-H2 3D model. For other proteins, we obtained a good correlation between measures and simulations ($R^2 = 0.79$, $\Delta G_{exp} = 0.24 \times \Delta G_{foldx}$ at 64 °C). As the alternative ancestral proteins drawn from the posterior sequence distribution are on average very close to the protein on which each FoldX model was built, we expect the results of FoldX's simulations to be very accurate to estimate variants' stabilities. The stabilities for the Hb of Anc1 were comparable to warm species (*P. mira*). Unfortunately, we could only measure a limited number of contemporary hemoglobins. Therefore, we simulated the stabilities of remaining ones. The hemoglobins of *P. grasslei* and *P. palmiformis* were simulated as being particularly stable, but this is not in agreement with the T_m obtained for these two proteins (see Fig. 2). At 64 °C, we should be above the melting temperature of these proteins (which are very close to one another with only one mutation), which were experimentally less stable than the Ancestors' proteins in this temperature range. Surprinsingly, we also simulated the stabilities of the second hemoglobins' family HB2 and most of these proteins were particularly unstable according to FoldX, even for warm species, which could hint that the two families have very different physiological implications.

	ΔHm (kJ/mol)	ΔCp (kJ/mol/K)	Tm (°C)
P. unidentata	563 ± 22	30.2 ± 6.2	60.2 ± 0.3
P. p. irlandei	392 ± 6	29.9 ± 4.2	57.9± 0.0
P. grasslei	–	–	52.8
P. palmiformis	–	–	51.3
A. pompejana	–	–	–
P. fijiensis	–	–	–
P. sulfincola	–	–	–
P. mira	412 ± 8	22.3 ± 3.5	63.9 ± 0.2
Anc1-H1	410 ± 6	21.4 ± 1.5	64.8 ± 0.2
Anc2-H1	–	–	–
Anc3-H1	472 ± 10	23.7 ± 2.1	64.8 ± 0.4
Anc4-H1	752 ± 19	53.1 ± 0.6	67.0 ± 0.2
Anc5-H1	–	–	–
Anc6-H1	–	–	–
Anc1-H2	476 ± 7	24.0 ± 1.0	62.5 ± 0.0
Anc2-H2	530 ± 17	22.8 ± 5.1	62.7 ± 0.0
Anc3-H2	472 ± 10	23.9 ± 2.1	64.8 ± 0.4
Anc4-H2	752 ± 19	53.1 ± 0.6	67.0 ± 0.2
Anc5-H2	–	–	–
Anc6-H2	–	–	–

Table 2 – Thermodynamic parameters measures for the unfolding of different proteins. ΔH, ΔC_p and T_m are the enthalpy variation, heat capacity variation and half-denaturation temperature.

4 discussion

Several proteins could not be measured by nanoDSF as the fluorescence signal showed very low variation on the temperature range. The Anc4-H1 protein, despite having a clear fluorescence signal, displays surprisingly high ΔH_m and ΔC_p compared to other hemoglobins of the family, and its result should be taken with caution. The T_m on the contrary is established with better confidence. From DSC measurements, it appears that several hemoglobins may have multiple unfolding transitions. In this case, the measures obtained by nanoDSF would not be appropriate. Because DSC measurements require much more protein to be carried, we were not able to obtain signals of suitable quality with this technique for the hemoglobin. In the future, the characterization of the whole family should be performed again by DSC.

The preliminary results obtained for the hemoglobin indicate that the proteins of the ancestor were as stable as the contemporary warm species *P. mira*. This would be in agreement with the results obtained for the SOD Cu/Zn and cMDH in chapter 2, indicating that the ancestor of the family was a warm species as well. It is worth noting that these *in vitro* measures may not fully reflect the *in vitro* stabilities, as discrepencies between stabilities of hemoglobin in intact red cells or in buffered solution are known, likely due to cytosolic osmolytes that may stabilize proteins of deep-sea animals (Hourdez and Weber, 2005).

We also identified three distinct families of intracellular hemoglobins in Alvinellidae, likely due to gene duplications at the birth of the lineage. Hemoglobins could provide valuable informations on the current and past environments of these species. For example, hemoglobins of deep-sea invertebrate and fish species are less affected by high pressure, compared to animals living at lower pressures (Hourdez and Weber, 2005). They also have brought insights in the evolution of diving mammals, considering the increased net charge on the protein's surface that is associated with higher hemoglobin concentrations in these animals as their diving capacity improves (Mirceta et al., 2013; Holm et al., 2016).

In *A. pompejana*, hemoglobins were reported to have high affinities for oxygen to cope with hypoxic environments between 30 and 40°C associated with a strong Bohr effect Hourdez et al. (2000); Hourdez and Jollivet (2020). Martineu et al. (1997) also reported that intracellular hemoglobins of *P. palmiformis* (but not for *A. pompejana* according to the study) have affinity for sulfides, which could be a gain of function for sulfide-detoxification in the blood. Indeed, globins as a whole show a range of functions other than oxygen transport, such as catalysis of redox reactions or NO detoxification (Hardison, 2012). The study of the three alvinellid

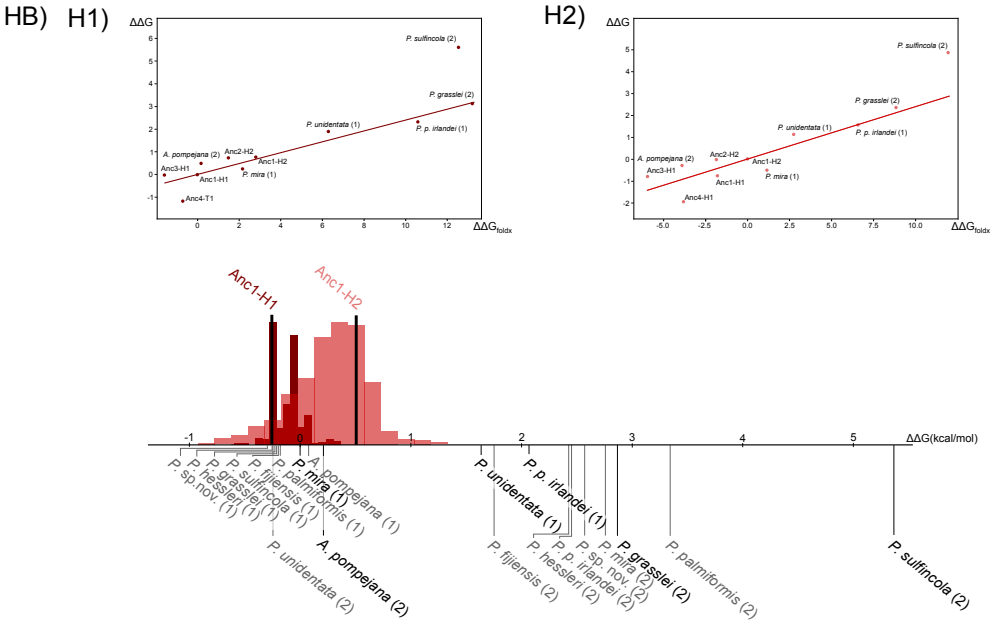


FIGURE 2 – Stability range for the potential ancestral proteins. 500 potential ancestral sequences (1000 for Hb-H2) corresponding to the LCA of the Alvinellidae are drawn from the posterior sequence distribution. The stability of the sequences are inferred with FoldX and compared to contemporary proteins of known species. The conversion between the FoldX prediction and the actual protein stability variation is obtained with a linear correlation calibrated with experimentally measured proteins. For contemporary species reported on the $\Delta\Delta G$ axis, measured proteins are in black while simulated ones are in grey. Proteins from the sub-family HB1 are labelled (1), while proteins from the subfamily HB2 are labelled (2). The evaluated ancestors for the hemoglobins correspond the sub-family one. Measured ML sequences for the different ancestors are also reported on the main axis.

hemoglobins families, which differ by a few conserved amino acids especially between the positions 66-68 and 94-107, would be particularly interesting to characterize potential gains or losses of functions after gene duplications, as well as parallel evolution. Moreover, even though the evolutionary relationships between the families should be elucidate first, the addition of closely-related sequences diverging *prior* to the last common ancestor of the Alvinellidae strengthens the confidence of ancestral reconstruction for this deep ancestor.

Whether these hemoglobins families have different biological roles is currently unknown. From FoldX simulations, it is likely that at least two families have very different range of temperature stabilities. In Fish, hemoglobin multiplicity is known in Notothenioidei facing variable temperatures (Di Prisco et al., 2007). Alvinellid worms are also exposed to high range of temperature shifts, due to sudden hot fluid emissions in vents (Jollivet and Hourdez, 2021), and during their life cycle. In *A. pompejana*, larvae are only able to grow between 10°C and 20°C, showing that the larvae develop outside the temperature ranges of adult colonies (Pradillon et al., 2005). Different hemoglobin expression patterns are already described in mammals between fetal and adult stages (Storz et al., 2011), although this hypothesis is less likely in our case as the transcripts used to obtain the contemporary hemoglobin sequences were harvested from adult specimens (with the exception of *A. pompejana*, for which the hemoglobins were predicted from an assembled genome). The study of regulatory elements for these hemoglobins in *A. pompejana* could be a first step in understanding the role of these proteins. Alvinellid hemoglobins' evolution could thus provide a rich model to explore further.

Références

Bunn, H. (1981). Evolution of mammalian hemoglobin function. *Blood*, 58(2) :189–197.

Di Prisco, G., Eastman, J. T., Giordano, D., Parisi, E., and Verde, C. (2007). Biogeography and adaptation of Notothenioid fish : Hemoglobin function and globin–gene evolution. *Gene*, 398(1-2) :143–155.

Dilly, G. F., Young, C. R., Lane, W. S., Pangilinan, J., and Girguis, P. R. (2012). Exploring the Limit of Metazoan Thermal Tolerance via Comparative Proteomics : Thermally Induced Changes in Protein Abundance by Two Hydrothermal Vent Polychaetes. *Proceedings of the Royal Society B : Biological Sciences*, 279(1741) :3347–3356.

Gagnière, N., Jollivet, D., Boutet, I., Brélivet, Y., Busso, D., Da Silva, C., Gaill, F., Higuet, D., Hourdez, S., Knoops, B., Lallier, F., Leize-Wagner, E., Mary, J., Moras, D., Perrodou, E., Rees, J.-F., Segurens, B., Shillito, B., Tanguy, A., Thierry, J.-C., Weissenbach, J., Wincker, P., Zal, F., Poch, O., and Lecompte, O. (2010). Insights into Metazoan Evolution from *Alvinella pompejana* cDNAs. *BMC Genomics*, 11(1) :634.

Hand, S. C. and Somero, G. N. (1983). Energy Metabolism Pathways of Hydrothermal Vent Animals : Adaptations to a Food-rich and Sulfide-rich Deep-sea Environment. *The Biological Bulletin*, 165(1) :167–181.

Hardison, R. C. (2012). Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor Perspectives in Medicine*, 2(12) :a011627–a011627.

Holm, J., Dasmeh, P., and Kepp, K. P. (2016). Tracking evolution of myoglobin stability in cetaceans using experimentally calibrated computational methods that account for generic protein relaxation. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1864(7) :825–834.

Hourdez, S. and Jollivet, D. (2020). Metazoan Adaptation to Deep-Sea Hydrothermal Vents. In di Prisco, G., Edwards, H. G. M., Elster, J., and Huiskes, A. H. L., editors, *Life in Extreme Environments*, pages 42–67. Cambridge University Press, 1 edition.

Hourdez, S. and Lallier, F. H. (2007). Adaptations to Hypoxia in Hydrothermal-Vent and Cold-Seep Invertebrates. *Reviews in Environmental Science and Bio/Technology*, 6(1-3) :143–159.

Hourdez, S., Lallier, F. H., De Cian, M., Green, B. N., Weber, R. E., and Toulmond, A. (2000). Gas Transfer System in *Alvinella pompejana* (Annelida Polychaeta, Terebellida) : Functional Properties of Intracellular and Extracellular Hemoglobins. *Physiological and Biochemical Zoology*, 73(3) :365–373.

Hourdez, S. and Weber, R. (2005). Molecular and functional adaptations in deep-sea hemoglobins. *Journal of Inorganic Biochemistry*, 99(1) :130–141.

Isogai, Y., Imamura, H., Nakae, S., Sumi, T., Takahashi, K.-i., Nakagawa, T., Tsuneshige, A., and Shirai, T. (2018). Tracing Whale Myoglobin Evolution by Resurrecting Ancient Proteins. *Scientific Reports*, 8(1) :16883.

Jollivet, D. and Hourdez, S. (2021). 7.7.4 Alvinellidae Desbruyères & Laubier, 1986. *Pleistoannelida, Sedentaria III and Errantia I*, 3 :18.

Jouin, C. and Gaill, F. (1990). Gills of Hydrothermal Vent Annelids : Structure, Ultrastructure and Functional Implications in Two Alvinellid Species. *Progress in Oceanography*, 24(1-4) :59–69.

Martineu, P., Juniper, S., Fisher, C., and Massoth, G. (1997). Sulfide Binding in the Body Fluids of Hydrothermal Vent Alvinellid Polychaetes. *Physiological Zoology*, 70(5) :578–588.

Mary, J., Rogniaux, H., Rees, J.-F., and Zal, F. (2010). Response of *Alvinella pompejana* to Variable Oxygen Stress : A Proteomic Approach. *Proteomics*, 10(12) :2250–2258.

Mirceta, S., Signore, A. V., Burns, J. M., Cossins, A. R., Campbell, K. L., and Berenbrink, M. (2013). Evolution of Mammalian Diving Capacity Traced by Myoglobin Net Surface Charge. *Science*, 340(6138) :1234192.

Pradillon, F., Le Bris, N., Shillito, B., Young, C. M., and Gaill, F. (2005). Influence of Environmental Conditions on Early Development of the Hydrothermal Vent Polychaete *Alvinella pompejana*. *Journal of Experimental Biology*, 208(8) :1551–1561.

Somero, G. N., Childress, J. J., and Anderson, A. E. (1989). Transport, Metabolism, and Detoxification of Hydrogen Sulfide in Animals from Sulfide-Rich Marine Environments. *Critical Reviews in Aquatic Sciences*, 1(4) :591–614.

Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2011). Phylogenetic diversification of the globin gene superfamily in chordates. *IUBMB Life*, 63(5) :313–322.

Chapitre IV

Reconstruction de séquences ancestrales

IV.1 Inférence des résidus d'acides aminés ancestraux

IV.1.1 Introduction

La reconstruction de séquences ancestrales (ASR) est une méthode probabiliste visant à inférer la séquence ancestrale à un noeud de l'arbre phylogénétique à partir d'un ensemble de séquences homologues observées dans le temps présent. L'idée a été initialement suggérée par Linus Pauling et Emile Zuckerkandl qui proposaient que l'observation des séquences contemporaines d'hémoglobines de vertébrés permettrait d'estimer la séquence commune la plus probable de l'ancêtre (Pauling et al., 1963). Cette approche permettrait non seulement de reconstruire la séquence ancestrale, mais également de déterminer dans quelles lignées ancestrales (branches internes de l'arbre) sont apparues les mutations observées dans les séquences contemporaines. Les auteurs projetaient que la synthèse des protéines reconstruites permettrait d'étudier leur fonctionnement passé, en donnant l'exemple de l'affinité de l'hémoglobine pour l'oxygène à différents pH, et également que cette méthode permettrait d'obtenir des informations importantes sur la biologie des ancêtres des organismes ne laissant pas ou peu de traces fossiles comme les animaux à corps mou, ce qui n'est pas sans rappeler le travail que nous avons mené au chapitre 2. La méthode d'ASR a été développée par la suite en se basant sur l'amélioration des méthodes d'inférence de phylogénie moléculaire, notamment par Koshi et Goldstein, 1996 qui ont proposé un algorithme de calcul des probabilités marginales des résidus ancestraux selon le critère de maximum de vraisemblance. L'ASR a trouvé de multiples applications en recherche. De façon évidente, la méthode a été utilisée pour reconstruire des protéines attribuées à des ancêtres communs parfois très anciens, jusqu'à 3,5 milliards d'années, afin de mesurer leurs conditions de fonctionnement et ainsi d'en déduire leurs conditions environnementales (Gaucher, Govindarajan et Ganesh, 2008 ; Hart et al., 2014 ; Blanquart, Groussin et al., 2021). D'un point de vue biochimique et fonctionnel, cette méthode permet également d'identifier

spécifiquement les mutations responsables de changement de repliement ou de fonction au sein d'une protéine, alors que l'alignement initial de protéines homologues ne permet pas forcément de suivre les effets synergiques des mutations qui se sont aléatoirement accumulées au cours du temps et qui peuvent affecter une protéine (Merkl et Sterner, 2016). Cette méthode a par exemple permis d'identifier les résidus responsables de la préférence de substrats entre deux protéines fortement apparentées que sont la Lactate déshydrogénase et la Malate déshydrogénase (Brochier-Armanet et Madern, 2021). L'ASR a également été utilisée dans le cadre de l'optimisation de protéines pour des applications industrielles, en permettant d'isoler des résidus d'importance particulière et d'effectuer de la mutagénèse au sein de protéines candidates, par exemple pour en augmenter la thermostabilité (Romero-Romero et al., 2016). Enfin, les méthodes développées pour l'ASR peuvent en soi trouver des applications dans des domaines plus éloignés comme retracer l'apparition d'un mutant dans des cas d'épidémiologie (Ishikawa et al., 2019), puisque fondamentalement la méthode est généralisable à tout graphe dont les noeuds sont liés entre eux par certaines lois de probabilités.

Les méthodes développées pour l'ASR découlent directement des modèles développés en phylogénie moléculaire. Elles nécessitent les mêmes éléments, à savoir un ensemble de séquences contemporaines alignées afin d'identifier les sites homologues qui partagent une histoire évolutive commune, un arbre phylogénétique qui spécifie les relations évolutives entre les séquences et les temps de divergence qui les séparent, ainsi qu'un modèle d'évolution qui tente d'expliquer de manière probabiliste comment un certain résidu change d'état au cours du temps (Merkl et Sterner, 2016). Les modèles évolutifs les plus usuellement utilisés font plusieurs hypothèses simplificatrices, potentiellement critiquables notamment lorsque l'on s'intéresse à l'évolution de séquences codantes ou du moins fonctionnelles comme dans notre cas. Ainsi, les sites de l'alignement sont considérés comme évoluant indépendamment les uns des autres (Yang, 2007) et ces sites évoluent de façon homogène selon une même loi de probabilité, décrite par une matrice de mutations instantanées, par exemple la matrice LG (S. Q. Le et O. Gascuel, 2008). Le processus évolutif est aussi considéré homogène dans le temps (les probabilités de mutation instantanées sont constantes le long des branches de l'arbre phylogénétique, bien que cette hypothèse soit couramment partiellement assouplie par l'introduction d'une loi de distribution Gamma censée accommoder la différence dans les taux d'évolution entre des sites sous différentes contraintes (Yang, 1994)), stationnaire (les fréquences attendues de chaque résidu sont constantes à chaque point de l'arbre phylogénétique) et, réversible, donc non directionnelle au cours de l'évolution (Naser-Khdour et al., 2019). Ces hypothèses ne sont évidemment pas réalistes, puisqu'elles prédisent que la séquence d'une protéine tendrait à évoluer vers un état homogène découlant d'un processus strictement stochastique. Néanmoins, elles permettent en pratique d'obtenir des résultats satisfaisants tout en simplifiant grandement les calculs de vraisemblance qui augmentent fortement avec le nombre de séquences étudiées et leur longueur. Ces modèles permettent facilement d'obtenir des probabilités pour chaque résidu à chaque noeud de l'arbre. Les deux méthodes de calcul les plus utilisées correspondent soit à une reconstruction jointe, auquel cas l'on cherche à estimer conjointement l'ensemble des résidus ancestraux à une position de l'alignement qui maximise la probabilité d'obtenir les résidus présents dans les séquences contemporaines à ce site (en cherchant le chemin évolutif le plus probable), soit une reconstruction marginale où la probabilité des résidus aux différents noeuds est obtenue en sommant tous les chemins évolutifs possibles. La recons-

truction jointe est par conséquent une approximation rapide de la reconstruction marginale. En outre, la reconstruction marginale a l'avantage d'exprimer une probabilité permettant d'évaluer la confiance que l'on donne à la reconstruction (Ishikawa et al., 2019), bien qu'il faille garder à l'esprit que la probabilité obtenue dépend du modèle évolutif utilisé, de l'arbre phylogénétique proposé et des hypothèses formulées, donc cette probabilité n'est pas une retranscription fidèle de la réalité biologique.

Il est notamment souvent suggéré que les méthodes d'ASR aboutissent à l'inférence de protéines ancestrales particulièrement thermostables, ce qui serait lié à un biais inhérent à la méthode (Trudeau, Kaltenbach et Tawfik, 2016). On peut immédiatement s'interroger sur l'effet des hypothèses usuelles de phylogénie lorsqu'on veut inférer des séquences ancestrales, qui est un cadre différent d'une phylogénie classique où l'on s'appuie sur la somme de l'information obtenue par des milliers voire centaines de milliers de sites. On essaie ici de déterminer avec précision la nature de chaque résidu individuel ancestral à chaque noeud de l'arbre. Il est notamment bien établi que tous les sites d'une protéine ou d'une séquence nucléotidique n'évoluent pas selon les mêmes lois de probabilités et dépendent notamment de la structure secondaire dans laquelle ils sont engagés, de leur état d'enfouissement au sein de la protéine (Koshi et Goldstein, 1995), ou du contexte nucléotidique, par exemple les îlots CpG (Baele, Van De Peer et Vansteelandt, 2010), ou encore de leur importance fonctionnelle (Sikosek et Chan, 2014 ; Arenas et Bastolla, 2020). En outre, les sites d'une même protéine peuvent être influencés par des relations d'épistasie, c'est-à-dire que la mutation d'un site vers un certain résidu peut permettre, ou au contraire restreindre, les possibilités de mutation vers un résidu à un autre site de la séquence (effets compensatoires) (Trudeau, Kaltenbach et Tawfik, 2016). Ces relations épistasiques peuvent influencer le profil mutationnel des sites et également rendre le processus évolutif non réversible, puisqu'une mutation qui apparaît dans une protéine, si elle n'est pas immédiatement contre-sélectionnée, a tendance à être accommodée par d'autres mutations au cours de l'évolution (Pollock, Thiltgen et Goldstein, 2012). D'autres biais de la méthode ont été suggérés dans la littérature. Par exemple, l'ASR souffrirait d'un effet consensus qui tendrait à accumuler dans les séquences ancestrales les résidus les plus communs à chaque site, ce qui augmenterait artificiellement la stabilité de la reconstruction (Trudeau, Kaltenbach et Tawfik, 2016). Même si la récurrence de certains résidus ayant évolué de façon convergente peut effectivement tendre à ce que ce même résidu soit proposé à l'état ancestral, il me semble que l'objectif d'une approche probabiliste est justement d'éviter ce biais. La nature du résidu ancestral ne dépend pas du résidu majoritaire dans les séquences descendantes, mais bien de la topologie de l'arbre qui relie les séquences entre elles. Aussi, même si un résidu récurrent a une probabilité non nulle d'être présent à l'état ancestral, il ne devrait pas être favorisé par rapport à un autre acide aminé qui serait un bon candidat pour expliquer l'évolution du résidu à ce site. Williams et al., 2006 ont proposé une interprétation plus pertinente de ce biais, en montrant sur des séquences simulées que le choix d'une séquence maximisant la vraisemblance de la reconstruction à chaque site tend à sélectionner des séquences biaisées vers une plus grande stabilité. En effet, l'ASR est une technique probabiliste qui donne des séquences ancestrales possibles, mais qui ne sont pas nécessairement les séquences ancestrales réelles. Une bonne utilisation de cette technique est donc de mesurer la variabilité du phénotype ancestral que l'on cherche à mesurer en fonction de l'incertitude sur les séquences reconstruites (Wheeler et al., 2016). Cette variabilité entre la séquence reconstruite et la séquence réelle, si l'on sé-

lectionne systématiquement la séquence reconstruite de maximum de vraisemblance, n'est pas symétrique. De façon générale les protéines plus stables sont souvent préférées aux protéines moins stables. En d'autres termes, bien que le choix de la séquence selon le critère du maximum de vraisemblance nous permette d'obtenir la séquence qui a le plus haut pourcentage d'identité avec la séquence ancestrale réelle, le maximum de vraisemblance nie la stochasticité du processus évolutif qui est plus à même d'introduire des mutations qui déstabilisent la protéine par rapport à son *optimum* de stabilité potentiel (Goldstein, 2011). Il a également été suggéré que les stabilités importantes observées dans la plupart des protéines ancestrales reconstruites pouvaient traduire le fait que les températures du Précambrien étaient plus élevées que les températures actuelles (Boussau et al., 2008 ; Wheeler et al., 2016), ou bien le fait que la machinerie cellulaire ancestrale était globalement moins perfectionnée avec des taux d'erreur de translation et de traduction des protéines plus importants. Les protéines ancestrales seraient donc optimisées pour être globalement plus stables que les protéines contemporaines, compensant les erreurs de translation/traduction (Trudeau, Kaltenbach et Tawfik, 2016). Ces hypothèses sont tout à fait crédibles. Néanmoins, au regard des nombreuses incertitudes concernant les modèles listées auparavant, il reste probable que des biais liés aux modèles utilisés viennent également affecter les résultats des reconstructions ancestrales. Il existe également d'autres sources d'incertitude inévitables, comme les branches longues qui témoignent d'une divergence importante entre deux noeuds de l'arbre et qui sont associées à une baisse du signal phylogénétique (Koshi et Goldstein, 1996). Ces longs temps de divergence sont particulièrement problématiques pour inférer les sites faiblement contraints et peu conservés (Eick et al., 2016). L'ambiguïté de la reconstruction à certains sites ne peut pas être facilement résolue, en particulier si l'on s'intéresse à la reconstruction de protéines très anciennes.

Pour palier ces incertitudes, plusieurs méthodes ont déjà été proposées. Une première approche est d'utiliser des modèles de phylogénie moins restrictifs, par exemple le modèle CAT implémenté dans Phylobayes (Blanquart et Lartillot, 2008) qui prévoit que les fréquences attendues des résidus à chaque position de l'alignement soient différentes. Les fréquences des acides aminés aux différentes positions sont optimisées par maximum de vraisemblance, et ce modèle va capturer une partie de la contrainte imposée aux sites par la sélection. Ce modèle a notamment été utilisé pour la reconstruction des malate déshydrogénases ancestrales d'archées halophiles (Blanquart, Groussin et al., 2021). En outre, le modèle CAT-BP, initialement présenté dans Blanquart et Lartillot, 2008, ajoute également la possibilité de "breakpoints" le long des branches où les fréquences par site peuvent varier. Le modèle est donc également hétérogène dans le temps. Toutefois, comme ce modèle contient beaucoup de paramètres, les temps de convergence peuvent être longs et la solution instable si les séquences ne sont pas assez nombreuses ou trop courtes. Ce modèle rend mieux compte de la réalité biologique de l'évolution de la séquence, mais peut être difficile à utiliser en pratique. D'autres auteurs ont proposé des algorithmes dans le même esprit, où le profil de fréquences à l'équilibre n'est pas optimisé au niveau de chaque site, mais sur l'alignement entier (W. Cai, Pei et Grishin, 2004). Ce compromis peut être une option mieux dimensionnée pour l'étude d'un alignement relativement court avec un nombre restreint de séquences. Ces options utilisent donc une seule matrice de mutation, qui est équilibrée de telle sorte de modifier les fréquences d'acides aminés obtenus à l'équilibre, soit en considérant un profil de fréquences site-spécifiques, soit en considérant un profil de fréquences sur l'ensemble de l'alignement. Une autre possibilité est d'utiliser plusieurs ma-

trices de mutations instantanées, obtenues de telle sorte qu'elles reflètent les dynamiques de mutation observées dans différentes régions de la protéine (par exemple les hélices α ou feuillets β, dans des régions exposées ou enfouies) (Koshi et Goldstein, 1995). Dans ce cas, l'utilisation des différentes matrices peut soit découler de connaissances de la structure de la protéine, soit être optimisée par maximum de vraisemblance sans connaissance de la structure de la protéine (Si Quang Le, Lartillot et Olivier Gascuel, 2008). Ce type de modèle permet d'améliorer grandement la vraisemblance des phylogénies obtenues (Si Quang Le et Olivier Gascuel, 2010 ; Moshe et Pupko, 2019), et potentiellement la fidélité des reconstructions ancestrales. Ces modèles ont toutefois été peu adoptés en pratique dans les articles utilisant l'ASR (Merkl et Sterner, 2016). Enfin certains modèles ont été spécifiquement développés dans le but d'améliorer les résultats obtenus par ASR en s'écartant du cadre phylogénétique habituel. Notamment, le programme protASR2 essaie d'inférer des résidus ancestraux considérés indépendants mais contraints selon l'effet qu'ils peuvent avoir sur la stabilité de la protéine (Arenas et Bastolla, 2020). En effet, on sait que les protéines maintiennent une stabilité marginale d'environ -10 kcal/mol dans leur environnement cellulaire (Taverna et Goldstein, 2002 ; Goldstein, 2011). L'idée de Arenas et Bastolla, 2020 est donc de modéliser la stabilité de la protéine et de la maintenir relativement constante au cours de l'évolution, afin notamment de corriger le biais de l'ASR qui consiste à proposer des protéines plus thermostables qu'elles ne l'étaient en réalité. L'idée est intéressante, mais repose sur l'hypothèse que la stabilité de la protéine doit rester constante. Or beaucoup d'études d'ASR visent à reconstruire des protéines parfois très anciennes, en s'interrogeant justement sur la température environnementale passée à partir de la variation de stabilité des protéines au cours du temps. Dans ce cas, prendre comme hypothèse que la stabilité de la protéine doit rester constante n'est sans doute pas judicieux pour reconstruire les protéines, notamment si la fonction ou l'abondance de la protéine a également pu varier sur l'échelle de temps considérée (Tokuriki et al., 2008 ; Serohijos et Eugene I Shakhnovich, 2014). Une dernière approche intéressante est de ne pas chercher à optimiser la reconstruction elle-même, mais l'évaluation de son incertitude. Évidemment, ce point est complémentaire au développement d'un modèle adapté à la reconstruction. Eick et al., 2016 ont comparé différentes méthodes visant à aider l'expérimentateur pour quantifier l'incertitude dans les résultats obtenus dans le cadre de l'ASR. Parmi elles, la méthode "Worst Plausible Case" est la plus facile à mettre en oeuvre. Dans ce cas, tous les résidus ambigus de la séquence (par exemple avec une probabilité inférieure à 0.8) sont remplacés par le résidu concurrent. Ceci permet d'obtenir une séquence alternative possible dans le pire des cas, donc qui maximise l'erreur qu'on aurait pu faire en ne synthétisant que la séquence de meilleure vraisemblance. Pour l'expérimentateur, cela amène à borner les résultats expérimentaux obtenus dans la caractérisation de la protéine tout en maintenant la faisabilité de la procédure. A l'inverse, une méthode statistiquement judicieuse mais difficile à mettre en oeuvre consiste à échantillonner un nombre élevé de séquences selon les probabilités associées à chaque résidu à chaque position, et de synthétiser et caractériser toutes les protéines obtenues. Concrètement, on est amené à ne synthétiser qu'un nombre restreint de ces séquences. Cette méthode est par exemple employée par (Gaucher, Govindarajan et Ganesh, 2008) avec la synthèse de cinq protéines ancestrales alternatives. En outre, le choix purement statistique dans l'échantillonnage des séquences peut souvent amener à la synthèse de protéines non fonctionnelles qu'on ne peut pas exprimer ou mesurer (Eick et al., 2016). Ceci démontre que l'ensemble réel des séquences biologiquement acceptables

n'est pas superposable aux séquences obtenues par de telles méthodes, probablement du fait de valeurs faussées dans les probabilités postérieures obtenues pour les différents résidus, puisque cette même méthode appliquée sur des données simulées donne effectivement des résultats conformes à l'attendu (Williams et al., 2006).

Dans ce chapitre, j'ai essayé de développer des modèles d'évolution moléculaire et de comparer leurs résultats avec les reconstructions obtenues par les méthodes usuellement utilisées. Pour l'évaluation des différentes méthodes, je me suis appuyé sur les familles de protéines utilisées chez les Alvinellidae dans le chapitre 2, afin notamment de constater si les reconstructions obtenues au chapitre 2 sont très dépendantes du modèle phylogénétique utilisé. Pour évaluer la qualité des modèles développés, j'ai également utilisé une phylogénie expérimentale obtenue par Randall et al., 2016 sur une protéine fluorescente RFP qui a l'intérêt de fournir les séquences ancestrales réelles à partir d'une phylogénie sans biais. J'ai préféré utiliser des données biologiques réelles à des données simulées pour ne pas biaiser l'évaluation des différents modèles. Les trois modèles développés et présentés dans ce chapitre sont pensés pour être appliqués dans le cadre de l'ASR. Par conséquent, ils profitent du fait qu'on étudie des protéines déjà caractérisées et pour lesquelles on dispose d'informations supplémentaires sur leur structures secondaire et tertiaire.

Trois modèles ont été développés :

— Ecoprior : ce modèle cherche à interpréter des différences dans les fréquences attendues entre acides aminés, selon la température de vie des organismes. En effet, il est connu que la température de vie des organismes a une influence directe sur la composition des protéomes, aussi bien chez les organismes thermophiles (Pack et Yoo, 2004) que psychrophiles (Metpally et Reddy, 2009). Ainsi, la matrice utilisée dans ce modèle pour décrire l'évolution de la séquence est modifiée (biaisée) par l'acquisition de mutations préférentielles selon l'environnement thermique des espèces. La pondération de ce biais est optimisée par maximum de vraisemblance sur chaque branche de l'arbre phylogénétique. Le modèle est par conséquent hétérogène dans le temps selon l'influence de tel ou tel environnement thermique sur le noeud considéré, et potentiellement non réversible car les mutations entre différentes paires d'acides aminés ne sont pas symétriques. Le concept est relativement similaire à celui développé par Foster, 2004 sur une petite phylogénie de procaryotes d'environnements thermiques différents. Cependant, dans notre cas, nous utilisons des matrices de mutations entre acides aminés que nous aurons préalablement optimisées sur des jeux de données comparant des organismes mésophiles et thermophiles, afin de faire ressortir les biais de mutation en acides aminés qui sont la conséquence de la température de vie des organismes ;

— Gempistasie : ce modèle s'appuie sur le travail de Laine, Karami et Carbone, 2019, qui a développé un outil pour quantifier l'effet délétère d'une mutation aléatoire dans une protéine connue. L'effet prédit par le logiciel, GEMME, est converti en probabilité d'observer chaque acide aminé aux différents sites de la protéine, et sert à adapter le modèle phylogénétique. Le modèle résultant n'est donc ni réversible, ni homogène dans le temps. GEMME utilise l'information contenue dans des séquences prises dans la banque NCBI et homologues aux séquences étudiées afin de prédire l'effet des mutations. GEMME regarde notamment si la mutation dont on cherche à prédire l'effet est fréquente dans des séquences présentant différents degrés de similarité avec les

séquences étudiées. Par conséquent, notre modèle de phylogénie intègre une information propre à la protéine. Ceci doit se traduire par des probabilités de mutation site-dépendant et contexte-dépendant. Il est toutefois difficile de connaître le niveau de complétude de cette information, qui dépend fondamentalement de la diversité des séquences homologues contenues dans la base de donnée utilisée (NCBI). Il est important de souligner que GEMME produit un calcul analytique de l'effet de la mutation grâce aux séquences homologues. Par conséquent, même si le modèle phylogénétique résultant est optimisé pour la protéine étudiée, il ne comporte pas plus de paramètres que les modèles phylogénétiques les plus simples, ce qui est très utile pour optimiser le modèle pour de petites protéines.

— Struct2 : ce dernier modèle utilise l'information de la structure secondaire des protéines afin d'appliquer des matrices de mutations instantanées différentes selon les contraintes structurelles qui s'opèrent sur la protéine. Au contraire des modèles de phylogénie utilisant ce type de matrices, Struct2 utilise la connaissance *a priori* du repliement de la protéine étudiée. La matrice d'évolution à utiliser n'est donc pas choisie par une optimisation de l'algorithme mais fixée *a priori* à partir de l'environnement protéique dans lequel se trouve le site analysé. Le modèle est donc contexte-dépendant et potentiellement non réversible si le repliement de la protéine est amené à changer au cours du temps ;

IV.1.2 Matériel et Méthodes

IV.1.2.1 Obtention des séquences contemporaines et hypothèses phylogénétiques

Trois protéines issues de la famille des Alvinellidae ont été utilisées pour mener les différents tests concernant les modèles de reconstruction : la Malate Deshydrogénase cytosolique (MDHc), la Superoxyde Dismutase Cuivre/Zinc (SOD) et une famille d'hémoglobine intracellulaire (Hb). Notre choix s'est porté sur ces protéines car elles correspondent aux protéines expérimentalement testées dans le chapitre 2 de ce manuscrit.

L'identification des gènes correspondant à ces différentes protéines a été réalisée par BLAST contre l'ensemble des transcrits présents en sortie d'assemblage (voir chapitre 1), dont le cadre de lecture est identifié à l'aide de Transdecoder v5.5.0. Nous avons commencé par chercher ces gènes dans le génome d'*A. pompejana*, assemblé et annoté par Richard Copley (numéro d'accession NCBI PRJEB46503, El Hilali et al., 2024). Ces séquences ont ensuite été cherchées dans les autres trancriptomes d'Alvinellidae, ainsi que dans les espèces constituant l'outgroup (Terebellidae gen. sp., *N. edwardsi*, *M. palmata*, *P. gouldii*, *Anobothrus* sp., *H. invalida*, *A. gunneriA. carldarei*). Les séquences protéiques obtenant la meilleure p-value ont ensuite été filtrées à la main et avec l'aide de PhyML pour grouper les séquences similaires (Guindon et al., 2010). Les séquences finales sont alignées avec ProbCons v1.12 (Do et al., 2005) et les alignements sont corrigés à la main si besoin.

La topologie d'arbre utilisée pour tester les différent modèles correspond à la phylogénie proposée par Desbruyères et Laubier pour les Alvinellidae (Desbruyères et Laubier, 1986). Dans cette hypothèse, les espèces *Paralvinella* forment un groupe monophylétique distinct des deux espèces *Alvinella*, et les espèces *P. p. irlandei* et *P. unidentata* forment un sous-

genre dénommé *Nautalvinella*. Cette topologie correspond par conséquent à la topologie T6 discutée au chapitre 1.

A ces trois jeux de protéines chez les Alvinellidae est ajouté un jeu de données utilisant la Red Fluorescent Protein (RFP). L'évolution de cette protéine a été conduite expérimentalement par Randall et al., 2016. Cette protéine peut servir de référence pour évaluer les méthodes de reconstruction développées dans ce chapitre, étant donné que les séquences ancestrales et la phylogénie sont connues avec certitude. Les séquences des RFP "contemporaines" ainsi que la topologie d'arbre utilisée pour la reconstruction proviennent de cet article.

IV.1.2.2 Calcul des séquences ancestrales

IV.1.2.2.1 Modèle "Ecoprior"

L'objectif de ce modèle est d'inclure aux matrices classiques de mutation des acides aminés (JTT, WAG ou LG par exemple, Jones, W. R. Taylor et J. M. Thornton, 1992 ; Whelan et Goldman, 2001 ; S. Q. Le et O. Gascuel, 2008) le biais dû à la sélection sur les acides aminés lié à la température dans l'inférence des acides aminés ancestraux. Nous considérons pour cela que les matrices de mutation classiques décrivent les probabilités d'évolution d'une séquence au cours du temps avec des contraintes très générales sur les mutations possibles, telles les propriétés physico-chimiques ou stériques des acides aminés. L'idée du modèle Ecoprior est de coupler ces matrices à deux matrices, HotJAMA (HJM) et ColdJAMA (CJM), qui enregistrent plus spécifiquement les pressions sélectives imposées par les températures de vie des organismes.

Si, au cours du temps, la probabilité de muter suit une loi décrit par $Pr = exp(P.t)$ dans un modèle classique (avec P pouvant être une matrice de mutation comme WAG ou LG), nous voulons construire un modèle où P est remplacée par :

$$M = a.P + b.HJM + c.CJM \tag{IV.1}$$

avec $a + b + c = 1$. L'idée est par conséquent que la divergence entre les séquences est dépendante du temps d'évolution t qui s'applique sur M, ainsi que de deux paramètre b et c qui décrivent deux biais de différentes amplitudes imposés à la matrice de mutation standard, et qui traduisent le fait qu'un organisme ait évolué vers un environnement plus chaud ou plus froid. L'organisme aurait aussi pu évoluer dans des environnements qui ont connu de multiples changements de température, impliquant qu'on observe globalement des substitutions plus fréquentes sur certaines classes d'acides aminés dont les fréquences sont connues pour être corrélées avec la température de l'environnement (Hickey et Singer, 2004).

Par conséquent, ce modèle n'est pas réversible puisque la matrice résultante M inclut par construction un biais de mutation non symétrique entre certains acides aminés. En outre, le modèle n'est pas non plus homogène puisque les paramètres b et c sont optimisés sur chaque branche.

IV.1.2.2.1.1 Optimisation des matrices de mutation biaisées par la température JAMA Deux matrices de mutations instantanées ont été optimisées à cette étape, la première afin de décrire l'évolution de séquences codantes d'organismes thermophiles, et la seconde d'organismes psychrophiles.

Deux bases de données ont été utilisées à cette fin, provenant de Szilágyi et Závodszky, 2000 et T. J. Taylor et Vaisman, 2010. Ces bases consistent en des références de protéines homologues issues d'organismes mésophiles, thermophiles ou hyperthermophiles, trouvées dans tous les domaines du vivant. Les deux articles s'appuient sur un total de 155 paires de protéines homologues. Nous avons ensuite récupéré les séquences depuis PDB, et les avons alignées avec ProbCons (Do et al., 2005). A cette étape, chaque groupe de protéines orthologues contient une séquence d'un organisme thermophile ou hyperthermophile (>42°C) et une ou plusieurs séquences d'organismes mésophiles (<37°C). L'optimisation des matrices se fait en regardant le biais à imposer à la matrice standard pour améliorer la vraisemblance de passer des séquences mésophiles aux séquences thermophiles (HJM) ou au contraire des séquences thermophiles vers les séquences mésophiles (CJM).

La matrice de mutation est entraînée sur ces séquences *via* un algorithme d'échantillonnage de Monte Carlo, décrit en détail dans les données supplémentaires de Sumanaweera, Allison et Konagurthu, 2022. Le principe de cet échantillonnage est d'itérer des modifications aléatoires de la matrice Q que l'on cherche à optimiser. A chaque itération k, nous calculons la vraisemblance L_k du modèle avec cette nouvelle matrice Q_k (soit la probabilité totale, sur l'ensemble des paires de séquences, d'observer les mutations présentes dans les alignements en utilisant la matrice de mutation Q_k). Si la vraisemblance est meilleure, la nouvelle matrice Q_k est conservée, et sert de nouveau point de départ pour l'itération suivante avec la matrice Q_{k+1}. Si la vraisemblance est moins bonne, la nouvelle matrice Q_k peut être conservée pour constituer Q_{k+1} avec une probabilité L_k/L_{k-1}. Autrement, la matrice Q_k est ignorée, et l'on prend $Q_k = Q_{k-1}$. Ainsi, pour l'itération $k+1$, c'est la matrice de l'itération $k-1$ qui est modifiée pour obtenir Q_{k+1}. L'objectif est donc d'apporter des petites modifications à Q tour après tour, et de ne conserver ces modifications que si elles améliorent le modèle, ou potentiellement si elles ne le dégradent pas trop. L_k est obtenue selon :

$$L_k = -\sum_{i=1}^{n}\sum_{j=1}^{s_i} log(Pr_{i_1,j} \times e^{Q_k.t_i} \times Pr_{i_2,j}) \tag{IV.2}$$

avec t_i le temps de divergence entre les séquences du groupe i, s_i la longueur de l'alignement du $i^{\text{ème}}$ groupe, $Pr_{i_1,j}$ et $Pr_{i_2,j}$ les vecteurs de probabilités des acides aminés observés des séquences du groupe i au site j (1 pour l'acide aminé présent, 0 sinon). Pour l'optimisation de HJM, Pr_{i_1} correspond à la séquence mésophile et Pr_{i_2} à la séquence thermophile, et inversement pour CJM.

Pour ce faire, à chaque itération k, une ligne de la matrice est sélectionnée au hasard selon une probabilité proportionnelle à la valeur diagonale de la matrice Q_{k-1} (les acides aminés avec la plus forte mutabilité sont proportionnellement plus souvent sélectionnés). Les 19 coefficients indépendants de cette ligne sont ensuite tirés au hasard selon une loi de Dirichlet ayant pour paramètres les valeurs de la ligne à l'itération $k-1$. La quantité L_k est calculée avec cette nouvelle matrice, et la matrice de l'itération $k-1$ est remplacée par

la matrice de l'itération k à la condition que

$$\alpha < exp(log(L_k) - log(L_{k-1}))$$

avec α un nombre aléatoire tiré uniformément sur $[0, 1]$. Les temps de divergence t_i entre les séquences doivent également être optimisés à chaque itération en fonction de la nouvelle matrice évaluée, en prenant son estimation par maximum de vraisemblance. Les positions des alignements présentant un gap sont exclus du calcul.

L'algorithme est initialisé avec une matrice de mutation simple, établie en comptant les paires d'acides aminés homologues entre les séquences. La matrice A obtenue par comptage donne Q_{init} selon Rivas, 2005 :

$$Q_{init} = \sum_{i=1}^{\infty} \frac{-1^{i+1}}{n} \times (A - I)^i \tag{IV.3}$$

Les séquences recueillies par Szilágyi et Závodszky, 2000 et T. J. Taylor et Vaisman, 2010 peuvent provenir d'organismes avec des températures de vie très différentes, et l'écart de température entre les organismes thermophiles et mésophiles au sein des groupes peut fortement varier. Pour cette raison, une seconde optimisation de matrice a été ajoutée par la suite. La même méthode est utilisée, à ceci près que l'on cherche à minimiser la vraisemblance :

$$L_k = -\sum_{i=1}^{n} \sum_{j=1}^{s_i} log(Pr_{i_1,j} \times exp[((1 - a_i).P + a_i.Q_k).t_i] \times Pr_{i_2,j}) \tag{IV.4}$$

P représente une matrice de mutation classique disponible dans la littérature, telle la matrice WAG ou JTT. Nous avons sélectionné la matrice PFASUM (Keul et al., 2017), car elle présente globalement de meilleures performances selon Sumanaweera, Allison et Konagurthu, 2022. Dans cette optimisation, à chaque itération, la matrice Q est échantillonnée selon la procédure décrite précédemment tandis que P est fixe. Les paramètres a_i et t_i sont optimisés par maximum de vraisemblance pour chaque alignement. a_i modèle la force du biais qui s'applique sur l'évolution de la séquence dû à la différence de température entre les séquences mésophile et thermophile, avec une transition entre une matrice non biaisée P qui modèle l'évolution stochastique due au temps de divergence, et une matrice Q_k qui enregistre le biais de mutation dû à l'effet de la température. Idéalement, on cherche à ce que les paramètres t_i et a_i soient orthogonaux bien qu'il existe une interaction évidente entre l'effet du différentiel de température sur l'évolution de l'organisme et la vitesse à laquelle cette différence de température lui a été imposée. En pratique, on optimise par ML une fonction du type

$$L_k = -\sum_{i=1}^{n} \sum_{j=1}^{v_i} log(Pr_{i_1,j} \times exp[\delta_i.P + \beta_i.Q_k)] \times Pr_{i_2,j}) \tag{IV.5}$$

ce qui implique que $t_i = \delta_i + \beta_i$ et $a_i = \beta_i/(\delta_i + \beta_i)$. Pour faciliter l'optimisation de δ_i et β_i par maximum de vraisemblance à chaque itération k, on peut obtenir la dérivée de la probabilité selon ces deux paramètres par la méthode des "blocs" décrite dans Mathias, 1996. En outre, les matrices P et Q_k sont mises à la même échelle de temps, c'est à dire que pour un temps de divergence de 0.01 imposé sur chacune des matrices, la probabilité de mutation d'un acide aminé pris au hasard est de 1%.

IV.1.2.2.1.2 Reconstruction des acides aminés ancestraux Les matrices HJM et CJM sont optimisées et fixées. Nous voulons les utiliser pour corriger une matrice standard (PFASUM dans notre cas) afin de tenir compte des biais de mutation liés la température de vie des organismes pour n'importe quelle phylogénie. Cela revient à considérer un modèle non homogène où les compositions en acides aminés attendues aux différents noeuds de l'arbre sont différentes. Nous ajoutons au modèle le paramètre Γ discrétisé en quatre classes pour simuler la variation du taux de mutation le long de la protéine. Les longueurs de branche ainsi que le paramètre Γ sont optimisés par maximum de vraisemblance. Les longueurs de branches, équivalentes aux paramètres a, b et c ci-dessous, sont optimisées pour maximiser la probabilité :

$$Pr = \prod_{i=1}^{s} \sum_{k=1}^{4} Pr_{i_1} \times exp[(a.P + b.HJM + c.CJM) \times g_k.t] \times Pr_{i_2} \tag{IV.6}$$

où s est la longueur de la séquence, Pr_{i_1} et Pr_{i_2} les probabilités des acides aminés au résidu i des séquences des noeuds encadrant la branche, g_k la valeur moyenne de la $k^{\text{ème}}$ classe du paramètre Γ. En pratique, on optimise la solution à l'aide d'une fonction :

$$Pr = \prod_{i=1}^{s} \sum_{k=1}^{4} Pr_{i_1} \times exp[g_k(\delta.P + \beta.HJM + \omega.CJM)] \times Pr_{i_2} \tag{IV.7}$$

Les dérivés de cette fonction selon δ, β et ω sont obtenues par la méthode des blocs.

Le modèle optimisé présente donc le paramètre Γ, les longueurs de branches ainsi que les poids relatifs des matrices PFASUM, HJM et CJM sur chaque branche. Les probabilités marginales des acides aminés aux différents noeuds de l'arbre, nécessairement calculées pendant l'optimisation du modèle, sont alors récupérées et enregistrées.

IV.1.2.2.2 Modèle "Gempistasy"

L'objectif du modèle Gempistasy est de construire un modèle de phylogénie qui tienne compte des contraintes mutationnelles sur chaque résidu de l'alignement. Ces contraintes sont obtenues grâce au logiciel GEMME développé par Laine, Karami et Carbone, 2019. Gempistasy se décompose selon ces étapes :

1. Alignement des séquences contemporaines ;
2. Etablissement du "portrait robot" de la protéine étudiée. Pour cela, les prédictions par GEMME des acides aminés attendus aux sites des protéines contemporaines de l'alignement sont moyennées par site ;
3. Optimisation des longueurs de branche de la phylogénie, en utilisant à la fois une matrice de mutation entre les acides aminés (type WAG, LG, PFASUM) et les prédictions du portrait-robot de la protéine ;
4. Obtention des probabilités marginales des acides aminés ancestraux à toutes les positions pour tous les noeuds de l'arbre, à partir de la phylogénie optimisée ;
5. Test de la fiabilité des acides aminés ancestraux inférés, sous différentes hypothèses de portrait-robot de la protéine. Les positions aux différents noeuds pour lesquelles

les acides aminés inférés sont stables sous toutes les hypothèses de portrait-robot sont considérées résolues. Les autres positions sont recalculées aux étapes suivantes ;

6. Perturbation des acides aminés aux positions/noeuds ambigus, afin de déterminer quelles autres positions/noeuds ambigus sont sous influence de la position perturbée. Cette étape permet de regrouper les positions qui sont sous influence les unes des autres (épistasie) ;
7. Résolution des acides aminés ancestraux aux positions/noeuds pour lesquelles on n'a pas détecté d'influence d'autres positions ambiguës de l'alignement. Pour résoudre ces positions/noeuds, nous n'utilisons plus le portrait-robot de la protéine (moyenne des prédictions GEMME) mais plutôt les prédictions GEMME des acides aminés attendus pour chaque séquence ancestrale individuellement. Le calcul est plus précis, mais aussi plus long ;
8. Résolution des acides aminés ancestraux aux positions/noeuds pour lesquelles on a détecté une influence d'autres positions ambiguës de l'alignement. Dans ce cas, nous résolvons les sites sous influence les uns des autres simultanément. Ces sites liés entre eux forment une petite séquence plus courte, et nous cherchons l'ensemble des petites séquences aux différents noeuds de l'arbre ayant la meilleure vraisemblance par échantillonnage de Monte Carlo.

Le modèle Gempistasy utilise la diversité des acides aminés dans les séquences homologues disponibles dans la banque NCBI afin de calibrer un modèle phylogénétique qui tient compte des mutations épistasiques. Il s'appuie sur le logiciel GEMME (Laine, Karami et Carbone, 2019), qui est un prédicteur de l'effet délétère d'une mutation sur une séquence protéique "query". Dans le principe, GEMME produit deux informations :

— une valeur de prédiction "indépendante" $Pred_{ind}$, qui consiste à regarder la fréquence des acides aminés à une position dans un large alignement constitué de plusieurs centaines de séquences homologues à la séquence query ;

— une valeur de prédiction "épistasique" $Pred_{epi}$, qui consiste à regarder la fréquence d'une mutation dans un large alignement de plusieurs centaines de séquences homologues à la séquence query, selon que ces séquences présentent une forte similarité avec la query ou qu'elles soient plus divergentes. Pour évaluer $Pred_{epi}$, GEMME établit un arbre neighbour-joining sur un alignement de séquences, puis calcule la distance minimale à parcourir sur l'arbre, à partir de la séquence query, pour trouver la mutation d'intérêt. Plus cette distance est grande, moins la mutation est probable (la séquence doit être accommodée d'un plus grand nombre de mutations à d'autres positions pour accepter la mutation d'intérêt).

Notre choix s'est porté sur GEMME pour la fiabilité de la méthode, qui fait parti des meilleures actuellement disponibles (basé sur la corrélation entre l'effet prédit pour une mutation par GEMME et l'effet réel mesuré dans des expériences de scans de mutations épistasiques de protéines, Laine, Karami et Carbone, 2019), ainsi que pour sa vitesse d'exécution. En effet, la reconstruction des séquences ancestrales passe par un processus itératif sur le calcul des probabilités (optimisation du modèle) et le test des séquences candidates ancestrales. Il est donc important de disposer d'une méthode fiable et suffisamment rapide pour être appliquée.

Notre objectif avec l'adaptation de ce modèle est de relâcher en partie l'hypothèse, courante en phylogénie, d'indépendance des résidus, et de résoudre l'ambiguïté de la nature

de certains remplacements d'acides aminés ancestraux grâce à l'information contenue dans le reste de la séquence (épistasie entre sites). La difficulté est que ces calculs peuvent rapidement devenir impossibles à effectuer, si l'on considère que les probabilités de mutation pour chaque séquence et chaque résidu dépendent de la nature de tous les autres résidus de cette séquence. En pratique, nous avons essayer de décomposer le problème en plusieurs étapes :

1. Initialiser le modèle en reconstruisant les séquences ancestrales de manière approchée afin d'avoir un "portrait robot" des séquences ancestrales à évaluer dans GEMME ;
2. Déterminer quels résidus peuvent être reconstruits sans ambiguïté et sont des variables fixes pour la recherche des mutations épistasiques ;
3. Déterminer quels groupes de résidus évoluent ensemble, et établir un critère permettant de constituer des séquences ne contenant que les résidus fortement dépendants les uns des autres afin de réduire le problème à une taille de séquence raisonnable.

IV.1.2.2.2.3 Initialisation du modèle Nous commençons par effectuer une prédiction par GEMME de toutes les protéines contemporaines étudiées, sur la base d'un blast de protéines homologues disponibles dans la banque NCBI. Les protéines contemporaines peuvent potentiellement être relativement divergentes et le résultat du blast NCBI peut sensiblement varier selon la protéine contemporaine utilisée pour le blast. Pour cette raison, nous établissons une matrice de distances entre les séquences contemporaines (selon leur pourcentage d'identité) et nous regroupons les séquences similaires par la méthode des k-means. Pour chacun de ces groupes, une séquence consensus est obtenue. Ce sont ces séquences consensus qui servent de référence pour les blasts, ainsi qu'au calcul de $Pred_{ind}$ dans la suite des calculs. Cela permet d'accélérer le fonctionnement de GEMME, considérant que $Pred_{ind}$ utilise uniquement l'information de la fréquence des résidus sur l'ensemble de l'alignement issu du blast et n'a plus besoin de la séquence exacte de la query.

A cette étape, nous disposons pour chaque acide aminé possible, à chaque position et pour chaque protéine, d'un score attribué par GEMME pour $Pred_{ind}$ et $Pred_{epi}$. Ces scores indiquent si une mutation de la séquence query à une position sur un acide aminé donné a un fort effet épistasique sur la protéine. Un score de zéro indique que la mutation a peu d'effet (on s'attend à ce qu'elle soit possible et plus fréquente) et un score fortement négatif a un effet plus fort (on s'attend à ce qu'elle soit délétère). Nous convertissons ces scores en fréquences à l'équilibre π_s attendues pour chaque position s des protéines. Pour cela, nous prenons :

$$\pi_{s,ind} = exp(Pred_{s,ind}) \tag{IV.8}$$

par construction selon l'article de Laine, Karami et Carbone, 2019, et :

$$\pi_{s,epi} = \frac{1}{1 + i_1 \times exp(i_2 \times Pred_{s,epi} + i_3)} - \frac{1}{1 + i_1 \times exp(i_3)} \tag{IV.9}$$

Cette courbe est choisie parce que son profil est très flexible en fonction de l'optimisation des paramètres. Elle peut ainsi s'approcher d'une relation linéaire, de puissance ou

sigmoïde selon les paramètres choisis. $Pred_{s,epi}$ est préalablement normalisé entre 0 et 1, avec 0 correspondant aux mutations avec le plus fort effet sur la protéine (peu fréquentes) et 1 les mutations avec les effets les plus faibles (plus courantes). i_1, i_2 et i_3 sont évalués une fois par ML puis fixés dans les analyses, puisqu'on voudrait que l'interprétation des valeurs prédites par GEMME soit une solution analytique de l'algorithme, et pas une optimisation selon la protéine. Enfin, la fréquence globale attendue est donnée par :

$$\pi_s = \pi_{s,epi} \times \alpha/(1+\alpha) + \pi_{s,ind} \times (1+\alpha) \tag{IV.10}$$

Dans l'article initial décrivant le fonctionnement de GEMME, la valeur de α est fixée à 0.6 par défaut, mais les auteurs soulignent que les prédictions "épistasiques" sont généralement meilleures que les prédictions "indépendantes". J'ai préféré considérer que α peut être variable, et donné par :

$$\alpha = r \times \frac{(\pi^p_{s,epi} - \pi^p_{s,ind})^2}{2 \times min[(\pi^p_{s,epi} - 1)^2 + \pi^{2p}_{s,ind}\ ,\ \pi^{2p}_{s,epi} + (\pi^p_{s,ind} - 1)^2]} \tag{IV.11}$$

L'objectif de cette transformation est que α est plus grand si les prédictions épistasiques et indépendantes sont très dissemblables, et petit si les deux prédictions sont similaires. r et p sont deux paramètres aussi évalués par ML, puis fixés dans la suite de l'analyse. Le paramètre r permet de savoir quel type de prédiction épistasique ou indépendante est préférentiellement avantagée, et p, entre 0 et 1, permet d'étirer le signal donné pour les faibles fréquences prédites. L'idée est que nous voulons équilibrer α selon la similitude de fréquence entre π_{ind} et π_{epi}, mais je ne veux pas, par exemple, qu'une mutation qui obtiendrait des scores respectivement de 0.8 et 0.3 soient considérés comme trop dissemblables, car une fréquence de 0.3 pour un acide aminé à une position peut déjà être considéré comme très élevée. En choisissant, dans cet exemple, $p = 0.5$, j'obtiens des fréquences transformées entre 0.9 et 0.55, réduisant ainsi l'écart relatif entre les deux estimations. Afin d'éviter d'obtenir des fréquences attendues à 0, on choisit également de contraindre les fréquences π_s obtenues au final entre 0.01 et 0.81.

Les matrices de mutation Q_s à chaque position s de l'alignement sont ensuite obtenues par :

$$Q_s = P \times \Pi_s \tag{IV.12}$$

avec P une matrice de mutation standard (type WAG, LG, PFASUM), Π_s la matrice diagonale dont les coefficients sont donnés par les fréquences π_s pour chaque acide aminé à la position étudiée. A cette étape, nous considérons simplement que les fréquences sont égales à la moyenne des fréquences attendues à chaque position pour l'ensemble des protéines contemporaines évaluées.

Ces étapes permettent d'obtenir des matrices de mutation site-dépendantes dont l'attendu dépend de la position du résidu dans la séquence. Ces matrices sont utilisées pour optimiser les longueurs de branche de la phylogénie étudiée et obtenir une première reconstruction marginale des résidus des séquences ancestrales.

IV.1.2.2.2.4 Déterminer les résidus reconstruits sans ambiguïté Dans un deuxième temps, nous déterminons la confiance dans la reconstruction des résidus ancestraux. Pour cela, nous effectuons une nouvelle prédiction par GEMME sur toutes les séquences ancestrales de maximum de vraisemblance obtenues à la première étape et nous considérons que ces séquences, additionnées aux séquences contemporaines, donnent une bonne approximation de la diversité potentielle des séquences que nous cherchons à reconstruire. Pour chaque ancêtre et chaque position de l'alignement, nous testons si l'acide aminé inféré à la première étape reste l'acide aminé le plus probable sous une hypothèse d'un effet épistasique fort. Nous effectuons pour cela la reconstruction en prenant le profil de fréquences à l'équilibre à cette position qui permet de maximiser le rapport

$$\frac{\pi_{aa_{max}}}{\pi_{aa_{test}}}$$

avec aa_{max} l'acide aminé inféré à l'étape 1, et aa_{test} l'acide aminé qu'on veut tester en candidat potentiel à l'épistasie. Maximiser ce rapport permet d'augmenter la probabilité de muter vers aa_{max}, ce qui permet d'augmenter la probabilité marginale de partir depuis un autre acide aminé à l'état ancestral.

Tester chaque acide aminé pour chaque ancêtre est une étape longue. Pour cette raison, nous commençons par tester si la quantité :

$$q = log(\frac{Pr_{aa_{test}}}{Pr_{aa_{max}}} \times \frac{\pi_{aa_{max}}}{\pi_{aa_{test}}} \times \frac{\pi_{aa_{test,init}}}{\pi_{aa_{max,init}}}) \tag{IV.13}$$

est inférieure à un certain seuil déterminé à -2. Cette quantité évalue simplement le ratio entre les probabilités marginales reconstruites à la première étape pour les deux acides aminés ($\frac{Pr_{aa_{test}}}{Pr_{aa_{max}}}$), le ratio entre les fréquences à l'équilibre sous l'approximation initiale ($\frac{\pi_{aa_{test,init}}}{\pi_{aa_{max,init}}}$) et le ratio entre les fréquences à l'équilibre sous la nouvelle hypothèse où $Pr_{aa_{test}}$ doit être maximisée ($\frac{\pi_{aa_{max}}}{\pi_{aa_{test}}}$). Cette quantité sert à éviter de calculer les cas triviaux, pour lesquels la probabilité marginale de aa_{max} est élevée et où le ratio $\frac{\pi_{aa_{max}}}{\pi_{aa_{test}}}$ ne change pas assez pour mettre en doute le résultat obtenu à la première étape. Si pour une position $q > -2$, alors nous testons rigoureusement les autres acides aminés candidats à cette position par le vrai calcul probabiliste, en commençant par les acides aminés ayant recueilli la probabilité marginale la plus élevée à la première étape.

Cette deuxième étape nous permet ainsi de dresser la liste des positions certaines et des positions ambiguës, mais aussi de dresser la liste des acides aminés candidats aux noeuds de l'arbre pour des positions ambiguës. Nous réduisons par conséquent considérablement l'espace des possibles à cette étape.

IV.1.2.2.2.5 Déterminer les groupes de résidus fortement dépendants les uns des autres Lors de la dernière étape, nous essayons de déterminer quelles positions doivent être inférées simultanément. En probabilités, deux événements A et B sont dits indépendants si $Pr(A \cap B) = Pr(A) \times Pr(B)$. On souhaite par conséquent déterminer pour quelles positions d'acides aminés on peut considérer que $Pr(aa_1 \cap aa_2) \simeq Pr(aa_1) \times Pr(aa_2)$. Dans notre modèle, cela équivaut à ce que les fréquences à l'équilibre de aa_1 ne dépendent pas des fréquences à l'équilibre de aa_2 et réciproquement.

On va donc effectuer plusieurs prédictions par GEMME sur les séquences ancestrales. En partant des séquences reconstruites à l'étape 1, nous allons perturber chaque position ambiguë en remplaçant l'acide aminé correspondant par l'un des acides aminés candidats déterminés à cette position. Nous voulons savoir si les profils de fréquences à l'équilibre sur les autres positions ambiguës sont sensiblement modifiés par cette mutation. Pour déterminer si deux profils de fréquence sont similaires, on calcule

$$d(1,2) = \sqrt{\sum_{n=1}^{20} (\pi_{n,1} - \pi_{n,2})^2} \tag{IV.14}$$

avec $\pi_{n,1}$ et $\pi_{n,2}$ les fréquences attendues de l'acide aminé à la position n selon deux mutations différentes à la position perturbée.

Idéalement, il faudrait regrouper ensemble (par transitivité) toutes les positions qui ont une influence les unes sur les autres. Si la position 1 influe sur 2, et que 2 influe sur 3, on constituerait par conséquent un groupe $(1, 2, 3)$. En pratique on peut rapidement arriver à regrouper beaucoup d'acides aminés ensemble et à constituer des groupes de plusieurs dizaines d'acides aminés. Pour des raisons calculatoires, on ne souhaite pas avoir des groupes de plus de 6 acides aminés et il faut déterminer une valeur seuil th tel que si $d(1,2) < th$, on considère que les profils sont globalement similaires (et donc que les mutations 1 et 2 n'influent pas sur le profil de fréquences de la position testée), et que si $d(1,2) \geq th$, les profils ne sont pas similaires. On va chercher th par dichotomie de manière à avoir le plus de regroupement possibles d'acides aminés, dans la limite d'un nombre de sites par groupe inférieur à 6 acides aminés.

Enfin, les probabilités marginales des acides aminés aux positions ambiguës sont recalculées. Pour les singletons qui ne dépendent de la nature d'aucune autre position, on calcule simplement la probabilité avec la longueur de branche optimisée à la première étape, et en prenant comme matrice de mutation sur chaque branche, à chaque position s,

$$Q_s = \frac{\Pi_{s,1} + \Pi_{s,2}}{2} \times P \tag{IV.15}$$

Q_s est par conséquent rééquilibrée selon la moyenne des fréquences à l'équilibre des noeuds à chaque extrémité de la branche.

Pour les positions regroupées, nous construisons un échantillonnage des acides aminés potentiels par l'algorithme de Monte Carlo. Ces positions regroupées constituent une sous-séquence (contrainte au maximum à 6 acides aminés). Ces sous-séquences ont des degrés variables d'ambiguïté pour chaque ancêtre. A certains noeuds et certaines positions, les résidus les plus probables sont stables, tandis qu'à d'autres noeuds, certaines positions de la sous-séquence peuvent avoir plusieurs solutions potentielles déterminées précédemment. Nous voulons savoir quels sont les acides aminés les plus probables, dans une évaluation conjointe de tous les résidus de la sous-séquence qui ont des influences épistasiques les uns sur les autres.

Les petites séquences sont initialisées avec les acides aminés ayant obtenu la meilleure probabilité à l'étape 1. Nous calculons la vraisemblance L_1 d'observer l'ensemble de ces sous-séquences. Ensuite, nous sélectionnons un des ancêtres au hasard selon une probabilité

qui dépend du nombre total d'acides aminés ambigus pour les ancêtres (plus la séquence d'un ancêtre a un nombre important d'acides aminés alternatifs possibles, plus elle a de chance d'être sélectionnée). La sous-séquence de cet ancêtre est modifiée en choisissant une nouvelle sous-séquence composée des acides aminés candidats à chaque position pour cet ancêtre. La vraisemblance avec cette nouvelle sous-séquence L_2 est calculée. Si

$$\alpha < exp(L_2 - L_1)$$

avec α une variable aléatoire tirée uniformément sur $[0, 1]$, alors les états ancestraux sont transférés sur cette nouvelle solution, et l'on passe à l'itération suivante. Ce processus est réitéré 20,000 fois. Les sous-séquences les plus probables pour chaque ancêtre sont obtenues en sous-échantillonnant dans les différents états possibles des sites candidats, en ignorant les 20% premières itérations, puis en retenant 100 éléments dans la distribution obtenue (un élément tous les 160, considéré comme un tirage indépendant).

Dans le calcul de la vraisemblance d'un ensemble de sous-séquences, nous avons besoin de calculer la probabilité de passer d'une sous-séquence s_i à une sous-séquence s_f le long d'une branche en prenant en compte les fréquences à l'équilibre de toutes les positions simultanément. Comme les mutations des positions influencent les fréquences à l'équilibre des autres positions, ceci n'est pas trivial.

Pour l'instant, la solution retenue est de diviser la branche de longueur t_f qui sépare s_i et s_f en plus petits fragments t sur lesquels la probabilité qu'une sous-séquence compte plus d'une mutation soit inférieure à 1%. Pour chaque pas de temps t_n, nous appelons l'ensemble des sous-séquences possibles s_n, pour lesquelles nous connaissons la probabilité. Initialement, nous partons de s_i, qui correspond à l'unique sous-séquence initiale et qui a une probabilité de 1. Ensuite, nous faisons évoluer les sous-séquences sur chaque fragment de temps $t_n \rightarrow t_{n+1}$ en calculant uniquement les probabilités de muter vers un ensemble de sous-séquences s_{n+1} à t_{n+1} qui ne présentent qu'une seule mutation par rapport aux sous-séquences possibles s_n à t_n. Ces probabilités sont calculées en prenant les fréquences à l'équilibre des acides aminés pour chaque sous-séquence possible de s_n. Nous obtenons donc un ensemble de sous-séquences s_{n+1} avec leurs probabilités $Pr_{s_{n+1},t_{n+1}}$. Le problème de cette méthode est qu'après quelques itérations, nous nous retrouvons avec un nombre très important de sous-séquences possibles, alors que la plupart ont des probabilité très faibles et dérivent vers des états qui s'éloignent de s_f et qui par conséquent ne sont pas pertinentes. Afin de limiter le nombre de sous-séquences générées à chaque itération, nous filtrons les sous-séquences à s_n pour ne conserver que celles qui ont une probabilité non négligeable d'aboutir à s_f sur le temps $t_f - t_n$ qu'il reste à parcourir sur la branche :

$$Pr = Pr_{s_n,t_n} \times Pr_{s_n \rightarrow s_f, t_f - t_n} \tag{IV.16}$$

avec $Pr_{s_n \rightarrow s_f, t_f - t_n}$ la probabilité approximative qui est obtenue en considérant les fréquences à l'équilibre attendues pour les acides aminés dans le contexte des sous-séquences retenues à s_n.

Le principe de ce calcul est donc de discrétiser l'évolution du système sur des petits pas de temps sur lesquels il est facile de calculer la probabilité d'évolution en connaissant la sous-séquence dans laquelle apparaît la mutation. Les sous-séquences obtenues sont ensuite filtrées en ne conservant que celles qui ont le plus de chance d'arriver sur la sous-séquence

s_f à t_{tot}, afin d'éviter que le nombre de sous-séquence simulées par GEMME ne deviennent trop important.

IV.1.2.2.3 Modèle "Struct2"

Le modèle Struct2 tient compte du contexte structural des acides aminés dans la reconstruction des protéines. Les structures secondaires des protéines ancestrales sont d'abord inférées sur la base des structures secondaires contemporaines. Ces informations sont utilisées dans un second temps pour calculer les probabilités d'apparition des acides aminés ancestraux dans un contexte structural donné. Ainsi, les probabilités de mutation des acides aminés diffèrent selon qu'ils appartiennent à une hélice α, un feuillet β ou un tour, et selon qu'ils soient enfouis ou exposés. La reconstruction des séquences ancestrales est donc effectuée selon le protocole suivant :

1. Alignement des séquences protéiques, et conversion des séquences protéiques dans la structure secondaire. Un résidu prend alors un état qui peut être "hélice α", "feuillet β", "tour", et qui peut être exposé ou enfoui. L'obtention des structures secondaires et de l'enfouissement associés à chaque résidu des protéines contemporaines est obtenu avec NetSurfP 2.0 (B. Petersen et al., 2009) ;
2. Reconstruction jointe des structures secondaires ancestrales sur la base des alignements de structures secondaires contemporaines. Les matrices utilisées pour cette reconstruction ont été obtenues dans le cadre de cette thèse. Elles consistent en 21 matrices, une pour chaque contexte structural immédiat possible flanquant la position reconstruite (par exemple hélice exposée-hélice exposée, hélice exposée-hélice enfouie, tour exposé-feuillet enfoui...) ;
3. Reconstruction marginale des acides aminés ancestraux. Dans ce cas, les matrices utilisées sont empruntées à Si Quang Le et Olivier Gascuel, 2010. Ces matrices décrivent les probabilités de mutation des acides aminés selon leur appartenance à un type de structure secondaire, exposé ou enfoui. La reconstruction des structures ancestrales secondaires à l'étape précédente permet de savoir quelle matrice de mutation d'acides aminés doit être utilisée à cette étape.

IV.1.2.2.3.6 Reconstruction des structures secondaires ancestrales Lai et al., 2020 ont constitué une base de données d'alignements de protéines dans le but de calculer une matrice de mutation entre structures secondaires (Supporting Information Table S2 du-dit article). Les auteurs assignent chaque position à une structure secondaire sur la base des structures cristallographiques disponibles dans la PDB : Beta-bridge, Beta-strand, Alpha-Helix, Helix-3, Helix-5, Bend, Turn. Nous réutilisons ces alignements d'acides aminés, mais les structures sont prédites à nouveau sur les reconstructions ancestrales grâce à NetSurfP 2.0, afin d'obtenir le caractère exposé/enfoui pour chaque acide aminé et de réduire le nombre de classes à trois (hélice α, feuillet β ou tour) au lieu des sept initialement utilisées. Cette réduction permet d'utiliser dans la seconde partie les matrices de mutation obtenues par Si Quang Le et Olivier Gascuel, 2010, qui décrivent les probabilités de mutation des acides aminés aux sein des feuillets, hélices et tours exposés ou enfouis.

Ces alignements nous ont permis de construire des matrices de mutations instantanées qui décrivent les probabilités de mutations entre structures secondaires enfouies ou exposées, selon la méthode employé par Jones, Taylor and Thornton (Jones, W. R. Taylor et J. M. Thornton, 1992). La méthode est détaillée de façon très précise dans les données supplémentaires de l'article de Lai et al., 2020.

Dans le comptage des mutations de structures secondaires, nous séparons les mutations au sein de 21 matrices différentes selon le contexte structural du résidu observé. Ces contextes correspondent au cas où le résidu est flanqué de part et d'autre d'une combinaison entre structures secondaires (hélice, feuillet ou tour) qui peuvent être soit exposées, soit enfouies. Par exemple, la matrice de mutations correspondant à une structure entourée par deux hélices exposées est différente de la matrice d'une structure entourée par une hélice enfouie et un tour exposé. Cette distinction a notamment pour objectif de distinguer les résidus en bordure de structure secondaire de ceux pris au sein d'une structure secondaire homogène. Chaque matrice contextuelle M est normalisée tel que :

$$Q = -\sum_{i=1}^{20} \pi_i a_{i,i} \times M$$

avec π_i la fréquence à l'équilibre pour la structure secondaire i, $a_{i,i}$ le $i^{\text{ème}}$ coefficient diagonal de la matrice M. Cette normalisation permet d'obtenir des matrices pour lesquelles une longueur de branche de 1 unité correspond en moyenne à une mutation de structure secondaire.

A partir de ces matrices, nous optimisons la longueur des branches de l'arbre phylogénétique par maximum de vraisemblance (ML) afin d'obtenir les probabilités marginales des structures secondaires à chaque position de l'alignement, à partir d'un alignement des structures secondaires des protéines contemporaines. Afin de déterminer quelle matrice de contexte structural utiliser à chaque position dans la reconstruction, nous commençons par initier la reconstruction en attribuant à chaque résidu, pour chaque noeud de l'arbre, une structure secondaire par un calcul rapide de reconstruction jointe selon l'algorithme décrit par Pupko et al., 2000, en prenant uniquement en considération le contexte structurel en 5' du résidu et sans optimisation des branches. L'initiation des contextes est utilisée ensuite pour calculer les probabilités marginales des structures secondaires de chaque résidu en prenant en compte le contexte structural en 5' et 3' du résidu. L'optimisation de la probabilité de muter à un site donné alterne entre une étape d'optimisation des longueurs de branche et une étape de détermination des probabilités jointes pour chaque résidu. Le calcul exact des probabilités des structures secondaires n'est pas possible en prenant seulement en compte le contexte à droite et à gauche de chaque résidu. En effet, cette condition impose une dépendance de toutes les probabilités les unes par rapport aux autres à la fois le long des séquences et entre les noeuds de l'arbre. Pour palier ce problème, nous considérons plus simplement qu'à chaque tour d'optimisation, la matrice contextuelle à utiliser pour calculer les probabilités de changement d'un résidu à un noeud est déterminé par les structures secondaires flanquant le résidu ayant les probabilités jointes maximum. Ainsi, l'état des structures secondaires flanquantes, à la différence de leur probabilité exacte, se stabilise rapidement et le calcul de la vraisemblance sur l'alignement est possible. Ensuite, le long d'une branche, la probabilité de mutation d'un vecteur de probabilités de structures

V depuis un résidu dans le contexte A vers un contexte B sur un temps t est donnée par :

$$Pr = V_i \times Q_A^{t/2} \times Q_B^{t/2} \tag{IV.17}$$

avec Q_A et Q_B les matrices de mutations instantanées des structures secondaires dans les contextes A et B. Si la probabilité jointe maximum des structures flanquantes diffère à une position par rapport à l'itération précédente, alors les probabilités jointes des voisins de cette position sont calculées à nouveau en prenant en compte l'information sur leur nouveau contexte, jusqu'à obtenir une solution stable. En pratique, cela arrive rapidement car de nombreuses positions de l'alignement sont conservées et ont donc des probabilités jointes élevées sur une unique solution de reconstruction ancestrale de la structure secondaire de la protéine. Ces positions d'invariants servent de point d'ancrage dans l'optimisation de la structure le long des séquences.

Nous faisons donc une approximation sur le calcul de ces probabilités, en considérant globalement que le long d'une branche, les structures secondaires des résidus mutent peu (Lai et al., 2020). Par conséquent, nous supposons que le changement potentiel de contexte flanquant autour d'un résidu se produit au plus une seule fois par branche, en moyenne au milieu de cette branche. Cette approximation est acceptable pour les protéines que l'on veut étudier, avec peu de variabilité dans les structures secondaires entre les différentes espèces et des temps de divergence courts. Si le modèle devait être appliqué à des protéines dont le repliement subit d'importants changements sur des temps longs, cette approximation pourrait ne plus être valable.

IV.1.2.2.3.7 Reconstruction des acides aminés ancestraux Les structures secondaires ancestrales inférées à la première étape permettent la reconstruction des acides aminés ancestraux en prenant en compte le contexte structural dans lequel ont évolué les acides aminés. La probabilité de mutation des acides aminés entre eux selon le contexte structural est obtenu grâce aux matrices de mutation contexte-dépendant établies par Si Quang Le et Olivier Gascuel, 2010. Le calcul des probabilités marginales des acides aminés pour chaque ancêtre à chaque noeud de l'arbre est effectué après optimisation des paramètres du modèle phylogénétique par maximum de vraisemblance. Ici, en plus de l'optimisation des branches, on ajoute un paramètre Γ afin de simuler l'hétérogénéité des vitesses d'évolution au sein des sites. Le paramètre est discrétisé en quatre classes, et la valeur moyenne du paramètre Γ au sein de chaque classe est utilisée, selon la proposition de Yang, 1994.

Le modèle comporte donc comme paramètres le nombre de branches ainsi que le paramètre Γ. L'optimisation des structures secondaires effectuée à la première étape n'intervient pas dans les paramètres du modèle, puisque les structures secondaires des séquences ancestrales optimisées à la première étape sont utilisées comme donnée fixe dans cette nouvelle optimisation.

De façon analogue à la première étape, les probabilités de mutation d'un site le long des branches sont obtenues en calculant :

$$Pr = \prod_{i=1}^{n} \sum_{k=1}^{4} V_i \times Q_{s_1,i}^{g_k.t/2} \times Q_{s_2,i}^{g_k.t/2} \tag{IV.18}$$

avec n la longueur des séquences, $Q_{s_1,i}$ et $Q_{s_2,i}$ les matrices de mutations instantanées d'un acide aminé dans les structures secondaires de la séquence mère et de la séquence fille à la position i, connues grâce à l'étape d'inférence des structures secondaires ancestrales. V_i représente le vecteur de probabilités de l'acide aminé en position i du noeud ancestral, et g_k la valeur moyenne de l'une des quatre classes du paramètre Γ. Les matrices de mutation des acides aminés selon différents contextes structuraux (hélice α, feuillet β, tour, soit exposé, soit enfoui) ainsi que les fréquences à l'équilibre attendues dans ces contextes ont été calculées par Si Quang Le et Olivier Gascuel, 2010 par optimisation de la vraisemblance sur un ensemble d'alignements. Le modèle optimisé présente donc le paramètre Γ et les longueurs de branches. Les probabilités marginales des acides aminés aux différents noeuds de l'arbre, nécessairement calculées pendant l'optimisation du modèle, sont alors récupérées et enregistrées.

IV.1.3 Résultats et Discussion

IV.1.3.1 Optimisation des matrices thermiques JAMA

Une première optimisation a été effectuée afin de définir une matrice de biais thermique pouvant décrire l'évolution d'une séquence d'un organisme mésophile vers un organisme thermophile à partir de la matrice PFASUM. Ceci est effectué en calculant l'équation IV.5, pour chacune des 155 paires de séquences mésophiles/thermophiles provenant de Szilágyi et Závodszky, 2000 et T. J. Taylor et Vaisman, 2010. Pr_{i_1} et Pr_{i_2} sont les probabilités des acides aminés dans les séquences 1 et 2 de la $i^{\text{ème}}$ paire de séquences, et δ_i et β_i deux valeurs, trouvées par maximum de vraisemblance pour maximiser Pr_i, qui correspondent au temps de divergence écoulé selon l'axe "mésophile" corrélé à la matrice de mutation P, et selon l'axe "thermophile" décrit par la matrice Q. Les résultats pour Q correspondant à cette optimisation pour chaque paire de séquences sont montrés en figure IV.1. Après optimisation, nous avons calculé le paramètre α qui correspond à l'influence de la matrice HJM dans la matrice de mutation M qui relie les séquences issues d'organismes mésophiles et les séquences issues d'organismes thermophiles, telle que $M = \delta.P + \beta.Q = [(1 - \alpha).P + \alpha.Q] \times t$. t est interprété comme le temps de divergence réel entre les séquences, et α comme le biais imposé à une matrice standard P qui permet de corriger les probabilités de mutation pour s'adapter à un contexte chaud.

On s'attend à ce que α soit corrélé à la température de vie des organismes thermophiles T_{max}, puisque le biais associé aux probabilités de mutation est plus fort pour les organismes vivants à des températures plus élevées. La figure IV.1a. montre cette corrélation (test de Pearson : $r = 0.29$, $p - value = 3.10^{-4}$), bien que certains organismes décrits comme faiblement thermophiles (température de vie inférieure à 60°C) montrent un coefficient α très élevé, pouvant aller jusqu'à 1. La figure IV.1b. peut apporter une explication à cela. Ici, le temps de divergence t est confronté à α. Normalement, on ne s'attend pas à trouver de corrélation forte entre le temps de divergence entre les séquences mésophiles et thermophiles (t), et le biais α associé à la température de vie de l'organisme thermophile. La corrélation est pourtant significative (Pearson $p - value = 0.02$), mais devient non significative si l'on retire les 27 paires de séquences pour lesquelles α est optimisé sur une valeur

extrême de 0 ou 1 (Pearson $p - value = 0.45$). Retirer ces séquences n'influence pas la significativité de la corrélation entre T_{max} et α ($p - value = 9.10^{-4}$). Cette différence de corrélation entre t et α est notamment imputable à quelques séquences que l'on peut voir sur la figure IV.1 qui proviennent d'organismes faiblement thermophiles, avec des temps de divergence courts mais des valeurs pour α maximum, à 1. Pour ces séquences, le faible temps de divergence pourrait influencer α, si la pression de sélection vers le chaud n'est pas très forte mais que les séquences évoluent loin de leur équilibre à cause du faible temps de divergence, ce qui biaise très fortement les remplacements d'acides aminés "froids" par des acides aminés "chauds". α semble se comporter comme un coefficient de sélection sur les mutations chaudes, qui peut être élevé soit si le court temps de divergence fait que les séquences évoluent encore loin de leur équilibre, soit si le temps de divergence est long mais que l'organisme vit à des températures très élevées.

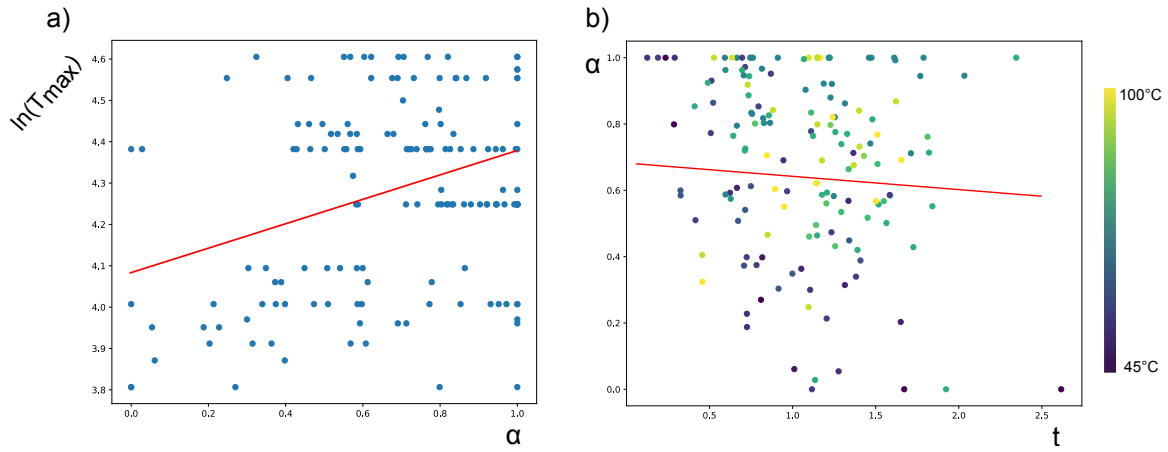


FIGURE IV.1 – Optimisation de la matrice thermique chaude HJM. (a) distribution du paramètre α, correspondant à l'influence de la matrice HJM dans la matrice de mutation M comparant les séquences issues d'organismes mésophiles à des séquences issues d'organismes thermophiles selon $M = (1 - \alpha) \times PFASUM + \alpha \times HJM$, en fonction d'un biais sur la température de vie estimée pour les organismes thermophiles. (b) distribution du temps de divergence t estimé entre les séquences par rapport à α. La couleur des points correspond à la gamme de température de vie des organismes thermophiles.

La figure IV.2 présente le résultat de cette optimisation sur une paire de séquences choisie dans le jeu de données, à savoir entre l'alcool déshydrogénase de la bactérie mésophile *Clostridium beijerinckii* et de l'archée hyperthermophile *Aeropyrum pernix* dont la température de croissance est estimée à 90°C. La figure montre la vraisemblance de l'alignement selon deux axes t_1 et t_2, calculée selon :

$$L = -\sum_{i=1}^{n} log(Pr_{i_1} \times e^{t_1.PFASUM+t_2.HJM} \times Pr_{i_2})$$

La figure IV.2a. montre la corrélation obtenue entre la matrice HJM optimisée par l'équation IV.2 et la matrice PFASUM, tandis que la figure b. montre l'évolution de cette corrélation lorsque HJM est optimisé par l'équation IV.5. On constate dans le panneau a. que les deux matrices sont fortement corrélées. En effet, la matrice HJM obtenue par l'équation IV.2 contient l'information de toutes les mutations aléatoires entre les deux séquences de la protéine. HJM contient PFASUM, et les deux matrices sont donc très redondantes. Sur le panneau IV.1b., on constate que le puits de vraisemblance est plus circulaire, ce qui traduit une décorrélation partielle des deux axes. La part associée à la divergence "thermophile", soit t_2, a augmenté au détriment de la part associée aux mutations "mésophiles" t_1. On note en outre que le champs de la vraisemblance se comporte bien, avec un puits unique de vraisemblance et l'absence de minima locaux. Ce comportement après optimisation selon l'équation IV.5 est observé pour toutes les séquences du jeu de données. Bien que l'on ne puisse pas complètement décorréler les deux matrices, la vraisemblance du modèle construite avec l'équation IV.5 est meilleure et la matrice HJM issue de ce modèle est conservé.

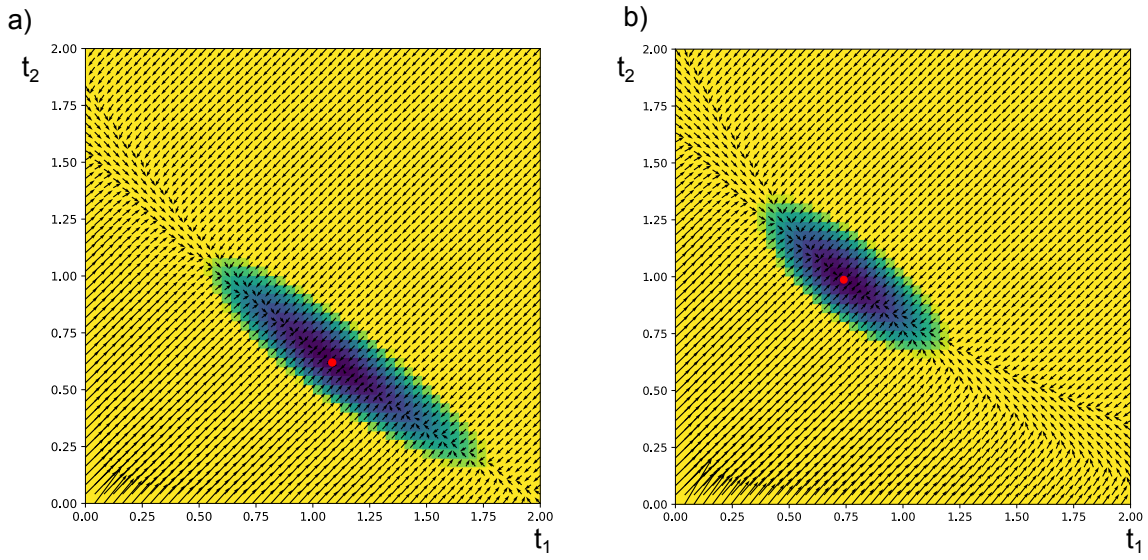


FIGURE IV.2 – Vraisemblance des temps de divergence t_1 et t_2 pour l'alcool déshydrogénase de l'archée hyperthermophile *Aeropyrum pernix* estimée vivre à 90°C, et la protéine mésophile à l'alcool déshydrogénase de *Clostridium Beijerinckii* estimée vivre à 37°C. t_1 correspond au temps de divergence associé à la matrice de mutation "mésophile" PFASUM, et t_2 à la matrice optimisée "thermophile" HJM. Les couleurs les plus sombres indiquent des vraisemblances plus élevées. Le point rouge correspond à la valeur de maximum de la vraisemblance pour le couple t_1, t_2. Les flèches correspondent au champs vectoriel de la dérivée de la vraisemblance. (a) Vraisemblance en utilisant une matrice HJM optimisée selon l'équation IV.2. (b) Vraisemblance en utilisant une matrice HJM optimisée selon l'équation IV.5.

Les mêmes opérations ont été effectuées pour obtenir une matrice CJM, qui décrit les mutations associées au refroidissement de l'environnement de vie des organismes. Comme nous n'avons pas trouvé de base de données similaire pour obtenir des séquences provenant d'organismes strictement psychrophiles, nous avons construit un modèle utilisant l'équation IV.5, mais en inversant la place des séquences thermophiles et mésophiles pour tenter de faire apparaître le biais de mutation à imposer à la matrice standard P pour rendre compte de la fréquence plus (ou moins) élevée de certaines mutations associées aux séquences des espèces vivant dans des milieux plus froids. La question s'est posée de savoir si cette optimisation devait être effectuée en utilisant la matrice PFASUM comme référence, ou bien une matrice HJM obtenue précédemment (soit par l'équation IV.2, soit par l'équation IV.5). Comme nous voulons obtenir une matrice CJM idéalement indépendante de CJM, et que les tests d'optimisations de CJM utilisant PFASUM ou HJM en référence ont donné des résultats très similaires, mais l'utilisation de HJM semble sur-paramétriser le modèle, j'ai préféré utiliser une combinaison PFASUM+CJM.

La figure IV.3 montre les fréquences à l'équilibre pour les matrices HJM et CJM. Ces fréquences correspondent à la composition moyenne en acides aminés attendue pour une séquence qui évoluerait un temps infiniment long à des températures extrêmes selon les probabilités données par la matrice en question. Les matrices HJM et CJM montrent une déviation importante des fréquences à l'équilibre par rapport aux matrices standard PFASUM et WAG (qui pour leur part sont très similaires l'une et l'autre). Ainsi, les fréquences attendues selon HJM sont notablement plus élevées pour les acides aminés Alanine (A), Arginine (R), Glutamate (E), Leucine (L), Proline (P), Valine (V), et plus faibles pour l'Asparagine (N), Aspartate (D), Cystéine (C), Glutamine (Q), Histidine (H), Sérine (S), Thréonine (T). La matrice CJM a des fréquences enrichies en Alanine (A) et Valine (V), et plus faibles en Arginine (R), Glutamate (E), Lysine (K), Proline (P), Tryptophane (W) et Tyrosine (Y). Bien que certains acides aminés (R, E, P) semblent être représentés de manière antagoniste entre les deux matrices, les matrices ne sont pas pour autant symétriques car certains acides aminés (A et V) sont enrichis dans les deux conditions. Globalement, la déviation dans les fréquences à l'équilibre est plus importante dans la matrice HJM par rapport à WAG ou PFASUM qu'elle ne l'est pour CJM. A ma connaissance, il n'existe pas de base de données importante contenant des séquences issues d'organismes psychrophiles comme il en existe pour les organismes thermophiles. L'étude la plus complète est celle de Metpally et Reddy, 2009. Elle utilise 2816 protéines psychrophiles, mais issues uniquement de six organismes ce qui pourrait biaiser certaines observations que l'on voudrait corréler avec la température. Pour cette raison, nous avons préféré essayé de tirer les mutations "froides" en forçant l'évolution de séquences thermophiles vers des séquences mésophiles. La matrice CJM obtenue reprend certaines caractéristiques relevées par Metpally et Reddy, 2009 (plus de petits acides aminés hydrophobes comme A et V, moins d'acides aminés chargés E, R, K et d'acides aminés aromatiques W et Y) mais l'effet n'est pas flagrant. Au contraire, la matrice HJM reprend bien les caractéristiques attendues des organismes (hyper)thermophiles (moins d'acides aminés polaires S, T, N, Q et plus d'acides aminés chargés positivement R et E, Szilágyi et Závodszky, 2000) qui dans le jeu de séquences initiales sont représentés par des archébactéries et bactéries. En outre, les fréquences d'acides aminés globalement surreprésentés chez les psychrophiles selon Metpally et Reddy, 2009 (A, D, S, T) ou moins représentés (E, L) sont inversées dans la matrice HJM. Par conséquent, la matrice HJM contient une forte information liée à la température de vie des organismes thermophiles,

mais pour la matrice CJM, si certains remplacements corrèlent avec l'attendu d'organismes psychrophiles, il est possible que son optimisation ne soit pas bien corrélée avec la température.

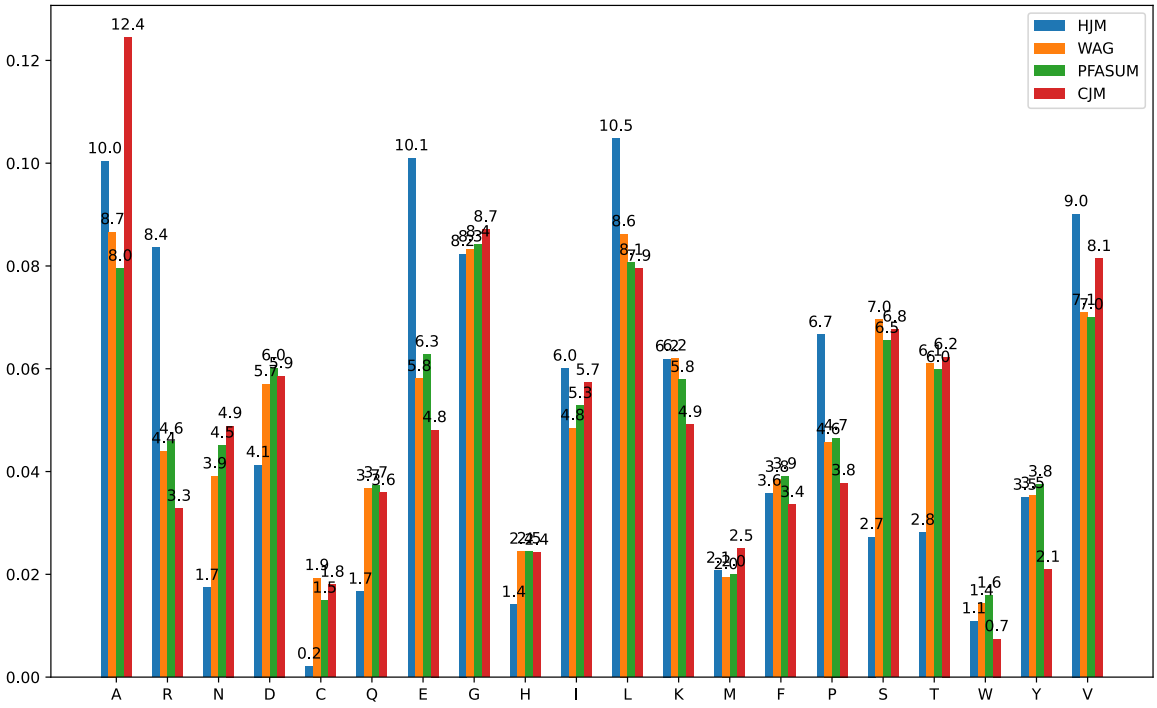


Figure IV.3 – Fréquences à l'équilibre des différents acides aminés pour les matrices HJM et CJM. Les fréquences à l'équilibre correspondant aux matrices PFASUM et WAG utilisées dans les différentes analyses sont également reportées pour référence.

La figure IV.4 montre les matrices de mutations instantanées HJM et CJM, qui sont comparées avec la matrice standard PFASUM. Les valeurs indiquées correspondent aux log_{odds}, c'est à dire à la probabilité d'observer d'un changement orienté entre deux acides aminés i et j par rapport à la probabilité d'observer cette mutation si les mutations sont aléatoires. En d'autres termes,

$$log_{odd,i,j} = log(\frac{\pi(i) \times P(i,j)}{\pi(i) \times \pi(j)})$$

avec π correspondant aux fréquences à l'équilibre pour l'acide aminé i. Nous utilisons les fréquences à l'équilibre correspondant à la matrice PFASUM, et les probabilités de mutation pour des séquences divergentes à 1% (temps évolutif relativement court).

La figure IV.4 montre que les matrice HJM et CJM diffèrent surtout par des différences dans les probabilités de mutation qui concernent des couples spécifiques d'acides aminés. Le plus flagrant est que les mutations vers le tryptophane (W) sont toutes plus probables selon la matrice HJM que selon CJM. Au contrainte, les mutations vers la cystéine (C) sont bien moins probables selon HJM que selon CJM. Un autre point important est que ces matrices ne sont pas optimisées de telle sorte que le modèle soit réversible dans le temps. Ainsi, pour CJM, les probabilités qu'un tryptophane mute vers un autre acide aminé ne sont pas très différentes de la matrice PFASUM, mais les probabilités qu'un acide aminé mute vers un tryptophane sont notablement plus faibles. C'est un point de vigilance dans l'optimisation du modèle dans le cadre des reconstructions ancestrales, puisqu'un certain nombre d'algorithmes (par exemple pour optimiser les longueurs de branche) utilisent usuellement la propriété que le modèle est réversible dans le temps pour permuter des branches et faciliter les calculs, ce que nous ne pouvons pas faire ici.

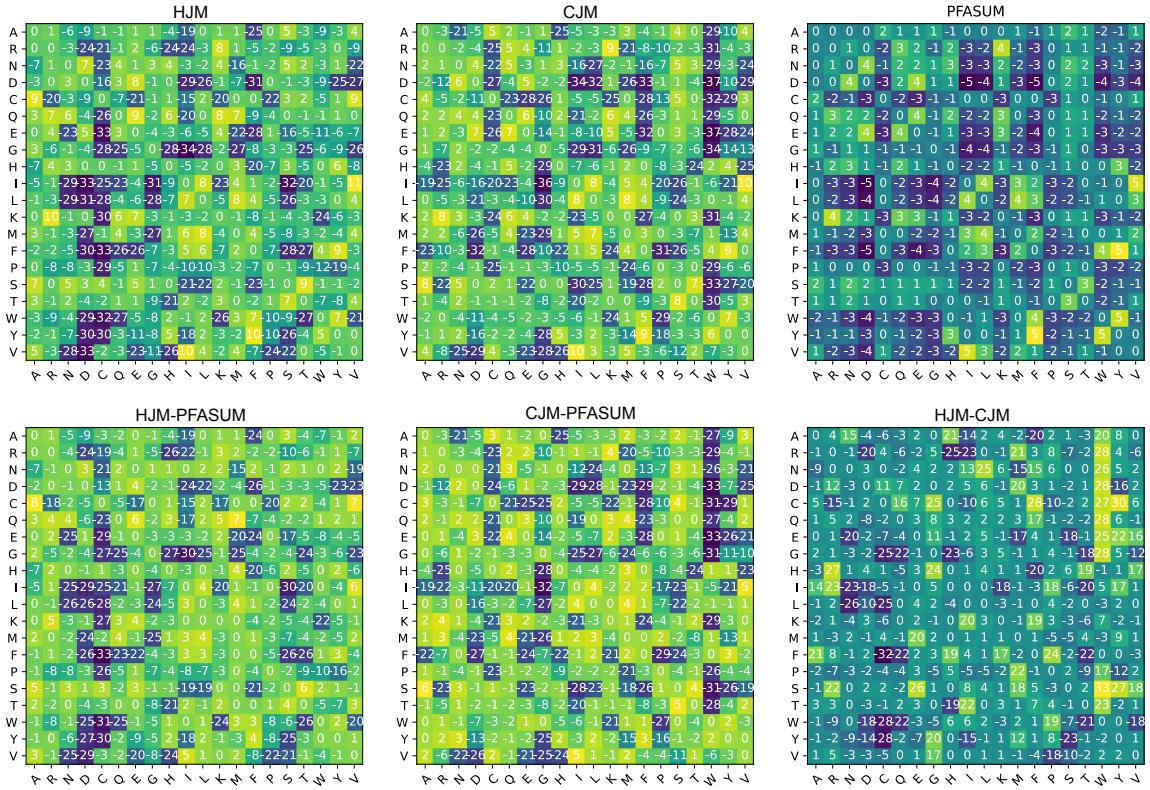


Figure IV.4 – Matrices HJM et CJM présentées sous forme de log_{odds}. Une valeur positive à la ligne i, colonne j, indique une probabilité plus élevée que l'attendu neutre d'observer une mutation de l'acide aminé i vers j, tandis qu'une valeur négative indique une probabilité plus faible. Sur la deuxième ligne, les contrastes entre les matrices HJM et CJM sont montrés par rapport à la matrice PFASUM, ou entre HJM et CJM. Dans ce cas, une valeur positive indique que la probabilité de mutation dans la matrice thermique est plus élevée que dans la matrice standard, et inversement si la valeur est négative. Pour la matrice HJM-CJM, il s'agit du même procédé où l'on compare les probabilités de muter selon HJM par rapport à CJM.

Ces matrices HJM et CJM ont ensuite été utilisées pour optimiser les longueurs de branches sur les trois protéines d'intérêt étudiées au chapitre 2, ainsi que la protéine RFP de référence dont l'évolution expérimentale a été mise en place et suivie par Randall et al., 2016. Cette dernière protéine n'a pas évolué dans des environnements thermiquement contrastés. Elle nous sert en quelque sorte de témoin négatif dans le cas où l'on est certain que la protéine a évolué dans des organismes à température constante. Les résultats de cette optimisation pour les quatre protéines sont montrés sur la figure IV.5. On constate tout d'abord que les branches sont globalement très biaisées vers l'une des matrices HJM ou CJM, ce qui indiquerait des changements de température de vie importants entre les différents ancêtres. Si l'on se concentre sur la RFP, on voudrait idéalement que toutes les branches soient noires, ce qui traduirait un processus d'évolution neutre au regard de la température. L'optimisation attribue une part de l'ordre de 30 à 50% de matrice chaude HJM aux différentes branches. Ceci pourrait traduire le fait que les *Escherichia coli* utilisées pour cette expérience ont été maintenues à 37°C, ce qui est déjà relativement élevé. L'optimisation du modèle donnée en annexe montre que même pour des espèces vivant à une température relativement faible, le poids de HJM sur la branche n'est pas nul, mais qu'il augmente (probablement de manière exponentielle, comme suggéré par la figure IV.1a.) avec la température de l'organisme. Si l'on compare le résultat obtenu pour la RFP avec les résultats obtenus pour les protéines issues d'Alvinellidae, il apparaît que l'optimisation de la RFP tend moins à imposer de forts contrastes thermiques. La plupart des branches de l'arbre basé sur la RFP ont environ 50% de leur longueur attribués à la matrice neutre PFASUM, ce qui n'est presque pas observé dans l'évolution des protéines chez les Alvinellidae. Aussi, la plupart des branches est optimisée de façon similaire, avec environ 30-50% attribués à la matrice HJM, 0-20% attribués à CJM. Ceci traduirait plutôt une température homogène au cours du processus évolutif, et nous renseigne également sur la manière d'interpréter ces résultats : la proportion de PFASUM, HJM et CJM dans une branche menant à un ancêtre montre les variations de température de l'organisme entre les noeuds de l'arbre.

Concernant les protéines issues d'Alvinellidae, on voudrait idéalement que les variations dans les températures estimées pour les branches soient corrélées entre les protéines, puisqu'elles sont sensées provenir des mêmes espèces contemporaines et ancestrales (bien que des incertitudes persistent sur la phylogénie et que les différents gènes peuvent ne pas avoir exactement la même histoire évolutive). Certaines branches semblent faire consensus : la branche menant à *P. gouldii* est intégralement attribuée à CJM pour la SOD1 et la cMDH, les branches menant à *M. palmata*, *Anobothrus* sp., *A. carldarei* et Terebellidae gen. sp. sont relativement similaires. Plus intéressant, la branche menant aux espèces *A. pompejana* et *A. caudata*, qui sont deux espèces pour lesquelles la thermotolérance est certaine et probablement la plus élevée de toutes les espèces incluses, est quasiment intégralement attribuée à la matrice HJM pour les quatre protéines observées (cMDH, SOD1 et les deux familles d'hémoglobines). La branche menant à l'ancêtre des Alvinellidae ne fait pas consensus entre les familles de protéines. En outre, on ne peut pas interpréter les résultats sur les branches trop courtes. En effet, les séquences étudiées ici font entre 141 et 341 acides aminés. Par conséquent, une branche de longueur 0.01 n'est que le reflet en moyenne de 1 à 4 mutations. On comprend facilement que sur des séquences si courtes avec peu de mutations enregistrées, l'optimisation vers une matrice chaude ou froide peut très vite changer du fait de l'incertitude statistique associée à la méthode. Une mesure de l'incertitude du modèle serait nécessaire dans ce cas, mais ce n'est pas l'objet d'une optimisation par maximum de

vraisemblance. Pour cette raison, une optimisation est donnée en annexe portant sur 10,000 sites de l'alignement de séquences des Alvinellidae. Sur ce plus long alignement, le modèle Ecoprior donne des résultats conformes à l'attendu par rapport aux variations de température de vie connues le long de la phylogénie (au regard de ce qu'on l'on sait de l'écologie des espèces contemporaines, ainsi que des résultats du chapitre 2 concernant la thermophilie de l'ancêtre des Alvinellidae). Le résultat de cette optimisation est discuté plus en détail dans l'annexe.

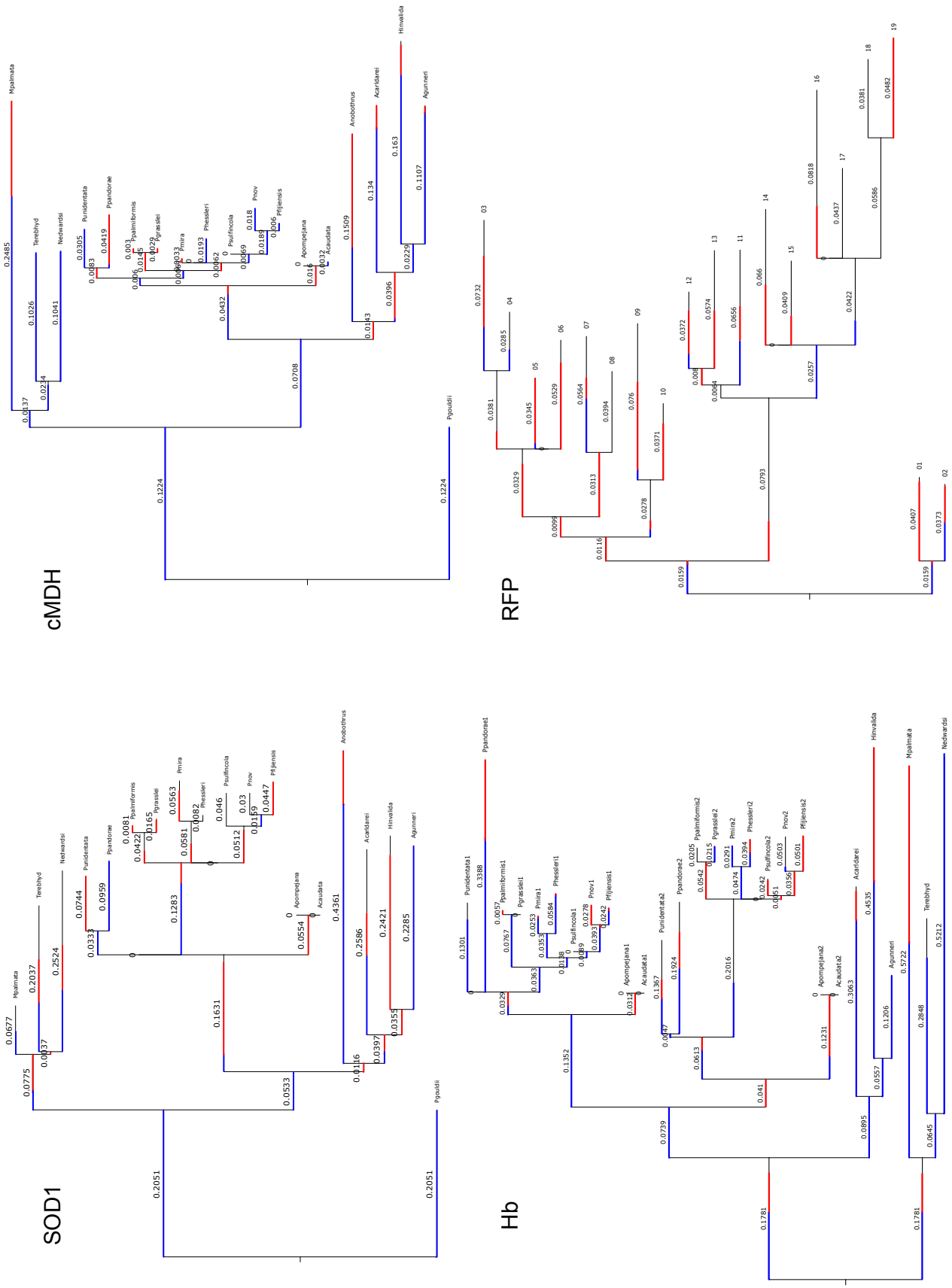


FIGURE IV.5 – Optimisation des vraisemblances appliquée à la MDHc, Hb, SOD1 et RFP par le modèle Ecoprior. Les longueurs de branche indiquent le nombre de remplacements d'acides aminés moyen attendu. La couleur des branches correspond aux poids optimisés des différentes matrices CJM (bleu), HJM (rouge) et PFASUM (noir) selon l'équation IV.1 pour chaque branche.

Une question importante est de savoir si la proportion des matrices HJM et CJM dans les branches de l'arbre phylogénétique témoignent bien de différences liées uniquement à la température. En effet, on voit que les fréquences en acides aminés à l'équilibre issues de ces matrices ont très probablement une partie du signal guidée par la température, notamment pour HJM, mais celles-ci restent partiellement corrélées avec le temps de divergence, comme montré en figure IV.2. Par conséquent, on ne peut pas exclure que la proportion de ces matrices attribuée aux branches de la phylogénie soit aussi influencée par certains coefficients de mutation dans les matrices qui ne sont pas strictement liés avec le différentiel de température. Une possibilité pour réduire cette incertitude serait de ne pas utiliser directement les matrices HJM et CJM, mais uniquement les profils de fréquence qui en sont issus et de considérer en variable explicative non pas la part de ces matrices le long des branches, mais la part des différents profils d'équilibre dans les branches (à l'instar des catégories de profils de fréquences proposés par H.-C. Wang et al., 2008, mais optimisés selon un gradient de température). Ceci permettrait de n'utiliser qu'une seule matrice de mutation avec les mêmes coefficients de mutation relatifs dans chaque colonne. Cependant, ce n'était pas l'objectif initial de l'optimisation des matrices HJM et CJM qui avait précisément pour but de déterminer des biais non réversibles au niveau des coefficients de mutation de la matrice. Ainsi, si le modèle Ecoprior permet bien de faire varier les compositions moyennes des séquences dans le temps et contient au moins partiellement le signal de la température de vie des organismes, il faut rester prudent sur le fait que ces changements de composition soient uniquement le reflet de différentiels de température notamment pour la part accordée à la matrice CJM.

IV.1.3.2 Evaluation du modèle Gempistasy

Le modèle Gempistasy s'appuie sur les prédictions effectuées par le logiciel GEMME afin d'améliorer la qualité du modèle phylogénétique utilisé pour les reconstructions ancestrales. La première étape consiste à faire correspondre les valeurs sélectives prédites par GEMME pour différentes mutations d'une protéine et fournir une quantité utilisable pour un modèle phylogénétique, comme les fréquences d'acides aminés attendues à un site.

J'ai choisi de convertir les valeurs données par GEMME, $Pred_{ind}$ et $Pred_{epi}$ en fréquences attendues *via* les équation IV.8, IV.9, IV.10 et IV.11. Nous cherchons à reproduire le comportement de GEMME développé par Laine, Karami et Carbone, 2019 qui associe un seuil α fixe pour équilibrer les prédictions $Pred_{ind}$ et $Pred_{epi}$. Ce seuil est fixé à 0,6 dans l'article original mais je cherche plutôt à le faire varier selon la dissimilitude entre les deux types de prédictions. Plus les prédictions sont dissemblables, plus on donne de poids à $Pred_{epi}$. L'idée étant qu'une mutation considérée très fréquente selon le modèle "indépendant", donc globalement très fréquente dans un alignement de protéines homologues, mais qui ne serait pas observée dans les séquences les plus similaires à la séquence d'intérêt, doit être délétère dans le contexte de la séquence d'intérêt. Au contraire, si la mutation est fréquente dans l'alignement ainsi qu'au voisinage de notre protéine dans des proportions similaires, alors la prédiction $Pred_{ind}$ est suffisante et potentiellement plus facile à interpréter en terme de fréquence attendue de la mutation.

Les paramètres intervenant dans les équations IV.9 et IV.11 sont optimisés par maxi-

mum de vraisemblance sur les phylogénies des différentes protéines étudiées. Le résultat est montré dans le tableau IV.1. L'optimisation est initialisée avec $r = 15$, $p = 0.5$, $i_1 = 1$, $i_2 = 10$, $i_3 = 5$, ce qui correspond à légèrement avantager la prédiction épistasique de façon générale par rapport à la prédiction indépendante, et à considérer que les prédictions de GEMME dans le modèle épistasique sont reliées aux fréquences attendues des acides aminés selon une courbe sigmoïde (donc les acides aminés obtenant des scores GEMME élevés ou bien faibles auront des fréquences similaires, avec une transition douce entre les deux groupes de prédictions). Le tableau IV.1 montre que l'optimisation de ces paramètres par ML permet d'améliorer très fortement la vraisemblance du modèle de phylogénie que nous voulons utiliser pour reconstruire les protéines ancestrales. Le paramètre p se stabilise à une valeur similaire pour les quatre protéines, autour de 0,27. Les autres paramètres sont en apparence plus variables, mais se stabilisent en réalité sur les mêmes comportements. Ainsi, le paramètre r est très élevé, entre 250 et 1000, mais son effet sur le calcul de α reste modéré car l'autre terme (cf. équation IV.11) varie beaucoup plus fortement. Ce paramètre r élevé indique que le modèle privilégie largement la prédiction épistasique par rapport à la prédiction indépendante. Les prédictions globalement meilleures du modèle épistasique de GEMME ont déjà été notées par les auteurs (Laine, Karami et Carbone, 2019), mais dans une moindre mesure. Les paramètres i_1, i_2 et i_3 sont sensiblement différents, mais il apparaît que dans tous les cas il y a une tendance à ce que $i_2 = i_3$. La figure IV.6 montre la forme des courbes obtenues pour ces différents paramètres avec les protéines testées. On constate que malgré des paramètres sensiblement différents, les courbes restent très similaires. Seules les mutations obtenant un score $Pred_{epi}$ supérieur à 0.96 ($Pred_{epi}$ étant ici normalisée entre 0 et 1 pour les différentes mutations testées) obtiennent une fréquence π_{epi} non nulle. La fréquence diminue très rapidement avec la baisse du score de GEMME. Ceci montre que les prédictions de GEMME, replacées dans un contexte phylogénétique, sont très pertinentes. En effet le modèle phylogénétique n'a pas besoin d'effectuer beaucoup d'ajustements sur les fréquences attendues, et l'on pourrait presque se limiter aux quelques acides aminés obtenant la meilleure prédiction GEMME pour ajuster les fréquences attendues à chaque position de l'alignement. En revanche, procéder de cette manière aurait pour conséquence de grandement restreindre l'espace exploré sur les séquences lors de la reconstruction des séquences ancestrales. J'ai donc préféré adoucir cette courbe, en considérant que les prédictions $Pred_{epi}$ au dessus de 0.7 devaient obtenir des fréquences non nulles. En revanche la nouvelle forme de la courbe est respectée.

Pour la suite de l'étude du modèle, j'ai fixé en conséquence ces paramètres aux valeurs $r = 250$, $p = 0.27$, $i_1 = 1$, $i_2 = 20$, $i_3 = 20$. L'objectif de cette étape était de trouver comment convertir les prédictions de GEMME de manière à être utilisables pour construire un modèle de phylogénie. Les paramètres utiles pour caractériser GEMME n'ont en revanche pas vocation à être ré-optimisés pour chaque phylogénie ultérieure, puisque le comportement de GEMME n'est pas censé être variable d'une protéine à une autre. Une dernière transformation sur les fréquences à l'équilibre a consisté à donner une fréquence minimum de 0,01 aux acides aminés pour lesquels la prédiction était de 0. Cela est simplement fait pour éviter des erreurs de calcul dans le modèle si des mutations prédites comme étant impossibles advenait à apparaître dans les séquences réelles. La fréquence maximale possible est par conséquent $1 - 0.01 \times 19 = 0.81$

protéine	*r*	*p*	*i1*	*i2*	*i3*	*Likelihood*
SOD	15	0.5	1	10	5	1646.88
	254.92	0.29	0.023	259.01	269.24	1337.40
Hb	15	0.5	1	10	5	2760.84
	1000*	0.23	23.67	452.88	454.32	2437.47
MDH	15	0.5	1	10	5	2797.00
	716.70	0.26	114.34	386.94	387.51	2145.45
RFP	15	0.5	1	10	5	2123.06
	361.77	0.28	0.354	147.79	153.30	1629.48

Table IV.1 – Optimisation des paramètres de GEMME. Ces différents paramètres permettent de convertir les valeurs prédites par GEMME en fréquences attendues à l'équilibre. (*) indique que l'optimisation a atteint la borne haute autorisée pour ce paramètre lors de l'optimisation.

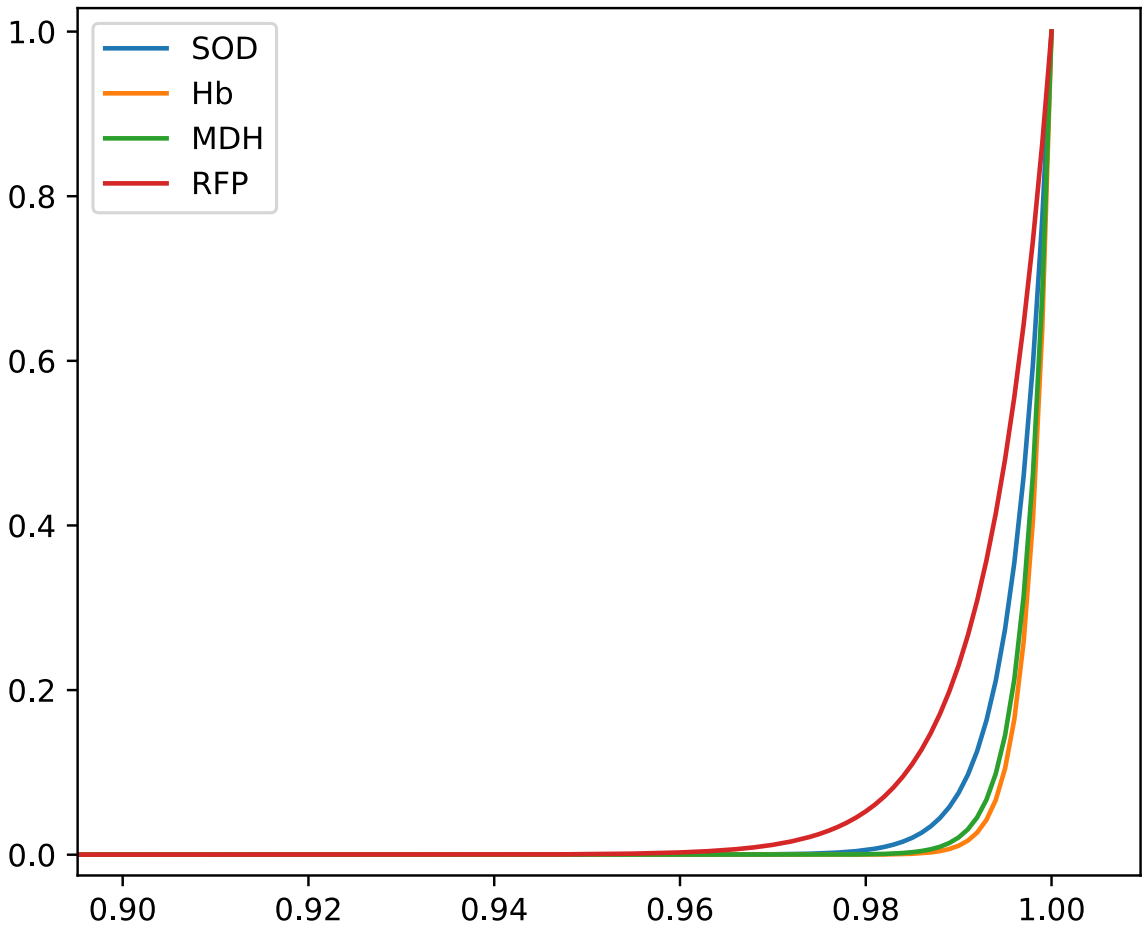


FIGURE IV.6 – Conversion de $Pred_{epi}$ en π_{epi}. Les prédictions obtenues pas GEMME, $Pred_{epi}$, sont ré-échelonnées entre 0 et 1. Cette donnée est ensuite convertie en π_{epi} *via* l'équation IV.10 dont les paramètres sont optimisés par maximum de vraisemblance. Les valeurs obtenues pour les différents acides aminés à une position sont ensuite ré-échelonnées pour que la somme des fréquences vaille 1. L'optimisation est donc techniquement effectuée sur les fréquences relatives attendues à l'équilibre.

Les figures IV.7 et IV.8 illustrent plus concrètement le résultat de ces opérations. La figure IV.7 montre, pour la MDH et la SOD, comment varie le paramètre α (qui définit les contributions de π_{epi} et π_{ind} selon les équations IV.10 et IV.11) en fonction des prédictions indépendantes et épistasiques. Chaque point représente les valeurs prédites pour les fréquences indépendantes ou épistasiques pour les différents acides aminés à toutes les positions de l'alignement. On peut considérer que les mutations proches de la bissectrice ont des prédictions similaires dans les deux modèles. Dans ce cas, α vaut 0 et on s'en remet aux fréquences prédites par le modèle indépendant. Si la prédiction est proche du point (0.01,0.81), alors la mutation est très fréquente parmi l'ensemble des séquences homologues mais peu fréquente dans les séquences les plus similaires de la séquence d'intérêt. La mutation est sans doute délétère dans ce contexte. Si la prédiction est proche du point (0.8,0.01), alors la mutation est virtuellement absente dans les séquences homologues, sauf dans les séquences plus similaires où elle est très présente. La mutation est sans doute permise dans ce contexte, ou bénéfique. Dans ces deux cas, α vaut 1 et la prédiction épistasique contribue d'avantage. Sur le graphique correspondant à la SOD, on voit également une ligne à la fréquence épistasique 0,4. Cela correspond à des positions pour lesquelles, selon le modèle épistasique, deux mutations ont exactement la même probabilité maximale d'être observées (donc $\pi = 0.81/2$). Ces positions en revanche sont réparties sur l'intégralité de l'axe des fréquences indépendantes. Le modèle indépendant peut par conséquent distinguer certaines mutations qui apparaissent similaires dans le modèle épistasique. C'est la raison pour laquelle les deux prédictions sont conservées dans le modèle global, bien que l'optimisation donne plus de poids à la prédiction du modèle épistasique. La figure IV.8 montre les quatre protéines utilisées pour tester notre modèle. Chaque colonne correspond à une position de la protéine, tandis que les lignes correspondent aux différents acides aminés. Les acides aminés les plus conservés, qui ont les fréquences à l'équilibre les plus élevées à une position, apparaissent en jaune. Les sites variables apparaissent avec différentes teintes de vert. Au contraire, les acides aminés qu'on n'observe jamais à une position sont en violet. Les différentes étapes de transformation des prédictions de GEMME en fréquence à l'équilibre permettent d'obtenir cette matrice, qui donne des attendus plus réalistes dans l'évolution des acides aminés aux différentes positions de la séquence qu'un modèle de phylogénie classique.

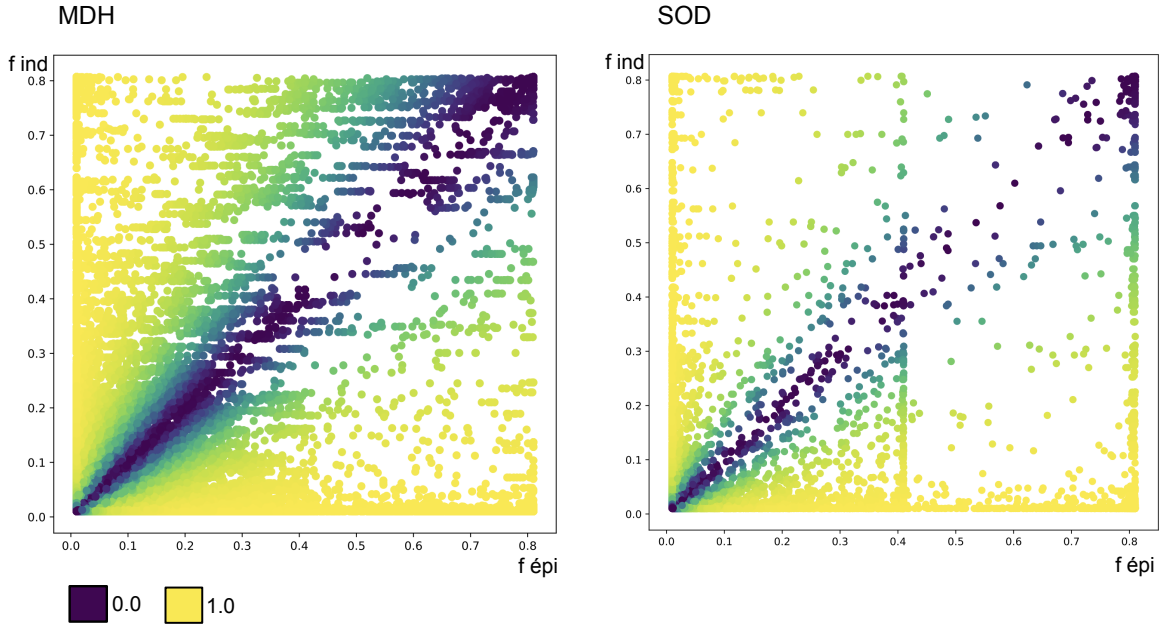


FIGURE IV.7 – Pondération entre les fréquences indépendantes et épistasiques. La valeur du paramètre α entre 0 et 1, qui détermine le poids donné à π_{epi} par rapport à π_{ind}, est représenté par couleur pour chaque mutation possible de la protéine. La valeur de α dépend de la position de chaque prédiction, plus précisément de sa distance relative entre la bissectrice et les points (0.8,0) ou (0,0.8).

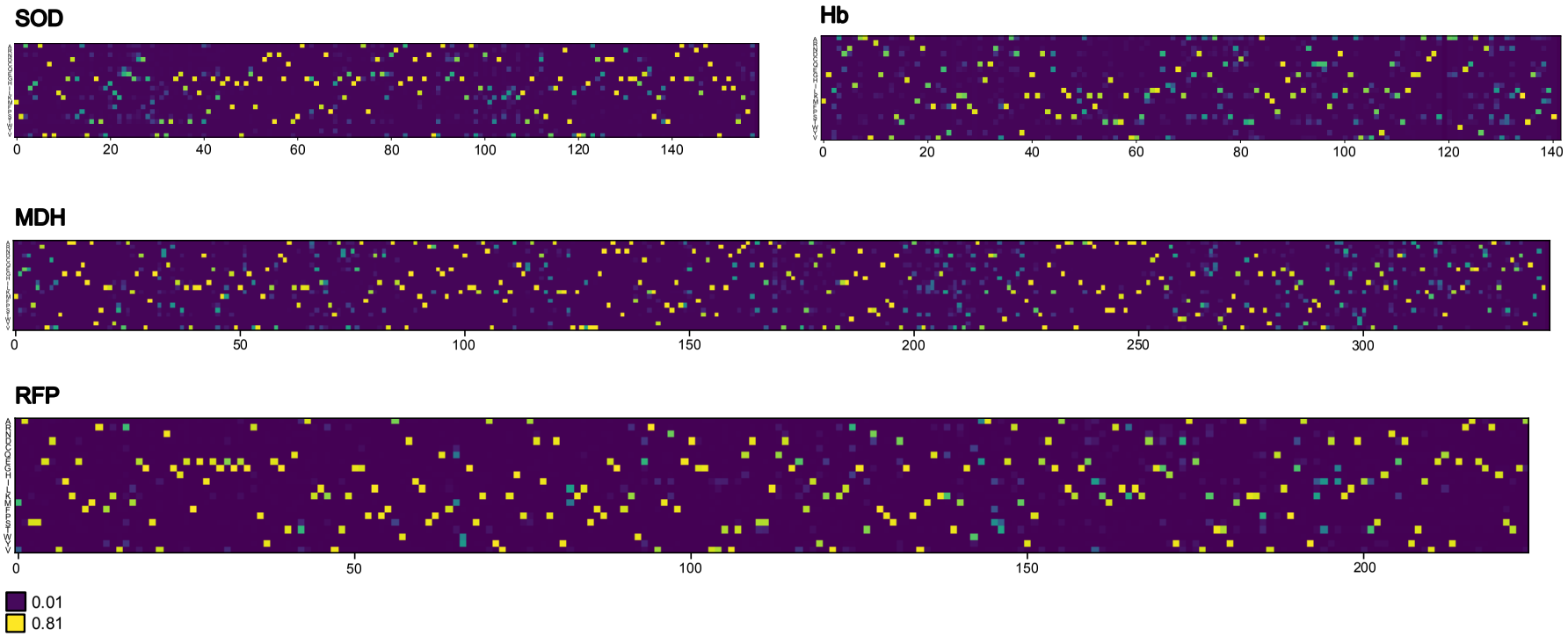


Figure IV.8 – Fréquences attendues des acides aminés aux positions des protéines étudiées. Chaque point de couleur représente la probabilité d'observer un acide aminé, selon l'ordonnée, à une position de la protéine, selon l'abscisse. Les acides aminés les plus conservés sont en jaune, les acides aminés impossibles sont en violet.

Après l'optimisation des fréquences des acides aminés, le modèle Gempistasy utilise ces fréquences moyennes calculées et présentées en figure IV.8 pour optimiser les longueurs de branche et effectuer une première reconstruction des séquences ancestrales. La confiance dans les résultats est ensuite évaluée pour chaque site à chaque noeud de l'arbre, et seuls les sites pour lesquels la solution est instable sont calculés à nouveau en utilisant un modèle plus précis qui tient alors compte des fréquences attendues pour chaque acide aminé dans chaque séquence (et non plus les fréquences moyennes par famille de protéines), ce qui est beaucoup plus long à calculer. La figure IV.9 montre la distribution des probabilités des acides aminés aux différents noeuds de l'arbre pour la MDHc. Pour chaque position à chaque noeud, nous avons calculé si les acides aminés non inférés dans la première reconstruction pouvaient devenir les plus vraisemblables à une position en changeant le profil de fréquences à l'équilibre des acides aminés possible à cette position. Dans ce cas, ils sont coloriés en vert. Autrement, ils sont coloriés en rouge. Nous cherchons donc à savoir si certaines hypothèses sur la fréquence peuvent déstabiliser le résultat de vraisemblance obtenu pour l'acide aminé testé. Comme ce calcul est très long (chaque acide aminé doit être testé avec différents profils de fréquences), nous construisons en premier une quantité q selon l'équation IV.13 pour essayer de déterminer *a priori* si la vraisemblance de l'acide aminé est stable sans avoir besoin d'effectuer le calcul rigoureusement. On constate que la quantité q arrive relativement bien à distinguer les acides aminés impossibles quelque que soit l'hypothèse sur les fréquences à l'équilibre (en rouge) des acides aminés possibles (en vert). Dans la suite, j'ai considéré que les sites pour lesquels $q < -2$ avaient une solution stable. Ces acides aminés ne sont plus recalculés par la suite. Pour les autres sites, les probabilités marginales seront recalculées de manière plus précise en considérant les fréquences attendues en fonction de l'intégralité de la séquence ancestrale inférée.

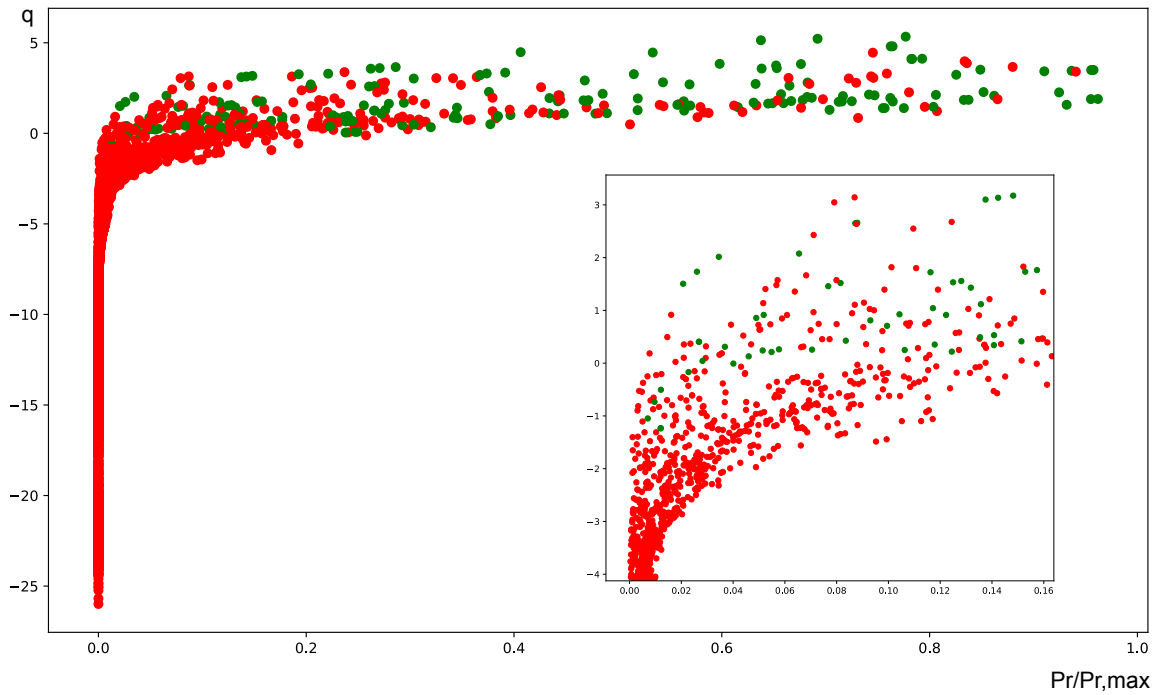


Figure IV.9 – Confiance dans les reconstruction ancestrales de la MDHc selon Gempistasy. Les points rouges correspondent à des acides aminés pour lesquels on a déterminé qu'ils ne pouvaient pas être présents à l'état ancestral, y compris sous des hypothèses de fréquences à l'équilibre visant à maximiser leur probabilité. Les points verts correspondent à des acides aminés non inférés pour l'état ancestral, mais qui pourraient être privilégiés sous des hypothèses avantageuses. L'abscisse $Pr/Pr, max$ correspond à la probabilité de l'acide aminé testé par rapport à l'acide aminé ayant obtenu le maximum de vraisemblance. la quantité q est obtenue par l'équation IV.13.

Dans une dernière étape, nous cherchons à déterminer quels sites évoluent de façon conjointe et doivent être en conséquence calculés simultanément. Pour cela, nous déstabilisons les séquences ancestrales approximatives obtenues à l'étape précédente au niveau des sites ambigus, pour voir comment les prédictions de GEMME sur d'autres sites de la séquence sont influencées. La figure IV.10a. montre comment un changement d'acide aminé à une position de la séquence de la MDHc peut, selon GEMME, influencer le profil de fréquences d'une autre position. Ici, nous avons modifié l'acide aminé en position 198 de la séquence de la MDH entre deux acides aminés potentiels de la reconstruction ancestrale, et nous regardons l'effet sur les fréquences attendues à la position 123. On constate que dans un contexte mutationnel donné (courbe orange), N est largement dominant (70% de fréquence attendue), tandis que S et T sont deux autres possibilités avec des fréquences attendues à moins de 10%. Une mutation sur l'acide aminé 198 modifie ce profil (courbe bleue), puisque maintenant la position 123 peut être occupée dans les mêmes proportions par N ou T (environ 33% de fréquence attendue) tandis que la probabilité d'observer S augmente également dans une moindre mesure (environ 12%). La mutation au site 198 apparaît alors, selon la prédiction GEMME, comme facilitant la présence de certains acides aminés à une autre position de la protéine. Nous effectuons ces prédictions pour tous les couples de positions ambiguës dans les différentes protéines ancestrales reconstruites jusqu'alors, et nous reportons ces résultats sur la figure IV.10b. Ici, nous observons la stabilité des profils de fréquences pour toutes les positions ambiguës de la MDHc. Pour chaque site, nous avons prédit les fréquences attendues en acides aminés dans différents contextes de séquences variant par une seule mutation, et nous avons conservé la dissimilarité maximale obtenue entre deux profils de fréquences (mesurée selon l'équationIV.14). Nous observons beaucoup de points proches de 0, ce qui indique des profils de fréquences globalement insensibles aux mutations à d'autres sites ambigus de la protéine. Ces positions peuvent être considérées indépendantes, et sont calculées sous l'hypothèse classique en phylogénie d'indépendance des résidus. D'autres sites au contraire montrent de fortes dissimilarités dans les fréquences attendues selon les mutations à d'autres sites de la séquence. Ces sites doivent être évalués conjointement. Nous fixons un seuil sur cette dissimilarité pour être considérée pertinente, ici à 0.07, qui correspond à la ligne rouge. Ce seuil est trouvé automatiquement sur des considérations purement calculatoires. Nous cherchons à former le plus de groupes possibles, sans toutefois que ces groupes ne contiennent trop de sites. Nous fixons la taille maximale des groupes à six sites, car nous ne sommes pas en mesure d'évaluer l'évolution conjointe de plus de sites à la fois. La figure IV.10c. localise les différents regroupements de sites obtenus sur la structure 3D de la MDH dimérique (MDHc d'*A. pompejana* modélisée avec SWISS-MODEL selon la structure de la MDHc humaine, PDB 7rm9, 1.65Å, 70% identité des résidus). Le programme a reconstitué trois groupes distincts (un groupe rouge de cinq sites, un cyan de cinq sites et un mauve de deux sites) composés de résidus qui montrent des traces d'épistasie importantes les uns avec les autres. Certains de ces sites sont proches dans la structure 3D, et liés entre aux par des liaisons hydrogènes au sein d'hélice ou de boucle. Il est facile dans ce cas de comprendre une interaction entre ces résidus. D'autres au contraire sont plus éloignés dans la structure, comme notamment les deux résidus du groupe mauve. Dans ce cas, il est difficile de savoir ce qui relie ces acides aminés. GEMME a prédit un changement de comportement réciproque pour ces deux résidus en fonction de leur nature mais nous ne pouvons pas expliquer cette interaction. Dans ce cas, seule l'expérience de mutants pourrait confirmer si cette prédiction est exacte ou si c'est un artefact

de la méthode utilisée lié par exemple à une diversité trop faible des séquences homologues récoltées dans la banque NCBI et qui sont le point de départ des prédictions de GEMME.

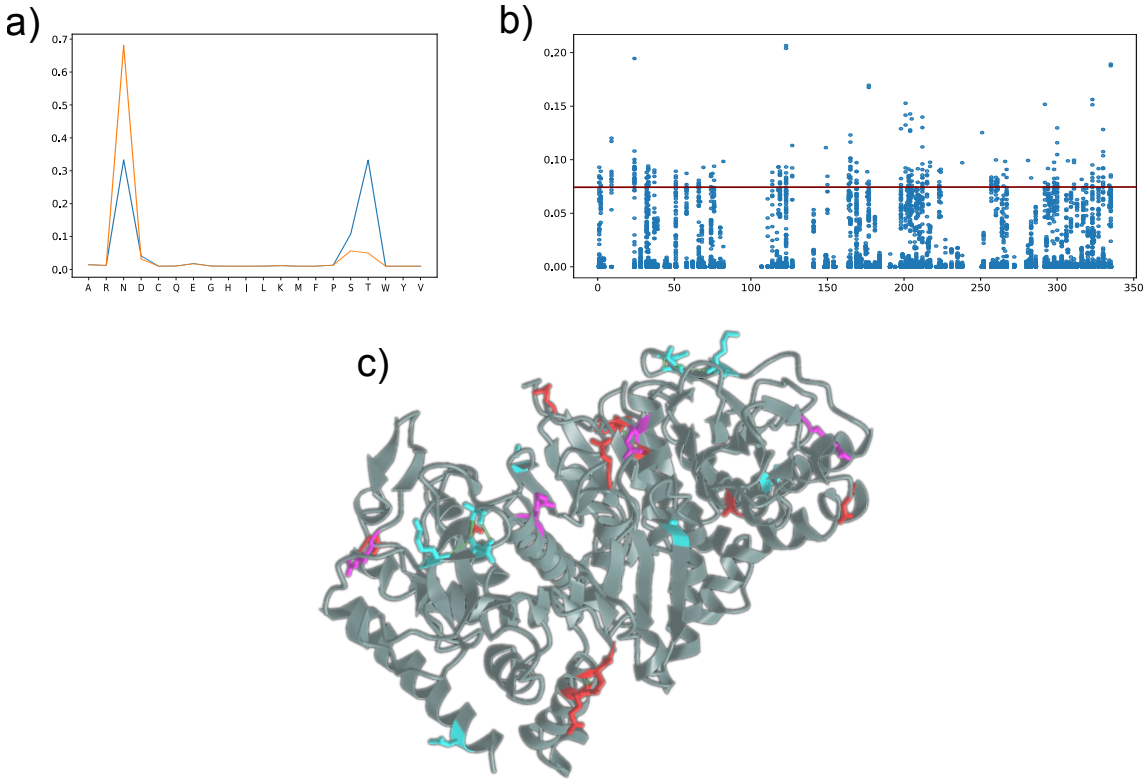


FIGURE IV.10 – Détection de l'épistasie entre les résidus de la MDHc. (a) Profils de fréquences attendus à la position 123 de la MDHc, selon la nature de l'acide aminé à la position 198. (b) Indépendance des sites au sein de la MDHc. Chaque point représente la dissimilarité des fréquences d'acide aminé attendues pour un site en fonction de mutations survenant à d'autres sites de la fréquence. Plus cette valeur est grande, plus le site est influencé par un autre site de la séquence. Une valeur seuil est représentée en rouge, qui sépare les dissimilarités que l'on considère significatives, des dissimilarités considérées suffisamment petites pour être ignorées et revenir à l'hypothèse classique d'indépendance des résidus. (c) Projection des résidus identifiés comme ayant une forte influence les uns sur les autres sur la MDHc d'*A. pompejana*, modélisée avec SWISS-MODEL en utilisant la structure de la MDHc humaine (PDB 7rm9, 1.65Å, 70% identité des résidus). Chaque couleur représente un groupe de sites dépendants. Le groupe rouge contient les sites 33, 114, 118, 165 et 265, le groupe cyan les sites 66, 292, 295, 297 et 324, et le groupe mauve les sites 196 et 255.

La figure IV.11 représente les localisations des sites déterminés comme covariants selon GEMME sur les trois protéines d'Alvinellidae étudiées. La RFP n'est pas représentée, car aucun regroupement d'acides aminés n'a été déterminé par l'algorithme. Entre un et trois groupes ont été déterminés pour chaque protéine. Pour la MDHc, au contraire du résultat précédant, seul un groupe a été déterminé cette fois. Il contient les sites 33, 114, 118, 165 et 265 correspondant au groupe rouge, figure IV.10c., additionné du site 123. A la différence du résultat antérieur, cette prédiction a été effectuée à partir d'un nouveau BLAST plus récent sur la base de données NCBI concernant les séquences homologues de MDHc, nécessaires à la prédiction de GEMME. Ceci a eu pour effet de déceler un nouveau site associé à ce groupe. Par conséquent, le seuil de 6 mutations à partir duquel on considère que des mutations dépendantes correspondent à un même groupe a eu pour effet de faire éclater les deux autres groupes dont les liens de dépendance sont plus faibles. Ceci est une limite de l'algorithme utilisé actuellement pour effectuer les regroupements et celui-ci devra être amélioré, puisque les groupes cyan et mauve pourraient être décelés et maintenus, même si les liens qui unissent ces acides aminés entre eux sont plus ténus que les liens qui unissent le septième acide aminé avec le groupe rouge.

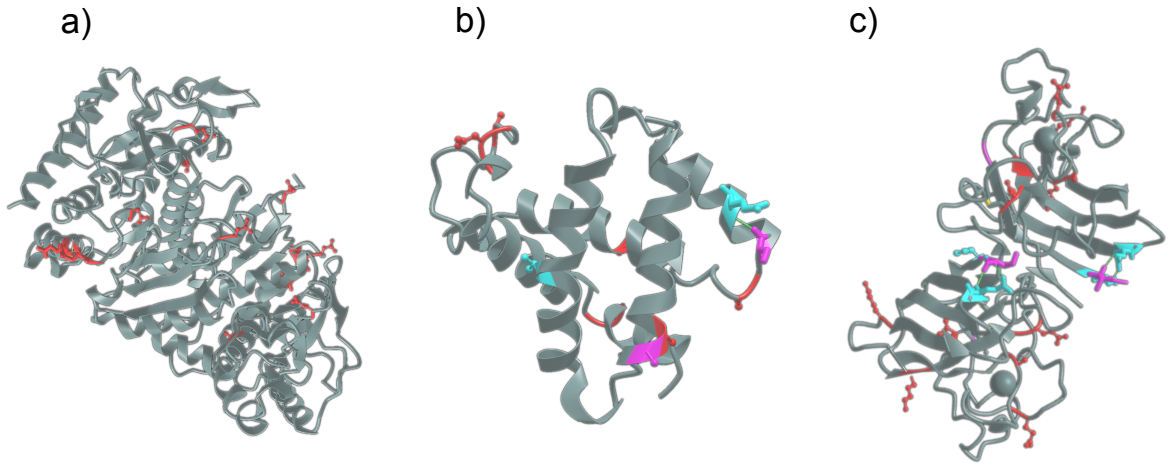


FIGURE IV.11 – Détection de l'épistasie entre les résidus des protéines d'Alvinellidae. (a) MDHc : protéine d'*A. pompejana*, modélisée avec SWISS MODEL sur la MDHc humaine (PDB 7rm9, 1.65Å, 70% identité des résidus). Un groupe est détecté, comprenant les sites 33, 114, 118, 123, 165 et 265. (b) Hb : protéine d'*A. pompejana*, modélisée avec SWISS-MODEL sur le patron prédit de l'hémoglobine d'*Arichlidon gathofi* (prédiction Uniprot A0A6M8AUJ7, 82% identité des résidus). Trois groupes sont détectés, comprenant les sites 11, 33, 36, 61, 71, 121, un deuxième groupe 73, 120, et un dernier 77, 103. (c) SOD : protéine d'*A. pompejana* (PDB 3f7k, 1.35Å). Trois groupes sont détectés, comprenant les sites 42, 48, 69, 80, 93, 118, un deuxième groupe 3, 141, et un dernier 4, 23.

Les positions d'épistasie potentielle au sein d'un groupe présentées sur la figure IV.11 peuvent être proches dans la structure tridimensionnelle de la protéine, mais la plupart d'entre elles sont plutôt éloignées les unes des autres. Dans des études empiriques de l'épistasie, le fait que ces mutations ne sont pas nécessairement regroupées est bien établi (Dettaï et al., 2008 ; Poelwijk, Socolich et Ranganathan, 2019 ; Vigué et al., 2022), et certaines mutations peuvent avoir des influences structurelles sur des parties éloignées de la protéine du fait de la transmission de l'encombrement stérique (Fraser et al., 2016). En outre, dans notre méthode, nous ne calculons que l'influence des sites pour lesquels il y a un doute sur la nature du résidu dans les séquences ancestrales, donc nous n'observons que partiellement l'information liée à l'épistasie sur la protéine entière.

IV.1.3.3 Inférence des séquences ancestrales

Maintenant que les paramètres de nos modèles ont été optimisés, nous pouvons nous intéresser au résultat pratique de ces différentes approches phylogénétiques sur les quatre protéines étudiées. Le tableau IV.2 récapitule ces résultats. Les modèles WAG et PFASUM correspondent à des standards afin de comparer la performance des nouveaux modèles. Sur ces deux modèles standards, la matrice PFASUM a systématiquement de moins bonnes performances que la matrice WAG, ce qui indique que le choix de la matrice WAG pour la construction des modèles Ecoprior et Gempistasie aurait été plus judicieux. Les résultats de la matrice PFASUM sont néanmoins conservés dans le tableau afin d'avoir un élément de comparaison plus direct avec les deux modèles dérivés dont ils sont issus.

Le modèle Ecoprior (PFASUM+HJM+CJM) ne produit qu'un gain faible de vraisemblance par rapport au modèle PFASUM seul. Ce gain est néanmoins intéressant, selon le critère BIC, pour les protéines Hb et MDHc. La MDHc est la plus longue des protéines étudiées, et l'arbre de l'hémoglobine contient plus de séquences *via* les deux sous-familles d'hémoglobine intracellulaire d'Alvinellidae trouvées. Comme on pouvait s'y attendre, la multiplication des paramètres du modèle Ecoprior tend à en faire un modèle sur-paramétrisé pour des petites séquences, mais devient plus pertinent lorsque les alignements sont plus longs. En pratique, on pourrait imaginer calibrer les matrices à utiliser sur chaque branche à partir d'une concaténation de gènes de grande taille, et d'utiliser cette donnée préalable pour effectuer la reconstruction des séquences ancestrales sur les séquences d'intérêt plus courtes pour bien estimer le poids relatif des matrice PFASUM, HJM et CJM sur chaque branche. Le modèle WAG a des vraisemblances similaires ou meilleures que le modèle Ecoprior. Il semble, sur ces protéines du moins, que l'entraînement des matrices HJM et CJM aurait bénéficié à être effectué à partir de la matrice WAG plutôt qu'à partir de la matrice PFASUM. Un point d'attention est que les vraisemblances obtenues pour le modèle Ecoprior sont presque identiques aux vraisemblances obtenues pour le modèle WAG, excepté pour la protéine RFP pour laquelle le modèle WAG est meilleur que le modèle Ecoprior. Ceci est plutôt en accord avec l'attendu, puisque le modèle Ecoprior ne semble pas pertinent pour décrire l'évolution de la protéine RFP, pour laquelle les organismes ont été expérimentalement maintenus à 37°C (Randall et al., 2016).

Le modèle Struct2 est le seul des trois modèles à ne pas utiliser la matrice PFASUM dans son évaluation, puisque les matrices de mutation utilisées sont celles établies par Si

Quang Le et Olivier Gascuel, 2010. Ce modèle représente systématiquement un gain de vraisemblance non négligeable par rapport aux modèles standards, ce qui est confirmé par le BIC.

Le modèle Gempistasie apporte un gain très important de vraisemblance, et est de loin le modèle qui rend le mieux compte des dynamiques évolutives des séquences. Un point notable est que ce modèle ne contient pas plus de paramètres que les modèles standards. Le paramètre Γ est également retiré du modèle, car nous faisons l'hypothèse que les différences de vitesse évolutive relevées entre les résidus d'une même protéines sont la conséquence de matrices de mutations non adaptées à aux dynamiques propres des différents sites, ce qui peut être compensé grâce au modèle Gempistasy site-dépendant (S. Q. Le, Dang et O. Gascuel, 2012). Si l'on se concentre seulement sur le BIC, ce modèle devrait sans ambiguïté aboutir à de meilleures reconstructions de séquences ancestrales.

		# paramètres	Taille échantillon	Γ	log(L)	BIC
MDH	Struct2	37	341	0.313	-3063.68	6221.07
	Ecoprior	109	341	0.338	-3196.84	6669.75
	Gempistasie	36	341	/	-2162.74	4416.66
	WAG	37	341	0.350	-3194.71	6483.13
	PFASUM	37	341	0.342	-3329.13	6732.97
SOD	Struct2	37	158	0.423	-1874.12	3829.59
	Ecoprior	109	158	0.449	-1929.77	4099.20
	Gempistasie	36	158	/	-1365.92	2910.99
	WAG	37	158	0.498	-1926.47	3934.29
	PFASUM	37	158	0.484	-1982.14	4045.63
Hb	Struct2	55	141	1.286	-3024.92	6168.05
	Ecoprior	163	141	1.186	-3069.82	6489.97
	Gempistasie	54	141	/	-2483.88	5083.82
	WAG	55	141	1.375	-3070.14	6258.49
	PFASUM	55	141	1.354	-3210.52	6539.25
RFP	Struct2	37	225	0.478	-2352.81	4792.65
	Ecoprior	109	225	0.512	-2440.02	5136.43
	Gempistasie	36	225	/	-1766.30	3617.28
	WAG	37	225	0.535	-2400.30	4887.63
	PFASUM	37	225	0.486	-2478.54	5044.11

Table IV.2 – Optimisation des modèles ASR sur différentes protéines. Le nombre de paramètres (nombre de branches et paramètre Γ) ainsi que la taille de l'échantillon (nombre de positions dans l'alignement) sont utilisés pour calculer le BIC score. Γ représente la valeur du paramètre Gamma dans les modèles l'utilisant. $log(L)$ représente la vraisemblance, qu'on veut maximiser. BIC est le Bayesian Information Criterion, qui mesure la pertinence du modèle en tenant compte de la vraisemblance, du nombre de paramètres ainsi que de la taille de l'échantillon. Un BIC plus faible est préférable.

Les séquences ancestrales finalement reconstruites par maximum de vraisemblance sont présentées en figure IV.12. Pour la protéine RFP, qui est issue d'une expérience de phylogénie expérimentale, les séquences ancestrales réelles sont connues (Randall et al., 2016). Pour cette protéine, seuls les ancêtres avec le plus fort taux d'erreur dans la reconstruction sont montrés sur la figure. La méthode WAG, également utilisée dans cet article, produit le même nombre d'erreurs aux mêmes noeuds de l'arbre, ce qui confirme que notre implémentation du modèle est similaire aux algorithmes employés par FastML ou PAML. Ce modèle produit au total 45 erreurs sur les noeuds considérés par rapport aux séquences réelles. La méthode PFASUM en produit 47, Struct2 45, et 46 pour Ecoprior et Gempistasie. Les acides aminés mal inférés sont globalement les mêmes pour toutes les méthodes, qui opèrent souvent le même choix erroné. Sur le comptage strict des acides aminés bien inférés, les modèles WAG et Struct2 ont obtenu le même résultat, tandis que le modèle Gempistasie a pu corriger un acide aminé par rapport au modèle PFASUM dont il est issu. Le modèle Ecoprior donne un résultat identique à Gempistasie, mais ce modèle n'a pas été imaginé pour améliorer la qualité des reconstructions de ce type de séquences. Les résultats des modèles sont plus intéressants si on regarde également le deuxième acide aminé le plus probable inféré par les différentes méthodes. En effet, l'objectif de ces modèles est d'évaluer correctement les probabilités des acides aminés ancestraux aux différents sites, mais dans le cas où deux acides aminés auraient des probabilités similaires on ne devrait pas considérer que le modèle est mauvais parce qu'il aboutit à choisir un acide aminé erroné par maximum de vraisemblance. Un chemin évolutif peut être moins probable mais correspondre à la réalité, et cette information ne peut pas être mieux approchée par le modèle que par les probabilités qu'il nous fournit. Ainsi, si l'on considère comme corrects les sites pour lesquels le modèle a donné une probabilité supérieure à 15% à l'acide aminé réel, le modèle Struct2 gagne 19 sites bien évalués, Ecoprior 11 sites, Gempistasie 12 sites, WAG 15 sites et PFASUM 13 sites. Dans cette vision, le modèle Struct2 apparaît comme étant plus à même de bien évaluer les probabilités des états ancestraux. En revanche, malgré une vraisemblance globale bien plus élevée du modèle Gempistasie, il ne parvient pas en pratique à surpasser les autres modèles dans les reconstructions. Si on le compare avec le modèle PFASUM dont il est issu, il permet de recouvrer le vrai site à la position 93 de l'alignement, ainsi qu'à la position 116 de l'ancêtre des lignées 18 et 29. En revanche, il introduit une erreur à la position 127 de l'ancêtre de l'arbre ainsi qu'à la position 116 des descendants de cet ancêtre. Ce modèle peut toutefois être facilement amélioré et gagnerait certainement à être associé aux matrices du modèle Struct2 plutôt que sur la matrice PFASUM pour augmenter sa performance, ainsi qu'à introduire un certain degré d'incertitude dans les fréquences à l'équilibre attendues aux sites des séquences ancestrales afin d'être moins rigide face à l'information donnée par les séquences contemporaines.

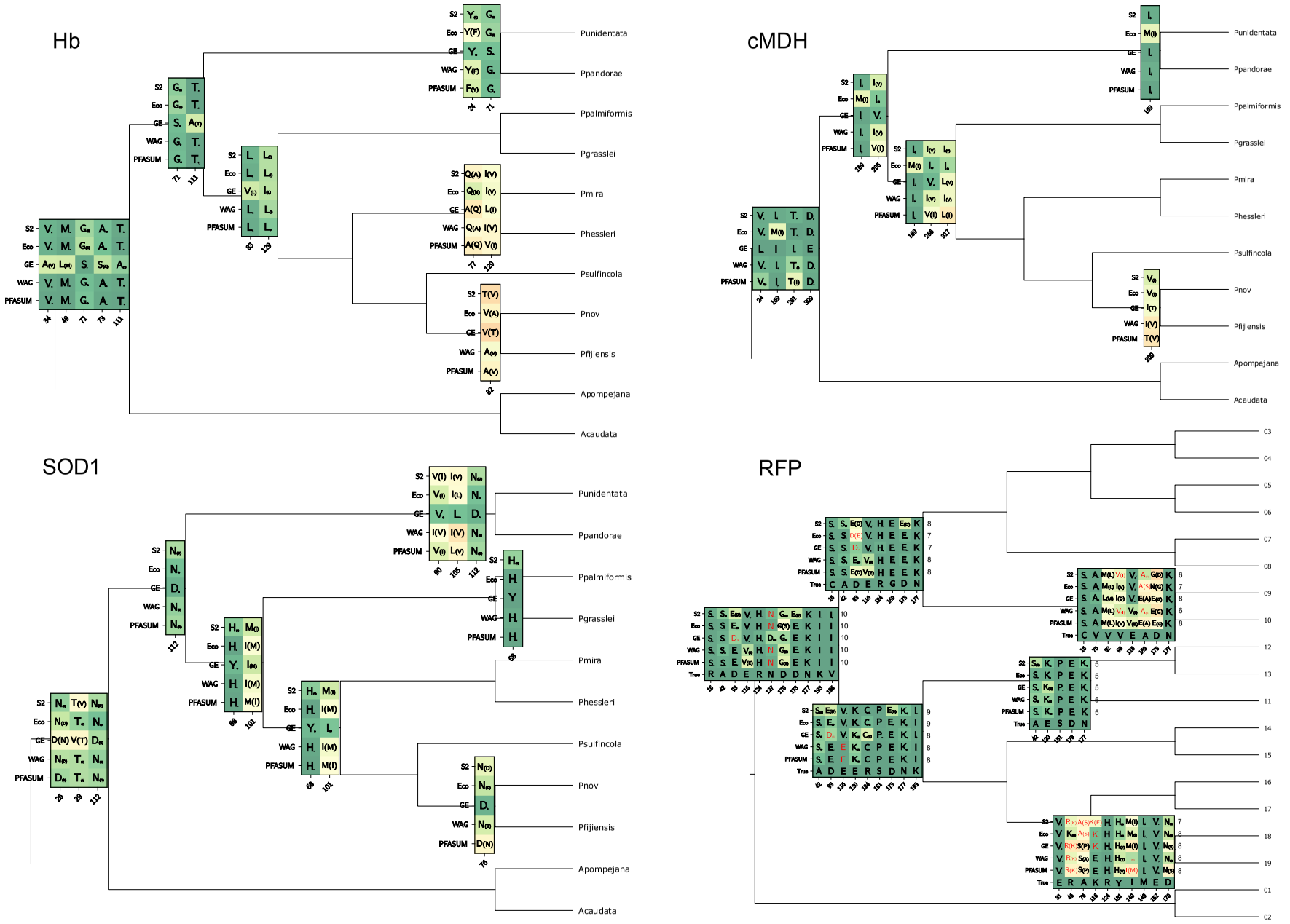


Figure IV.12 – Séquences ancestrales reconstruites par les différents modèles. Seuls les acides aminés ancestraux variables sont montrés avec les positions dans l'alignement. Les cases X(Y) donnent les deux acides aminés les plus vraisemblables, et leur taille donne leur probabilité relative. La case est verte si $Pr(X) = 1$, rouge si $Pr(X) = 0$. S2 : Struct2, Eco : Ecoprior, GE : Gempistasie. Pour la RFP, seuls les noeuds les plus problématiques sont reportés. la séquence réelle de l'ancêtre est notée "True", les acides aminés corrects sont en rouge.

Un autre point de vigilance est que certains sites sont mal inférés par toutes les méthodes, et ce avec une très forte probabilité d'accepter le mauvais acide aminé (par exemple les sites 16, 177, 193, 196 de l'ancêtre des RFP). Comme expliqué ci-dessus, la méthode d'ASR est essentiellement probabiliste et l'erreur n'est pas un problème en soi car l'acide aminé réel était peut-être effectivement improbable. Toutefois, le fait qu'aucun des modèles ne soit capable d'identifier l'acide aminé réel comme étant une possibilité crédible pose question. Plus que la recherche de la séquence ancestrale exacte, il serait utile de se concentrer sur une bonne estimation des probabilités des résidus ancestraux. En effet, ces probabilités sont indispensables pour mesurer l'incertitude que l'on a sur la reconstruction (Williams et al., 2006 ; Eick et al., 2016). Or l'optimisation de la vraisemblance sur les modèles Struct2 et Gempistasie de ce chapitre montre que l'amélioration de la vraisemblance du modèle n'aboutit pas nécessairement à l'estimation de meilleures probabilités sur les résidus ancestraux de la protéine RFP. Il faudrait essayer d'optimiser un modèle qui maximise les probabilités obtenues pour les véritables résidus ancestraux. La difficulté est que mises à part certaines phylogénies expérimentales comme celle de Randall et al., 2016 que nous utilisons ici, les séquences ancestrales vraies ne sont jamais connues et l'optimisation d'un tel modèle n'est pas possible. Pour l'instant, on ne peut avoir accès qu'à des données simulées mais la génération de ces données et leur résolution selon des modèles similaires induit une circularité. Il serait très utile de reproduire des phylogénies expérimentales comme celle-ci pour essayer d'optimiser ces modèles sur des données biologiques plus proches de la réalité, et sans doute beaucoup plus complexes. En outre, nous nous sommes exclusivement concentrés sur la modélisation des séquences ancestrales. Pour l'ASR, il faut garder à l'esprit que pour un expérimentateur, la synthèse et la caractérisation des protéines ancestrales ainsi que la quantification de leur fiabilité restent la finalité. Outre la synthèse de certains variants choisis, d'autres méthodes existent comme par exemple l'intégration de l'incertitude directement *via* la dégénérescence de certains sites dans la synthèse des gènes afin d'exprimer un ensemble de protéines contenant toute l'incertitude dans les reconstructions (Chang, Ugalde et Matz, 2005). Poelwijk, Socolich et Ranganathan, 2019 ont également suggéré que l'expression d'un certain nombre de mutants choisis permettrait de reconstruire toute la variabilité phénotypique des mutants potentiels du fait de l'épistasie relativement limitée entre les résidus de la séquence. Dans cette expérience, les auteurs estiment que les 2^{13} mutants étudiés de la protéine fluorescente de *Entacmaea quadricolor* peuvent être intégralement caractérisés par seulement 6 à 11% de ces mutants. Cela représente quand même 500 à 800 protéines pour seulement 13 résidus variables initialement, mais des expressions en haut débit couplées à des simulations de certains mutants pourraient peut-être permettre la caractérisation de l'ensemble des variants potentiels d'une protéine dans des études ASR.

Enfin, pour les protéines étudiées au chapitre 2 de ce manuscrit, on constate sur la figure IV.12 que les différents modèles donnent là encore des résultats en général assez similaires si l'on regarde les deux acides aminés les plus crédibles à chaque site, mais les probabilités entre les deux acides aminés alternatifs peuvent s'inverser. Le modèle Gempistasie, en particulier, a souvent tendance à privilégier le deuxième choix affiché par les autres méthodes. Les modèles WAG et Struct2, qui sont les deux meilleurs modèles selon les reconstructions effectuées sur la RFP, donnent pratiquement les mêmes résultats à tous les noeuds observés. Les séquences obtenues par maximum de vraisemblance dans les différents modèles et correspondant à l'ancêtre des Alvinellidae sont données en annexe. Si on les compare avec

les protéines exprimées dans le chapitre 2, le modèle Struct2 et Ecoprior donnent les mêmes résultats (sauf pour la position 45 de l'hémoglobine et 169 de la cMDH) à l'ancêtre Anc1 exprimé sous la topologie H1 (voir chapitre 2), ce qui confirme la fiabilité des protéines obtenues au deuxième chapitre. En revanche, le modèle Gempistasie est le plus différent, et introduit entre 3 et 5 mutations par protéine par rapport à l'ancêtre Anc1. Pour l'instant, bien que Gempistasie ait de très bonnes vraisemblances sur les reconstructions phylogénétiques, nous ne pouvons pas conclure si les protéines reconstruites par Gempistasie plus réalistes mais cela devrait faire l'objet de futures investigations.

IV.2 Inférence d'événements d'insertion et de délétion ancestraux

IV.2.1 Introduction

Dans cette seconde partie, j'ai voulu optimiser un modèle évolutif pour inférer les événements d'insertion et délétion (indels) dans les séquences ancestrales. La motivation initiale était en premier lieu que les modèles d'ASR comme FastML reconstruisent les états ancestraux des résidus par un modèle probabiliste, mais utilisent un algorithme parcimonieux pour ce qui est des indels (Ashkenazy, Penn et al., 2012). Par conséquent, je souhaitais essayer de reconstruire les indels ancestraux par un modèle de vraisemblance également. Comme la création de ce modèle implique de calculer la vraisemblance sur l'ensemble des arbres de gènes en tenant également compte de la divergence entre les séquences, j'ai étendu l'idée initiale pour un algorithme d'évaluation de topologies d'arbre par maximum de vraisemblance, sur la base des indels pris comme des caractères phylogénétiques informatifs. La recherche heuristique de l'arbre optimal n'est pas envisagée, l'algorithme cherche uniquement l'optimisation des paramètres du modèle (taux d'insertion et de délétion, distributions de taille des indels) et des longueurs de branches sur un arbre obtenu par d'autres méthodes phylogénétiques basées sur les acides aminés ou les nucléotides.

L'intérêt de l'utilisation des indels en phylogénie est d'augmenter l'information disponible pour l'inférence des arbres. Luan et al., 2013 ont montré sur une phylogénie de 16 espèces d'Arctoidea que les arbres obtenus sur des introns nucléaires sans prise en compte des indels aboutissaient à une ambiguïté entre la monophylie de certains groupes (les Pinnipidiea pouvant être groupés avec les Ursidae ou les Musteloidea), alors que la prise en compte de 6843 indels permettait de résoudre la monophylie des Pinnipidiea avec les Ursidea. Les auteurs ont essayé trois méthodes de codage des indels pour établir leur homologie, ensuite traduits en caractères binaires présence/absence afin d'obtenir ce résultat. L'utilisation des indels en phylogénie reste relativement marginale, mais les articles ayant essayé de les intégrer utilisent en général une méthodologie similaire (Paśko, Ericson et Elzanowski, 2011 ; Ashkenazy, Cohen et al., 2014). Les deux types de codage des indels les plus utilisés sont le Simple Indel Coding (SIC), et le Complex Indel Coding (CIC, Simmons et Ochoterena, 2000). Le principe est que deux gaps présents dans deux séquences sont considérés homologues s'ils partagent un début en 5' et une fin en 3' de la séquence identiques. Autrement, ils sont considérés comme étant différents. Le Complex Indel Coding assouplit un peu ces exi-

gences en permettant l'homologie pour seulement une des deux extrémités, sous certaines conditions. Pour ces deux méthodes, un gap est considéré comme un ensemble de positions consécutives dans l'alignement pour lesquelles il manque un résidu dans au moins une des séquences. Il est donc borné par des positions de l'alignement pour lesquelles toutes les séquences ont la même base ou le même résidu, et les gaps peuvent être traités indépendamment les uns des autres. C'est cette définition du gap que j'utilise dans la suite de cette partie. L'encodage selon le SIC ou CIC permet de créer une matrice présence/absence du caractère dans les séquences, qui peut être utilisée dans les programmes de phylogénie habituels qui autorisent l'utilisation de matrices de mutations importées par l'utilisateur, suivant par exemple une loi de probabilité exponentielle pour quantifier la probabilité de l'insertion ou de la délétion d'une séquence, sans tenir compte de sa taille. D'autres méthodes ont également été développées dans certaines études, soit reposant fondamentalement sur un algorithme de parcimonie (bi-partition des séquences selon l'indel le plus parcimonieux, Donath et Stadler, 2018) ou alors avec des versions modifiées du CIC pour compter parcimonieusement le nombre de transformations nécessaires pour passer d'un gap à un autre entre séquences prises deux à deux (Müller, 2006). En outre, les indels peuvent trouver leur utilité dans d'autres domaines de la biologie que la phylogénie. Löytynoja et Goldman, 2008a a par exemple montré que l'inférence des taux de substitutions dans les séquences et la périodicité des indels (triplets de nucléotides) peut servir à l'identification de séquences codantes.

Toutefois, la difficulté principale des indels est que ces caractères ne sont pas directement observables dans les séquences. Ainsi, ils sont uniquement inférés à partir d'au moins deux séquences de tailles différentes, ce qui implique que l'addition de séquences dans un alignement modifie le nombre de caractères exploitables. Dès lors, une importance cruciale est accordée au programme utilisé pour réaliser l'alignement des séquences. Il est par exemple connu que plusieurs aligneurs ont tendance à surestimer le nombre de délétions par rapport aux insertions, une fois que celles-ci sont replacées sur l'arbre phylogénétique (Löytynoja et Goldman, 2008b ; Löytynoja et Goldman, 2008a ; Löytynoja et Goldman, 2010). Ceci découle du fait que ces aligneurs ont tendance à être conservatifs en essayant d'augmenter l'homologie entre les résidus observés dans les séquences contemporaines plutôt que d'introduire un gap. Cet excès artéfactuel d'homologie entraîne la présence du résidu homologue également dans les séquences ancestrales, qui doit être compensé par de multiples délétions indépendantes sur les branches terminales de l'arbre (Löytynoja et Goldman, 2008b). L'alignement des gaps, idéalement, doit se faire en distinguant insertions et délétions et utiliser un arbre guide adéquat. C'est ce que des aligneurs comme PRANK essaient de réaliser (Löytynoja et Goldman, 2008b). Aucun des aligneurs n'est toutefois parfait. Notamment, les résultats des aligneurs distinguant insertions et délétions ont des résultats variables selon la qualité de l'arbre guide utilisé pour orienter les indels, généralement obtenu par neighbour-joining (Löytynoja et Goldman, 2005 ; Löytynoja et Goldman, 2008b). Ashkenazy, Cohen et al., 2014 ont par exemple montré que le même alignement, en sous-échantillonnant les séquences, aboutissait à des résultats sensiblement différents avec une proportion de gaps stables et fiables compris entre 10 et 25% . Certains aligneurs semblent quand même donner de meilleurs résultats en moyenne, notamment MAFFT et potentiellement PRANK (Dessimoz et Gil, 2010 ; Ashkenazy, Cohen et al., 2014). Pour contourner cette difficulté, deux stratégies ont été envisagées : soit filtrer les indels selon leur fiabilité, par exemple selon leur stabilité ou leur homoplasie potentielle dans des

alignements successifs des mêmes séquences (Ashkenazy, Cohen et al., 2014 ; Donath et Stadler, 2018), ou alors ré-estimer l'alignement et la phylogénie de façon jointe par une méthode bayésienne. Dans ce cas, l'alignement et les états ancestraux potentiels sont échantillonnés en même temps que les autres paramètres comme l'arbre phylogénétique pour établir le modèle, et l'incertitude dans les gaps est directement prise en compte (Redelings et Suchard, 2007 ; Westesson et al., 2012). En théorie, cette procédure est la plus rationnelle, toutefois elle introduit un modèle très lourd dans l'inférence phylogénétique global et les temps de calcul peuvent devenir extrêmement conséquents si l'on utilise beaucoup de séquences avec des fragments assez courts, comme on peut en avoir dans une phylogénie basée sur des transcriptomes par exemple.

Ainsi, la modélisation des indels est effectuée dès l'étape de l'alignement des séquences. En conséquence, plusieurs méthodes pour modéliser les indels ont été développées, généralement plutôt dans l'objectif d'améliorer la qualité des alignements que d'être réellement intégrer en phylogénie par la suite. Ces méthodes peuvent être soit un véritable modèle d'évolution des séquences, soit une approximation mathématique jugée raisonnable pour expliquer les indels. Le modèle le plus simple est de considérer que le gap représente un état supplémentaire au résidu (généralement appelé "5-th character" pour un alignement nucléotidique, "A", "C", "G", "T", "-", Rivas, 2005). Cette simple solution a le défaut de considérer que les gaps successifs sont indépendants et réversibles, impliquant une distribution géométrique de la taille des indels qui diminue rapidement. Ce modèle va avoir tendance à fragmenter les gaps et à produire des morceaux de séquence alignés et intercalés entre des gaps courts. Une alternative populaire est de considérer une fonction affine pour modéliser les gaps (Rivas et Eddy, 2015). Dans ce cas, il existe des probabilités différentes pour l'ouverture d'un gap et pour son extension, ce qui permet d'obtenir des indels en blocs plus longs, et ces probabilités peuvent être modulées selon la divergence des séquences. C'est par exemple la solution utilisée par l'aligneur ProbCons. Historiquement, les modèles évolutifs les plus populaires sont le modèle TKF91, dans lequel les sites successifs appartenant au même indel sont considérés indépendants, et son extension par le modèle TKF92 où les sites sont cette fois considérés comme appartenant à des fragments eux même insécables et indépendants, et la longueur de ces sites est distribuée selon une loi géométrique (Holmes, 2017). Un des modèles les plus récemment développé est le Cumulative Indel Model (De Maio, 2021). Ce modèle repose sur trois paramètres, à savoir les probabilités relatives des insertions et délétions instantanées, associées à des paramètres définissant leur distribution de taille. Cette distribution de taille des indels est modélisée selon une loi géométrique dépendant du temps de divergence, bien que cette distribution de taille ne décrive pas bien le comportement des indels pour les temps longs de divergence. En revanche, l'intégration numérique des équations différentielles obtenues permettent de retrouver de bonnes propriétés sur des indels simulés par INDELible (Fletcher et Yang, 2009), au moins pour des temps de divergence relativement courts. Un autre modèle récent est le Long Indel Model (Miklos, 2003). Ce modèle intègre la probabilité de passer d'un indel à un autre selon tous les chemins évolutifs possibles. Les auteurs développent un "trajectory likelihood algorithm", qui permet de limiter l'espace possible des états de transition, potentiellement infini, à un espace plus restreint d'états vraisemblables. Ce modèle considère également que les indels ont des tailles distribuées géométriquement, et le modèle est réversible. Bien que donnant de bons résultats, ce modèle a l'inconvénient de ne pas donner d'expression analytique de la solution, et d'être lourd et complexe à mettre en oeuvre, surtout si l'on veut

établir un algorithme de reconstruction phylogénétique qui en tiendrait compte.

Mon objectif était par conséquent d'établir un modèle d'évolution des indels qui soit suffisamment rapide mais réaliste pour servir dans une inférence phylogénétique de maximum de vraisemblance. Obtenir une solution analytique pour le calcul des probabilités d'un indel entre deux séquences selon un temps de divergence reste hors de portée, voire impossible (Holmes, 2017), cependant on peut envisager certaines hypothèses simplificatrices qui donnent un résultat, on l'espère, équivalent. En outre, comme la délétion d'un site homologue ne peut pas déboucher sur l'insertion de ce même site homologue par la suite, je souhaite que ce modèle soit non réversible. Enfin, plusieurs auteurs ont souligné que la distribution empirique des tailles des indels était mieux approchée par une loi de puissance que par une loi géométrique, et que même dans le cas d'une distribution instantanée géométrique des indels, la forme de la distribution n'est pas nécessairement géométrique pour des temps de divergence longs (Gu et Li, 1995 ; Fletcher et Yang, 2009 ; Holmes, 2017 ; De Maio, 2021). Une distribution géométrique est très pratique, notamment parce qu'elle permet des simplifications dans les calculs, mais je souhaite conserver une distribution de puissance pour les distributions instantanées de taille d'indels. Idéalement, produire un modèle de maximum de vraisemblance pour les gaps pourrait permettre d'effectuer un calcul intégralement probabiliste sur les alignements observés, y compris sur les substitutions entre sites homologues si tenté que l'on puisse pondérer l'importance des indels par rapport à celle des substitutions (Sumanaweera, Allison et Konagurthu, 2022).

Ce modèle devrait aussi pouvoir nous permettre d'évaluer certaines caractéristiques des indels :

— la fiabilité de différents aligneurs sur l'inférence des indels à partir de données empiriques, sans connaissance *a priori* des relations phylogénétiques entre les séquences ;

— les paramètres d'évolution des indels dans des séquences codantes, à savoir leur taux par rapport aux substitutions, le rapport relatif entre la fréquence des insertions et des délétions, ainsi que la distribution instantanée des tailles d'insertions et de délétions ;

— l'utilité des indels pour résoudre une phylogénie empirique ambiguë, naturellement en prenant la phylogénie des Alvinellidae. Nous préférons confronter le modèle à des données empiriques plutôt qu'à des simulations, car la modélisation des gaps fait intervenir beaucoup d'incertitude (Donath et Stadler, 2018). Ces incertitudes concernent notamment la dynamique des différents mécanismes à l'origine des gaps observés (duplication de séquences, répétitions en tandem, recombinaisons par crossing over inégal, Müller, 2006) et l'inférence des états ancestraux potentiels. Aussi, je ne suis pas certain que la simulation des indels donne une image fidèle de la réalité (Donath et Stadler, 2018) et je préfère confronter le modèle à des données empiriques directement.

IV.2.2 Matériel et Méthodes

IV.2.2.1 Obtention des gènes orthologues, alignements et phylogénie

Afin d'évaluer la méthode, j'ai décidé d'utiliser des gènes orthologues bien identifiés et annotés avec une phylogénie connue sans ambiguïté, issus de la base de données OrthoDB 9 (Zdobnov et al., 2021). Les gènes présents en copie simple et présents dans un ensemble d'espèces de mammifères et d'invertébrés ont été sélectionnés (*Homo sapiens*, *Ursus maritimus*, *Gorilla gorilla gorilla*, *Macaca mulatta*, *Canis lupus familiaris*, *Mus musculus*, *Rattus norvegicus*, *Camelus dromedarius*, *Camelus ferus*, *Cavia porcellus*, *Drosophila melanogaster*, *Drosophila virilis*, *Drosophila suzukii*, *Capitella teleta*, *Ixodes scapularis*, *Helobdella robusta*, *Crassostrea gigas*, *Tetranychus urticae*, *Caenorhabditis elegans*, *Caenorhabditis japonica*). Le premier jeu de donnée inclut donc 10 espèces de mammifères et 10 espèces d'invertébrés et contient 507 gènes encodés en acides aminés. Le set "mammifères" contient toutes les espèces de mammifères de la banque et l'outgroup *C. japonica*, tandis que le set "invertébrés" contient les espèces invertébrés avec *C. ferus* en outgroup pour un même nombre de gènes. Un nouveau jeu de données a aussi été réutilisé pour les Alvinellidae. Il correspond aux mêmes contraintes imposées à l'alignement dans le chapitre 1 (gènes en simple copie et présent chez tous les lophotrocozoaires d'ODB9) identifiés chez au moins 20 transcriptomes Alvinellidae+outgroup. En outre, les espèces *Paralvinella pandorae irlandei*, *Paralvinella unidentata*, une espèce d'*Alvinella* et une autre espèce de *Paralvinella* doivent être présentes pour chaque gène. En revanche, il n'y a pas de filtre sur l'hétérogénéité en composition des acides aminés, puisque la composition à l'équilibre en acides aminés n'intervient pas dans les hypothèses du modèle d'évolution des indels développé ici. Aussi, le logiciel Orthofinder utilisé au chapitre 1 pour identifier les gènes orthologues issus de OrthoDB (Zdobnov et al., 2021) concatène les différents fragments trouvés qui appartiennent au même gêne. Nous avons pris soin dans ce cas de récupérer uniquement le fragment les plus long correspondant aux transcrits identifiés, sans concaténation qui aboutirait à la création de faux gaps. Ce nouveau jeu de données pour les Alvinellidae contient 888 gènes encodés en acides aminés.

Afin de comparer la fiabilité des résultats obtenus, les gènes des mammifères et invertébrés ont été alignés avec MAFFT FFT-NS-2 (Katoh et Standley, 2013), PRANK (Löytynoja et Goldman, 2010) et ProbCons (Do et al., 2005). MAFFT et PRANK sont généralement considérés comme des aligneurs ayant de bonnes performances dans l'alignement des indels (Dessimoz et Gil, 2010 ; Ashkenazy, Cohen et al., 2014 ; Donath et Stadler, 2018) et ProbCons est l'aligneur utilisé au chapitre 1. Les gènes d'Alvinellidae ont été alignés uniquement avec MAFFT, puisque l'aligneur est réputé à la fois rapide et fiable pour les indels. Les paramètres par défaut ont systématiquement été utilisés, afin de ne pas optimiser l'usage des aligneurs et reproduire le cas d'un utilisateur qui n'aurait pas de connaissance *a priori* sur la phylogénie à obtenir.

Ces alignements de gènes obtenus avec MAFFT pour les jeux de données mammifère+invertébrés sont concaténés en deux supergènes. La phylogénie correspondant aux séquences de mammifères et invertébrés est obtenue avec IQ-TREE, selon une unique partition, en utilisant ModelFinder pour la sélection du modèle évolutif (Minh et al., 2020 ; Kalyaanamoorthy et al., 2017). Pour les Alvinellidae, l'évaluation du maximum de vrai-

semblance des topologies 1 à 15 présentées au chapitre 1 pour ce nouvel alignement est effectuée avec IQ-TREE selon une partition unique.

Les alignements par gène sont ensuite convertis en séquences indel. Pour cela, les acides aminés sont encodés par la lettre "N", et les indels par le caractère "-". Les régions en 5' et 3' UTR des alignements initiaux sont encodés comme des gaps ('-') mais sont dus à la fragmentation des séquences et non pas à la présence d'indels. Aussi, ces régions sont encodées par le caractère "X", comme fragments manquants.

Les alignements d'indels obtenus sont ensuite concaténés dans un supergène. Pour le jeu de données mammifères (104,013 sites) et le jeu de données invertébrés (37,679 sites), la limite minimum du nombre de séquences présentes pour chaque gène est fixée à 11 espèces. Seules les séquences alignées avec MAFFT ont été concaténées. Pour le jeu de données mammifères+invertébrés, la limite est fixée à 15 espèces. Les séquences alignées avec PRANK (5,417 sites), MAFFT (5,860 sites) et ProbCons (9,619 sites) ont été concaténées séparément. Pour les Alvinellidae, la limite est fixée à 20 transcriptomes (176,458 sites) à partir des séquences alignées par MAFFT.

IV.2.2.2 Modèle de maximum de vraisemblance

Le modèle utilise la topologie d'arbre raciné donnée en entrée pour optimiser un certain nombre de paramètres, tels que la longueur des branches de l'arbre comptée en nombre moyen d'événements d'insertions ou de délétions qui se sont produits entre chaque nucléotide au cours de l'évolution des espèces. La première étape consiste à établir la liste des états ancestraux possibles à chaque site indel. Un site indel (SID) est représenté par l'ensemble des positions contiguës pour lesquelles au moins une séquence présente un caractère gap ("-"). Un SID est donc borné par deux positions de l'alignement pour lesquelles toutes les séquences présentes possèdent un résidu, et les SIDs sont considérés comme indépendants les uns des autres. Nous utilisons le terme SID pour désigner le gap aligné dans les séquences contemporaines, et qui représente caractère phylogénétique sur lequel nous construisons le modèle. Un SID est différent du terme "indel" qui se réfère à l'événement évolutif d'insertion ou de délétion dans une séquence. Les états ancestraux à un SID sont potentiellement infinis, si on se place dans un cadre probabiliste rigoureux. Cependant, certaines possibilités sont bien moins probables que d'autres. Pour constituer les états ancestraux possibles, un exemple est montré en figure IV.13 : les séquences contemporaines sont alignées, chaque transition vers un résidu "N" ou un gap "-" définit une frontière, et les états ancestraux que l'on considère possibles sont l'ensemble des séquences pour lesquelles les fragments entre les frontières peuvent être présents ou absents. L'objectif de cette décomposition est d'avoir tous les chemins parcimonieux qui peuvent permettre de passer d'une séquence potentielle ancestrale à une autre.

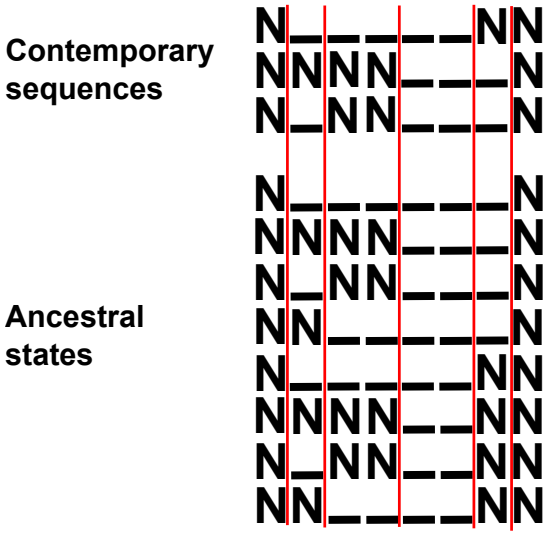


Figure IV.13 – Etats possibles pour les indels ancestraux.

Les états ancestraux sont également définis par la présence des résidus homologues dans les séquences contemporaines. En effet, si un résidu est présent dans au moins deux des trois lignées partant d'un noeud, dans ce cas le résidu doit également être présent à ce noeud et dans tous les noeuds traversés puisque la délétion d'un résidu homologue ne peut pas permettre son insertion ultérieure (irréversibilité du processus). En pratique, cette condition permet de grandement limiter le nombre de séquences ancestrales potentielles, mais cela implique que l'indel tel que reconstruit par l'aligneur doit être particulièrement fiable, car l'aligneur a une grande influence sur le résultat.

La probabilité de transition entre deux séquences, qui est à la base du calcul de la vraisemblance, est détaillée dans la sous-section suivante.

Le calcul de la vraisemblance est effectué entre les sites en prenant en compte les états contemporains observés et les états ancestraux parcimonieux considérés comme possibles. A la racine de l'arbre, on donne à toutes les séquences potentielles ancestrales une probabilité de fréquence uniforme (donc on normalise simplement la vraisemblance de chaque site d'indel par le nombre d'états potentiels de la séquence de l'ancêtre). L'optimisation du modèle de vraisemblance est effectuée avec les paramètres suivants :

— longueur des branches, exprimé en nombre moyen d'insertions ou de délétions par résidu d'acide aminé entre deux noeuds de l'arbre ;

— σ, qui permet la variation du taux de mutation par rapport à la longueur de la branche, similaire au rôle du paramètre Γ classiquement (Yang, 1994). Ainsi, on considère que le taux d'indels varie selon 10^x, où x suit une loi normale de paramètres $(0, \sigma)$. x est discrétisé en 4 classes, correspondant à la moyenne de x pour les quatre quartiles de sa distribution ;

— pi, qui correspond à la probabilité qu'un site change de classes entre deux branches selon le modèle du covarion (Galtier, 2001). Le changement de classe est aléatoire et uniforme entre les 4 classes définies pour x ;

— I, qui donne la proportion de sites invariables.

Enfin, j'ai rajouté une étape préalable qui permet le filtre des gaps selon leur degré d'ambiguité, du fait des limitations des aligneurs. Ce filtre peut se faire en amont selon trois paramètres : L, la taille maximum autorisée pour un SID en nombre de résidus, n le nombre maximum d'événements d'insertions et de délétions possibles pour expliquer la diversité actuelle des SID, en considérant que les états ancestraux possibles correspondent à la définition précédente, et n/L, le nombre maximum d'événements autorisés selon la taille du SID. Ce dernier paramètre vise à autoriser plus d'indels pour un SID plus long, et probablement plus complexe. A noter que le nombre minimum d'événements est toujours d'au moins 1 pour un SID, quel que soit la valeur du paramètre n/L.

Dans le cas de la comparaison de plusieurs topologies entre elles, le nombre de SIDs qui se conforment à ces paramètres pourrait être différent selon la topologie. Dans ce cas, pour être inclus, il faut que le SID passe ces trois critères pour au moins l'une des topologies comparée, autrement il n'est pas considéré dans le calcul de la vraisemblance.

IV.2.2.3 Probabilité d'une insertion et d'une délétion

IV.2.2.4 Simulations

Afin de reproduire la dynamique évolutive des indels, une première étape de simulations a été effectuée. On se place dans le cadre classique des chaînes de Markov, où l'évolution du système au temps $t+1$ dépend uniquement de l'état du système au temps t. Pour passer de t à $t+1$, une matrice de mutations instantanées Q a été construite, qui simule les probabilités relatives d'insertion ou de délétion dans une séquence. Pour cela, je suppose qu'entre deux générations, au moment de la réplication, une insertion ou une délétion peut apparaître après un résidu et que la taille de ces insertions ou délétions est distribuée selon une loi de puissance. La matrice est donc paramétrée avec trois paramètres : r, qui donne le ratio de probabilités entre une insertion et une délétion instantanée, s_1, qui donne la forme de la distribution des insertions (selon x^{-s_1}, x étant la taille de l'insertion) et s_2, qui donne la forme de la distribution des délétions (selon x^{-s_2}, x étant la taille des délétions). La matrice Q est construite selon les hypothèses suivantes :

- Il peut y avoir autant d'indels à un site qu'il y a de génération, et ces indels peuvent se superposer. Entre deux générations en revanche, la probabilité d'observer deux indels simultanément est nulle (par exemple, une séquence qui passe de 1 à 4 résidus est due à une insertion de 3 résidus nécessairement, et ne peut pas résulter d'une insertion de 5 résidus couplée à une délétion de 2 résidus) ;
- Pour une séquence de longueur j, la probabilité instantanée d'une insertion l'amenant à une longueur k est

$$Pr_{ins}(j,k) = (j+1) \times r \times \frac{(k-j)^{-s_1}}{\zeta(s_1)} \tag{IV.19}$$

- Pour une séquence de longueur j, la probabilité instantanée d'une délétion l'amenant à une longueur k est

$$Pr_{del}(j,k) = (k+1) \times \frac{(j-k)^{-s_2}}{\zeta(s_2)} \tag{IV.20}$$

- Il est aussi possible qu'une délétion apparaisse dans la séquence de longueur j et qu'elle soit plus longue que la séquence. Cette délétion est irréversible, aboutissant à une longueur de séquence que je note "-1", puisqu'elle supprime des sites situés en dehors de la séquence permise d'évoluer. Cet événement se produit avec une probabilité

$$Pr_{del}(j,-1) = j+1 - \sum_{n=1}^{j}(k+1) \times \frac{(j-k)^{-s_2}}{\zeta(s_2)} \tag{IV.21}$$

- les coefficients diagonaux, $P_o(j)$, équilibrent les lignes de la matrice à 0.

Ces équations font intervenir la fonction ζ définie par $\zeta(s) = \sum_{n=1}^{+\infty} n^{-s}$ pour $s > 1$. Cette matrice Q, dont la représentation est donnée ci-dessous avec j les lignes et k les colonnes (j et k sont étendus jusqu'à 1200 dans les simulations effectuées), permet d'obtenir les probabilités de changement entre deux tailles de séquences pour n'importe quel temps

avec e^{Qt}. Cependant, le problème est que le nombre de tailles de séquence qui puisse théoriquement être visité est infini, puisque les insertions peuvent s'additionner indéfiniment. Pour avoir une solution correcte, il faut utiliser une matrice beaucoup plus large que les SIDs les plus longs observés dans l'alignement, afin d'inclure des tailles de fragments pour lesquelles la probabilité d'être visités soit suffisamment faible pour ne plus intervenir dans l'évolution du système. Dans ce cas, le temps de calcul peut être très long et ce n'est pas très satisfaisant.

$$\begin{array}{c} \\ -1 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} -1 & 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ P_{del}(0,-1) & P_o(0) & Pr_{ins}(0,1) & Pr_{ins}(0,2) & Pr_{ins}(0,3) & Pr_{ins}(0,4) \\ P_{del}(1,-1) & Pr_{del}(1,0) & P_o(1) & Pr_{ins}(1,2) & Pr_{ins}(1,3) & Pr_{ins}(1,4) \\ P_{del}(2,-1) & Pr_{del}(2,0) & Pr_{del}(2,1) & P_o(2) & Pr_{ins}(2,3) & Pr_{ins}(2,4) \\ P_{del}(3,-1) & Pr_{del}(3,0) & Pr_{del}(3,1) & Pr_{del}(3,2) & P_o(3) & Pr_{ins}(3,4) \\ P_{del}(4,-1) & Pr_{del}(4,0) & Pr_{del}(4,1) & Pr_{del}(4,2) & Pr_{del}(4,3) & P_o(4) \end{pmatrix} \end{array} \tag{IV.22}$$

J'ai toutefois constaté qu'en faisant évoluer cette matrice pour n'importe quelle période de temps, les distributions de longueur des insertions ($P_{ins}(j,k)$ pour j fixé et k variant de $j+1$ à $+\infty$) et des délétions ($P_{del}(j,k)$ pour k fixé et j variant de $k+1$ à $+\infty$) étaient systématiquement bien décrites par deux fonctions du type $A(t)(x+\alpha(t))^{-l_1(t)}$ et $B(t)(x+\beta(t))^{-l_2(t)}$, $r^2 = 1$. Initialement, $A(0) = 0$, $B(0) = 0$, $\alpha(0) = 0$, $\beta(0) = 0$, $l_1(0) = s_1$, $l_2(0) = s_2$,

Ces fonctions ayant trois paramètres dépendant du temps, le calcul des probabilités de trois longueurs x est suffisant pour caractériser toute la distribution. J'ai donc écrit les probabilités de $P_{ins,n}(x,t)$, $P_{del,n}(x,t)$ et $P_{o,n}(t)$, qui correspondent respectivement aux probabilités d'insérer ou de supprimer x résidus à partir d'une séquence de longueur n dans le cas d'une insertion ou d'une délétion, ou d'obtenir une séquence de longueur n à partir d'une séquence de longueur n, à un temps t. Les équations sont écrites pour $n = 0, 1, 2, 3$ et $x = 1, 2, 3$ (avec $x \leq n$ dans le cas des délétions). Par énumération, on obtient les dérivées suivantes en fonction du temps :

$$\begin{aligned} Pr'_{ins,n}(x,t) = & \sum_{z=1}^{n}(n-z+1)Pr_{del,n}(z,t) \times A'(0)(x+z)^{-s_1} \\ & + (n+1)Pr_{o,n}(t) \times A'(0)x^{-s_1} \\ & + \sum_{z=1}^{x-1}(n+z+1)Pr_{ins,n}(z,t) \times A'(0)(x-z)^{-s_1} \\ & - (n+x+1)Pr_{ins,n}(x,t) \\ & + \sum_{z=x+1}^{+\infty}(n+x+1)Pr_{ins,n}(z,t) \times B'(0)(z-x)^{-s_2} \end{aligned} \tag{IV.23}$$

$$
\begin{aligned}
Pr'_{del,n}(x,t) = & \sum_{z=1}^{x-1}(n-x+1)Pr_{del,n}(z,t) \times B'(0)(x-z)^{-s_2} \\
& - (n-x+1)Pr_{del,n}(x,t) \\
& + \sum_{z=1}^{+\infty}(n-x+1)Pr_{ins,n}(z,t) \times B'(0)(z+x)^{-s_2} \\
& + A'(0) \sum_{z=x+1}^{n}(1+n-z)Pr_{del,n}(z,t)(z-x)^{-s_1} \\
& + (n+1-x) \times Pr_{o,n}(t) \times B'(0)x^{-s_2}
\end{aligned}
\tag{IV.24}
$$

$$
\begin{aligned}
Pr'_{o,n}(t) = & \sum_{z=1}^{n}(n-z+1)Pr_{del,n}(z,t) \times A'(0)z^{-s_1} \\
& - (n+1)Pr_{o,n}(t) \\
& + \sum_{z=1}^{+\infty}(n+1)Pr_{ins,n}(z,t) \times B'(0)(z)^{-s_2}
\end{aligned}
\tag{IV.25}
$$

En outre, comme on a définit

$$r = \sum_{x=1}^{+\infty} Pr'_{ins,0}(x) / \sum_{x=1}^{+\infty} Pr'_{del,x}(x)$$

et

$$dt \times [\sum_{x=1}^{+\infty} Pr'_{ins,0}(x) + \sum_{x=1}^{+\infty} Pr'_{del,x}(x)] = 1 - P_{o,0}(t) \approx dt$$

on obtient :

$$B'(0) = \frac{1}{(1+r)\zeta(s_2)} \tag{IV.26}$$

$$A'(0) = \frac{r}{(1+r)\zeta(s_1)} \tag{IV.27}$$

Ces équations permettent de définir l'ensemble du système des probabilités à tout temps t, par la méthode de Runge-Kutta d'ordre 4. Je référence ces équations comme définissant le modèle 3P à trois paramètres. Le calcul est suffisamment précis et plus rapide. Néanmoins, il reste assez lent puisque le système d'équations différentielles ne peut pas être résolu numériquement. C'est la difficulté principale de conserver des distributions de taille des indels qui suivent des lois de puissance, car les sommes infinies présentes dans les équations n'ont pas de solution analytique contrairement à la somme d'une suite géométrique. Lorsque les probabilités du système ont été calculées pour un temps t, on utilise spécifiquement les

probabilités $P_{ins,0}(x,t)$, $P_{del,x}(x,t)$ et $P_{o,0}(t)$ dans le calcul de la vraisemblance du modèle évolutif. Ces probabilités correspondent aux probabilités de voir apparaître entre deux résidus une insertion d'une longueur x à un temps t, de voir une délétion de longueur x exactement entre deux résidus à un temps t, et qu'un site soit resté à son état initial à un temps t (bien que des insertions aient pu survenir entre temps, mais elles ont été supprimées par des délétions ultérieures).

Du fait de la difficulté des calculs, j'introduis un système d'équations plus simple, nommé modèle 1P à un paramètre. Dans ce cas, on considère que le rapport entre le nombre d'insertions et de délétions observées est indépendant du temps, et que la distribution de taille des insertions et des délétions est également invariable au cours du temps. Ce modèle n'est valable que pour des temps de divergence courts, lorsque les SIDs n'ont connus qu'un seul événement d'insertion ou de délétion. Toutefois, nous verrons dans les résultats que le modèle 1P permet en pratique de trouver les paramètres optimaux du modèle de vraisemblance de façon quasi-exacte, au moins sur les exemples étudiés. Le modèle 1P permet donc d'accélérer grandement la maximisation de la vraisemblance. Le modèle 1P, en considérant que le ratio entre insertions et délétions est constant, implique que l'on n'a plus à se soucier de décrire le système de probabilités lié aux délétions qui est en réalité la véritable difficulté de ces modèles puisque ce sont les délétions de sites homologues qui introduisent l'irréversibilité du processus. Le modèle est alors réduit à une simple matrice :

$$Q = \begin{pmatrix} \gamma & \psi \\ \omega & \tau \end{pmatrix} \tag{IV.28}$$

où

$$\begin{aligned} \gamma =& -1 \\ \psi =& B'(0)\zeta(s_1+s_2) \\ \omega =& A'(0) \\ \tau =& 2 \times [-1 + B'(0) \sum_{n=1}^{+\infty} (n+1)^{-s1} n^{-s2}] \end{aligned} \tag{IV.29}$$

Dans ce cas, on obtient :

$$\begin{aligned} P_{o,0}(t) =& \begin{pmatrix} 1 & 0 \end{pmatrix} \times exp(Qt) \times \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ P_{ins,0}(x,t) =& \begin{pmatrix} 0 & 1 \end{pmatrix} \times exp(Qt) \times \begin{pmatrix} 1 \\ 0 \end{pmatrix} \times x^{-s_1} \\ P_{del,x}(x,t) =& P_{ins,0}(1,t) \times \frac{\zeta(s_1)}{r \times \zeta(s_2)} \times x^{-s_2} \end{aligned} \tag{IV.30}$$

La matrice Q et les équations qui en découlent ont l'avantage d'être rapide à calculer et facilement dérivables. On note aussi que dans toutes ces équations, j'ai utilisé le terme "résidu". Ces équations pourraient tout aussi bien s'appliquer à des séquences nucléotidiques

que protéiques, mais les séquences utilisées dans cette partie sont toutes en acides aminés. Enfin, les "chop zones" (De Maio, 2021), qui sont les zones pour lesquelles deux séquences alignées présentent un SID constitué de deux gaps successifs dans chacune des séquences, sont calculés dans cette approche par la probabilité d'avoir une insertion et une délétion de façon indépendante.

IV.2.3 Résultats et Discussion

IV.2.3.0.0.1 Répartition des tailles et complexité des indels Le choix de l'aligneur influence la qualité des indels obtenus. Aussi, pour comparer cet effet, la figure IV.14 montre la qualité d'assignation des SIDs sur la phylogénie métazoaire selon les trois programmes d'alignement utilisés, PRANK, ProbCons et MAFFT. ProbCons est l'aligneur produisant le plus de SIDs, avec 501 sites formés, suivi de PRANK (347) et MAFFT (233). Les trois aligneurs produisent pourtant des résultats qualitativement similaires au niveau de la distribution des tailles et de la complexité des SIDs. ProbCons tend néanmoins à produire des SIDs un peu plus longs et moins parcimonieux (plus de 4 insertions ou délétions sur un site, panneau n de la figure). Il apparaît que MAFFT est généralement plus conservatif, produisant moins de SIDs et des SIDs plus courts que les autres aligneurs, mais PRANK est l'aligneur qui produit les SIDs les plus parcimonieux. Notamment, le panneau n/L montre le nombre d'indels minimum par site selon la longueur du site. Le pic à 1 correspond à des indels d'un seul acide aminé dans une séquence. Ici, PRANK et MAFFT ne produisent pratiquement aucun site au delà de 1, au contraire de ProbCons. ProbCons infère par conséquent beaucoup plus de petits sites d'indels d'un ou deux acides aminés très peu parcimonieux une fois mis sur la phylogénie. Rappelons que les alignements sont effectués sans utiliser de guide tree fourni par l'utilisateur, afin de mimer le cas où l'on analyserait des données pour lesquelles on n'aurait aucune connaissance *a priori* sur la phylogénie à obtenir.

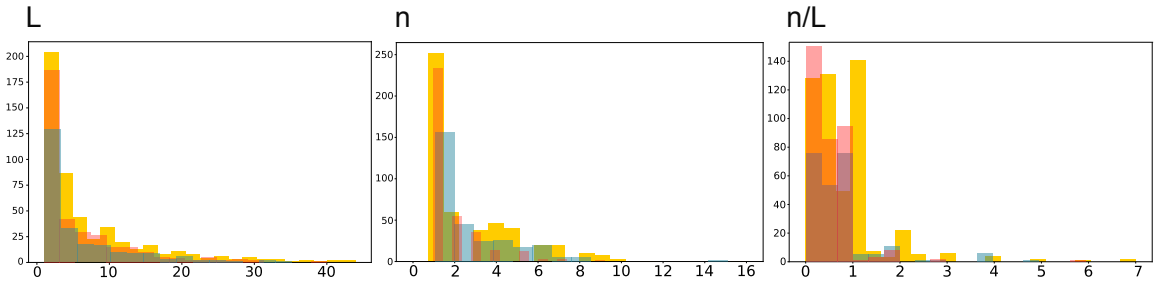


FIGURE IV.14 – Distribution de la complexité des indels sur la phylogénie des métazoaires. Les résultats produits par PRANK sont en rouge, en jaune pour ProbCons et bleu pour MAFFT. L représente la longueur des sites d'indels en nombre de positions dans l'alignement, n le nombre minimum d'événements d'insertion ou de délétion pour expliquer la diversité actuelle du site, et n/L le ratio entre ces deux mesures.

Afin de comparer les résultats obtenus, nous choisissons trois niveaux de filtre des SIDs : une condition "no filter" pour laquelle les données brutes sont utilisées, une condition "soft filter" pour laquelle les SIDs de longueur à plus de 20 positions (L), ou plus de 5 événements en parcimonie (n), ou n/L supérieur à 0.4, sont ignorés, et enfin une condition "hard filter" pour laquelle les SIDs d'une longueur (L) de plus de 10 positions, 2 évènements parcimonieux (n) ou n/L supérieur à 0.3 sont ignorés. Rappelons que le paramètre n/L autorise quand même systématiquement au moins un événement de création d'indel pour un SID, même si le SID en question est très court. La figure IV.3 récapitule les optimisations obtenues pour les différentes conditions de filtre sur les phylogénies de métazoaires. Le tableau A. montre qu'augmenter la force du filtre sur les indels a une forte influence sur les paramètres obtenus pour caractériser les indels. Notamment, un filtre plus fort diminue la variabilité des taux d'insertion et de délétion (paramètre σ), tend à raccourcir les insertions et délétions (s_1 et s_2) et surtout modifie de manière drastique le ratio entre la création d'insertions et de délétions instantanées, r. La diminution de la taille des indels n'est pas surprenante puisque les filtres tendent à raccourcir la longueur des sites. En revanche, la diminution de r avec l'ajout de nouveaux indels est nécessairement un artefact de la méthode. Il s'agit d'un problème relevé plusieurs fois par Löytynoja et Goldman, 2008a ou la création de fausses d'homologies entre des résidus sur l'alignement a pour conséquence d'augmenter le nombre de résidus homologues également dans les ancêtres communs de ces séquences. Ces résidus faussement imposés aux séquences ancestrales doivent être compensés par plus de délétions dans certaines branches descendantes, impliquant une surestimation du nombre de délétions. La condition "hard filter" serait potentiellement celle où le biais d'estimation de ce paramètre est le moins fort, puisqu'on se concentre sur les régions les moins variables et les moins complexes à résoudre, comme indiqué par le score de vraisemblance moyen par site qui est le plus élevé dans cette condition et similaire entre aligneurs. Dans ce cas, l'image est différente puisque les insertions sont environ deux fois plus fréquentes que les délétions, et les trois programmes sont relativement en accord sur cette estimation. Au contraire, MAFFT apparaît comme un aligneur très conservateur dans la condition "no filter", avec près de quatre fois plus de délétions que d'insertions, ce qui s'explique sûrement par la fusion d'indels entre eux. Dans la littérature, les valeurs de r sont généralement établies entre 0.25 et 0.8 (Cheynier et al., 2001 ; Paśko, Ericson et Elzanowski, 2011 ; Donath et Stadler, 2018). Ainsi le biais en faveur des délétions est bien établi, mais dans les alignements étudiés dans ce chapitre en se concentrant sur les sites d'indels les moins complexes, ce biais n'est pas évident et peut même apparaître inversé en faveur des insertions.

Au niveau des paramètres instantanés sur la taille des indels, s_1 est généralement plus petit que s_2, ce qui implique des insertions un peu plus courtes que les délétions. Selon nos optimisations, on aurait donc une dynamique globale d'insertions plus fréquentes que les délétions, mais plus courtes. Pour ces protéines, on pourrait en outre s'attendre à ce qu'il y ait un équilibre entre le nombre moyen de résidus insérés entre deux générations et le nombre moyen de délétions, puisque la taille des protéines devrait être relativement constante. On ne peut pas réellement estimer cette différence ici, parce que la taille d'insertion ou de délétion avec des paramètres s inférieurs à 2 implique des moyennes à l'infini. Au départ, j'avais limité le paramètre s à une gamme de valeurs qui évite ce cas qui n'est évidemment pas pertinent biologiquement. Toutefois, des paramètres s entre 1.5 et 2 ont également été mesurés par Gu et Li, 1995 et Cartwright, 2009 sur un ensemble de gènes et pseudo-gènes. L'estimation est donc probablement correcte, mais d'autres paramètres

devraient être pris en compte dans la modélisation des indels qui échappent au modèle actuel (comme probablement différents mécanismes d'évolution des indels qui concernent des gammes de taille différentes, Creer, 2007 ; Paśko, Ericson et Elzanowski, 2011). Pour les indels courts (moins de 20-30 nucléotides, soit 10 acides aminés), l'utilité d'une loi zéta pour décrire les distributions comme utilisée ici est bien établie (Cartwright, 2009 ; Paśko, Ericson et Elzanowski, 2011).

Le paramètre I quantifie la part de sites invariables dans les protéines, qui s'établit entre 70 et 90% des séquences en prenant les conditions "no filter", ce qui montre, comme on pouvait s'y attendre, que les indels sont contre sélectionnés dans la majorité des séquences codantes. Cela implique également que les indels ont tendance à se regrouper dans des "hotspots" qui rendent la reconstruction plus difficile du fait de leur superposition potentielle (Redelings et Suchard, 2007).

Enfin, les longueurs de branches sont très corrélées entre l'optimisation obtenue à partir des indels et l'optimisation dans le modèle de vraisemblance s'appuyant sur les substitutions des acides aminés (voir figure supplémentaire 1). Pour les conditions no filter, le ratio moyen entre longueurs de branches en acides aminés et longueurs de branches en indels s'établit à 0.008 pour PRANK, 0.0053 pour MAFFT et 0.0102 pour ProbCons. Ceci équivaut à des taux de création d'indels par substitution entre 0.5 et 1%. Considérant le paramètre I, cela implique que dans les zones variables uniquement, qui représentent 15 à 25% de l'alignement, les taux d'indels par substitution s'établissent à 3% pour PRANK, 4% pour MAFFT et 5% pour ProbCons. Ces ordres de grandeur sont en bon accord avec ceux qui ont été relevés dans la littérature : entre 15 et moins de 10% sur des gènes entiers (Cartwright, 2009 ; Westesson et al., 2012), entre 7 et 12% entre la souris et l'humain sur le génome (Donath et Stadler, 2018), taux d'indels environ 10 fois moindre dans l'ADN codant par rapport à l'ADN non codant (Ashkenazy, Cohen et al., 2014).

A

		nbr. gaps	σ	π	I	s_1	s_2	r	ML
PRANK	no filter	347	0.50	0.50	0.73	1.64	1.78	1.03	14.68
	soft filter	271	0.10	0.50	0.74	1.65	1.74	1.58	12.30
	hard filter	224	0	0	0.58	1.73	1.82	1.96	10.02
ProbCons	no filter	501	0.43	0.02	0.78	1.62	1.88	0.47	20.88
	soft filter	278	0	0	0.78	1.68	1.72	1.55	12.36
	hard filter	231	0	0	0.68	1.87	1.97	2.14	10.09
MAFFT	no filter	233	0.42	0	0.86	1.67	1.86	0.26	19.66
	soft filter	148	0.05	0	0.87	1.62	1.78	0.63	13.26
	hard filter	116	0	0	0.61	1.83	1.91	1.67	10.17

B

		nbr. gaps	σ	π	I	s_1	s_2	r	ML
mammals	no filter	2555	0.10	0	0.71	1.65	1.54	0.97	12.64
	soft filter	1905	0	0	0.78	1.76	1.63	1.38	8.57
	hard filter	1561	0	0	0.82	1.83	1.67	1.47	7.62
invertebrates	no filter	878	0.53	0	0.81	1.60	1.78	0.36	20.12
	soft filter	538	0	0	0.64	1.71	1.68	0.86	11.98
	hard filter	438	0	0	0.51	1.84	1.87	1.04	9.89

Table IV.3 – Caractéristiques des indels sur les phylogénies de métazoaires. A. phylogénie de l'ensemble des métazoaires, selon les trois programmes d'alignement et les trois conditions de filtrage. B. phylogénie "mammals" ou "invertebrates", selon le programme MAFFT dans les trois conditions de filtre. La vraisemblance (ML) est normalisée par le nombre de SIDs retenus dans chaque condition.

Deux question persistent : quel aligneur et quelle condition de filtre choisir ? Dans le choix de l'aligneur, PRANK apparaît comme celui ayant les estimations les plus stables entre différentes conditions, notamment pour le paramètre r bien qu'il décroisse également avec l'ajout de SIDs. Les vraisemblances par SID sont également notablement meilleures dans la condition "no filter" que pour les autres aligneurs. Pour la suite, j'ai toutefois choisi l'aligneur MAFFT parce qu'il serait supérieur à PRANK dans la littérature (Dessimoz et Gil, 2010 ; Ashkenazy, Cohen et al., 2014 ; Donath et Stadler, 2018). Dans le futur, je souhaiterais estimer une comparaison similaire de ces aligneurs sur des procaryotes ou champignons pour voir si les conclusions de mon algorithme restent les mêmes.

Concernant la condition, on serait tenté de choisir la condition "hard filter" qui est probablement celle qui donne les meilleures estimations de paramètres. Cependant, le problème de ce filtre est qu'il retire tous les SIDs complexes qui ont un signal phylogénétique profond. La figure supplémentaire 1 compare les optimisations de branche obtenues sur cette phylogénie avec les trois conditions, et l'on constate que les branches profondes disparaissent en augmentant la force du filtre. Par conséquent, dans la suite, j'ai considéré que "soft filter" était le meilleur compromis entre information phylogénétique et fiabilité des indels. D'autres méthodes ont été suggérées pour filtrer les indels : Donath et Stadler, 2018 ignorent les gaps de taille 1, considérant que l'homoplasie est importante sur ces sites. Ashkenazy, Cohen et al., 2014 développent une méthode RELINDEL, qui consiste à évaluer la stabilité des gaps selon un bootstrap sur l'alignement de séquences. Leur méthode, notamment, permet d'établir que 23% des indels obtenus par MAFFT sont fiables, et seulement 3% pour PRANK. Ici nous avons utilisé un filtre simple basé sur la parcimonie des indels, mais ces suggestions mériteraient d'être explorées. D'autres auteurs ont également suggéré de conduire l'alignement des séquences en même temps que la phylogénie (Redelings et Suchard, 2007). Comme notre méthode permet également d'estimer des probabilités rapides sur les insertions et les délétions (*via* le modèle 1P), cette option pourrait être envisagée mais elle nécessite dans ce cas d'estimer conjointement les probabilités de substitution sur les acides aminés, car le modèle indel présenté ici ignore complètement cet aspect de l'alignement.

Le tableau IV.3B. compare les paramètres obtenus pour les deux sous-phylogénies centrées sur les mammifères ou les invertébrés. L'objectif était de comparer si les paramètres r, s_1 et s_2, étaient différents dans les deux groupes, considérant que ces paramètres sont ceux qui décrivent la dynamique instantanée des indels dans les protéines. Pour toutes les conditions, le paramètres r suggère que la proportion d'insertions est plus importante chez les mammifères, avec des insertions potentiellement un peu plus courtes que chez les invertébrés, et des délétions plus longues que chez les invertébrés. On aurait donc une image selon laquelle les insertions courtes sont fréquentes par rapport à de longues délétions chez les mammifères, et des délétions courtes mais plus fréquentes chez les invertébrés. Il faut cependant remarquer que la vraisemblance par site est plus élevé chez les mammifères, pour lesquels les temps de divergence sont plus courts que chez les invertébrés considérés ici. Le modèle utilisé en vraisemblance est supposé prendre en compte la dynamique des indels sur les temps longs, mais ce résultat pourrait être biaisé par cette différence. Si l'on compare des conditions similaires en terme de vraisemblance (par exemple hard filter invertebrates et soft filter mammals), il semble tout de même y avoir une différence de dynamique dans la proportion des délétions et leur taille. L'étude des dynamiques d'indels dans les diffé-

rentes lignées serait très intéressant pour mesurer dans quelle mesure elle varie et si elle est potentiellement sélectionnée.

IV.2.3.0.0.2 algorithme de maximisation de la vraisemblance Quelques points spécifiques dans le développement de l'algorithme de maximisation de la vraisemblance sur les SIDs méritent d'être abordés. Tout d'abord, quelle est la fiabilité des modèles 1P et 3P, dans lesquels les distributions de taille des indels et le ratio entre insertions et délétions sont considérés soit constants (1P), soit variables au cours du temps (3P). La figure IV.15 montre les probabilités calculées pour les insertions de taille 1, 2 ou 3 résidus et les délétions de 1, 2 ou 3 résidus, ainsi que la probabilité de ne pas observer d'indel entre deux résidus, après un temps t. Le calcul des probabilités dans le modèle 3P reproduit bien les attendus des simulations. Les probabilités obtenues pour le modèle 1P reproduisent la bonne dynamique mais dévient rapidement des vraies probabilités, après un temps d'évolution d'environ 0,2. En pratique, les vraisemblances des phylogénies ont été optimisées d'abord avec le modèle 1P qui est beaucoup plus rapide à calculer. Comme les temps de divergence pour les indels sont très courts, le changement entre le modèle 1P et 3P n'induit aucun changement dans les paramètres estimés dans les conditions "hard filter". Les vraisemblances calculées dans les deux modèles diffèrent de moins de 1 unité log, sans doute parce que les SIDs complexes sur lesquels les unités d'insertion ou de délétion se superposent ont été retirés. Sur la condition "soft filter", la vraisemblance est meilleure avec le modèle 3P, sans changement dans l'estimation des paramètres du modèle. Dans la condition "no filter", le gain de vraisemblance est assez important (entre 50 et 100 unités log), et la plupart des paramètres estimés sont identiques, exceptés pour les paramètres s_1 et s_2 qui sont jusqu'à 0.1 plus grands dans le modèle 3P que dans le modèle 1P. Ceci témoigne du fait que les indels plus longs sont mieux expliqués dans le modèle 3P par la superposition de plusieurs indels courts. Ainsi, une bonne stratégie pour l'optimisation de la vraisemblance est de trouver les meilleurs paramètres grâce au modèle 1P, puis éventuellement de corriger les paramètres s_1, s_2 et r avec le modèle 3P. Dans les phylogénies avec des séquences codantes peu divergentes, les deux modèles permettent d'estimer les paramètres de façon très similaire.

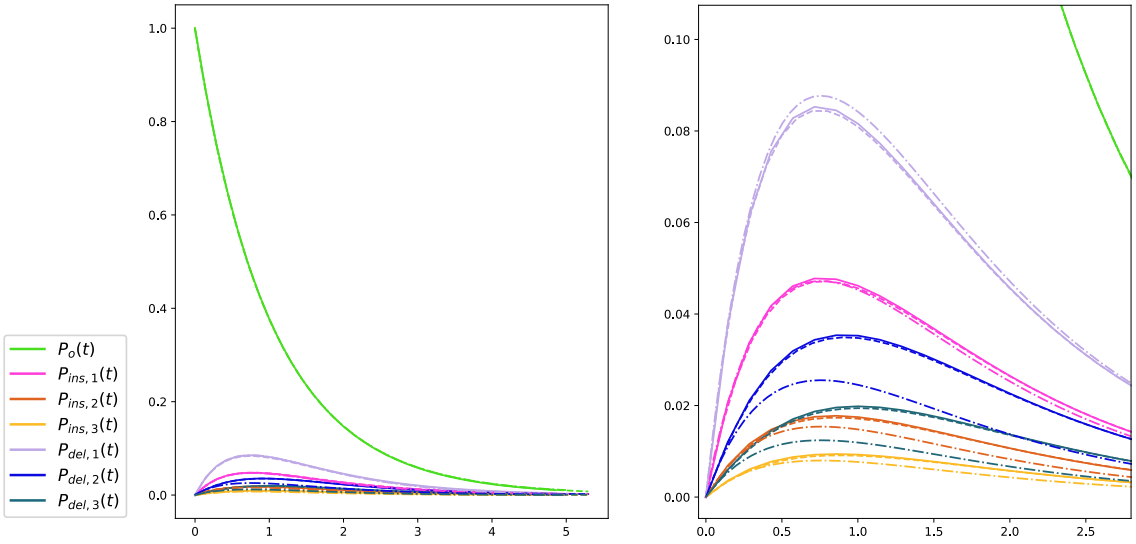


FIGURE IV.15 – Probabilités calculées par le modèle 1P ou 3P. La probabilité de ne pas observer d'indel entre deux résidus à un temps t, $P_o(t)$, est en vert. Les probabilités d'observer une insertion de taille x à t entre deux résidus sont notées $P_{ins,x}(t)$, et les probabilités d'observer une délétion de taille x à t entre deux résidus sont notées $P_{del,x}(t)$. Le panneau de droite est un zoom du panneau de gauche. Les probabilités simulées sont les lignes continues. Le modèle 1P est représenté par les lignes semi-pointillées (•-•-•) et le modèle 3P est tracé en pointillés (- - -). Les courbes sont obtenues en utilisant les paramètres $s_1 = 1.62$, $s_2 = 1.78$, $r = 0.63$, ce qui correspond aux paramètres estimées pour la phylogénie de métazoaires avec les conditions MAFFT-soft filter.

Enfin, un taux variable d'insertions et de délétions est modélisé par le paramètre σ. La figure IV.16 montre la distribution des vitesses évolutives des indels à partir des longueurs de branches estimées par le modèle de vraisemblance sur l'alignement métazoaires no filter. Pour obtenir cette figure, nous avons calculé pour chaque SID la probabilité d'obtenir les séquences contemporaines selon une gamme de vitesse qui varie selon 10^x avec x pris entre -2 et 2. Chaque SID est ensuite associé avec la catégorie de x qui a la meilleure probabilité d'expliquer la vitesse observée. On constate que la distribution de la vitesse d'apparition des indels selon x s'accorde assez bien avec une distribution normale, ce qui a motivé le choix d'un paramètre σ qui donne la variance d'une loi normale. Pourtant, après optimisation des paramètres du modèle (et notamment des longueurs de branches), on constate que x n'est pas centré sur 0 dans la condition hard filter, mais sur 0.5 environ. En regardant dans le détail la distribution des taux de mutation, il apparaît que la plupart des SIDs s'accordent sur des longueurs de branche environ $10^{0.5}$ soit environ trois fois plus longues que celles optimisées par maximum de vraisemblance. En triplant manuellement la longueur des branches, on améliore effectivement la vraisemblance sur les SIDs, mais on détériore la vraisemblance sur l'alignement total du fait de la contrainte des sites invariants, qui ne peut pas être compensée par le paramètre I. Ainsi, dans la condition no filter, beaucoup de gaps ont des vitesses de mutation qui sont trop élevées par rapport au calcul des probabilités de mutation que j'obtiens à l'aide des modèles 1P et 3P. Ceci pourrait être dû soit à une sous-estimation du nombre d'indels sur les portions invariantes de l'alignement, soit à une surestimation du nombre d'événements (n) dans les SIDs reconstruits. Ce problème disparaît en revanche lorsqu'on filtre les SIDs pour conserver uniquement les moins complexes (conditions soft filter et hard filter). Dans ces conditions, σ vaut 0 et les SIDs ont des probabilités de mutation compatibles avec la probabilité des sites invariants de ne pas muter. Je pense que ce problème de SIDs trop rapides par rapport à la taille optimale des branches est lié au fait d'utiliser des alignements qui ne reposent pas sur un modèle d'évolution moléculaire, mais sur des pénalités arbitraires d'ouverture et allongement de gaps. Si on ne veut pas filtrer les gaps, il serait plus judicieux d'aligner les séquences en prenant déjà en compte des probabilités issues d'un modèle d'évolution moléculaire comme les modèles 1P ou 3P. L'autre possibilité est que le modèle repose sur la mauvaise hypothèse de vouloir unifier le comportement des indels dans l'alignement, alors qu'on devrait distinguer le processus de mutation, qui concerne l'intégralité de la séquence, du processus de sélection différentiel entre des régions où les indels sont systématiquement contre-sélectionnées et des régions qui autorisent des indels, et qui en conséquence sont autorisées à muter.

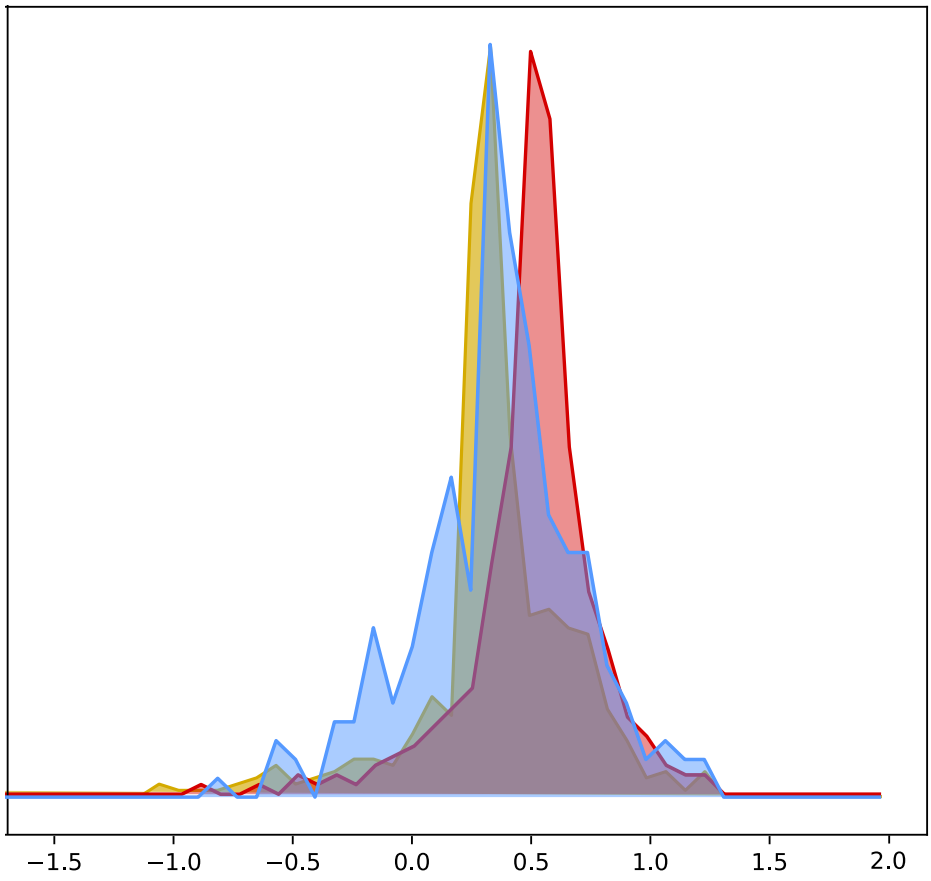


FIGURE IV.16 – Distribution des vitesses évolutives des indels sur l'alignement métazoaires-no filter. PRANK est en rouge, ProbCons en jaune et MAFFT en bleu.

Avec les estimations disponibles des paramètres, il devrait être possible de construire un modèle qui optimise la vraisemblance à la fois sur les résidus des séquences et sur les gaps. Les paramètres s_1, s_2 et r pourraient être fixés selon le modèle 1P qui donne des performances satisfaisantes au regard de son coût en calculs, à l'instar d'une matrice empirique de mutations des acides aminés. Quelques unes de ces matrices pré-optimisées sont données en figure supplémentaire 2. Les longueurs de branches à partir des matrices de substitution en acides aminés ou de création d'indels pourraient être unifiées en considérant le ratio trouvé d'environ 1% entre le nombre d'indels par rapport au nombre de mutations de résidus. Deux points pourraient être notamment améliorés dans le modèle indel :

— La reconstruction des états ancestraux des indels. Pour l'instant, nous considérons des possibilités d'états ancestraux en s'inspirant du Complex Indel Coding (Simmons et Ochoterena, 2000). Plus récemment, certains auteurs ont proposé des algorithmes plus complexes qui retracent l'ensemble des chemins évolutifs possibles et augmentent graduellement la taille des états ancestraux possibles selon leurs probabilités (Miklos, 2003 ; Westesson et al., 2012 ; Ishikawa et al., 2019). Cette option semble plus intéressante, mais il faudrait étudier sa faisabilité dans un cas où l'on doit optimiser plusieurs centaines d'indels de manière itérative ;

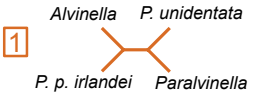
— La fiabilité des alignements autour des indels. Nous ne nous intéressons pas ici aux algorithmes d'alignements eux-mêmes. Cependant, un point qui m'interroge, bien que je ne l'ai pas étudié spécifiquement, est la question du poids donné à un SID mal aligné dans le modèle. Un tel SID devrait avoir une vraisemblance très faible du fait de la multiplication des événements d'insertion et délétion nécessaires à l'expliquer. Par conséquent, il est possible que les gaps mal alignés soient responsables de la majorité du signal phylogénétique mesuré sur les alignements. Il faudrait une méthode pour normaliser ce bruit et l'empêcher d'enfler. On sait que la prise en compte de la taille des gaps et de la complexité des indels est importante pour obtenir un bon signal phylogénétique (Müller, 2006 ; Simmons, Müller et Norton, 2007), donc la normalisation du signal pour chaque indel n'est sûrement pas la meilleure solution. Plutôt que de travailler sur un supergène, un bon compromis serait peut être de construire d'abord les arbres de gène selon un modèle hybride acides aminés/indel, puis de réconcilier ces arbres de gène. Ainsi, même si certains gaps mal alignés induisent des erreurs dans certains arbres de gène, la vraisemblance extrêmement mauvaise qu'ils pourraient obtenir n'influencerait pas le résultat obtenu sur d'autres arbres. Si l'on souhaite établir une phylogénie uniquement basée sur les indels sans prendre en compte les remplacements en acides aminés, il n'y a probablement pas assez de signal phylogénétique contenu dans les indels pour établir une topologie complète pour chaque arbre de gène. Dans ce cas, on pourrait établir des sous-arbres de gène qu'on pourrait effectivement définir grâce aux indels présents dans chaque gène, puis trouver l'arbre des espèces qui soit le plus en accord avec les sous-arbres obtenus.

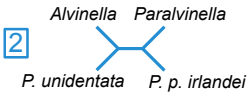
IV.2.3.0.0.3 Application à la phylogénie des Alvinellidae La figure IV.17 montre les résultats obtenus par le modèle de vraisemblance indel sur la phylogénie des Alvinellidae, en considérant un seul supergène issu de la concaténation des 888 gènes. Pour la comparaison des topologies, les alignements de MAFFT sont utilisés sous la condition "soft filter". Pour que les SIDs retenus ne diffèrent pas d'une phylogénie à l'autre, nous impo-

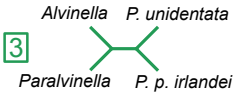
sons qu'un indel doit entrer dans les conditions imposées par "soft filter" pour au moins une topologie (donc, en pratique, qu'au moins une des topologies doit pouvoir expliquer un SID de manière suffisamment parcimonieuse). Les 15 topologies envisagées au chapitre 1 ont été comparées, et ont convergé vers les mêmes paramètres optimaux : 874 SIDs sont retenus, avec $\sigma = 0$, $\pi = 0$, $I = 0.92$. Les paramètres correspondant à la dynamique des indels sont $s_1 = 2.28$, $s_2 = 1.67$, $r = 0.66$. Le paramètre s_2, qui quantifie les tailles des délétions, est identique à celui observé pour les délétions chez les invertébrés (IV.3B.). Le paramètre r est légèrement plus faible chez les Alvinellidae. Ce paramètre est très sensible à la qualité de l'alignement, néanmoins nous avons établi que les erreurs dans l'alignement avaient tendance à diminuer r. Pour les Alvinellidae on s'attendrait à ce qu'il y ait moins d'erreurs puisque les séquences sont plus similaires. Il est donc possible que les délétions soient plus fréquentes dans la famille, mais il faudrait confirmer ce résultat. s_1 en revanche, est beaucoup plus grand dans la phylogénie des Alvinellidae, chose qui n'était observée dans aucune des phylogénies optimisées précédemment. Ce résultat semble bien montrer que les insertions sont en moyenne beaucoup plus courtes dans cette famille. On aurait donc une dynamique avec moins d'insertions, et plus courtes chez les Alvinellidae, ce qui est compatible avec l'hypothèse d'une lignée globalement sous sélection pour des températures chaudes pour lesquelles les protéines sont souvent plus courtes, notamment par la réduction des boucles (Thompson et Eisenberg, 1999). Ma phylogénie des Alvinellidae inclut toutefois des outgroups, et rigoureusement il faudrait tester si cette dynamique particulière d'indels est bien spécifique de la famille et n'est pas tirée par certaines espèces extérieures particulières.

Concernant les vraisemblances des différentes topologies, les différences sont très faibles pour le modèle indel. En effet, seuls 874 SIDs sont retenus, et les branches qui nous intéressent, très courtes, ne se distinguent que par quelques indels diagnostiques. En moyenne, la vraisemblance d'un SID dans ces phylogénies vaut environ 13 unités log. Ainsi, ces phylogénies ne sont peut être distinguées que par trois ou quatre SIDs. Il est toutefois intéressant de constater que les topologies ayant les meilleurs scores sont la T13, T3, T6, T9, T14 et T1. Les trois premières topologies correspondent toutes au scénario 3, dans lequel les espèces *P. unidentata* et *P. p. irlandei* sont jumelles par rapport aux autres espèces Alvinellidae. Les topologies T9, T14 et T1 correspondent au scénario 2, pour lequel *P. unidentata* et les espèces *Alvinella* sont jumelles. Les topologies avec les moins bons scores sont T12, T4, T2, et correspondent toutes au scénario 1 pour lesquels *P. p. irlandei* et *Alvinella* sont jumelles. Ces résultats rappellent beaucoup ceux déjà obtenus au chapitre 1 du manuscrit. Évidemment, les résultats ne sont pas indépendants et les indels responsables des meilleurs scores des topologies T13, T3 et T6 sont vraisemblablement les mêmes, tout comme doivent l'être les indels responsables des moins bons scores de T12, T4 et T2. Ce résultat devrait être confirmé par beaucoup plus d'indels, et en l'occurrence cette méthode n'est peut être pas adaptée à la résolution de la phylogénie pour des branches aussi courtes. Néanmoins, cela confirme que la prise en compte des indels conjointement avec celle des remplacements d' acides aminés peut améliorer la qualité du signal phylogénétique.

	amino acids	gaps	scenario
T1	602	19	2
T2	662	36	1
T3	308	4	3
T4	413	37	1
T5	449	34	2
T6	64	13	3
T7	101	34	1
T8	75	25	3
T9	0	18	2
T10	659	34	2
T11	475	28	3
T12	677	37	1
T13	366	0	3
T14	516	19	2
T15	774	24	1







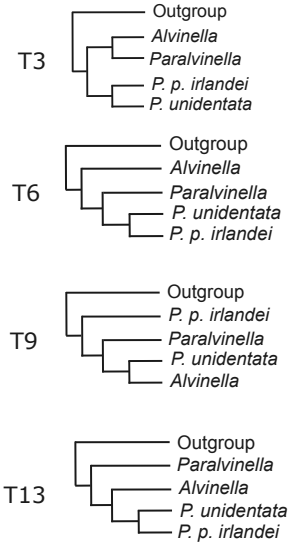


Figure IV.17 – Évaluation des phylogénies Alvinellidae sur la base des indels. La vraisemblance obtenue par le modèle LG+Γ sur le supergène concaténé en acides aminés des Alvinellidae est indiqué sous "amino acids", relativement à la vraisemblance obtenue par la topologie T9. La vraisemblance des modèles indels est indiquée relativement à la topologie T13. Les scénarios (cf. figure 3, chapitre 1) sont rappelés, ainsi que les 4 topologies issues du chapitre 1 ayant obtenues les meilleures vraisemblances d'après l'analyse des indels, T3, T6, T9, T13.

La figure IV.18 compare la topologie T9, qui obtient le meilleur résultat selon le modèle LG basé sur les remplacements d'acides aminés, avec la même topologie T9 optimisée à partir des indels. On constate que les longueurs de branche sont globalement très corrélées. Les branches internes tendent à être plus courtes, mais cet artefact était prévisible car lié à la filtration des indels avec "soft filter" ou "hard filter" (voir figure supplémentaire 1). On remarque cependant que les branches menant aux espèces *P. sulfincola* et *P. p. irlandei* sont beaucoup plus longues. Ces deux espèces sont caractérisées par des transcriptomes de moins bonne qualité dans le jeu de données actuel des Alvinellidae. En particulier, les séquences sont beaucoup fragmentées. L'algorithme de maximum de vraisemblance indel est codé pour ignorer les séquences manquantes, donc la fragmentation en soi ne devrait pas être problématique. En revanche, au vu de ce résultat, il est probable que cela induise un biais au moment de l'alignement avec la création de gaps spécifiques à l'espèce. En revanche, le transcriptome de *P. p. irlandei* est moins fragmenté que celui de *P. sulfincola*, et nous avons pris soin à ce que les séquences de *P. p. irlandei* soient de la meilleure qualité possible en amont des alignements. Par conséquent, même si la branche menant à *P. p. irlandei* est probablement trop longue par rapport à la réalité, il n'est pas exclu que cette lignée ait effectivement connu un taux plus élevé d'indels. Ceci serait en accord avec l'observation que le transcriptome de *P. p. irlandei*, en compositions d'acides aminés, est également très différent de tous les autres Alvinellidae et des outgroups. L'espèce *P. p. irlandei* a sûrement connu un taux d'évolution plus fort que les autres espèces d'Alvinellidae, pour des raisons qui restent pour l'instant à élucider.

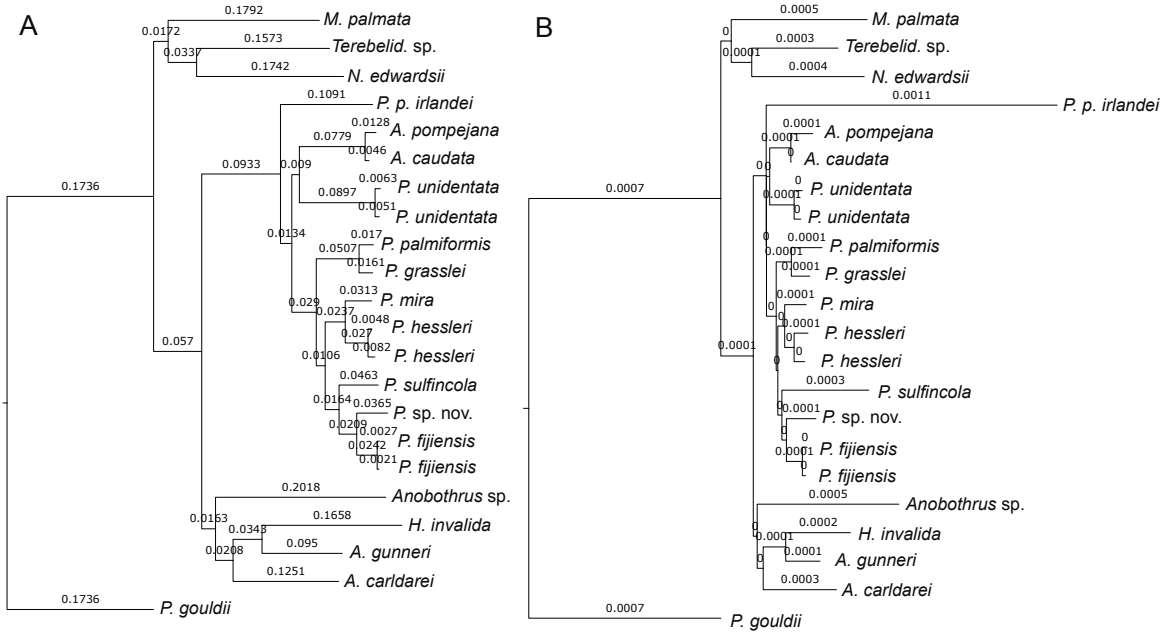


Figure IV.18 – Comparaison des topologies T9 optimisées par le modèle LG et le modèle indel. A. optimisation des longueurs de branches selon la matrice LG et le paramètre Γ. B. optimisation des longueurs de branches par les indels.

IV.3 Annexes

Annexe

Modèle Ecoprior

Le modèle Ecoprior, optimisé sur des alignements de gènes dans le chapitre 3, montre des variations importantes dans les parts attribuées à chaque branche de la phylogénie aux matrices thermiques chaude HJM et froide CJM, ainsi qu'à la matrice PFASUM.
Pour établir la validité de ce modèle, l'optimisation a été effectuée à nouveau sur des alignements plus longs. A partir de l'alignement de 888 gènes utilisés dans la partie Indels du chapitre 3, un échantillonnage aléatoire de 10,000 sites a été effectué pour lesquels tous les transcriptomes utilisés dans l'analyse présentent un résidu d'acide aminé. Ainsi, cet alignement ne présente pas de gaps, et n'a pas été filtré pour réduire le biais de variation de composition des séquences en acides aminés. L'optimisation du modèle Ecoprior est effectuée sur la topologie 6 du chapitre 1. Le résultat est montré en figure 1.

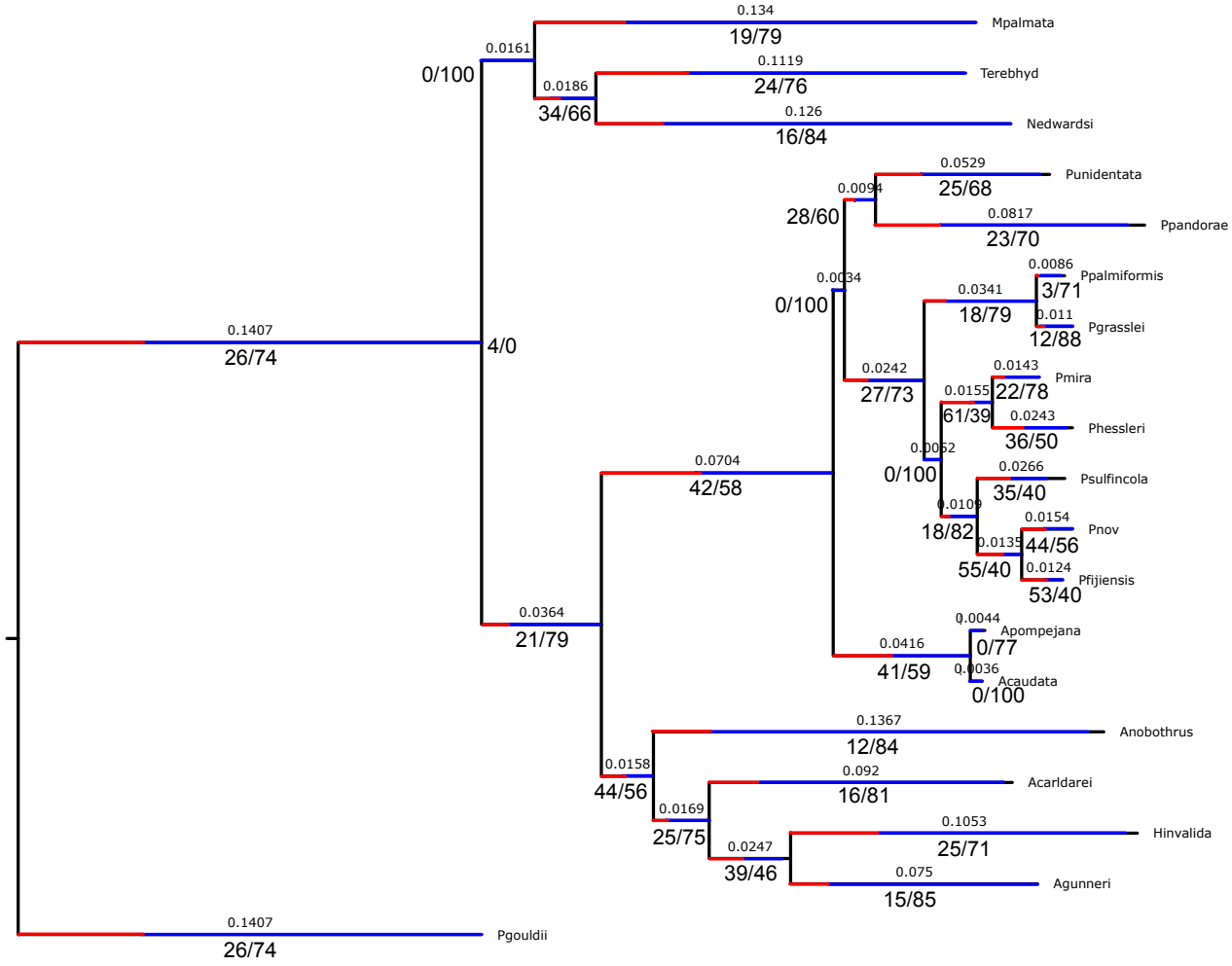


Figure 1. Optimisation du modèle Ecoprior+G sur un alignement de 10,000 sites contenant les espèces de la famille des Alvinellidae et des espèces outgroup. Les longueurs de branche indiquent le nombre attendu de mutations par résidu. Les chiffres X/Y sous les branches indiquent les parts respectives, en pourcent, des matrices HJM et CJM dans chaque branche, qui sont également représentées en rouge et bleu sur les branches. L'optimisation des fréquences à la racine de l'arbre est indiquée de la même manière.

Comme observé en figure 1, la part de la matrice PFASUM est généralement réduite à 0% dans l'optimisation du modèle sur la phylogénie. L'optimisation se fait donc globalement sur un axe unique entre la matrice froide CJM et la matrice chaude HJM. Dans l'objectif initial du modèle, nous voulions que la part de la matrice PFASUM sur la branche corresponde au temps de divergence, tandis que les matrices HJM et CJM auraient été deux biais, non nécessairement symétriques, pour expliquer des profils de mutations en accord avec des organismes vivant à différentes températures. Comme vu au chapitre 3, la décorrélation des matrices HJM et CJM de la dimension temporelle n'est pas totale, ce qui s'est traduit ici par le fait que la matrice PFASUM n'est pas utile à l'optimisation du modèle sur un nombre important de sites. En revanche, au niveau de la racine de l'arbre, la matrice PFASUM occupe 96% de l'optimisation des fréquences. Ce noeud toutefois est particulier, puisque seules les fréquences à l'équilibre y sont optimisées et non pas la matrice de mutation comptant l'intégralité des 400 coefficients de mutation. Comme on le voit en figure 2, l'information portée par la matrice PFASUM et CJM est très similaire. Dans une future version de ce modèle, la dimension PFASUM pourrait certainement être supprimée, réduisant ainsi le nombre de paramètres du modèle pour une vraisemblance équivalente.

La part de la matrice chaude HJM sur les différentes branches varie entre 0 et 61%. Cela se traduit notamment par des compositions en acides aminés très variables le long des branches de la phylogénie et aux différents noeuds. Pour interpréter ce signal, la figure 2 présente le biais CvP (acides aminés chargés contre polaires) déduit de ces compositions en acides aminés. Par exemple, l'ancêtre des Alvinellidae se situe à l'extrémité d'une branche 42/58 (voir figure 1), soit une branche dont la matrice de mutation est composée à 42% de la matrice HJM, 58% de la matrice CJM, et 0% de la matrice PFASUM. Reporté sur le diagramme de la figure 2, cette composition implique des fréquences à l'équilibre qui sont marquées par le point « Alvi+Alvinella », soit un biais CvP entre -6 et -8%. On constate également sur la figure 2 que le biais maximum possible entre une branche portant 100% du signal sur la matrice CJM ou 100% du signal sur la matrice HJM couvre les biais de composition les plus extrêmes observés chez les organismes psychrophiles et hyperthermophiles.
L'indicateur CvP est admis comme étant un bon prédicteur de la thermotolérance des organismes. Nous observons ainsi que la racine de la phylogénie est assimilée à un biais fortement négatif d'environ -13%, ce qui équivaut à des compositions en acides aminés d'organismes psychrophiles. Ce biais est similaire pour l'ancêtre commun des Alvinellidae et Ampharetidae « Alvi+Ampha ». En revanche, il augmente notablement le long de la branche menant à l'ancêtre des Alvinellidae, jusqu'à -7%, ce qui correspond à des organismes faiblement thermophiles. Selon l'optimisation de la figure 1, le biais CvP diminue dans les ancêtres des *Paralvinella* pour s'établir sur des biais d'organismes mésophiles à -11%. Selon cette prédiction, les espèces *Paralvinella pandorae irlandei* et *Paralvinella unidentata* seraient plus thermotolérantes que les espèces *Paralvinella grasslei* et *Paralvinella palmiformis*, les espèces les plus psychrophiles de la famille.
Au contraire, certaines branches notamment chez les *Miralvinella* (*P. mira* et *P. hessleri*) et les espèces *Paralvinella* sp. nov. et *P. fijiiensis*, montrent des biais chauds très importants, ce qui témoigne d'une bascule rapide dans les compositions attendues des protéines.

A noter également que nous avons simplifié l'information obtenue dans le diagramme en choisissant de représenter les biais de composition à l'équilibre issus des matrices de mutation optimisées pour chaque branche. Le changement de composition des protéines dépend toutefois non seulement du biais thermique de la matrice de mutation, mais aussi du temps de divergence entre les nœuds. Par exemple, les branches menant aux espèces *A. pompejana* et *A. caudata* n'utilise pas du tout la matrice HJM dans leur optimisation.

Toutefois, comme ces branches sont également très courtes (0.4% de divergence), la composition des protéines chez ces espèces est équivalente à celle de leur ancêtre commun. Une représentation plus juste de la variation du biais CvP sur la phylogénie intégrerait non seulement la force du biais présentée en figure 2, mais aussi la durée pendant laquelle cette force opère sur les séquences, équivalent au temps de divergence.

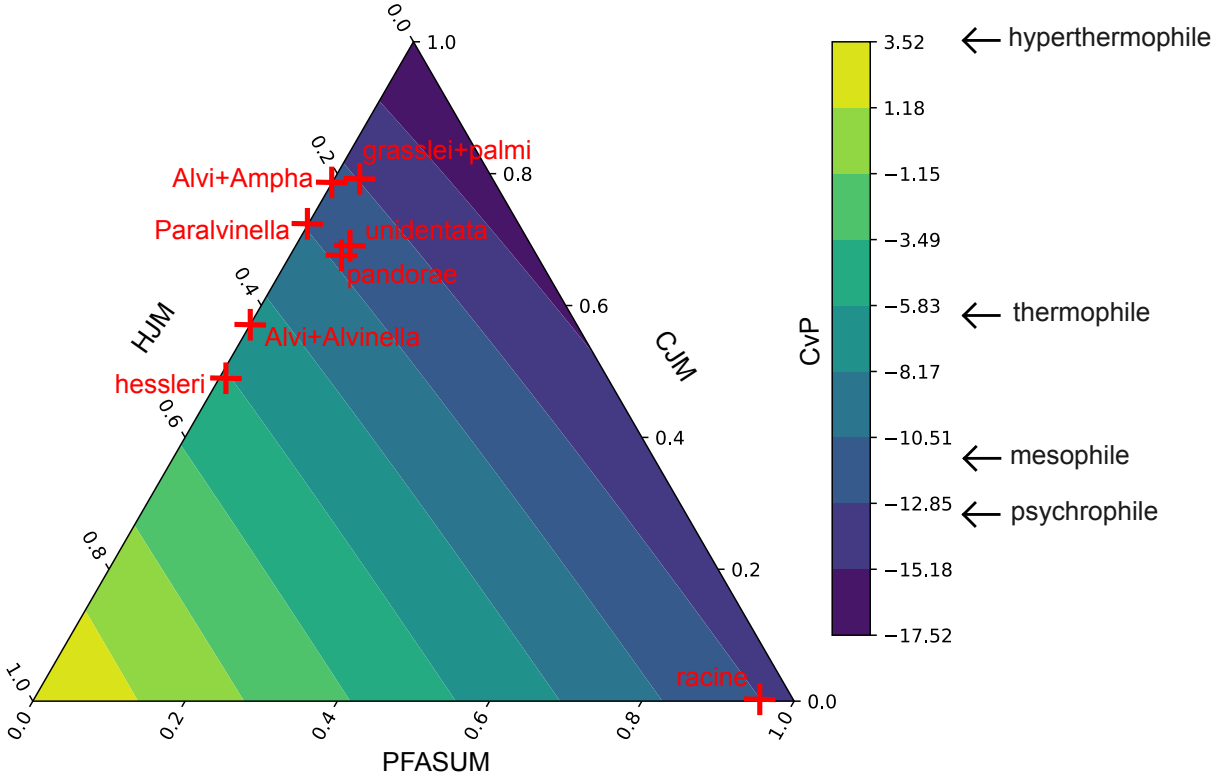
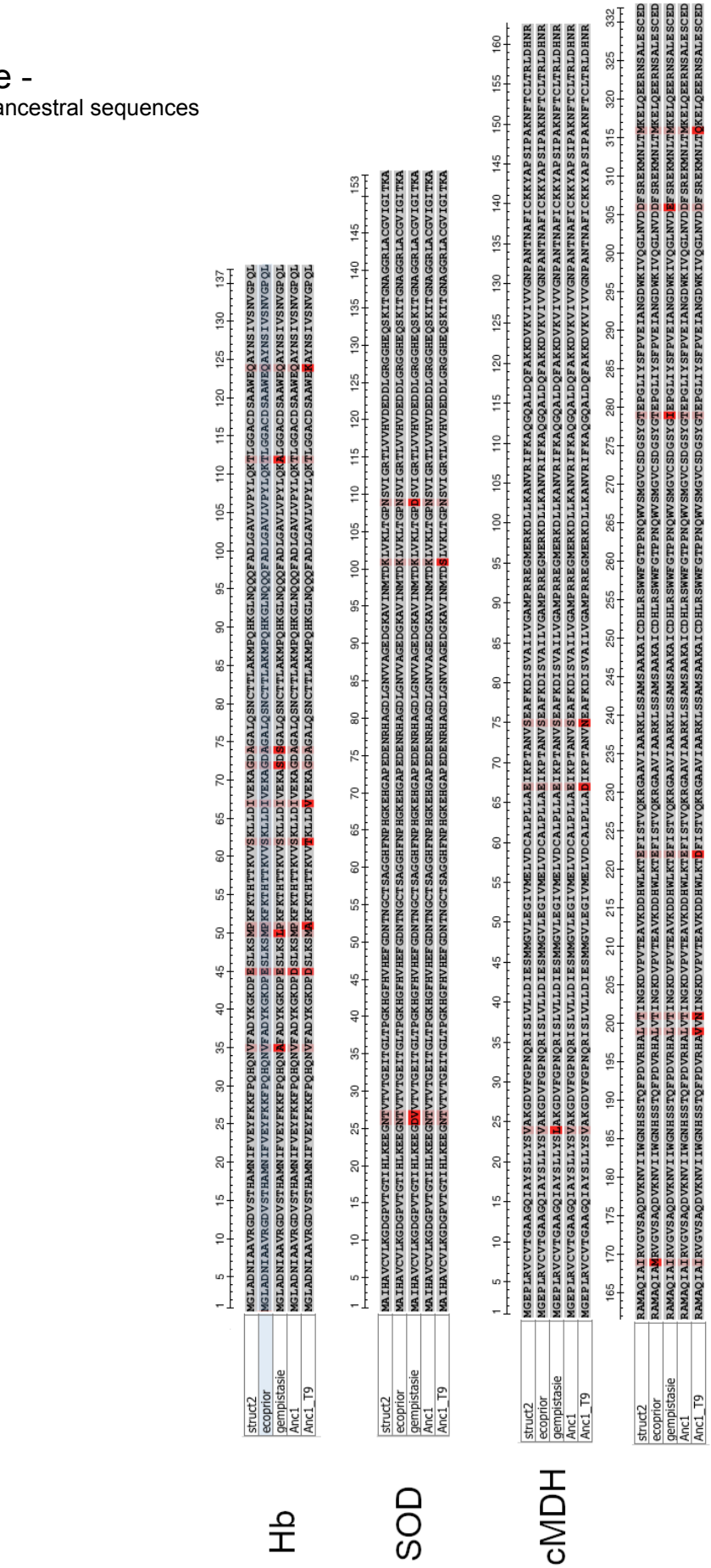


Figure 2. Projection des fréquences à l'équilibre attendues en différents acides aminés selon la part des matrices HJM, CJM et PFASUM dans la matrice de mutations. Les fréquences attendues sont récapitulées selon le biais CvP (acides aminés DEKR contre GHNPQST). « Alvi+Ampha » : biais CvP de l'ancêtre des Alvinellidae et Ampharetidae. « Alvi + Alvinella » : biais CvP de l'ancêtre des Alvinellidae et de l'ancêtre des *Alvinella*. « Paralvinella » : biais CvP de l'ancêtre des *Paralvinella*, excluant les branches *P. unidentata* et *P. pandorae irlandei*. Certaines valeurs de CvP type pour des organismes psychrophiles, mésophiles, thermophiles et hyperthermophiles sont tirées de Metpally et Reddy (2009) ainsi que Szilágyi et Závodszky (2000).

Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey, A. Szilágyi et P. Závodszky, Structure, 8-5, 2000
Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins, R. Metpally et B. Reddy, BMC Genomics, 10-1, 2009

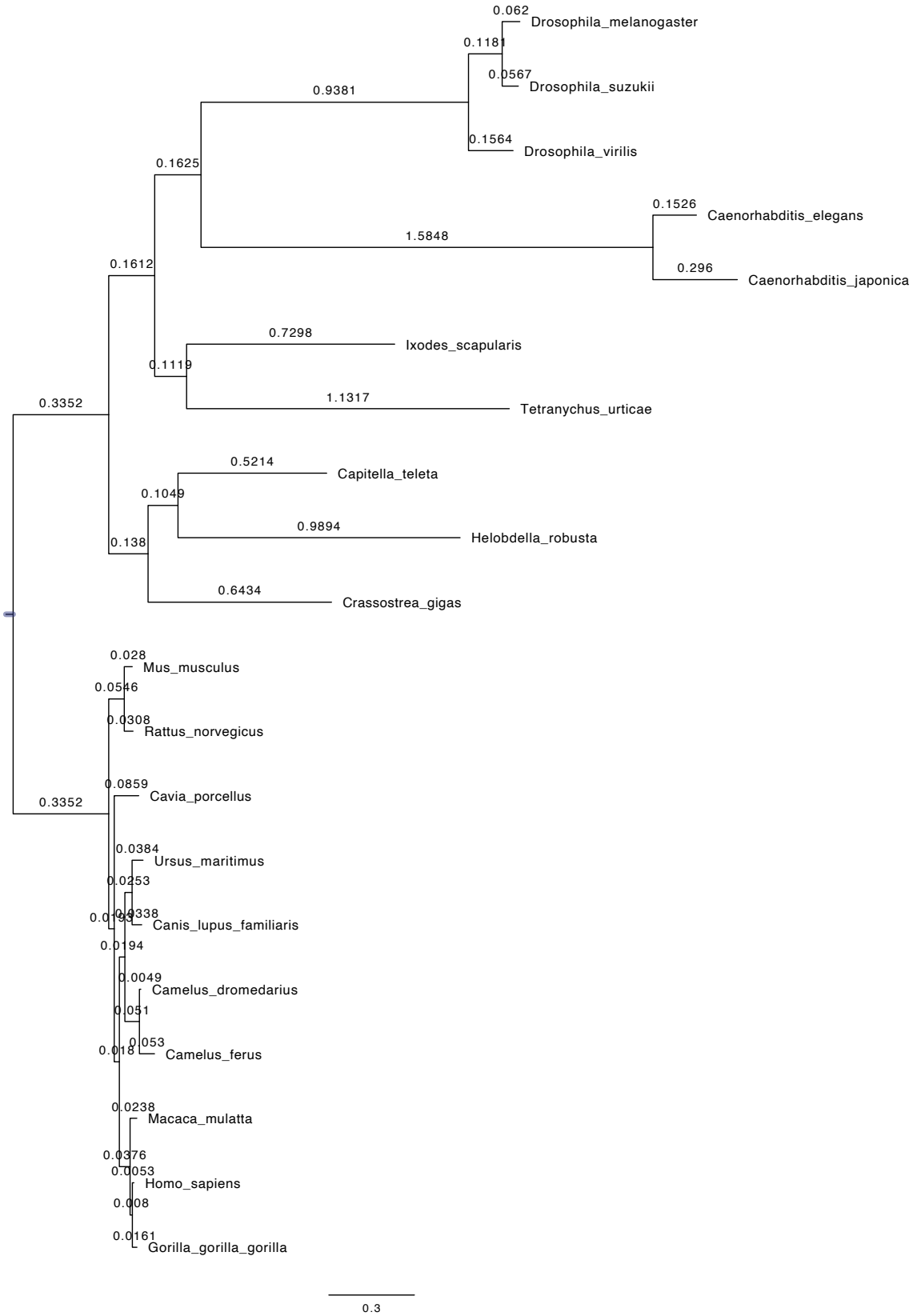
Annexe -

Alvinellid ancestral sequences

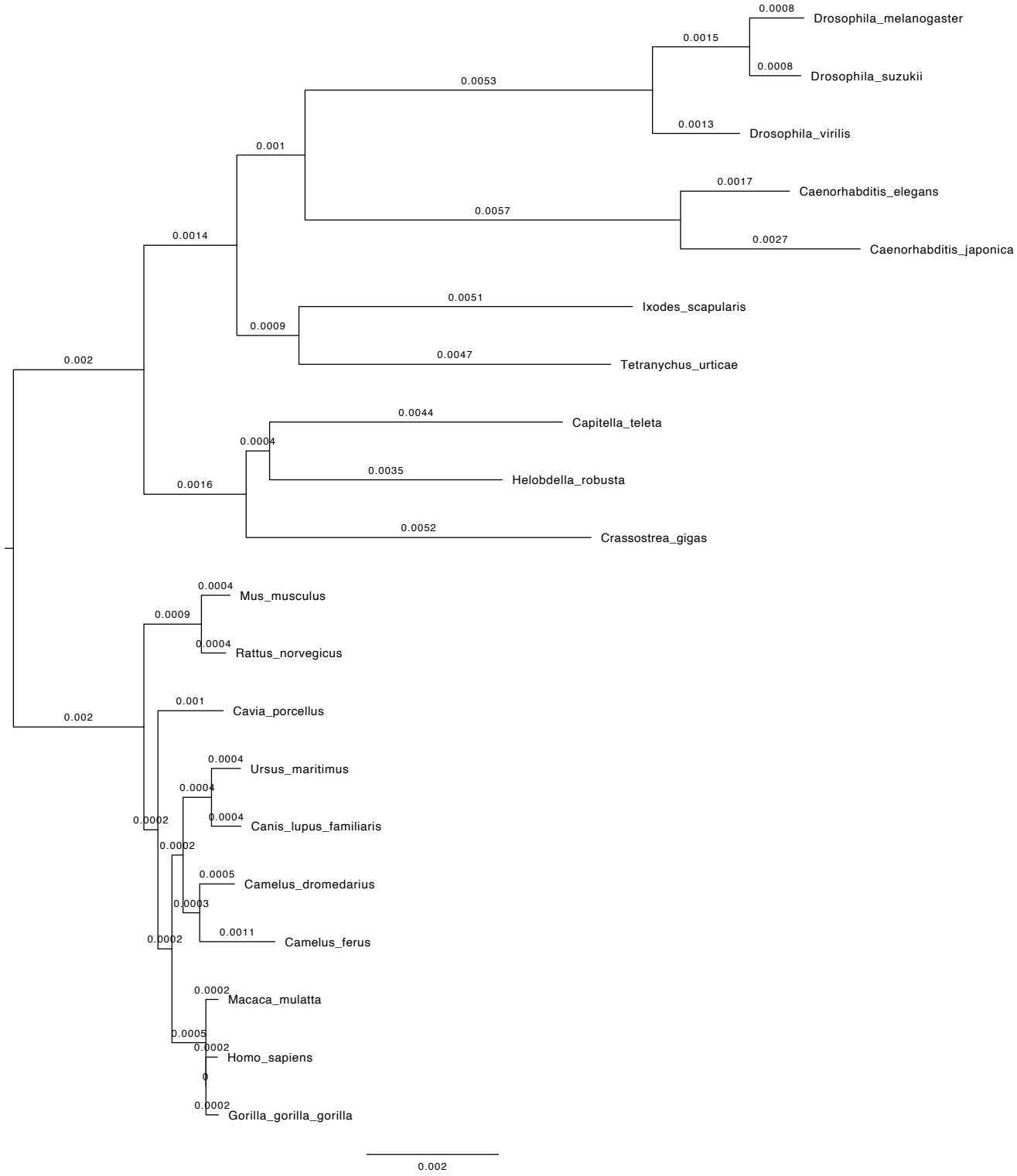


Reconstruction par maximum de vraisemblance des protéines de l'ancêtre des Alvinellidae selon les différents modèles ASR. Les ancêtres Anc1 et Anc1_T9 correspondent aux ancêtre exprimés et caractérisés au chapitre 2, sous l'hypothèse des phylogénies H1 et H2. Les reconstructions sous les différents modèles ASR sont effectuées sous l'hypothèse H1.

Annexes - reconstruction indels ancestraux

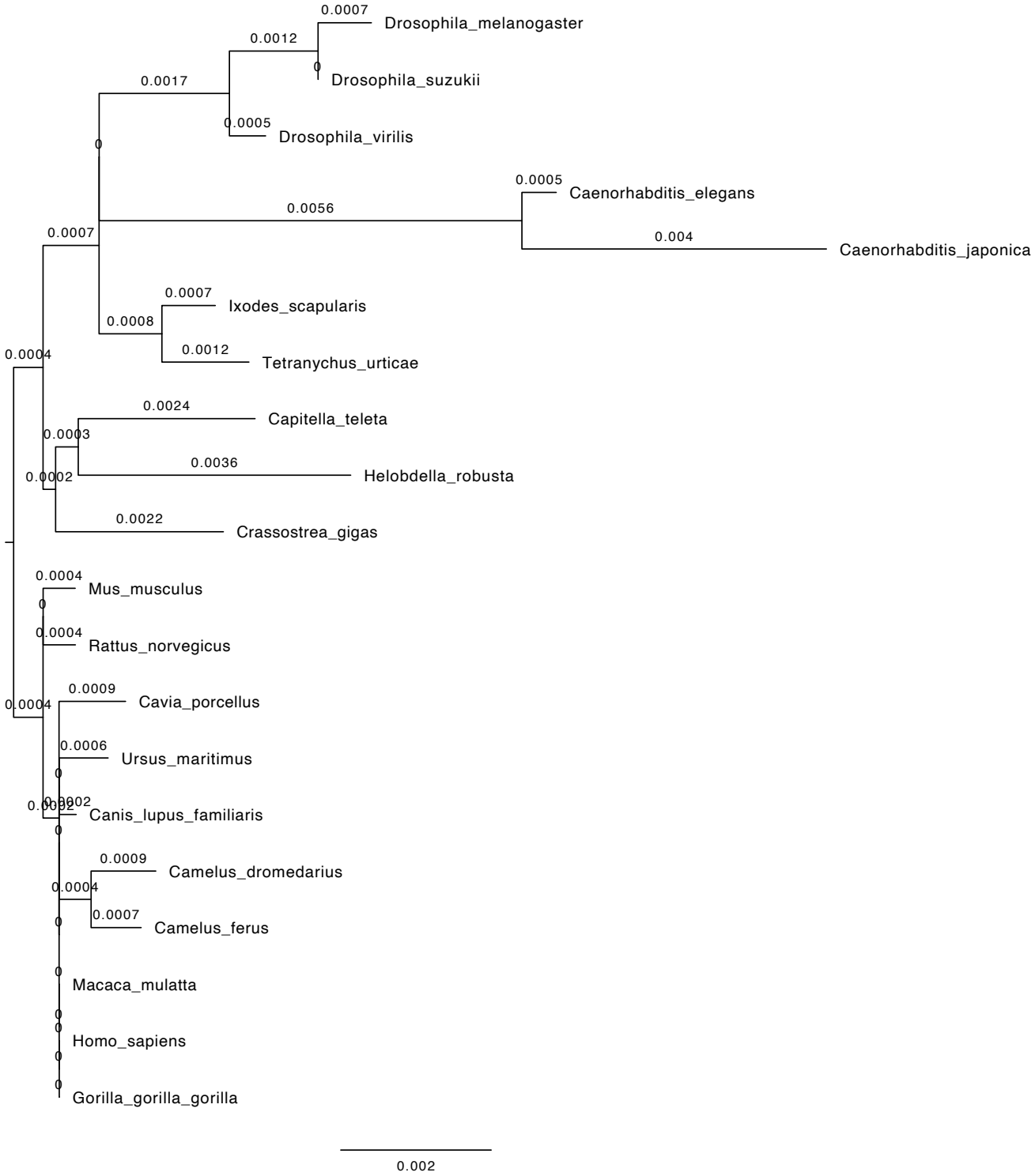


A. Optimisation des longueurs de branches de la phylogénie métazoaire selon la matrice de mutation des acides aminés LG (+G)



B. Optimisation des longueurs de branches de la phylogénie métazoaire selon les indels alignés par MAFFT (no filter)

C. Optimisation des longueurs de branches de la phylogénie métazoaire selon les indels alignés par MAFFT (soft filter)



D. Optimisation des longueurs de branches de la phylogénie métazoaire selon les indels alignés par MAFFT (hard filter)

Figure supplémentaire 1. Comparaison de l'optimisation de la phylogénie des métazoaires selon le type d'information de séquences (acides aminés ou indels) alignés avec MAFFT, sous différents niveau de filtration des indels.

A.

$$\begin{pmatrix} -1 & 0.223 \\ 0.285 & -1.845 \end{pmatrix}$$

B.

$$\begin{pmatrix} -1 & 0.365 \\ 0.173 & -1.741 \end{pmatrix}$$

C.

$$\begin{pmatrix} -1 & 0.218 \\ 0.348 & -1.865 \end{pmatrix}$$

Figure supplémentaire 2. Matrices d'indels pré-optimisées correspondant à différentes conditions. A. Soft filter PRANK-ProbCons (s_1 = 1.66, s_2 = 1.73, r = 1.56). B. Soft filter MAFFT (s_1 = 1.62, s_2 = 1.78, r = 0.63). C. Hard filter PRANK-ProbCons-MAFFT (s_1 = 1.8, s_2 = 1.9, r = 1.9).

Chapitre V

Discussion générale

V.1 Un ancêtre des Alvinellidae déjà adapté aux températures chaudes

En introduction de cette thèse, nous avions posé les questions suivantes : L'ancêtre des Alvinellidae était-il déjà une espèce thermotolérante des sources hydrothermales de l'Océan Pacifique ?

- Quelle est l'histoire évolutive des Alvinellidae ? Quel scénario a vraisemblablement permis la colonisation ancestrale des sources hydrothermales du Pacifique par les Alvinellidae ? Cette histoire souscrit-elle à l'origine abyssale de la faune moderne hydrothermale ?
- Les espèces contemporaines montrent-elles réellement des signes forts d'adaptation à la température ? La stabilité des protéines est-elle une bonne approche pour évaluer cela ?
- La thermotolérance ou la perte de thermotolérance des espèces contemporaines est-elle apparue plusieurs fois au cours de l'évolution de la lignée ? Et si oui, les mécanismes moléculaires mis en œuvre sont-ils similaires ?
- Si l'ancêtre des Alvinellidae était déjà une espèce thermophile, peut-on affirmer que l'espèce colonisait déjà le pôle chaud de l'habitat hydrothermal (murs des cheminées), ou y a-t-il d'autres hypothèses plausibles pour expliquer ce résultat ?

Le chapitre 1 a permis de montrer que **l'ancêtre de la lignée était probablement une espèce endémique de la Ride Est Pacifique (EPR) de l'océan Pacifique, datant de la fin du Crétacé entre 70 et 90 millions d'années** (chapitre 1 - Figure 5). La radiation des Alvinellidae sur cette période a été rapide, entraînant l'apparition de plusieurs lignées menant aux actuelles espèces *Alvinella*, *Paralvinella* et *Nautalvinella*. La divergence entre les Alvinellidae et les Ampharetidae est quant à elle estimée à 130 millions d'années. Il a été proposé qu'entre l'Aptian et le Cenomanian, des masses émergées du Pacifique aient pu servir de refuge aux espèces hydrothermales, durant de longues périodes d'anoxie de l'océan profond (Jacobs et Lindberg, 1998). Sous cette hypothèse, la période estimée pour la recolonisation du milieu profond par l'ancêtre des Alvinellidae est compatible avec

celles de plusieurs autres lignées hydrothermales datées à la fin du Crétacé (Vrijenhoek, 2013). L'EPR est proposée pour avoir joué un rôle important dans la dispersion des espèces hydrothermales (Bachraty, Legendre et Desbruyères, 2009), notamment *via* deux rides désormais éteintes qui reliaient le Pacifique Est et Ouest, la ride de Kula et une deuxième passant par la Guinée du nord, actives jusqu'à 50 millions d'années et illustrées en figure V.1 (Hessler et P. F. Lonsdale, 1991). L'histoire que l'on tirerait de ce premier chapitre s'accorde bien avec ces événements géologiques. Dans ce cas, **la colonisation du Pacifique Ouest se serait opérée il y a environ 50 millions d'années depuis l'EPR, puis la famille se serait vraisemblablement étendue dans l'océan Indien par la ride Antarctique après l'ouverture de la mer de Chine**, dans la continuité de la ride centrale Indienne (voir figure V.1).

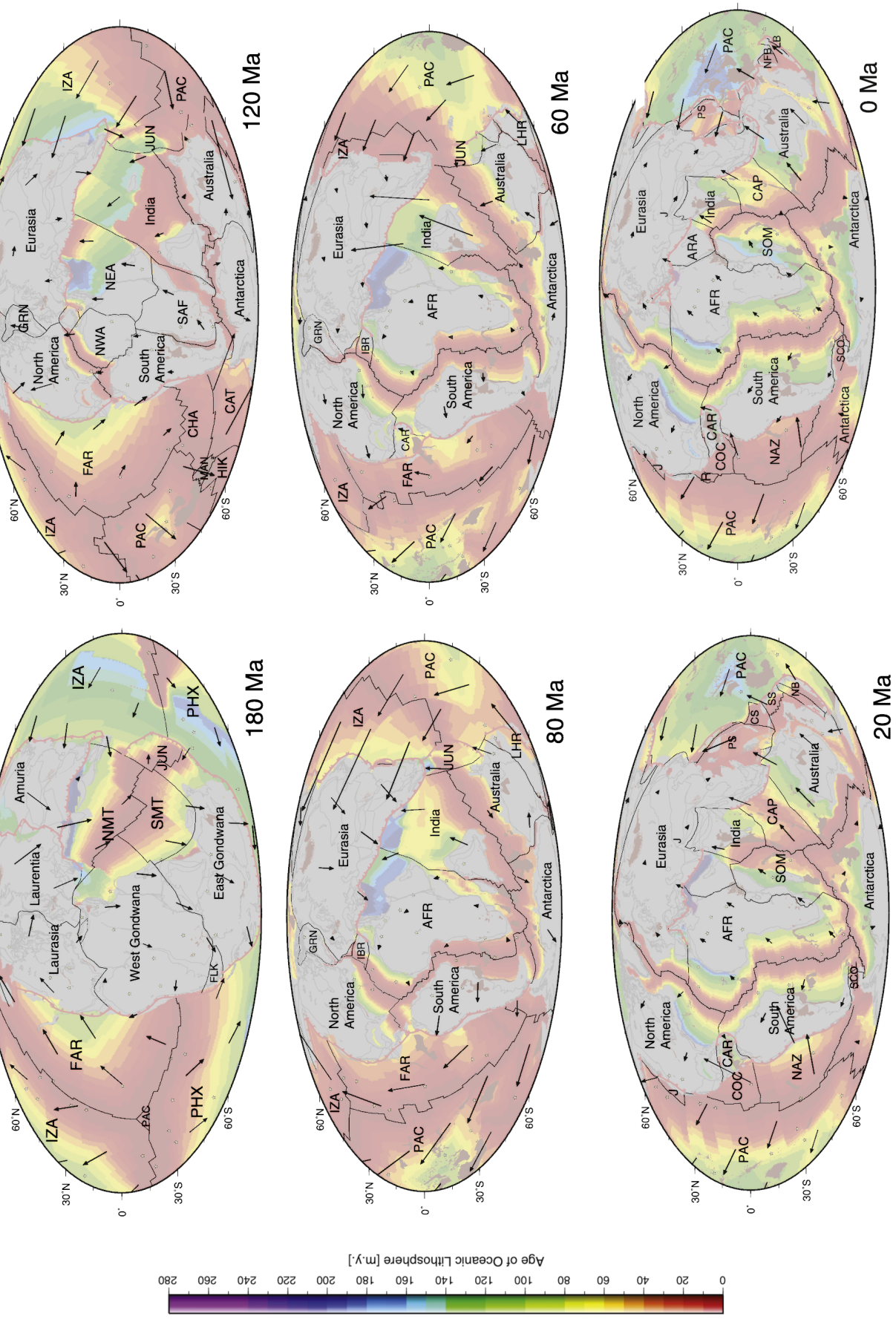


Figure V.1 – Mouvements tectoniques globaux depuis le Crétacé. L'ouverture de la plaque Pacifique est datée de 180 millions d'années. La subduction de la plaque Izanagi, qui isole les parties Est et Ouest de l'océan Pacifique, est datée entre 40 et 50 millions d'années. La subduction de la plaque Farallon, qui isole les populations de Juan de Fuca et de l'EPR, est datée à 30 millions d'années. La fermeture du bassin des Philippines est daté entre 60 et 40 millions d'années. Adapté de M. Seton et al., 2012.

Les résultats de ce chapitre nous apprennent également que la radiation des Alvinellidae s'est probablement opérée en quelques millions d'années, ce qui a entraîné du tri incomplet de lignées (ILS) entre les premières lignées observable sur l'ensemble du transcriptome. Outre l'ILS, le déséquilibre dans les phylogénies des gènes (chapitre 1 - Figure 3) indique de forts taux d'introgression interspécifique entre les lignées menant aux actuelles espèces *Paralvinella unidentata*, *Paralvinella pandorae* et *Alvinella*, si importants que le signal statistique ne permet pas de distinguer entre la topologie vraie de l'arbre des espèces et la topologie alternative correspondant à l'effet de l'introgression sur les lignées ancestrales. Les gènes de l'espèce *Paralvinella unidentata* sont tantôt regroupés en espèce soeur avec *Paralvinella pandorae*, tantôt en espèce soeur des espèces *Alvinella*. Il est alors remarquable que la description initiale de cette espèce par Desbruyères et Laubier ait déjà noté le caractère ambivalent de *P. unidentata*, avec d'un côté l'absence de tentacules buccaux, la forme des branchies en peigne et l'absence de lobes notopodiaux digitiformes qui la rapprochent de *P. pandorae*, tandis que la forme des filaments secondaires des branchies en lamelles est un trait exclusivement partagé avec les *Alvinella* (Desbruyères et Laubier, 1993). On peut alors proposer le scénario d'une population ancestrale d'Alvinellidae s'étant sub-divisée en clades écologiques partageant certains traits par de l'hybridation adaptative pendant 3 à 10 millions d'années et potentiellement dispersées le long d'une ride médio-Pacifique regroupant les actuels Juan de Fuca et EPR, entre lesquelles des hybridations ponctuelles auraient amené à des transferts de gènes important entre les lignées. L'idée d'un isolement lié en premier lieu à une spéciation écologique précédant une isolation des lignées ancestrales des Alvinellidae a notamment été proposée par Jollivet, Desbruyères et al., 1995. Le scénario de plusieurs populations s'hybridant pendant quelques millions d'années avant leur isolement dans le contexte hydrothermal est-il possible ?

Actuellement, plusieurs espèces d'Alvinellidae montrent déjà des signes de variations morphologiques intra-spécifiques importantes et de larges zones d'expansion géographique (à l'instar de *P. hessleri* et *P. unidentata* sur l'ensemble du Pacifique Ouest), ce qui implique des processus complexes d'isolement partiel des populations (Didier Jollivet, communication personnelle). Le reséquençage de génomes (WGS) d'*Alvinella pompejana* à la suite de l'assemblage récent de son génome au niveau chromosomique montre que les populations Nord EPR et Sud EPR préalablement isolées l'une de l'autre autour du 7ème parallèle sud ont des individus fortement divergents sur plusieurs chromosomes mais avec pour certains de larges blocs de gènes introgressés pouvant être le signe de remaniements chromosomiques. A l'inverse d'autres chromosomes sont faiblement différenciés, comme illustré en figure V.2. Cette forte hétérogénéité du flux génique est attendue dans les phases intermédiaires de la spéciation lorsque la barrière génétique est encore faible (Wu, 2001). Jang et al., 2016 ont estimé à partir de 11 marqueurs nucléaires et mitochondriaux que la séparation entre les populations d'*Alvinella pompejana* de l'EPR et de la ride Pacifique-Antarctique datait de 4,2 millions d'années, tandis que la séparation entre les populations Nord et Sud de l'EPR remonterait à environ 1 million d'années (figure V.3). Sur la base de la divergence du COI, Thomas-Bulle et al., 2022 estiment la séparation entre les populations Nord et Sud EPR à 3 millions d'années, avec 3% de gènes très divergents, attribués à de la sélection positive ou à de l'introgression d'allèles issus de la population Pacifique-Antarctique (thèse A. Bioy, données non publiées).

L'exemple d'*Alvinella pompejana*, sans aucun doute une espèce très proche du portrait

que l'on se fait de l'ancêtre des Alvinellidae (espèce hydrothermale profonde de l'Est Pacifique associée aux cheminées hydrothermales) montre que l'image d'une population large, répartie le long de l'EPR, au sein de laquelle une divergence importante a débuté il y a 4,5 Ma entre le Nord et le Sud de l'EPR avec cependant le maintien d'un génome en mosaïque, est tout à fait envisageable pour expliquer la radiation initiale des Alvinellidae et le maintien d'héritages génétiques entre les espèces en formation. Pour confirmer cette hypothèse, il sera nécessaire toutefois d'observer si différentes fenêtres génomiques présentent des histoires de gènes différentes, et si le signal observé au niveau du transcriptome n'est pas dû à un biais non identifié. Pour l'instant, nous disposons au total de 888 gènes partagés par toutes les lignées d'intérêt dans les jeux complets de données transcriptomiques. Nous pourrions essayer de positionner ces gènes sur le génome assemblé d'*A. pompejana*, mais la densité de ces gènes, une fois pris en compte l'incertitude topologique propre à chaque gène, ne permet pas de bien distinguer l'architecture des proximités phylogénétiques des différentes lignées initiales le long du génome. Pour cela, nous aurions besoin d'au moins un génome pour chacune des lignées d'intérêt, à savoir *Alvinella*, *P. pandorae*, *P. unidentata* et une autre espèce *Paralvinella*. La bonne nouvelle est que des données de séquençage du génome existent pour *Paralvinella palmiformis* (Perez et al., 2023) ainsi qu'un premier séquençage combinant Illumina and ion Torrent pour *P. pandorae irlandei* effectué au laboratoire, bien que la couverture soit relativement faible et que nous n'avons pas complètement exploré ces données pour l'instant. Il manque toutefois un génome pour *P. unidentata*. En outre, les études portant sur les fenêtres génomiques s'intéressent généralement à des espèces ayant divergé récemment et pour lesquelles la structure du génome est conservée. En effet, des remaniements chromosomiques vont faire perdre le signal de fenêtres portant des histoires évolutives cohérentes. En revanche, dans le cas du génome des Alvinellidae, des études de synténie incluant des espèces de siboglinidae, annélides hydrothermaux assez éloignés, montrent que la structure des chromosomes est bien conservée pour au moins la moitié du génome (El Hilali et al., 2024). Par conséquent, cette étude devrait être réalisable, sous couvert d'obtenir les données génomiques manquantes.

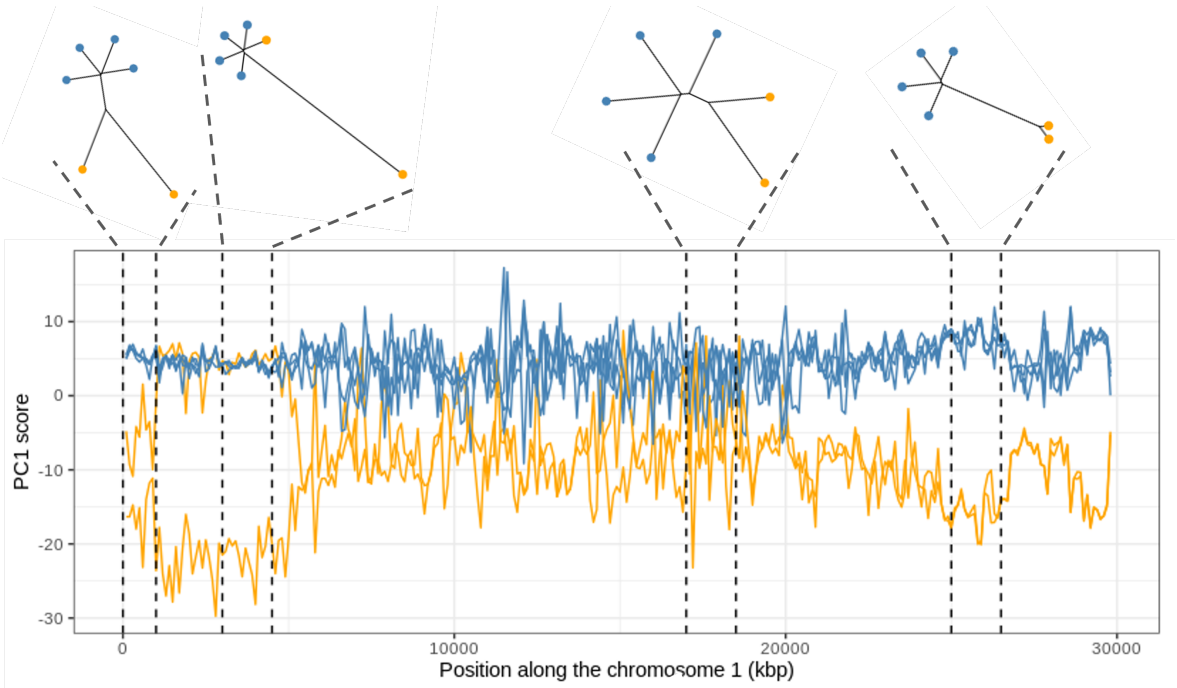


FIGURE V.2 – Régions génomiques sur le chromosome 1 d'*Alvinella pompejana* entre les populations Nord et Sud 7°S/EPR. Les régions sont obtenues à partir de fragments de 1 à 1.5 Mpb. Les points bleus correspondent à des individus Nord EPR, et les points oranges Sud EPR. Les phylogénies montrent de gauche à droite une région d'introgression où un individu Sud EPR est situé à mi-distance des individus Sud et Nord EPR, une région d'introgression où un individu Sud EPR est similaire aux individus Nord EPS, une région de faible divergence sans structure des populations notable et une région de forte divergence entre les individus Nord et Sud EPR (EL HILALI et al., 2024).

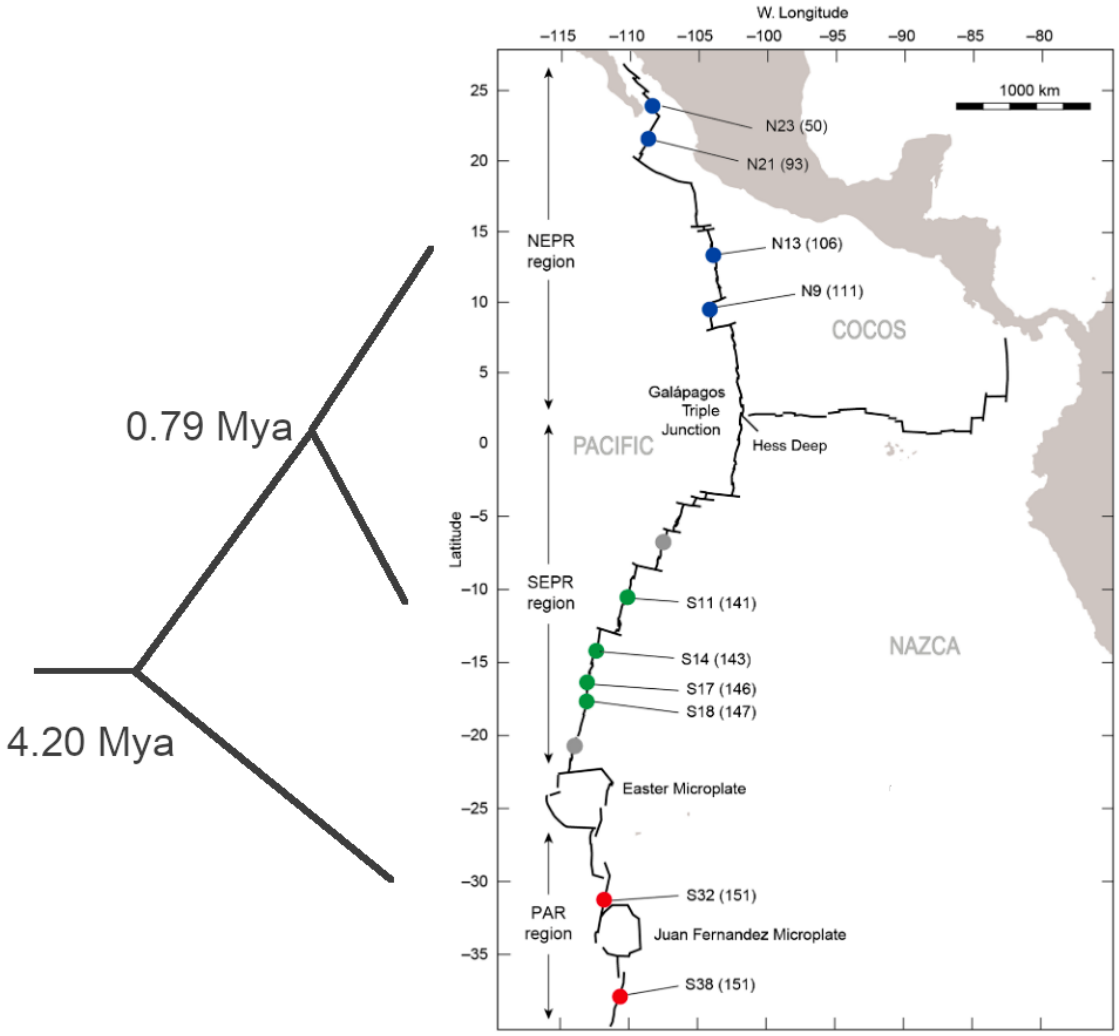


Figure V.3 – Isolation partielle des populations actuelles d'*Alvinella pompejana*. La séparation entre les différentes populations est datée en millions d'années. Adapté de Jang et al., 2016.

L'obtention d'une phylogénie des Alvinellidae nous a permis d'effectuer des reconstructions de protéines pour plusieurs ancêtres de la famille des Alvinellidae. Considérant que la stabilité des protéines est supposée corréler avec la température physiologique de l'organisme, nous voulions caractériser la stabilité des protéines de l'ancêtre de la lignée afin de déterminer si celui-ci était déjà adapté aux températures chaudes du Crétacé.

Outre ce test d'hypothèse, l'étude de la stabilité de certaines protéines (Superoxyde dismutase Cu/Zn et Malate déshydrogénase cytosolique) nous renseigne également sur l'écologie probable des espèces contemporaines. En effet, les espèces d'Alvinellidae présentent des préférences thermiques contrastées mais certaines espèces, notamment froides, sont relativement eurythermales (Cottin et al., 2008) et peuvent survivre sur des gammes de température de plusieurs dizaines de degrés. **Le résultat de l'expression des protéines recombinantes montre effectivement que les protéines des espèces considérées chaudes** (*A. pompejana*, *Paralvinella mira*, *Paralvinella sulfincola*, *Paralvinella fijiensis*) **sont plus stables que celles issues des espèces froides** (*Paralvinella grasslei*, *Paralvinella palmiformis*, *Paralvinella unidentata* et *Paralvinella pandorae*), ce qui confirme le résultat d'autres études menées sur des protéines d'Alvinellidae qui notaient une différence de stabilité entre espèces chaudes et froides (Jollivet, Desbruyères et al., 1995 ; Rinke et Lee, 2009). Comme illustré dans la figure 2 du chapitre 2, la différence de stabilité entre les protéines les plus stables des espèces froides et les moins stables des espèces chaudes peut toutefois être ténue. La variation dans la composition des protéomes détectée avec le modèle Ecoprior et présentée en annexe du chapitre 2 suggère que les espèces *P. grasslei* et *P. palmiformis* sont les espèces les plus froides de la famille, ce qui est en bon accord avec les mesures expérimentales obtenues sur les protéines recombinantes. Les espèces montrent par conséquent de fortes différences d'adaptation à la température, mais l'étude de plusieurs marqueurs simultanément est nécessaire pour discerner les espèces à la limite entre différents régimes thermiques. En outre, notre étude se focalise sur la stabilité de protéines exprimées *in vitro*, mais nous ne savons pas si la stabilité intracellulaire des protéines peut être également modulée, par exemple par des protéines chaperonnes ou bien des osmolytes présents dans le cytosol (Hourdez et Weber, 2005 ; Dilly et al., 2012).

Les protéines reconstruites les plus probables des ancêtres des Alvinellidae montrent une thermostabilité élevée, comparable à celle des espèces thermophiles contemporaines, comme montré en figure 3 du chapitre 2. La reconstruction de ces séquences repose sur plusieurs hypothèses, notamment l'histoire évolutive des gènes reconstruits (qui peut varier selon les régions du génome) ainsi que le modèle probabiliste phylogénétique en lui-même utilisé pour la reconstruction. Notre étude a montré que les résultats obtenus étaient fiables malgré ces différentes incertitudes. L'hypothèse sur la phylogénie des gènes pourrait certainement être tranchée grâce à la résolution de la phylogénie sur des fenêtres génomiques comme suggéré précédemment. En effet, en supposant que les régions proches du génome partagent une même histoire, l'allongement des séquences alignées par cette méthode donnerait plus de confiance sur l'une ou l'autre des topologies possibles. Concernant la fiabilité des modèles phylogénétiques, il a été suggéré que le choix de la séquence par maximum de vraisemblance, comme nous l'avons fait dans le chapitre 2, peut biaiser les résultats vers la sélection d'une protéine plus thermostable qu'elle ne l'était réellement (Williams et al., 2006). Ceci a amené plusieurs auteurs à choisir d'exprimer et de caractériser la meilleure séquence obtenue par maximum de vraisemblance mais de regarder aussi plusieurs sé-

quences alternatives de moindre probabilité, afin d'évaluer l'incertitude sur le phénotype mesuré (Gaucher, Govindarajan et Ganesh, 2008 ; Eick et al., 2016). Nous avons choisi de profiter du nombre important de protéines exprimées et mesurées, ancestrales comme contemporaines, pour calibrer un modèle permettant d'évaluer la stabilité thermique relative des séquences alternatives par rapport à la reconstruction ayant obtenu le maximum de vraisemblance. Cette approche s'est montrée efficace pour caractériser l'ensemble de la distribution des variants possibles et confirmer le résultat principal de cette thèse, à savoir que **l'ancêtre de la lignée des Alvinellidae était une espèce thermotolérante de l'océan Pacifique**. Ce résultat est bien confirmé par l'expression de la MDHc et de la Cu/Zn SOD. Bien que les mesures pour l'hémoglobine intracellulaire aient été moins fructueuses, nous avons quand même pu obtenir des stabilités pour les hémoglobines de l'ancêtre de la lignée qui ont montré une stabilité similaire à celle de l'espèce chaude *P. mira*. Les mesures relatives à cette protéine méritent d'être achevées, mais devraient confirmer les résultats déjà obtenus.

Dans ce cas, **l'adaptation aux environnements froids serait apparue de manière indépendante à plusieurs reprises** dans les lignées *P. unidentata*, *P. pandorae* et *P. palmiformis+P. grasslei*. En outre, les stabilités mesurées pour les protéines ancestrales des *Paralvinella* (excluant *P. unidentata* et *P. pandorae*) indiquent un ancêtre toujours chaud, mais l'analyse des compositions en résidus du protéome (annexe du chapitre 3) soutiennent plutôt un relâchement de la contrainte sélective et des ancêtres plus froids, avec des changements brusques des compositions d'acides aminés dans les branches menant aux *Miralvinella* et aux espèces *Paralvinella* sp. nov et *P. fijiensis*. Sous cette dernière hypothèse, l'acquisition de la thermotolérance aurait été partiellement perdue, puis acquise à nouveau plus récemment dans la lignée "chaude" au sein des *Paralvinella*. Pour cet ancêtre, nous n'avons pas évalué la confiance dans la stabilité mesurée de la protéine, dont seule la séquence retenue par maximum de vraisemblance a été caractérisée. Une histoire complexe de changements de régimes thermiques dans la lignée impliquant de multiples convergences écologiques serait particulièrement intéressante, et ces résultats devraient être éclaircis. En outre, d'un point de vue structural, nous n'avons pas eu l'opportunité d'étudier comment les mutations des protéines de la famille agissent sur la stabilité des protéines. Les différentes enthalpies associées à des protéines de stabilités similaires (tableau 2 du chapitre 2) suggèrent en effet que les mécanismes de stabilisation des variants ancestraux sont différents. Le lien entre le chemin adaptatif pris par les protéines montrant des convergences phénotypiques, potentiellement dépendant du contexte de la séquence primaire des protéines ancestrales, serait également une piste d'étude intéressante par l'étude de l'épistasie entre les mutations impliquées dans la stabilisation (Pollock, Thiltgen et Goldstein, 2012 ; Hart et al., 2014).

Peut-on alors conclure que l'ancêtre des Alvinellidae était déjà une espèce associée aux cheminées hydrothermales de l'océan Pacifique, à l'instar des espèces contemporaines *Alvinella pompejana* ou *Paralvinella sulfincola* ? Comme montré en figure V.4, les températures de l'océan Pacifique profond étaient plus élevées il y a 90 millions d'années, date à laquelle la radiation des Alvinellidae est estimée. Au temps actuel, ces températures sont d'environ 3-5°C, mais pouvaient vraisemblablement atteindre 10 à 15°C au Crétacé. Les températures des eaux de surface du Pacifique étaient quant-à elles un peu plus chaudes, entre 15 et 20°C (Savin, 1977). Les estimations que nous avons obtenues au chapitre 2 pour la température

de vie de l'ancêtre sont néanmoins supérieures à cet écart. Les espèces froides contemporaines des Alvinellidae ont déjà des températures de confort entre 10 et 20°C (Cottin et al., 2008). Par conséquent, l'environnement hydrothermal de ces animaux est situé dans des gammes de température au delà des températures anciennes de l'océan profond et la température inférée pour l'ancêtre ne s'explique pas uniquement par des conditions climatiques anciennes plus chaudes.

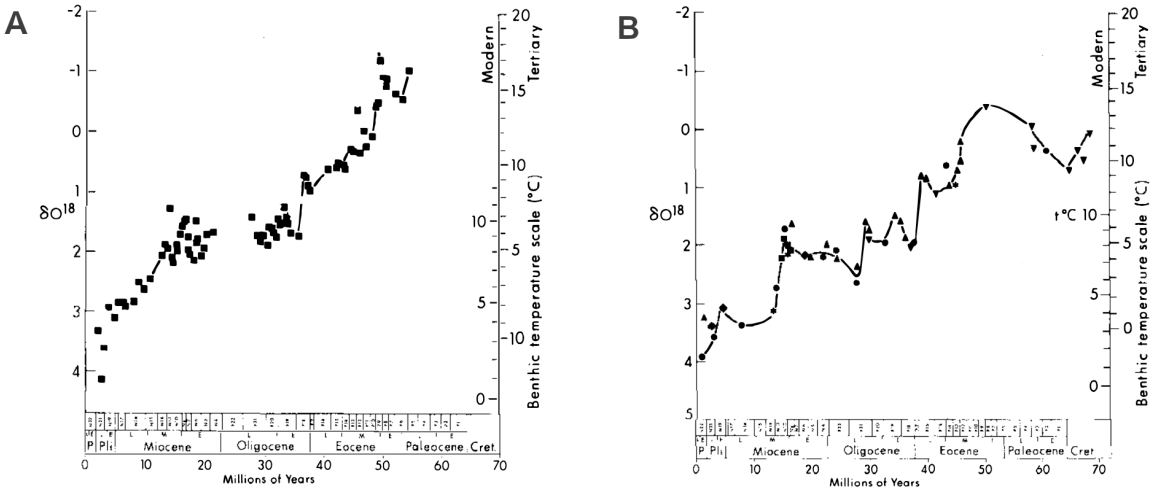


Figure V.4 – Températures benthiques de l'océan Pacifique depuis le Crétacé. Les températures déduites de l'anomalie en δO^{18} relevée chez les sédiments de foraminifères benthiques sont montrées A. pour le Pacifique Nord, B. pour le Pacifique Sud sub-Antarctique. L'échelle des températures est calculée selon deux méthodes, "modern" et "tertiary", en prenant respectivement comme hypothèses des valeurs pour δO^{18} de -0.08 et -1.00 pour mille selon deux hypothèses de fonte des calottes glaciaires. Adapté de Savin, 1977.

Toutefois, **il faudrait réunir plus de preuves pour conclure définitivement que l'ancêtre était déjà endémique des sources hydrothermales profondes**. C'est cette question qui nous avait initialement guidé dans le choix de la cMDH, SOD et hémoglobine intracellulaire, qui sont des protéines présentant un aspect fonctionnel que l'on aurait pu corréler avec d'autres paramètres environnementaux particuliers des sources hydrothermales comme l'hypoxie, les conditions oxydo-réductrices du milieu et potentiellement la pression (Dahlhoff et Somero, 1991). Cette étude fonctionnelle n'a pas pu être menée pour l'instant mais nous avons montré que l'expression et la caractérisation de ces protéines étaient possibles. L'étude fonctionnelle de la SOD peut être plus délicate si l'on veut l'effectuer sur une gamme de températures, étant donné que la plupart des protocoles utilise une enzyme rapportrice dont l'activité ne doit pas être inhibée avant celle de la SOD. Pour cela, il faut que l'activité de cette première enzyme soit thermostable. Paoletti et al., 1986 proposent cependant un protocole qui ne requiert pas d'autres enzymes, en suivant l'inhibition de l'oxydation du NADH. Ce protocole pourrait être envisagé. D'autres enzymes auraient pu être très intéressantes à étudier, notamment les métallothionéines qui sont impliquées dans la séquestration de métaux comme le cuivre et le zinc (Stéphane Hourdez et Jollivet, 2020). Ces protéines n'ont cependant pas été trouvées dans tous les transcriptomes à notre disposition, peut-être parce qu'elles ne sont pas exprimées fortement dans les tissus sur lesquels les extractions d'ARN ont été effectuées (généralement les branchies). Nous notons toutefois qu'un gène correspondant n'a pu être identifié chez *Alvinella pompejana*, alors même que la complétude du génome et de la prédiction des gènes ont montré de bons résultats. Nous savons pourtant que cette protéine existe chez *A. pompejana* (Stéphane Hourdez et Jollivet, 2020). Ainsi, la caractérisation de lignées ancestrales pour des protéines impliquant d'autres paramètres environnementaux que la température permettrait de confirmer l'image cohérente d'un ancêtre endémique des cheminées hydrothermales issu d'un environnement plus chaud que le milieu abyssal classique. Aussi, nous avons borné notre étude aux seuls Alvinellidae, mais le laboratoire dispose de nombreux transcriptomes issus de polynoidae plus ou moins éloignés phylogénétiquement. Ceci constitue une ressource précieuse pour déterminer par contraste quels traits adaptatifs peuvent être attribués à l'environnement hydrothermal, et lesquels ne lui sont sont pas spécifiques.

Dans ce cas, une difficulté importante qui se pose est la question de la fiabilité dans les reconstructions de protéines ancestrales. La stratégie que nous avons adopté dans le chapitre 2, qui allie simulations informatiques et confirmation/calibration par des mesures expérimentales, s'est montrée efficace pour déterminer la stabilité des protéines pour laquelle plusieurs logiciels comme FoldX (Schymkowitz et al., 2005) donnent des résultats satisfaisants. En revanche, cette approche n'est plus possible pour la caractérisation fonctionnelle des protéines. Si nous voulions par exemple mesurer les affinités pour l'oxygène des hémoglobines ancestrales, nous devrions effectuer des mesures expérimentales pour plusieurs variants de protéines et redéfinir notre stratégie. La question de la fiabilité de la séquence obtenue par maximum de vraisemblance, mais également de la fiabilité des séquences alternatives, se pose. Nous avons vu dans le chapitre 3 que les modèles actuels réussissaient à donner une bonne image des résidus ancestraux ambigus, mais qu'ils pouvaient également se tromper dans l'estimation des probabilités de ces résidus selon le modèle choisi (figure 12 du chapitre 3). Un simple raffinement de ces modèles, comme avec le modèle Struct2 qui intègre de façon plus précise les dynamiques évolutives propres aux différentes structures secondaires de la protéine, permet d'améliorer la probabilité de plusieurs résidus de la pro-

téine RFP qui nous sert de témoin, même si le résultat n'est toujours pas parfait. La question est d'autant plus importante qu'on pourrait croire que les résidus les plus ambigus sont nécessairement les plus variables dans l'alignement, et qu'ils interviennent davantage dans la stabilité globale de la protéine que sur ses caractéristiques fonctionnelles qui doivent être globalement plus conservées. Pourtant, Eick et al., 2016 ont montré que même les caractéristiques fonctionnelles pouvaient quantitativement varier beaucoup d'une reconstruction à l'autre. Ces effets peuvent être dus aux interactions entre les résidus qui, même quand ils se situent dans une région qui n'est pas associée à la fonction de la protéine, participent à la stabilité globale de l'édifice et peuvent être accommodés par d'autres mutations dans la séquence (Tokuriki et al., 2008 ; Pollock, Thiltgen et Goldstein, 2012). Une bonne pratique pour cette discipline serait de continuer à travailler sur l'amélioration de ces modèles et d'intégrer le plus d'informations possible concernant les protéines étudiées : mutations contexte-dépendantes, biais de mutation liés au *prior* de température inféré sur de longs alignements, épistasie entre les résidus de la protéine étudiée. Les modèles développés au chapitre 3, en effet, peuvent être unifiés dans le même modèle, et tentent de résoudre de façon pragmatique des questions différentes mais complémentaires de la dynamique évolutive des protéines.

L'approche Gempistasie, basée sur le logiciel GEMME (Laine, Karami et Carbone, 2019), est à mon sens la plus prometteuse. L'objectif d'obtenir des profils de mutations site-dépendant pour une protéine donnée serait un gain non négligeable dans la qualité du modèle évolutif. Bien que le logiciel GEMME ne modélise pas explicitement les interactions épistasiques entre les résidus ni leur rôle dans la protéine, les profils obtenus par son intermédiaire, basés sur des homologies de séquences puisées depuis le NCBI, ont apporté des gains très importants de vraisemblance (tableau 2, chapitre 2). Notamment, la figure 6 du chapitre 2 donne une forme d'exponentielle décroissante pour relier le score obtenu par GEMME pour un résidu à une position avec sa probabilité d'apparaître dans la séquence. Ce résultat est très intriguant parce qu'il suggère une distribution similaire à une statistique de Maxwell-Boltzmann de la forme $exp(-\beta E)/Z(T)$. Cette distribution est suggérée par Kleinman et al., 2010 pour relier un certain potentiel statistique pour un acide aminé (défini par les auteurs comme un système de score énergétique établissant la compatibilité l'acide aminé à une position et la conformation de la protéine) avec les probabilités d'observer cet acide aminé. Toutefois, on remarque que le gain de vraisemblance du modèle pour toutes les protéines ne permet pas de changer de manière drastique le résultat obtenu sur les positions mal inférées, notamment pour la RFP. Une question qui se pose est de savoir si ce modèle permet d'améliorer significativement l'estimation des probabilités sur les résidus ambigus, ce qui est l'objectif, ou si le gain de vraisemblance se fait avant tout sur les résidus conservés et quasi-certains, en se contentant d'augmenter encore leur probabilité.

Enfin, nous avons sélectionné des protéines pour lesquelles il n'y avait pas d'événements d'insertion ou de délétion important d'une espèce à une autre. Pourtant, ces événements ne sont ni rares, ni anodins, et sont largement ignorés par ces reconstruction ou par les modèles d'évolution des protéines, notamment en terme de stabilité moléculaire. Dans la famille 2 des hémoglobines d'Alvinellidae par exemple, la protéine de *P. p. irlandei* montre une délétion de quatre acides aminés entre les positions 72 et 75, ce qui peut avoir une influence sur la stabilité de la protéine que l'on n'est pas capable de simuler avec un bon niveau de confiance. La dynamique des indels est complexe à décrire et certains programmes

de reconstruction ancestrale, comme FastML utilisé au chapitre 2, préfèrent utiliser un algorithme de parcimonie pour leur inférence (Ashkenazy, Penn et al., 2012). Dans le chapitre 3, nous avons essayé de caractériser les dynamiques évolutives des indels dans deux groupes de métazoaires et avons constaté que les alignements de séquences obtenus par des logiciels populaires s'accordaient mal avec la définition d'un modèle probabiliste d'évolution de ces séquences, une fois replacés dans un contexte de maximum de vraisemblance (figure 4, chapitre 3). Il ne s'agit que d'étapes préliminaires dans l'étude des indels, mais des approches de ce type sont nécessaires pour reconstruire des séquences issues d'organismes plus distants et qui ont accumulé plus d'indels.

V.2 Perspectives

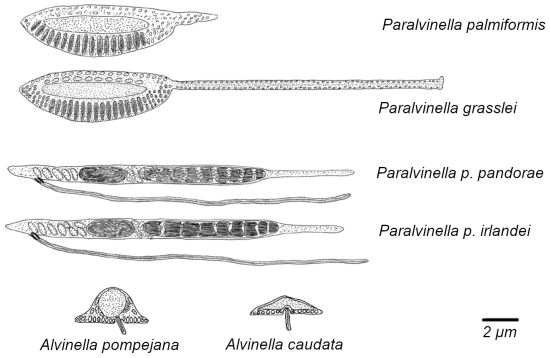
V.2.1 Biologie des Alvinellidae

L'aspect le plus lacunaire dans la connaissance des Alvinellidae concerne la physiologie de ces animaux. Pourtant, comme on l'a établi précédemment, cette famille représente un modèle d'étude très intéressant justement parce que la physiologie de ses différents membres est variée, sur le plan de la température mais pas uniquement car certains de ces vers forment des associations symbiotiques (Jollivet et Stéphane Hourdez, 2020). Approfondir la connaissance de leur physiologie permettrait de bien mieux interpréter certains résultats. Cependant, nous nous heurtons évidemment à la difficulté principale de ce modèle, à savoir notre capacité limitée d'observer ces espèces *in situ* et de les échantillonner. Le maintien à long-terme et l'expérimentation *in vivo* d'individus en aquarium pressurisé est pour l'instant impossible (ou difficilement réalisable), bien que l'on note qu'*Alvinella pompejana* ne résiste pas à la décompression (Ravaux et al., 2013) tandis que les espèces *Paralvinella sulfincola* et *Paralvinella grasslei* ont l'air de mieux supporter la décompression, ce qui facilite leur échantillonage et l'expérimentation (Girguis et Lee, 2006).

Un aspect qui me semblerait particulièrement intéressant à étudier seraient la structure des tubes sécrétés par les Alvinellidae. En effet, ces structures montrent une grande diversité au sein de la famille, et leur rôle présumé de protection est facile à caractériser *via* leur résistance au pH ou à la température. Les Alvinellidae passent la plupart de leur temps dans leurs tubes, n'exposant que leurs branchies aux eaux environnantes (Desbruyères et Laubier, 1991). Chez les espèces *Alvinella*, mais aussi par convergence probable chez l'espèce *Paralvinella* sp. nov., le tube est parcheminé, composé de couches fibrillaires de glycoprotéines empilées comme du contreplaqué dans lesquelles sont incorporés des composés inorganiques (cendres, sédiments, magnésium, calcium, phosphate, fer) ainsi que des bactéries (Françoise Gaill et Hunt, 1986 ; Vovelle et Francoise Gaill, 1986). Chez les *Paralvinella*, le tube est mieux décrit par un mucus moins dense que les animaux sécrètent dans les anfractuosités de la roche pour constituer leur cocon (Desbruyères et Laubier, 1986 ; Desbruyères et Laubier, 1991). Au sein du genre, la diversité des cocons est encore grande, entre des espèces chaudes (*P. hessleri*, *P. sulfincola*, *P. mira*) qui produisent des cocons épais de plusieurs couches pouvant intégrer des particules minérales (Tunnicliffe, Desbruyères et al., 1993 ; Han et al., 2021), et des espèces froides qui produisent un voile

fin de mucus blanchâtre (*P. grasseil*, *P. palmiformis*, *P. pandorae*) voire pas de mucus du tout (*P. bactericola*, Desbruyères et Laubier, 1991). Ainsi, parce qu'ils constituent le premier micro-environnement de ces animaux, ces tubes sont dignes d'intérêt mais leur reconstitution au laboratoire est impossible. Néanmoins, plusieurs protéines ont été récemment mises en évidence à partir de l'annotation du génome d'*A. pompejana*, et des comparaisons inter-spécifiques de ces protéines pourraient être faites dans un futur proche (D. Jollivet, communication personnelle).

Un point physiologique annexe mais que je trouve intriguant est l'étude des spermatozoïdes. En effet, les spermatozoides des polychaetes ont des structures extrêmement variées (Jamieson et Rouse, 1989) et les Alvinellidae ne font pas exception. L'ultrastructure des spermatozoides de six espèces (*P. palmiformis*, *P. grasslei*, *P. p. irlandei*, *P. p. pandorae*, *A. caudata* et *A. pompejana*) montrent de grandes différences de morphologie souvent associées à une motilité plus réduite pour ces formes relativement modifiées (Jouin-Toulmond, Mozzo et Stéphane Hourdez, 2002). Comment la morphologie des gamètes s'accorde avec la phylogénie des espèces ? Le fonctionnement de ces gamètes est-il également adapté à l'environnement hydrothermal, si l'on considère que ces cellules sont de petites structures relativement autonomes ? On note toutefois que les Alvinellidae s'engagent dans une pseudo-copulation avec un transfert direct des gamètes qui fait intervenir les tentacules modifiés des mâles. Les gamètes mâles sont ensuite stockés dans des spermathèques chez les femelles, trait propre aux Alvinellidae parmi les Terebelloformia (Jouin-Toulmond, Mozzo et Stéphane Hourdez, 2002). Pourtant, les espèces de *P. pandorae* ainsi que *P. unidentata* ne présentent pas de tentacules chez les mâles. Pour *P. pandorae*, des cocons ont été observés contenant des mâles, femelles et juvéniles (Jollivet, pers. obs). L'expulsion des spermatozoïdes se ferait dans le milieu extérieur et le cocon empêcherait leur dispersion dans l'eau de mer (Jollivet et Stéphane Hourdez, 2020). On constate malheureusement que la plupart des études amorçant des réflexions sur la physiologie des Alvinellidae date de la fin des années 80 - début 2000. Il est dommage que l'effort récent sur la génétique moléculaire de ces animaux se fasse en pratique au détriment d'études portant sur leur physiologie, alors même que les traits d'histoire de vie des Alvinellidae sont encore si peu connus.



Concernant l'espèce *P. p. irlandei*, que nous avons étudiée au cours de cette thèse, nous remarquons qu'elle est surprenante à bien des égards. Les compositions en acides aminés et purine chez cette espèce sont notamment biaisées par rapport aux autres Terebelloformia

inclus dans la phylogénie moléculaire (figure supplémentaire 5, chapitre 1) et le taux d'indels semble particulièrement élevé (figure 6, chapitre 3). D'un point de vue morphologique, son mode de reproduction est différent des autres Alvinellidae (absence de tentacules buccaux) et elle présente une période de reproduction continue au long de l'année, couvant des juvéniles qui ont peu de capacité de dispersion ce qui serait également propre à cette espèce (McHugh, 1989 ; Jollivet et Stéphane Hourdez, 2020). Enfin, Lelièvre et al., 2018 et Portail et al., 2016 ont montré que le rapport isotopique $\delta^{15}N$ était plus faible chez cette espèce que celui relevé chez *P. sulfincola*, *P. palmiformis*, *P. grasslei*, *P. dela* et *P. bactericola*, ce qui suggère que cette espèce se nourrit d'une source microbienne différente des autres Alvinellidae mesurés. Les deux espèces soeurs de *Paralvinella pandorae* montrent une forte divergence des autres Alvinellidae (figure 5, chapitre 1) et semblent avoir acquis des caractères physiologiques distincts qui les rendent curieuses. L'espèce *Alvinella pompejana* est emblématique de la famille et de loin la plus décrite dans les différents articles relatifs aux Alvinellidae, mais des études comparatives entre *A. pompejana* et *P. pandorae* seraient à mon sens du plus grand intérêt.

Un aspect physiologique également relativement ignoré est la dimension dynamique de la vie à différentes températures chez les Alvinellidae. Une espèce comme *Alvinella pompejana* montre par exemple une véritable "métamorphose thermique" entre la forme larvaire et adulte, la première ne pouvant pas se développer au dessus de 20°C tandis que la seconde montre des signes de stress thermique en dessous de 20°C (Pradillon et al., 2005 ; Ravaux et al., 2013). Cela rejoint potentiellement l'observation de différents allèles de la phosphoglucomutase chez cette espèce montrant des stabilités différentes, et potentiellement associés à des colonisation des cheminées hydrothermales plus ou moins jeunes, et par conséquent plus ou moins chaudes (Piccino et al., 2004). La diversité allélique de certaines protéines montrant des stabilités parfois très différentes a aussi été notée chez d'autres espèces comme *P. sulfincola* ou *P. palmiformis* (Jollivet, Desbruyères et al., 1995). Au cours de cette thèse, nous avons pu modéliser la stabilité de familles d'hémoglobines intracellulaires apparues lors de duplications spécifiques de la lignée des Alvinellidae qui, selon les simulations, auraient des stabilités thermiques très différentes et pourraient potentiellement être attribuées à des régimes variables de température. Cela pose de nombreuses questions sur la capacité d'adaptation des lignées à des environnements fluctuants dans le temps (la température de respiration de l'animal en eau froide n'étant pas la même que la température de nutrition à l'intérieur de son tube), et à la possibilité que le maintien de variants alléliques avec des valeurs sélectives très différentes puisse expliquer la forte introgression proposée entre les lignées ancestrales des Alvinellidae. En outre, nous avons également établi au chapitre 1 que la radiation rapide des Alvinellidae a entraîné un fort taux de tri incomplet de lignées (ILS), sans toutefois le quantifier du fait du bruit induit par les erreurs dans l'estimation des arbres de gène individuels. Si fort taux d'ILS il y a eu au début de la radiation, il est nécessairement la conséquence d'une diversité allélique importante dans les ancêtres de la lignée. Nous pourrions peut-être élucider ces questions par des études de phylogénie qui tiennent compte la variabilité allélique contemporaine (Andermann et al., 2018), bien que ces méthodes ne soient pas pour l'instant utilisées pour résoudre des phylogénies profondes.

V.2.2 Reconstruction de protéines ancestrales

La problématique des modèles d'évolution moléculaire utilisés dans le cadre de l'ASR est aussi très intéressante. Notamment, le cadre thermodynamique récapitulé en introduction ainsi que les résultats obtenus au chapitre 2 m'ont amené à m'interroger sur les liens entre biologie évolutive et biologie structurelle. Un point qui a retenu particulièrement mon attention est la possibilité de donner une valeur quantifiée à la valeur sélective (fitness), concept beaucoup manipulé en biologie de l'évolution mais rarement définie mathématiquement. Dans le cadre de l'évolution des protéines, nous pouvons en effet poser l'hypothèse que la valeur sélective de la protéine est définie par sa stabilité et sa fonction (Serohijos et Eugene I Shakhnovich, 2014), et que la stabilité de la protéine va en moyenne fluctuer selon son abondance cellulaire, sa fonction, la taille de la population et la température de l'organisme. Dans ce cas, il devient possible de quantifier la force de sélection qui s'exerce sur une mutation aléatoire de la séquence par son effet sur $\Delta G^o(T)$, soit l'énergie libre associée à la réaction de dénaturation de la protéine. Bien sûr, l'hypothèse que la valeur sélective dépend uniquement de la stabilité de la protéine est une simplification de la réalité mais cela permet de poser un cadre théorique intéressant pour construire des modèles d'évolution moléculaires en phylogénie qui intègrent des paramètres relatifs aux traits d'histoire de vie bien identifiés en génétique des populations.

Concernant le lien entre sélection et séquence, Kleinman et al., 2010 ont proposé un modèle très intéressant dans lequel les auteurs essaient de prédire la probabilité d'observer une séquence d'acides aminés à partir de la conformation de la protéine (ce qui est en quelque sorte l'inverse d'une prédiction de structure à partir de la séquence). Ce modèle repose notamment sur les interactions immédiates des résidus entre eux, et des contraintes/torsions qu'ils induisent dans le squelette de la protéine. L'objectif final était de produire un modèle d'évolution de séquences, dans lequel l'effet aléatoire des mutations est séparé de l'effet de la sélection qui est prédit selon le modèle développé dans l'article. Les résultats sont encourageants et certaines positions des protéines étudiées sont très bien prédites, mais l'ensemble de la séquence n'est globalement pas suffisamment réaliste pour être utilisé dans le modèle de phylogénie. Les auteurs avancent que la défaillance du modèle peut être due à plusieurs effets non pris en compte, notamment la sélection qui s'opère sur les cinétiques de repliement de la protéine, sur des acides aminés impliqués dans la fonction de la protéine ou encore sur des interactions particulières liées par exemple à des changements de conformation dans la protéine fonctionnelle.

Dans cette optique, l'utilité d'algorithmes comme GEMME (Laine, Karami et Carbone, 2019) prend beaucoup de valeur. Contrairement à l'approche développée par Kleinman et al., 2010, ce genre d'algorithme ne cherche pas à modéliser directement la physique derrière les mutations observées mais utilise des alignements empiriques pour prédire l'épistasie entre les mutations. L'approche pragmatique de GEMME s'est montrée très efficace au chapitre 3 pour attribuer de bonnes valeurs prédictives aux différents acides aminés de la séquence selon leur position. L'utilisation de protéines homologues permet certainement de capturer certains aspects qui échappent à la modélisation physique, notamment les contraintes liées à la fonctionnalité de certains résidus. D'autres auteurs comme Vigué et al., 2022 ont proposé des algorithmes relativement similaires qui cherchent à déterminer l'épistasie entre des mutations à partir d'alignements avec des protéines homologues.

Dans ce second article, les auteurs utilisent un nombre important de génomes d'*Escherichia coli* avec très peu de mutations d'une séquence à l'autre pour pouvoir finement mesurer l'épistasie entre des domaines homologues de protéines. Certaines précautions prises dans cet article, notamment concernant le contrôle de l'indépendance phylogénétique des séquences utilisées dans l'alignement, seraient utiles à incorporer dans GEMME. Toutefois, la méthode donne de bons résultats au prix d'un besoin en séquences homologues très élevé, et est en soit moins générale que GEMME dans l'utilisation que l'on cherche à obtenir du logiciel.

Au delà de l'aspect théorique de ces modèles, une difficulté reste celle de leur validation empirique. En effet, la reconstruction de protéines ancestrales s'attaque à des questions très fines d'évolution moléculaire, mais se heurte au problème que les séquences ancestrales ne peuvent jamais être observées pour valider l'exactitude du modèle que l'on évalue par une mesure de vraisemblance calculée selon certaines hypothèses arbitraires. On est par conséquent amené à tester les modèles sur des données simulées, mais la réalité de l'évolution des séquences est évidemment plus complexe. D'autant plus si l'on s'intéresse aux questions d'épistasie dans les protéines qu'on ne sait pas simuler de manière réaliste. Idéalement, pour mettre à l'épreuve les modèles ASR, nous aurions besoin de plus de données empiriques que nous pourrions obtenir par évolution expérimentale, à l'instar de la phylogénie expérimentale de Randall et al., 2016 utilisée au chapitre 3. L'obtention de réplicats de phylogénies expérimentales serait donc particulièrement précieux dans ce domaine.

V.3 Conclusion

Dans cette thèse, nous avons montré que des protéines ancestrales reconstruites de la lignées des Alvinellidae avaient une stabilité similaire à celle des espèces contemporaines thermotolérantes. Bien que ces protéines puissent avoir des différences avec les protéines ancestrales réelles, ce résultat est obtenu avec un bon niveau de confiance, ce qui nous amène à conclure que l'ancêtre de la lignée des Alvinellidae était une espèce déjà adaptée aux températures chaudes. Les conditions paléoclimatiques du Crétacé, bien que notablement plus chaudes dans l'océan Pacifique profond qu'aux temps actuels, ne suffisent pas à expliquer cette thermotolérance élevée. Il est probable que l'ancêtre ait alors déjà été une espèce endémique des cheminées des sources hydrothermales, mais cette affirmation mériterait d'être confirmée notamment par une étude fonctionnelle des protéines ancestrales reconstruites.

Or l'étude des protéines ancestrales repose sur deux domaines de la biologie, à savoir la biologie structurelle et la biologie évolutive, qui gagneraient à être mieux intégrés. L'amélioration des modèles de reconstruction des protéines par l'intégration de paramètres structuraux permettrait en particulier de mieux identifier les forces évolutives qui peuvent avoir une influence sur les variations de stabilité constatées entre les protéines anciennes et contemporaines.

Les Alvinellidae, qui montrent une histoire complexe de convergences phénotypiques permettant aux espèces contemporaines de coloniser des milieux de températures contrastées, représentent une famille idéale pour appliquer ces concepts. L'acquisition de la psy-

chrophilie chez certaines espèces serait apparue trois fois de manières indépendante, et la thermotolérance a potentiellement été gagnée à nouveau plus récemment dans certaines espèces du genre *Paralvinella*. En particulier, nous avons identifié trois familles d'hémoglobines intracellulaires spécifiques des Alvinellidae et qui semblent être spécialisées pour fonctionner dans des régimes thermiques différents. La multiplicité de ces hémoglobines serait la conséquence de duplications successives intervenues lors de la radiation de la lignée. Elles pourraient constituer un modèle d'étude exceptionnel pour comprendre ces différentes convergences, considérant que ces protéines doivent avoir évolué de façon parallèle dans les différentes branches de la lignée.

V.4 Soutenance

La soutenance publique de la thèse (29 avril 2024) est disponible à l'adresse suivante : https://www.youtube.com/live/GKjzCAM-aIQ.

Bibliographie

Akanuma, Satoshi (août 2017). “Characterization of Reconstructed Ancestral Proteins Suggests a Change in Temperature of the Ancient Biosphere”. en. In : *Life* 7.3, p. 33. doi : 10.3390/life7030033.

Andermann, Tobias et al. (mai 2018). “Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements”. en. In : *Systematic Biology*. Sous la dir. de Susanne Renner. doi : 10.1093/sysbio/syy039.

Arenas, Miguel et Ugo Bastolla (fév. 2020). “ProtASR2 : Ancestral reconstruction of protein sequences accounting for folding stability”. en. In : *Methods in Ecology and Evolution* 11.2. Sous la dir. d’Emmanuel Paradis, p. 248-257. doi : 10.1111/2041-210X.13341.

Artemieva, Irina M. (jan. 2023). “Back-arc basins : A global view from geophysical synthesis and analysis”. en. In : *Earth-Science Reviews* 236, p. 104242. doi : 10.1016/j.earscirev.2022.104242.

Ashkenazy, Haim, Ofir Cohen et al. (déc. 2014). “Indel Reliability in Indel-Based Phylogenetic Inference”. en. In : *Genome Biology and Evolution* 6.12, p. 3199-3209. doi : 10.1093/gbe/evu252.

Ashkenazy, Haim, Osnat Penn et al. (juill. 2012). “FastML : a Web Server for Probabilistic Reconstruction of Ancestral Sequences”. en. In : *Nucleic Acids Research* 40.W1, W580-W584. doi : 10.1093/nar/gks498.

Bachraty, Charleyne, Pierre Legendre et Daniel Desbruyères (août 2009). “Biogeographic Relationships Among Deep-Sea Hydrothermal Vent Faunas at Global Scale”. en. In : *Deep Sea Research Part I : Oceanographic Research Papers* 56.8, p. 1371-1378. doi : 10.1016/j.dsr.2009.01.009.

Baele, Guy, Yves Van De Peer et Stijn Vansteelandt (juill. 2010). “Using Non-Reversible Context-Dependent Evolutionary Models to Study Substitution Patterns in Primate Non-Coding Sequences”. en. In : *Journal of Molecular Evolution* 71.1, p. 34-50. doi : 10.1007/s00239-010-9362-y.

Barnes, Harold (1991). *Oceanography and Marine Biology*. en. OCLC : 1027145927. Milton : Taylor & Francis.

Beauchamp, Charles et Irwin Fridovich (nov. 1971). “Superoxide dismutase : Improved assays and an assay applicable to acrylamide gels”. en. In : *Analytical Biochemistry* 44.1, p. 276-287. doi : 10.1016/0003-2697(71)90370-8.

Beaulieu, Stace E. et al. (nov. 2013). “An authoritative global database for active submarine hydrothermal vent fields”. en. In : *Geochemistry, Geophysics, Geosystems* 14.11, p. 4892-4905. doi : 10.1002/2013GC004998.

Becktel, Wayne J. et John A. Schellman (nov. 1987). “Protein stability curves”. en. In : *Biopolymers* 26.11, p. 1859-1877. doi : 10.1002/bip.360261104.

Bioy, Alexis et al. (jan. 2022). “Balanced Polymorphism at the Pgm-1 Locus of the Pompeii Worm Alvinella pompejana and Its Variant Adaptability Is Only Governed by Two QE Mutations at Linked Sites”. en. In : *Genes* 13.2, p. 206. doi : 10.3390/genes13020206.

Blanquart, Samuel, Mathieu Groussin et al. (août 2021). “Resurrection of Ancestral Malate Dehydrogenases Reveals the Evolutionary History of Halobacterial Proteins : Deciphering Gene Trajectories and Changes in Biochemical Properties”. en. In : *Molecular Biology and Evolution* 38.9. Sous la dir. de Miriam Barlow, p. 3754-3774. doi : 10.1093/molbev/msab146.

Blanquart, Samuel et Nicolas Lartillot (mai 2008). “A Site- and Time-Heterogeneous Model of Amino Acid Replacement”. en. In : *Molecular Biology and Evolution* 25.5, p. 842-858. doi : 10.1093/molbev/msn018.

Bock, Thomas et al. (déc. 2014). “An Integrated Approach for Genome Annotation of the Eukaryotic Thermophile *Chaetomium thermophilum*”. en. In : *Nucleic Acids Research* 42.22, p. 13525-13533. doi : 10.1093/nar/gku1147.

Boël, Grégory et al. (jan. 2016). “Codon Influence on Protein Expression in *E. coli* Correlates with mRNA Levels”. en. In : *Nature* 529.7586, p. 358-363. doi : 10.1038/nature16509.

Bordo, Domenico, Kristina Djinovic et Martino Bolognesi (mai 1994). “Conserved Patterns in the Cu,Zn Superoxide Dismutase Family”. en. In : *Journal of Molecular Biology* 238.3, p. 366-386. doi : 10.1006/jmbi.1994.1298.

Boulart, Cédric et al. (mars 2022). “Active Hydrothermal Vents in the Woodlark Basin May Act as Dispersing Centres for Hydrothermal Fauna”. en. In : *Communications Earth & Environment* 3.1, p. 64. doi : 10.1038/s43247-022-00387-9.

Boussau, Bastien et al. (déc. 2008). “Parallel Adaptations to High Temperatures in the Archaean Eon”. en. In : *Nature* 456.7224, p. 942-945. doi : 10.1038/nature07393.

Bouvignies, Guillaume et al. (sept. 2011). “Solution structure of a minor and transiently formed state of a T4 lysozyme mutant”. en. In : *Nature* 477.7362, p. 111-114. doi : 10.1038/nature10349.

Brochier-Armanet, Céline et Dominique Madern (déc. 2021). “Phylogenetics and biochemistry elucidate the evolutionary link between l-malate and l-lactate dehydrogenases and disclose an intermediate group of sequences with mix functional properties”. en. In : *Biochimie* 191, p. 140-153. doi : 10.1016/j.biochi.2021.08.004.

Brun, Pierre-Guillaume et al. (juill. 2023). *A step in the deep evolution of Alvinellidae (Annelida : Polychaeta) : a phylogenomic comparative approach based on transcriptomes*. en. preprint. Evolutionary Biology. doi : 10.1101/2023.07.24.550320.

Cai, Liming et al. (avr. 2021). “The Perfect Storm : Gene Tree Estimation Error, Incomplete Lineage Sorting, and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade, Malpighiales”. en. In : *Systematic Biology* 70.3. Sous la dir. de Mark Fishbein, p. 491-507. doi : 10.1093/sysbio/syaa083.

Cai, Wei, Jimin Pei et Nick V Grishin (2004). “Reconstruction of Ancestral Protein Sequences and its Applications”. en. In : *BMC Evolutionary Biology* 4.1, p. 33. doi : 10.1186/1471-2148-4-33.

Cartwright, Reed A. (fév. 2009). “Problems and Solutions for Estimating Indel Rates and Length Distributions”. en. In : *Molecular Biology and Evolution* 26.2, p. 473-480. doi : 10.1093/molbev/msn275.

Chang, Belinda S.W., Juan A. Ugalde et Mikhail V. Matz (2005). “Applications of Ancestral Protein Reconstruction in Understanding Protein Function : GFP-Like Proteins”. en. In : *Methods in Enzymology*. T. 395. Elsevier, p. 652-670. doi : 10.1016/S0076-6879(05)95034-9.

Chausson, Fabienne, Christopher R. Bridges et al. (déc. 2001). “Structural and Functional Properties of Hemocyanin from *Cyanagraea praedator*, a Deep-Sea Hydrothermal Vent Crab”. en. In : *Proteins : Structure, Function, and Genetics* 45.4, p. 351-359. doi : 10.1002/prot.10014.

Chausson, Fabienne, Sarah Sanglier et al. (jan. 2004). “Respiratory Adaptations to the Deep-Sea Hydrothermal Vent Environment : the Case of *Segonzacia mesatlantica*, a Crab from the Mid-Atlantic Ridge”. en. In : *Micron* 35.1-2, p. 31-41. doi : 10.1016/j.micron.2003.10.010.

Chen, Shifu et al. (sept. 2018). “Fastp : an Ultra-Fast All-in-One FASTQ Preprocessor”. en. In : *Bioinformatics* 34.17, p. i884-i890. doi : 10.1093/bioinformatics/bty560.

Chernomor, Olga, Arndt Von Haeseler et Bui Quang Minh (nov. 2016). “Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices”. en. In : *Systematic Biology* 65.6, p. 997-1008. doi : 10.1093/sysbio/syw037.

Chevaldonné, Pierre, Charles R. Fisher et al. (2000). “Thermotolerance and the ’Pompeii worms’”. en. In : *Marine Ecology Progress Series* 208, p. 293-295. doi : 10.3354/meps208293.

Chevaldonné, Pierre et Didier Jollivet (1993). “Videoscopic Study of Deep-Sea Hydrothermal Vent Alvinellid Polychaete Populations : Biomass Estimation and Behaviour”. en. In : *Marine Ecology Progress Series* 95, p. 251-262. doi : 10.3354/meps095251.

Chevaldonné, Pierre, Didier Jollivet et al. (2002). “Sister-Species of Eastern Pacific Hydrothermal Vent Worms (Ampharetidae, Alvinellidae, Vestimentifera) Provide New Mitochondrial COI Clock Calibration”. en. In : *Cahiers de Biologie Marine* 43.3-4, p. 367-370.

Cheynier, Rémi et al. (juill. 2001). “Insertion/deletion frequencies match those of point mutations in the hypervariable regions of the simian immunodeficiency virus surface envelope gene”. en. In : *Journal of General Virology* 82.7, p. 1613-1619. doi : 10.1099/0022-1317-82-7-1613.

Cottin, Delphine et al. (juill. 2008). “Thermal Biology of the Deep-Sea Vent Annelid *Paralvinella grasslei* : *In Vivo* Studies”. en. In : *Journal of Experimental Biology* 211.14, p. 2196-2204. doi : 10.1242/jeb.018606.

Creer, Simon (jan. 2007). “Choosing and Using Introns in Molecular Phylogenetics”. en. In : *Evolutionary Bioinformatics* 3, p. 117693430700300. doi : 10.1177/117693430700300011.

Dahlhoff, Elizabeth, John O’Brien et al. (nov. 1991). “Temperature Effects on Mitochondria from Hydrothermal Vent Invertebrates : Evidence for Adaptation to Elevated and Variable Habitat Temperatures”. en. In : *Physiological Zoology* 64.6, p. 1490-1508. doi : 10.1086/physzool.64.6.30158226.

Dahlhoff, Elizabeth et George N. Somero (sept. 1991). “Pressure and Temperature Adaptation of Cytosolic Malate Dehydrogenases of Shallowand Deep-Living Marine Invertebrates : Evidence for High Body Temperatures in Hydrothermal Vent Animals”. en. In : *Journal of Experimental Biology* 159.1, p. 473-487. doi : 10.1242/jeb.159.1.473.

Day, John H. (1964). “A Review of the Family Ampharetidae (Polychaeta)”. In : *Annals of the South African Museum. Annale van die Suid-Afrikaanse Museum* 48, p. 97-120.

De Maio, Nicola (fév. 2021). “The Cumulative Indel Model : Fast and Accurate Statistical Evolutionary Alignment”. en. In : *Systematic Biology* 70.2. Sous la dir. d’Edward Susko, p. 236-257. doi : 10.1093/sysbio/syaa050.

Deckert, Gerard et al. (mars 1998). “The Complete Genome of the Hyperthermophilic Bacterium *Aquifex aeolicus*”. en. In : *Nature* 392.6674, p. 353-358. doi : 10.1038/32831.

Desbruyères, Daniel, Pierre Chevaldonné et al. (jan. 1998). “Biology and Ecology of the “Pompeii worm” (*Alvinella pompejana* Desbruyères and Laubier), a Normal Dweller of an Extreme Deep-Sea Environment : A Synthesis of Current Knowledge and Recent Developments”. en. In : *Deep Sea Research Part II : Topical Studies in Oceanography* 45.1-3, p. 383-422. doi : 10.1016/S0967-0645(97)00083-0.

Desbruyères, Daniel, Jun Hashimoto et Marie-Claire Fabri (2006). “Composition and Biogeography of Hydrothermal Vent Communities in Western Pacific Back-Arc Basins”. en. In : *Geophysical Monograph Series*. Sous la dir. de David M. Christie et al. T. 166. Washington, D. C. : American Geophysical Union, p. 215-234. doi : 10.1029/166GM11.

Desbruyères, Daniel et Lucien Laubier (1980). “*Alvinella pompejana* gen. sp. nov., Ampharetidae Aberrant des Sources Hydrothermales de la Ride Est-Pacifique”. In : *Oceanologica Acta* 3, p. 267-274.

— (1982). “*Paralvinella grasslei*, new Genus, new Species of Alvinellinae (Polychaeta : Ampharetidae) from the Galapagos Rift Geothermal Vents”. In : *Proceedings of the Biological Society of Washington* 95, p. 484-494.

— (oct. 1986). “Les Alvinellidae, une Famille Nouvelle d’Annélides Polychètes Inféodées aux Sources Hydrothermales Sous-Marines : Systématique, Biologie et Ecologie”. en. In : *Canadian Journal of Zoology* 64.10, p. 2227-2245. doi : 10.1139/z86-337.

— (1989). "*Paralvinella hessleri*, New Species Of Alvinellidae (Polychaeta) from the Mariana Back-Arc Basin Hydrothermal Vents". In : *Proceedings of the Biological Society of Washington* 102, p. 761-767.
— (1991). "Systematics, Phylogeny, Ecology and Distribution of the Alvinellidae (Polychaeta) from Deep-Sea Hydrothermal Vents". In : *Ophelia Supplement* 5, p. 31-45.
— (1993). "New Species of Alvinellidae (Polychaeta) from the North Fiji Back-Arc Basin Hydrothermal Vents (Southwestern Pacific)". In : *Proceedings of the Biological Society of Washington* 106, p. 225-236.

Dessimoz, Christophe et Manuel Gil (2010). "Phylogenetic assessment of alignments reveals neglected tree signal in gaps". en. In : *Genome Biology* 11.

Detinova, N.N. (1988). "New Species of Polychaetous Annelids from Hydrothermal Vents of the Juan-de-Fuca Ridge (Pacific-Ocean)". In : *Zoologichesky Zhurnal* 67.6, p. 858-864.

Dettaï, Agnes et al. (2008). "Inferring Evolution of Fish Proteins : The Globin Case Study". en. In : *Methods in Enzymology*. T. 436. Elsevier, p. 539-570. doi : 10.1016/S0076-6879(08)36030-3.

Dilly, Geoffrey F. et al. (août 2012). "Exploring the Limit of Metazoan Thermal Tolerance via Comparative Proteomics : Thermally Induced Changes in Protein Abundance by Two Hydrothermal Vent Polychaetes". en. In : *Proceedings of the Royal Society B : Biological Sciences* 279.1741, p. 3347-3356. doi : 10.1098/rspb.2012.0098.

Do, Chuong B. et al. (fév. 2005). "ProbCons : Probabilistic Consistency-Based Multiple Sequence Alignment". en. In : *Genome Research* 15.2, p. 330-340. doi : 10.1101/gr.2821705.

Donath, Alexander et Peter F. Stadler (déc. 2018). "Split-inducing indels in phylogenomic analysis". en. In : *Algorithms for Molecular Biology* 13.1, p. 12. doi : 10.1186/s13015-018-0130-7.

Drummond, Alexei J. et al. (mars 2006). "Relaxed Phylogenetics and Dating with Confidence". en. In : *PLoS Biology* 4.5. Sous la dir. de David Penny, e88. doi : 10.1371/journal.pbio.0040088.

Eick, Geeta N. et al. (oct. 2016). "Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty". en. In : *Molecular Biology and Evolution*, msw223. doi : 10.1093/molbev/msw223.

Eilertsen, Mari H. et al. (déc. 2017). "Do Ampharetids Take Sedimented Steps Between Vents and Seeps ? Phylogeny and Habitat-Use of Ampharetidae (Annelida, Terebelliformia) in chemosynthesis-Based Ecosystems". en. In : *BMC Evolutionary Biology* 17.1, p. 222. doi : 10.1186/s12862-017-1065-1.

El Hilali, Sami et al. (juin 2024). *Chromosome-scale genome assembly and gene annotation of the hydrothermal vent annelid Alvinella pompejana yield insight into animal evolution in extreme environments*. en. doi : 10.1101/2024.06.25.600561.

Erauso, Gaël et al. (nov. 1993). "Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent". en. In : *Archives of Microbiology* 160.5. doi : 10.1007/BF00252219.

Farias, Sávio T et Maria Christina M Bonato (2003). “Preferred amino acids and thermostability”. en. In : *Genetics and Molecular Research*.

Feng, Shaohong et al. (mai 2022). “Incomplete lineage sorting and phenotypic evolution in marsupials”. en. In : *Cell* 185.10, 1646-1660.e18. doi : 10.1016/j.cell.2022.03.034.

Fields, Jeremy H.A. et James F. Quinn (jan. 1981). “Some theoretical considerations on cytosolic redox balance during anaerobiosis in marine invertebrates”. en. In : *Journal of Theoretical Biology* 88.1, p. 35-45. doi : 10.1016/0022-5193(81)90327-1.

Fletcher, Willian et Ziheng Yang (août 2009). “INDELible : A Flexible Simulator of Biological Sequence Evolution”. en. In : *Molecular Biology and Evolution* 26.8, p. 1879-1888. doi : 10.1093/molbev/msp098.

Fontanillas, Eric et al. (jan. 2017). “Proteome Evolution of Deep-Sea Hydrothermal Vent Alvinellid Polychaetes Supports the Ancestry of Thermophily and Subsequent Adaptation to Cold in Some Lineages”. en. In : *Genome Biology and Evolution*, evw298. doi : 10.1093/gbe/evw298.

Foster, Peter G. (juin 2004). “Modeling Compositional Heterogeneity”. en. In : *Systematic Biology* 53.3. Sous la dir. de Ted Schultz, p. 485-495. doi : 10.1080/10635150490445779.

Fraser, Nicholas J. et al. (juin 2016). “Evolution of Protein Quaternary Structure in Response to Selective Pressure for Increased Thermostability”. en. In : *Journal of Molecular Biology* 428.11, p. 2359-2371. doi : 10.1016/j.jmb.2016.03.014.

Gagnière, Nicolas et al. (déc. 2010). “Insights into Metazoan Evolution from *Alvinella pompejana* cDNAs”. en. In : *BMC Genomics* 11.1, p. 634. doi : 10.1186/1471-2164-11-634.

Gaill, Francoise et al. (fév. 1995). “Structural Comparison of Cuticle and Interstitial Collagens from Annelids Living in Shallow Sea-water and at Deep-sea Hydrothermal Vents”. en. In : *Journal of Molecular Biology* 246.2, p. 284-294. doi : 10.1006/jmbi.1994.0084.

Gaill, Françoise et S. Hunt (1986). “Tubes of Deep Sea Hydrothermal Vent Worms *Riftia pachyptila* (Vestimentifera) and *Alvinella pompejana* (Annelida)”. en. In : *Marine Ecology Progress Series* 34, p. 267-274. doi : 10.3354/meps034267.

Galtier, Nicolas (mai 2001). “Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model”. en. In : *Molecular Biology and Evolution* 18.5, p. 866-873. doi : 10.1093/oxfordjournals.molbev.a003868.

Gaucher, Eric A., Sridhar Govindarajan et Omjoy K. Ganesh (fév. 2008). “Palaeotemperature trend for Precambrian life inferred from resurrected proteins”. en. In : *Nature* 451.7179, p. 704-707. doi : 10.1038/nature06510.

Gehring, W J et R Wehner (mars 1995). “Heat shock protein synthesis and thermotolerance in Cataglyphis, an ant from the Sahara desert.” en. In : *Proceedings of the National Academy of Sciences* 92.7, p. 2994-2998. doi : 10.1073/pnas.92.7.2994.

Genard, B et al. (mai 2013). “Living in a hot redox soup : antioxidant defences of the hydrothermal worm Alvinella pompejana”. en. In : *Aquatic Biology* 18.3, p. 217-228. doi : 10.3354/ab00498.

Gerday, Charles et al. (oct. 1997). “Psychrophilic enzymes : a thermodynamic challenge”. en. In : *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1342.2, p. 119-131. doi : 10.1016/S0167-4838(97)00093-9.

Girguis, Peter R. et Raymond W. Lee (avr. 2006). “Thermal Preference and Tolerance of Alvinellids”. en. In : *Science* 312.5771, p. 231-231. doi : 10.1126/science.1125286.

Glasby, Christopher J., Patricia A. Hutchings et Kathryn Hall (oct. 2004). “Assessment of Monophyly and Taxon Affinities within the Polychaete Clade Terebelliformia (Terebellida)”. en. In : *Journal of the Marine Biological Association of the United Kingdom* 84.5, p. 961-971. doi : 10.1017/S0025315404010252h.

Goldstein, Richard A. (mai 2011). “The evolution and evolutionary consequences of marginal thermostability in proteins”. en. In : *Proteins : Structure, Function, and Bioinformatics* 79.5, p. 1396-1407. doi : 10.1002/prot.22964.

Grabherr, Manfred G. et al. (juill. 2011). “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome”. en. In : *Nature Biotechnology* 29.7, p. 644-652. doi : 10.1038/nbt.1883.

Gu, Xun et Wen-Hsiung Li (avr. 1995). “The Size Distribution of Insertions and Deletions in Human and Rodent Pseudogenes Suggests the Logarithmic Gap Penalty for Sequence Alignment”. en. In : *Journal of Molecular Evolution* 40.4, p. 464-473. doi : 10.1007/BF00164032.

Guindon, Stéphane et al. (mars 2010). “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0”. en. In : *Systematic Biology* 59.3, p. 307-321. doi : 10.1093/sysbio/syq010.

Han, Yuru et al. (mai 2021). “Out of the Pacific : A New Alvinellid Worm (Annelida : Terebellida) From the Northern Indian Ocean Hydrothermal Vents”. en. In : *Frontiers in Marine Science* 8, p. 669918. doi : 10.3389/fmars.2021.669918.

Hand, Steven C. et George N. Somero (août 1983). “Energy Metabolism Pathways of Hydrothermal Vent Animals : Adaptations to a Food-rich and Sulfide-rich Deep-sea Environment”. en. In : *The Biological Bulletin* 165.1, p. 167-181. doi : 10.2307/1541362.

Hanson-Smith, Victor, Bryan Kolaczkowski et Joseph W. Thornton (sept. 2010). “Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty”. en. In : *Molecular Biology and Evolution* 27.9, p. 1988-1999. doi : 10.1093/molbev/msq081.

Hart, Kathryn M. et al. (2014). “Thermodynamic system drift in protein evolution”. In : *PLoS Biol* 12.11, e1001994.

Haymon, Rachel M., Randolph A. Koski et Colin Sinclair (mars 1984). “Fossils of Hydrothermal Vent Worms from Cretaceous Sulfide Ores of the Samail Ophiolite, Oman”. en. In : *Science* 223.4643, p. 1407-1409. doi : 10.1126/science.223.4643.1407.

Henscheid, Kristy L. et al. (mars 2005). “The Splicing Factor U2AF65 is Functionally Conserved in the Thermotolerant Deep-Sea Worm *Alvinella pompejana*”. en. In : *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1727.3, p. 197-207. doi : 10.1016/j.bbaexp.2005.01.008.

Hessler, Robert R. et Peter F. Lonsdale (fév. 1991). “Biogeography of Mariana Trough Hydrothermal vVnt Communities”. en. In : *Deep Sea Research Part A. Oceanographic Research Papers* 38.2, p. 185-199. doi : 10.1016/0198-0149(91)90079-U.

Hickey, Donal A. et Gregory A.C. Singer (2004). “Genomic and Proteomic Adaptations to Growth at High Temperature”. en. In : *Genome Biology*, p. 7.

Hoang, Diep Thi et al. (fév. 2018). “UFBoot2 : Improving the Ultrafast Bootstrap Approximation”. en. In : *Molecular Biology and Evolution* 35.2, p. 518-522. doi : 10.1093/molbev/msx281.

Hoff, Katharina J. et Mario Stanke (juill. 2013). “WebAUGUSTUS - a Web Service for Training AUGUSTUS and Predicting Genes in Eukaryotes”. en. In : *Nucleic Acids Research* 41.W1, W123-W128. doi : 10.1093/nar/gkt418.

Holder, Thomas et al. (déc. 2013). “Deep Transcriptome-Sequencing and Proteome Analysis of the Hydrothermal Vent Annelid *Alvinella pompejana* Identifies the CvP-Bias as a Robust Measure of Eukaryotic Thermostability”. en. In : *Biology Direct* 8.1, p. 2. doi : 10.1186/1745-6150-8-2.

Holm, Jeppe, Pouria Dasmeh et Kasper P. Kepp (juill. 2016). “Tracking evolution of myoglobin stability in cetaceans using experimentally calibrated computational methods that account for generic protein relaxation”. en. In : *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1864.7, p. 825-834. doi : 10.1016/j.bbapap.2016.04.004.

Holmes, Ian H. (déc. 2017). “Solving the master equation for Indels”. en. In : *BMC Bioinformatics* 18.1, 255, s12859–017–1665-1. doi : 10.1186/s12859-017-1665-1.

Hourdez, S et R Weber (jan. 2005). “Molecular and functional adaptations in deep-sea hemoglobins”. en. In : *Journal of Inorganic Biochemistry* 99.1, p. 130-141. doi : 10.1016/j.jinorgbio.2004.09.017.

Hourdez, Stéphane et Didier Jollivet (oct. 2020). “Metazoan Adaptation to Deep-Sea Hydrothermal Vents”. en. In : *Life in Extreme Environments*. Sous la dir. de Guido di Prisco et al. 1re éd. Cambridge University Press, p. 42-67. doi : 10.1017/9781108683319.004.

Hourdez, Stéphane et François H. Lallier (août 2007). “Adaptations to Hypoxia in Hydrothermal-Vent and Cold-Seep Invertebrates”. en. In : *Reviews in Environmental Science and Bio/Technology* 6.1-3, p. 143-159. doi : 10.1007/s11157-006-9110-3.

Hourdez, Stéphane, François H. Lallier et al. (mai 2000). “Gas Transfer System in *Alvinella pompejana* (Annelida Polychaeta, Terebellida) : Functional Properties of Intracellular and Extracellular Hemoglobins”. en. In : *Physiological and Biochemical Zoology* 73.3, p. 365-373. doi : 10.1086/316755.

Huang, Xiaoqiu et Anup Madan (sept. 1999). “CAP3 : A DNA Sequence Assembly Program”. en. In : *Genome Research* 9.9, p. 868-877. doi : 10.1101/gr.9.9.868.

Hudson, Richard R. et Jerry A. Coyne (août 2002). “Mathematical Consequences Of The Genealogical Species Concept”. en. In : *Evolution* 56.8, p. 1557-1565. doi : 10.1111/j.0014-3820.2002.tb01467.x.

Ishikawa, Sohta A et al. (sept. 2019). “A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios”. en. In : *Molecular Biology and Evolution* 36.9. Sous la dir. de Tal Pupko, p. 2069-2085. doi : 10.1093/molbev/msz131.

Isogai, Yasuhiro et al. (déc. 2018). “Tracing Whale Myoglobin Evolution by Resurrecting Ancient Proteins”. en. In : *Scientific Reports* 8.1, p. 16883. doi : 10.1038/s41598-018-34984-6.

Jacobs, David K. et David R. Lindberg (août 1998). “Oxygen and Evolutionary Patterns in the Sea : Onshore/Offshore Trends and Recent Recruitment of Deep-Sea Faunas”. en. In : *Proceedings of the National Academy of Sciences* 95.16, p. 9396-9401. doi : 10.1073/pnas.95.16.9396.

Jaenicke, Rainer et Gerald Böhm (déc. 1998). “The Stability of Proteins in Extreme Environments”. en. In : *Current Opinion in Structural Biology* 8.6, p. 738-748. doi : 10.1016/S0959-440X(98)80094-8.

Jamieson, Barrie G. M. et Greg W. Rouse (mai 1989). “The Spermatozoa of the Polychaeta (Annelida) : An Ultrastructural Review”. en. In : *Biological Reviews* 64.2, p. 93-157. doi : 10.1111/j.1469-185X.1989.tb00673.x.

Jang, Sook-Jin et al. (déc. 2016). “Population Subdivision of Hydrothermal Vent Polychaete *Alvinella pompejana* across Equatorial and Easter Microplate Boundaries”. en. In : *BMC Evolutionary Biology* 16.1, p. 235. doi : 10.1186/s12862-016-0807-9.

Jarmuszkiewicz, Wieslawa et al. (juin 2015). “Temperature Controls Oxidative Phosphorylation and Reactive Oxygen Species Production through Uncoupling in Rat Skeletal Muscle Mitochondria”. en. In : *Free Radical Biology and Medicine* 83, p. 12-20. doi : 10.1016/j.freeradbiomed.2015.02.012.

Jiang, Jingjing et al. (août 2016). “Hydrogen Sulfide—Mechanisms of Toxicity and Development of an Antidote”. en. In : *Scientific Reports* 6.1, p. 20831. doi : 10.1038/srep20831.

Johnson, Kenneth S. et al. (mars 1986). “In Situ Measurements of Chemical Distributions in a Deep-Sea Hydrothermal Vent Field”. en. In : *Science* 231.4742, p. 1139-1141. doi : 10.1126/science.231.4742.1139.

Jollivet, Didier, Daniel Desbruyères et al. (1995). “Evidence for Differences in the Allozyme Thermostability of Seep-Sea Hydrothermal Vent Polychaetes (Alvinellidae) : a Possible Selection by Habitat”. en. In : *Marine Ecology Progress Series* 123, p. 125-136. doi : 10.3354/meps123125.

Jollivet, Didier et Stéphane Hourdez (2020). “7.7.4 Alvinellidae Desbruyères & Laubier, 1986”. en. In : *Pleistoannelida, Sedentaria III and Errantia I* 3, p. 18.

Jollivet, Didier, Jean Mary et al. (fév. 2012). “Proteome Adaptation to High Temperatures in the Ectothermic Hydrothermal Vent Pompeii Worm”. en. In : *PLoS ONE* 7.2. Sous la dir. de David Liberles, e31150. doi : 10.1371/journal.pone.0031150.

Jones, David T., William R. Taylor et Janet M. Thornton (1992). “The Rapid Generation of Mutation Data Matrices from Protein Sequences”. en. In : *Bioinformatics* 8.3, p. 275-282. doi : 10.1093/bioinformatics/8.3.275.

Jouin, Claude et Françoise Gaill (jan. 1990). “Gills of Hydrothermal Vent Annelids : Structure, Ultrastructure and Functional Implications in Two Alvinellid Species”. en. In : *Progress in Oceanography* 24.1-4, p. 59-69. doi : 10.1016/0079-6611(90)90019-X.

Jouin-Toulmond, Claude, Masina Mozzo et Stéphane Hourdez (2002). “Ultrastructure of Spermatozoa in Four Species of Alvinellidae (Annelida : Polychaeta)”. en. In : *Cahiers de Biologie Marine* 43, p. 5.

Kalyaanamoorthy, Subha et al. (juin 2017). “ModelFinder : fast model selection for accurate phylogenetic estimates”. en. In : *Nature Methods* 14.6, p. 587-589. doi : 10.1038/nmeth.4285.

Karshikoff, Andrey, Lennart Nilsson et Rudolf Ladenstein (oct. 2015). “Rigidity versus flexibility : the dilemma of understanding protein thermal stability”. en. In : *The FEBS Journal* 282.20, p. 3899-3917. doi : 10.1111/febs.13343.

Kashiwagi, Sayo et al. (2010). “Characterization of a Y-Family DNA Polymerase eta from the Eukaryotic Thermophile *Alvinella pompejana*”. en. In : *Journal of Nucleic Acids* 2010, p. 1-13. doi : 10.4061/2010/701472.

Katoh, K. et D. M. Standley (avr. 2013). “MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability”. en. In : *Molecular Biology and Evolution* 30.4, p. 772-780. doi : 10.1093/molbev/mst010.

Kaule, Gunhild et al. (juill. 1998). “Prolyl Hydroxylase Activity in Tissue Homogenates of Annelids from Deep Sea Hydrothermal Vents”. en. In : *Matrix Biology* 17.3, p. 205-212. doi : 10.1016/S0945-053X(98)90059-2.

Keul, Frank et al. (déc. 2017). “PFASUM : a substitution matrix from Pfam structural alignments”. en. In : *BMC Bioinformatics* 18.1, p. 293. doi : 10.1186/s12859-017-1703-z.

Kleinman, Claudia L. et al. (juill. 2010). “Statistical Potentials for Improved Structurally Constrained Evolutionary Models”. en. In : *Molecular Biology and Evolution* 27.7, p. 1546-1560. doi : 10.1093/molbev/msq047.

Kocot, Kevin M. et al. (sept. 2016). “Phylogenomics of Lophotrochozoa with Consideration of Systematic Error”. en. In : *Systematic Biology*, syw079. doi : 10.1093/sysbio/syw079.

Koshi, Jeffrey M. et Richard A. Goldstein (juill. 1995). “Context-Dependent Optimal Substitution Matrices”. en. In : *Protein Engineering Design and Selection* 8.7, p. 641-645. doi : 10.1093/protein/8.7.641.

— (fév. 1996). “Probabilistic reconstruction of ancestral protein sequences”. en. In : *Journal of Molecular Evolution* 42.2, p. 313-320. doi : 10.1007/BF02198858.

La Touche, C.J. (jan. 1950). “On a thermophile species of Chaetomium”. en. In : *Transactions of the British Mycological Society* 33.1-2, 94-IN7. doi : 10.1016/S0007-1536(50)80051-7.

Lai, Jhih-Siang et al. (avr. 2020). “Evolutionary Model of Protein Secondary Structure Capable of Revealing New Biological Relationships”. en. In : *Proteins* 88, p. 9.

Laine, Elodie, Yasaman Karami et Alessandra Carbone (août 2019). “GEMME : A Simple and Fast Global Epistatic Model Predicting Mutational Effects”. en. In : *Molecular Biology and Evolution*, p. 16.

Lallier, François H. et Jean-Paul Truchot (avr. 1997). “Hemocyanin Oxygen-Binding Properties of a Deep-Sea Hydrothermal Vent Shrimp : Evidence for a Novel Cofactor”. en. In : *The Journal of Experimental Zoology* 277.5, p. 357-364. doi : 10.1002/(SICI)1097-010X(19970401)277:5<357::AID-JEZ1>3.0.CO;2-O.

Lartillot, Nicolas, Henner Brinkmann et Hervé Philippe (2007). “Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model”. en. In : *BMC Evolutionary Biology* 7.Suppl 1, S4. doi : 10.1186/1471-2148-7-S1-S4.

Lartillot, Nicolas et Hervé Philippe (juin 2004). “A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process”. en. In : *Molecular Biology and Evolution* 21.6, p. 1095-1109. doi : 10.1093/molbev/msh112.

Lazou, A. et al. (jan. 1987). “Purification, catalytic and regulatory properties of malate dehydrogenase from the foot of Patella caerulea (L.)” en. In : *Comparative Biochemistry and Physiology Part B : Comparative Biochemistry* 88.4, p. 1033-1040. doi : 10.1016/0305-0491(87)90002-2.

Le, S. Q., C. C. Dang et O. Gascuel (oct. 2012). “Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates”. en. In : *Molecular Biology and Evolution* 29.10, p. 2921-2936. doi : 10.1093/molbev/mss112.

Le, S. Q. et O. Gascuel (avr. 2008). “An Improved General Amino Acid Replacement Matrix”. en. In : *Molecular Biology and Evolution* 25.7, p. 1307-1320. doi : 10.1093/molbev/msn067.

Le, Si Quang et Olivier Gascuel (mars 2010). “Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial”. en. In : *Systematic Biology* 59.3, p. 277-287. doi : 10.1093/sysbio/syq002.

Le, Si Quang, Nicolas Lartillot et Olivier Gascuel (déc. 2008). “Phylogenetic Mixture Models for Proteins”. en. In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 363.1512, p. 3965-3976. doi : 10.1098/rstb.2008.0180.

Le Bris, Nadine et Françoise Gaill (jan. 2007). “How Does the Annelid *Alvinella pompejana* Deal with an Extreme Hydrothermal Environment ?” en. In : *Reviews in Environmental Science and Bio/Technology* 6.1-3, p. 197-221. doi : 10.1007/s11157-006-9112-1.

Le Bris, Nadine, Magali Zbinden et Françoise Gaill (juin 2005). “Processes Controlling the Physico-Chemical Micro-Environments Associated with Pompeii Worms”. en. In : *Deep Sea Research Part I : Oceanographic Research Papers* 52.6, p. 1071-1083. doi : 10.1016/j.dsr.2005.01.003.

Leibrock, E., P. Bayer et H.-D. Lüdemann (avr. 1995). “Nonenzymatic hydrolysis of adenosinetriphosphate (ATP) at high temperatures and high pressures”. en. In : *Biophysical Chemistry* 54.2, p. 175-180. doi : 10.1016/0301-4622(94)00134-6.

Lelièvre, Yann et al. (mai 2018). “Biodiversity and Trophic Ecology of Hydrothermal Vent FFauna Associated with Tubeworm Assemblages on the Juan de Fuca Ridge”. en. In : *Biogeosciences* 15.9, p. 2629-2647. doi : 10.5194/bg-15-2629-2018.

Lepage, Thomas et al. (juin 2007). “A General Comparison of Relaxed Molecular Clock Models”. en. In : *Molecular Biology and Evolution* 24.12, p. 2669-2680. doi : 10.1093/molbev/msm193.

Lepock, J R, H E Frey et R A Hallewell (déc. 1990). “Contribution of conformational stability and reversibility of unfolding to the increased thermostability of human and bovine superoxide dismutase mutated at free cysteines.” en. In : *Journal of Biological Chemistry* 265.35, p. 21612-21618. doi : 10.1016/S0021-9258(18)45784-5.

Liu, Liang, Lili Yu et Scott V. Edwards (2010). “A Maximum Pseudo-Likelihood Approach for Estimating Species Trees under the Coalescent Model”. en. In : *BMC Evolutionary Biology* 10.1, p. 302. doi : 10.1186/1471-2148-10-302.

Lonsdale, Peter (sept. 1977). “Clustering of suspension-feeding macrobenthos near abyssal hydrothermal vents at oceanic spreading centers”. en. In : *Deep Sea Research* 24.9, p. 857-863. doi : 10.1016/0146-6291(77)90478-7.

Löytynoja, Ari et Nick Goldman (juill. 2005). “An algorithm for progressive multiple alignment of sequences with insertions”. en. In : *Proceedings of the National Academy of Sciences* 102.30, p. 10557-10562. doi : 10.1073/pnas.0409137102.

— (déc. 2008a). “A model of evolution and structure for multiple sequence alignment”. en. In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 363.1512, p. 3913-3919. doi : 10.1098/rstb.2008.0170.

— (juin 2008b). “Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis”. en. In : *Science* 320.5883, p. 1632-1635. doi : 10.1126/science.1158395.

— (déc. 2010). “webPRANK : a phylogeny-aware multiple sequence aligner with interactive alignment browser”. en. In : *BMC Bioinformatics* 11.1, p. 579. doi : 10.1186/1471-2105-11-579.

Luan, Peng-tao et al. (mars 2013). “Incorporating indels as phylogenetic characters : Impact for interfamilial relationships within Arctoidea (Mammalia : Carnivora)”. en. In : *Molecular Phylogenetics and Evolution* 66.3, p. 748-756. doi : 10.1016/j.ympev.2012.10.023.

Mammerickx, Jacqueline, Ellen M. Herron et Leroy Dorman (1980). “Evidence for Two Fossil Spreading Ridges in the Southeast Pacific”. en. In : *Geological Society of America Bulletin* 91.5, p. 263. doi : 10.1130/0016-7606(1980)91<263:EFTFSR>2.0.CO;2.

Mann, Karlheinz et al. (août 1996). “Glycosylated Threonine but not 4-Hydroxyproline Dominates the Triple Helix Stabilizing Positions in the Sequence of a Hydrothermal Vent Worm Cuticle Collagen”. en. In : *Journal of Molecular Biology* 261.2, p. 255-266. doi : 10.1006/jmbi.1996.0457.

Margesin, R. et F. Schinner (1999). *Cold-Adapted Organisms - Ecology, Physiology, Enzymology and Molecular Biology*. Springer-Verlag.

Margulies, Marcel et al. (sept. 2005). “Genome Sequencing in Microfabricated High-Density Picolitre Reactors”. en. In : *Nature* 437.7057, p. 376-380. doi : 10.1038/nature03959.

Marie, Benjamin et al. (nov. 2006). “Effect of Ambient Oxygen Concentration on Activities of Enzymatic Antioxidant Defences and Aerobic Metabolism in the Hydrothermal Vent Worm, *Paralvinella grasslei*”. en. In : *Marine Biology* 150.2, p. 273-284. doi : 10.1007/s00227-006-0338-9.

Martin, Simon H et Steven M Van Belleghem (mai 2017). “Exploring Evolutionary Relationships Across the Genome Using Topology Weighting”. en. In : *Genetics* 206.1, p. 429-438. doi : 10.1534/genetics.116.194720.

Martineu, Pascale et al. (sept. 1997). “Sulfide Binding in the Body Fluids of Hydrothermal Vent Alvinellid Polychaetes”. en. In : *Physiological Zoology* 70.5, p. 578-588. doi : 10.1086/515864.

Mary, Jean et al. (avr. 2010). “Response of *Alvinella pompejana* to Variable Oxygen Stress : A Proteomic Approach”. en. In : *Proteomics* 10.12, p. 2250-2258. doi : 10.1002/pmic.200900394.

Mathias, Roy (juill. 1996). “A Chain Rule for Matrix Functions and Applications”. en. In : *SIAM Journal on Matrix Analysis and Applications* 17.3, p. 610-620. doi : 10.1137/S0895479895283409.

Matsumura, Masazumi, Shigeyoshi Yasumura et Shuichi Aiba (sept. 1986). “Cumulative effect of intragenic amino-acid replacements on the thermostability of a protein”. en. In : *Nature* 323.6086, p. 356-358. doi : 10.1038/323356a0.

McHugh, Damhnait (oct. 1989). “Population Structure and Reproductive Biology of Two Sympatric Hydrothermal Vent Polychaetes, *Paralvinella pandorae* and *P. palmiformis*”. en. In : *Marine Biology* 103.1, p. 95-106. doi : 10.1007/BF00391068.

Merkl, Rainer et Reinhard Sterner (jan. 2016). “Ancestral Protein Reconstruction : Techniques and Applications”. en. In : *Biological Chemistry* 397.1, p. 1-21. doi : 10.1515/hsz-2015-0158.

Metpally, Raghu et Boojala Reddy (2009). “Comparative proteome analysis of psychrophilic versus mesophilic bacterial species : Insights into the molecular basis of cold adaptation of proteins”. en. In : *BMC Genomics* 10.1, p. 11. doi : 10.1186/1471-2164-10-11.

Miklos, I. (déc. 2003). “A "Long Indel" Model For Evolutionary Sequence Alignment”. en. In : *Molecular Biology and Evolution* 21.3, p. 529-540. doi : 10.1093/molbev/msh043.

Minárik, P et al. (2002). “Malate Dehydrogenases – Structure and Function”. en. In : *General Physiology and Biophysics* 21, p. 257-265.

Minh, Bui Quang et al. (mai 2020). “IQ-TREE 2 : New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. en. In : *Molecular Biology and Evolution* 37.5. Sous la dir. d’Emma Teeling, p. 1530-1534. doi : 10.1093/molbev/msaa015.

Mirceta, Scott et al. (juin 2013). “Evolution of Mammalian Diving Capacity Traced by Myoglobin Net Surface Charge”. en. In : *Science* 340.6138, p. 1234192. doi : 10.1126/science.1234192.

Moalic, Yann et al. (jan. 2012). “Biogeography Revisited with Network Theory : Retracing the History of Hydrothermal Vent Communities”. en. In : *Systematic Biology* 61.1, p. 127. doi : 10.1093/sysbio/syr088.

Molloy, Erin K. et Tandy Warnow (mars 2018). “To Include or Not to Include : The Impact of Gene Filtering on Species Tree Estimation Methods”. en. In : *Systematic Biology* 67.2, p. 285-303. doi : 10.1093/sysbio/syx077.

Monaco, André et Patrick Prouzet (oct. 2015). “Hydrothermal Vents : Oases at Depth”. en. In : *Marine Ecosystems*. Sous la dir. d’André Monaco et Patrick Prouzet. Hoboken, NJ, USA : John Wiley & Sons, Inc., p. 225-292. doi : 10.1002/9781119116219.ch6.

Moshe, Asher et Tal Pupko (août 2019). “Ancestral Sequence Reconstruction : Accounting for Structural Information by Averaging over Replacement Matrices”. en. In : *Bioinformatics* 35.15. Sous la dir. de Russell Schwartz, p. 2562-2568. doi : 10.1093/bioinformatics/bty1031.

Müller, Kai (mars 2006). “Incorporating information from length-mutational events into phylogenetic analysis”. en. In : *Molecular Phylogenetics and Evolution* 38.3, p. 667-676. doi : 10.1016/j.ympev.2005.07.011.

Nakamura, Kentaro et al. (mars 2012). “Discovery of New Hydrothermal Activity and Chemosynthetic Fauna on the Central Indian Ridge at 18°–20°S”. en. In : *PLoS ONE* 7.3. Sous la dir. de Joel M. Schnur, e32965. doi : 10.1371/journal.pone.0032965.

Naser-Khdour, Suha et al. (déc. 2019). “The Prevalence and Impact of Model Violations in Phylogenetic Analysis”. en. In : *Genome Biology and Evolution* 11.12. Sous la dir. de David Bryant, p. 3341-3352. doi : 10.1093/gbe/evz193.

Nicholson, H., W. J. Becktel et B. W. Matthews (déc. 1988). “Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles”. en. In : *Nature* 336.6200, p. 651-656. doi : 10.1038/336651a0.

Noort, Vera van et al. (déc. 2013). “Consistent Mutational Paths Predict Eukaryotic Thermostability”. en. In : *BMC Evolutionary Biology* 13.1, p. 7. doi : 10.1186/1471-2148-13-7.

Nute, Michael et al. (mai 2018). “The Performance of Coalescent-Based Species Tree Estimation Methods under Models of Missing Data”. en. In : *BMC Genomics* 19.S5, p. 286. doi : 10.1186/s12864-018-4619-8.

Pack, Seung Pil et Young Je Yoo (août 2004). “Protein thermostability : structure-based difference of amino acid between thermophilic and mesophilic proteins”. en. In : *Journal of Biotechnology* 111.3, p. 269-277. doi : 10.1016/j.jbiotec.2004.01.018.

Paoletti, Francesco et al. (mai 1986). “A sensitive spectrophotometric method for the determination of superoxide dismutase activity in tissue extracts”. en. In : *Analytical Biochemistry* 154.2, p. 536-541. doi : 10.1016/0003-2697(86)90026-6.

Parker, E.S. et William K. Gealey (mars 1985). “Plate Tectonic Evolution of the Western Pacific-Indian Ocean Region”. en. In : *Energy* 10.3-4, p. 249-261. doi : 10.1016/0360-5442(85)90045-3.

Paśko, Łukasz, Per G.P. Ericson et Andrzej Elzanowski (déc. 2011). “Phylogenetic utility and evolution of indels : A study in neognathous birds”. en. In : *Molecular Phylogenetics and Evolution* 61.3, p. 760-771. doi : 10.1016/j.ympev.2011.07.021.

Pauling, Linus et al. (1963). “Chemical Paleogenetics. Molecular "Restoration Studies" of Extinct Forms of Life.” en. In : *Acta Chemica Scandinavica* 17 supl. P. 9-16. doi : 10.3891/acta.chem.scand.17s-0009.

Pease, James B. et al. (mars 2018). “Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life”. en. In : *American Journal of Botany* 105.3, p. 385-403. doi : 10.1002/ajb2.1016.

Perez, Maeva et al. (août 2023). “Third-Generation Sequencing Reveals the Adaptive Role of the Epigenome in Three Deep-Sea Polychaetes”. en. In : *Molecular Biology and Evolution* 40.8. Sous la dir. de Sophie Von Der Heyden, msad172. doi : 10.1093/molbev/msad172.

Petersen, Bent et al. (déc. 2009). “A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions”. en. In : *BMC Structural Biology* 9.1, p. 51. doi : 10.1186/1472-6807-9-51.

Petersen, Malte et al. (déc. 2017). “Orthograph : a Versatile Tool for Mapping Coding Nucleotide Sequences to Clusters of Orthologous Genes”. en. In : *BMC Bioinformatics* 18.1, p. 111. doi : 10.1186/s12859-017-1529-8.

Piccino, Patrice et al. (nov. 2004). “Thermal Selection of PGM Allozymes in Newly Founded Populations of the Thermotolerant Vent Polychaete *Alvinella pompejana*”. en. In : *Proceedings of the Royal Society of London. Series B : Biological Sciences* 271.1555, p. 2351-2359. doi : 10.1098/rspb.2004.2852.

Poelwijk, Frank J., Michael Socolich et Rama Ranganathan (sept. 2019). “Learning the pattern of epistasis linking genotype and phenotype in a protein”. en. In : *Nature Communications* 10.1, p. 4213. doi : 10.1038/s41467-019-12130-8.

Pollock, David D., Grant Thiltgen et Richard A. Goldstein (mai 2012). “Amino acid coevolution induces an evolutionary Stokes shift”. en. In : *Proceedings of the National Academy of Sciences* 109.21. doi : 10.1073/pnas.1120084109.

Portail, Marie et al. (sept. 2016). “Food-Web Complexity in Guaymas Basin Hydrothermal Vents and Cold Seeps”. en. In : *PLOS ONE* 11.9. Sous la dir. d’Elena Gorokhova, e0162263. doi : 10.1371/journal.pone.0162263.

Pörtner, Hans-Otto (août 2002). “Climate Variations and the Physiological Basis of Temperature-Dependent Biogeography : Systemic to Molecular Hierarchy of Thermal Tolerance in Animals”. en. In : *Comparative Biochemistry and Physiology Part A : Molecular & Integrative Physiology* 132.4, p. 739-761. doi : 10.1016/S1095-6433(02)00045-4.

Pörtner, Hans-Otto et al. (juin 1999). “Intracellular pH and energy metabolism in the highly stenothermal Antarctic bivalve Limopsis marionensis as a function of ambient temperature”. In : *Polar Biology* 22.1, p. 17-30. doi : 10.1007/s003000050386.

Pradillon, Florence et al. (avr. 2005). “Influence of Environmental Conditions on Early Development of the Hydrothermal Vent Polychaete *Alvinella pompejana*”. en. In : *Journal of Experimental Biology* 208.8, p. 1551-1561. doi : 10.1242/jeb.01567.

Pucci, Fabrizio, Jean Marc Kwasigroch et Marianne Rooman (nov. 2017). “SCooP : an accurate and fast predictor of protein stability curves as a function of temperature”. en. In : *Bioinformatics* 33.21. Sous la dir. d’Alfonso Valencia, p. 3415-3422. doi : 10.1093/bioinformatics/btx417.

Pupko, Tal et al. (juin 2000). “A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences”. en. In : *Molecular Biology and Evolution* 17.6, p. 890-896. doi : 10.1093/oxfordjournals.molbev.a026369.

Randall, Ryan N. et al. (nov. 2016). “An Experimental Phylogeny to Benchmark Ancestral Sequence Reconstruction”. en. In : *Nature Communications* 7.1, p. 12847. doi : 10.1038/ncomms12847.

Ranwez, Vincent et al. (sept. 2011). “MACSE : Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons”. en. In : *PLoS ONE* 6.9. Sous la dir. de William J. Murphy, e22594. doi : 10.1371/journal.pone.0022594.

Ravaux, Juliette et al. (mai 2013). “Thermal Limit for Metazoan Life in Question : *In Vivo* Heat Tolerance of the Pompeii Worm”. en. In : *PLoS ONE* 8.5. Sous la dir. de Nikolas Nikolaidis, e64074. doi : 10.1371/journal.pone.0064074.

Razban, Rostam M (sept. 2019). “Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance–Evolutionary Rate Correlation Seen in Proteins”. en. In : *Molecular Biology and Evolution* 36.9. Sous la dir. de Jianzhi Zhang, p. 1955-1963. doi : 10.1093/molbev/msz119.

Razvi, Abbas et J. Martin Scholtz (juill. 2006). “Lessons in stability from thermophilic proteins”. en. In : *Protein Science* 15.7, p. 1569-1578. doi : 10.1110/ps.062130306.

Redelings, Benjamin D et Marc A Suchard (2007). “Incorporating indel information into phylogeny estimation for rapidly emerging pathogens”. en. In : *BMC Evolutionary Biology* 7.1, p. 40. doi : 10.1186/1471-2148-7-40.

Rhee, Jae-Sung et al. (2011). “Expression of superoxide dismutase (SOD) genes from the copper-exposed polychaete, Neanthes succinea”. en. In : *Marine Pollution Bulletin* 63.5-12, p. 277-286. doi : 10.1016/j.marpolbul.2011.04.023.

Rinke, Christian et Raymond W. Lee (avr. 2009). “Pathways, Activities and Thermal Stability of Anaerobic and Aerobic Enzymes in Thermophilic Vent Paralvinellid Worms”. en. In : *Marine Ecology Progress Series* 382, p. 99-112. doi : 10.3354/meps07980.

Rivas, Elena (2005). “Evolutionary models for insertions and deletions in a probabilistic modeling framework”. en. In : *BMC Bioinformatics* 6.1, p. 63. doi : 10.1186/1471-2105-6-63.

Rivas, Elena et Sean R. Eddy (déc. 2015). “Parameterizing sequence alignment with an explicit evolutionary model”. en. In : *BMC Bioinformatics* 16.1, p. 406. doi : 10.1186/s12859-015-0832-5.

Roch, Sebastien et Tandy Warnow (juill. 2015). “On the Robustness to Gene Tree Estimation Error (or Lack thereof) of Coalescent-Based Species Tree Methods”. en. In : *Systematic Biology* 64.4, p. 663-676. doi : 10.1093/sysbio/syv016.

Rodriguez, Jorge A. et al. (mai 2002). “Familial Amyotrophic Lateral Sclerosis-associated Mutations Decrease the Thermal Stability of Distinctly Metallated Species of Human Copper/Zinc Superoxide Dismutase”. en. In : *Journal of Biological Chemistry* 277.18, p. 15932-15937. doi : 10.1074/jbc.M112088200.

Röhl, Ursula et James G. Ogg (oct. 1996). “Aptian-Albian Sea Level History from Guyots in the Western Pacific”. en. In : *Paleoceanography* 11.5, p. 595-624. doi : 10.1029/96PA01928.

Romero-Romero, M. Luisa et al. (oct. 2016). “Engineering ancestral protein hyperstability”. en. In : *Biochemical Journal* 473.20, p. 3611-3620. doi : 10.1042/BCJ20160532.

Rousset, Vincent, Fredrik Pleijel et al. (fév. 2007). “A Molecular Phylogeny of Annelids”. en. In : *Cladistics* 23.1, p. 41-63. doi : 10.1111/j.1096-0031.2006.00128.x.

Rousset, Vincent, Greg W. Rouse et al. (mars 2003). “Molecular and Morphological Evidence of Alvinellidae Relationships (Terebelliformia, Polychaeta, Annelida)”. en. In : *Zoologica Scripta* 32.2, p. 185-197. doi : 10.1046/j.1463-6409.2003.00110.x.

Russell, N. J. (avr. 2000). “Toward a molecular understanding of cold activity of enzymes from psychrophiles”. In : *Extremophiles* 4.2, p. 83-90. doi : 10.1007/s007920050141.

Sanders, Nancy K., Alissa J. Arp et James J. Childress (jan. 1988). “Oxygen Binding Characteristics of the Hemocyanins of two Deep-Sea Hydrothermal Vent Crustaceans”. en. In : *Respiration Physiology* 71.1, p. 57-67. doi : 10.1016/0034-5687(88)90115-6.

Saunders, Neil F.W. et al. (juill. 2003). “Mechanisms of Thermal Adaptation Revealed From the Genomes of the Antarctic *Archaea Methanogenium frigidum* and *Methanococcoides burtonii*”. en. In : *Genome Research* 13.7, p. 1580-1588. doi : 10.1101/gr.1180903.

Savin, Samuel M. (mai 1977). “The History of the Earth’s Surface Temperature During the Past 100 Million Years”. en. In : *Annual Review of Earth and Planetary Sciences* 5.1, p. 319-355. doi : 10.1146/annurev.ea.05.050177.001535.

Scandurra, Roberto et al. (nov. 1998). “Protein thermostability in extremophiles”. en. In : *Biochimie* 80.11, p. 933-941. doi : 10.1016/S0300-9084(00)88890-2.

Schellart, W.P., G.S. Lister et V.G. Toy (juin 2006). “A Late Cretaceous and Cenozoic reconstruction of the Southwest Pacific region : Tectonics controlled by subduction and slab rollback processes”. en. In : *Earth-Science Reviews* 76.3-4, p. 191-233. doi : 10.1016/j.earscirev.2006.01.002.

Schymkowitz, J. et al. (juill. 2005). “The FoldX web server : an online force field”. en. In : *Nucleic Acids Research* 33.Web Server, W382-W388. doi : 10.1093/nar/gki387.

Sepkoski Jr., J. John (2002). “A Compendium of Fossil Marine Animal Genera”. In : *Bulletins of American Paleontology*, p. 1-560.

Serohijos, Adrian WR et Eugene I Shakhnovich (juin 2014). “Merging molecular mechanism and evolution : theory and computation at the interface of biophysics and evolutionary population genetics”. en. In : *Current Opinion in Structural Biology* 26, p. 84-91. doi : 10.1016/j.sbi.2014.05.005.

Seton, M. et al. (juill. 2012). “Global continental and ocean basin reconstructions since 200Ma”. en. In : *Earth-Science Reviews* 113.3-4, p. 212-270. doi : 10.1016/j.earscirev.2012.03.002.

Seton, Maria et al. (oct. 2020). “A Global Data Set of Present-Day Oceanic Crustal Age and Seafloor Spreading Parameters”. en. In : *Geochemistry, Geophysics, Geosystems* 21.10, e2020GC009214. doi : 10.1029/2020GC009214.

Shimodaira, Hidetoshi (mai 2002). “An Approximately Unbiased Test of Phylogenetic Tree Selection”. en. In : *Systematic Biology* 51.3. Sous la dir. de Nick Goldman, p. 492-508. doi : 10.1080/10635150290069913.

Shin, David S., Michael DiDonato, David P. Barondeau, Greg L. Hura, J. Andrew Berglund et al. (2010). “Superoxide Dismutase Structures, Stability, Mechanism and Insights into the Human Disease Amyotrophic Lateral Sclerosis from Eukaryotic Thermophile *Alvinella pompejana*”. en. In : p. 39.

Shin, David S., Michael DiDonato, David P. Barondeau, Greg L. Hura, Chiharu Hitomi et al. (fév. 2009). “Superoxide Dismutase from the Eukaryotic Thermophile *Alvinella pompejana* : Structures, Stability, Mechanism, and Insights into Amyotrophic Lateral Sclerosis”. en. In : *Journal of Molecular Biology* 385.5, p. 1534-1555. doi : 10.1016/j.jmb.2008.11.031.

Si Quang, Le, Olivier Gascuel et Nicolas Lartillot (oct. 2008). “Empirical Profile Mixture Models for Phylogenetic Reconstruction”. en. In : *Bioinformatics* 24.20, p. 2317-2323. doi : 10.1093/bioinformatics/btn445.

Sicot, Francois-Xavier et al. (sept. 2000). “Molecular adaptation to an extreme environment : origin of the thermal stability of the pompeii worm collagen”. en. In : *Journal of Molecular Biology* 302.4, p. 811-820. doi : 10.1006/jmbi.2000.4505.

Siddiqui, Khawar Sohail et Ricardo Cavicchioli (juin 2006). “Cold-Adapted Enzymes”. en. In : *Annual Review of Biochemistry* 75.1, p. 403-433. doi : 10.1146/annurev.biochem.75.103004.142723.

Siebenaller, Joseph et George N. Somero (juill. 1978). “Pressure-Adaptive Differences in Lactate Dehydrogenases of Congeneric Fishes Living at Different Depths”. en. In : *Science* 201.4352, p. 255-257. doi : 10.1126/science.208149.

Sikosek, Tobias et Hue Sun Chan (nov. 2014). “Biophysics of protein evolution and evolutionary protein biophysics”. en. In : *Journal of The Royal Society Interface* 11.100, p. 20140419. doi : 10.1098/rsif.2014.0419.

Siliakus, Melvin F., John van der Oost et Servé W. M. Kengen (juill. 2017). “Adaptations of Archaeal and Bacterial Membranes to Variations in Temperature, pH and Pressure”. en. In : *Extremophiles* 21.4, p. 651-670. doi : 10.1007/s00792-017-0939-x.

Simmons, Mark P., Kai Müller et Andrew P. Norton (août 2007). “The relative performance of indel-coding methods in simulations”. en. In : *Molecular Phylogenetics and Evolution* 44.2, p. 724-740. doi : 10.1016/j.ympev.2007.04.001.

Simmons, Mark P. et Helga Ochoterena (2000). “Gaps as Characters in Sequence-Based Phylogenetic Analyses”. en. In : *Systematic Biology*, p. 13.

Smith, Alan D. (avr. 2003). “A Reappraisal of Stress Field and Convective Roll Models for the Origin and Distribution of Cretaceous to Recent Intraplate Volcanism in the Pacific Basin”. en. In : *International Geology Review* 45.4, p. 287-302. doi : 10.2747/0020-6814.45.4.287.

Stiller, Josefin, Vincent Rousset et al. (mars 2013). “Phylogeny, Biogeography and Systematics of Hydrothermal Vent and Methane Seep *Amphisamytha* (Ampharetidae, Annelida), with Descriptions of Three New Species”. en. In : *Systematics and Biodiversity* 11.1, p. 35-65. doi : 10.1080/14772000.2013.772925.

Stiller, Josefin, Ekin Tilic et al. (avr. 2020). “Spaghetti to a Tree : A Robust Phylogeny for Terebelliformia (Annelida) Based on Transcriptomes, Molecular and Morphological Data”. en. In : *Biology* 9.4, p. 73. doi : 10.3390/biology9040073.

Suhre, Karsten et Jean-Michel Claverie (mai 2003). “Genomic Correlates of Hyperthermostability, an Update”. en. In : *Journal of Biological Chemistry* 278.19, p. 17198-17202. doi : 10.1074/jbc.M301327200.

Sumanaweera, Dinithi, Lloyd Allison et Arun S Konagurthu (juin 2022). “Bridging the gaps in statistical models of protein alignment”. en. In : *Bioinformatics* 38.Supplement_1, p. i229-i237. doi : 10.1093/bioinformatics/btac246.

Szilágyi, András et Péter Závodszky (mai 2000). “Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits : results of a comprehensive survey”. en. In : *Structure* 8.5, p. 493-504. doi : 10.1016/S0969-2126(00)00133-7.

Takai, Ken et al. (août 2008). “Cell proliferation at 122°C and isotopically heavy CH_4 production by a hyperthermophilic methanogen under high-pressure cultivation”. en. In : *Proceedings of the National Academy of Sciences* 105.31, p. 10949-10954. doi : 10.1073/pnas.0712334105.

Taverna, Darin M. et Richard A. Goldstein (jan. 2002). “Why are proteins marginally stable ?” en. In : *Proteins : Structure, Function, and Bioinformatics* 46.1, p. 105-109. doi : 10.1002/prot.10016.

Taylor, Todd J et Iosif I Vaisman (2010). “Discrimination of thermophilic and mesophilic proteins”. en. In.

Thomas, Ellen (2007). “Cenozoic mass extinctions in the deep sea : What perturbs the largest habitat on Earth ?” en. In : *Large Ecosystem Perturbations : Causes and Consequences*. Geological Society of America. doi : 10.1130/2007.2424(01).

Thomas-Bulle, Camille et al. (sept. 2022). “Genomic patterns of divergence in the early and late steps of speciation of the deep-sea vent thermophilic worms of the genus Alvinella”. en. In : *BMC Ecology and Evolution* 22.1, p. 106. doi : 10.1186/s12862-022-02057-y.

Thompson, Michael J. et David Eisenberg (juill. 1999). “Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability”. en. In : *Journal of Molecular Biology* 290.2, p. 595-604. doi : 10.1006/jmbi.1999.2889.

Thorne, Jeffrey L., Hirohisa Kishino et Ian S. Painter (déc. 1998). “Estimating the Rate of Evolution of the Rate of Molecular Evolution”. en. In : *Molecular Biology and Evolution* 15.12, p. 1647-1657. doi : 10.1093/oxfordjournals.molbev.a025892.

Tokuriki, Nobuhiko et al. (fév. 2008). “How Protein Stability and New Functions Trade Off”. en. In : *PLoS Computational Biology* 4.2. Sous la dir. de David Eisenberg, e1000002. doi : 10.1371/journal.pcbi.1000002.

Toulmond, André et al. (déc. 1990). “Extracellular Hemoglobins of Hydrothermal Vent Annelids : Structural and Functional Characteristics in Three Alvinellid Species”. en. In : *The Biological Bulletin* 179.3, p. 366-373. doi : 10.2307/1542329.

Trudeau, Devin L., Miriam Kaltenbach et Dan S. Tawfik (oct. 2016). “On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins”. en. In : *Molecular Biology and Evolution* 33.10, p. 2633-2641. doi : 10.1093/molbev/msw138.

Tunnicliffe, Verena (avr. 1988). “Biogeography and Evolution of Hydrothermal-Vent Fauna in the Eastern Pacific Ocean”. en. In : *Proceedings of the Royal Society of London. Series B. Biological Sciences* 233.1272, p. 347-366. doi : 10.1098/rspb.1988.0025.

Tunnicliffe, Verena, Daniel Desbruyères et al. (fév. 1993). “Systematic and Ecological Characteristics of *Paralvinella sulfincola* Desbruyères and Laubier, a New Polychaete (Family Alvinellidae) from Northeast Pacific Hydrothermal Vents”. en. In : *Canadian Journal of Zoology* 71.2, p. 286-297. doi : 10.1139/z93-041.

Tunnicliffe, Verena, S. Kim Juniper et M. Sibuet (2003). “Reducing environments of the deep-sea floor”. In : *Ecosystems of the deep oceans*. Elsevier, p. 81-110.

Valenti, Anna et al. (fév. 2011). “Positive Supercoiling in Thermophiles and Mesophiles : of the Good and Evil”. en. In : *Biochemical Society Transactions* 39.1, p. 58-63. doi : 10.1042/BST0390058.

Vigué, Lucile et al. (juill. 2022). “Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes”. en. In : *Nature Communications* 13.1, p. 4030. doi : 10.1038/s41467-022-31643-3.

Vihinen, Mauno (1987). “Relationship of protein flexibility to thermostability”. en. In : *"Protein Engineering, Design and Selection"* 1.6, p. 477-480. doi : 10.1093/protein/1.6.477.

Vovelle, Jean et Francoise Gaill (jan. 1986). “Donnees Morphologiques, Histochimiques et Microanalytiques sur l’Elaboration du Tube Organomineral d’*Alvinella pompejana*, Polychete des Sources Hydrothermales, et leurs Implications Phylogenetiques”. en. In : *Zoologica Scripta* 15.1, p. 33-43. doi : 10.1111/j.1463-6409.1986.tb00206.x.

Vrijenhoek, Robert C. (août 2013). “On the Instability and Evolutionary Age of Deep-Sea Chemosynthetic Communities”. en. In : *Deep Sea Research Part II : Topical Studies in Oceanography* 92, p. 189-200. doi : 10.1016/j.dsr2.2012.12.004.

Wang, Guang-Zhong et Martin J. Lercher (2010). “Amino Acid Composition in Endothermic Vertebrates is Biased in the Same Direction as in Thermophilic Prokaryotes”. en. In : *BMC Evolutionary Biology* 10.1, p. 263. doi : 10.1186/1471-2148-10-263.

Wang, Huai-Chun et al. (déc. 2008). “A Class Frequency Mixture Model that Adjusts for Site-Specific Amino Acid Frequencies and Improves Inference of Protein Phylogeny”. en. In : *BMC Evolutionary Biology* 8.1, p. 331. doi : 10.1186/1471-2148-8-331.

Westesson, Oscar et al. (avr. 2012). “Accurate Reconstruction of Insertion-Deletion Histories by Statistical Phylogenetics”. en. In : *PLoS ONE* 7.4. Sous la dir. d’Art F. Y. Poon, e34572. doi : 10.1371/journal.pone.0034572.

Wheeler, Lucas C et al. (juin 2016). “The thermostability and specificity of ancient proteins”. en. In : *Current Opinion in Structural Biology* 38, p. 37-43. doi : 10.1016/j.sbi.2016.05.015.

Whelan, Simon et Nick Goldman (mai 2001). “A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach”. en. In : *Molecular Biology and Evolution* 18.5, p. 691-699. doi : 10.1093/oxfordjournals.molbev.a003851.

Wickstrom, Conrad E. et Richard W. Castenholz (sept. 1973). “Thermophilic Ostracod : Aquatic Metazoan with the Highest Known Temperature Tolerance”. en. In : *Science* 181.4104, p. 1063-1064. doi : 10.1126/science.181.4104.1063.

Williams, Paul D et al. (juin 2006). “Assessing the Accuracy of Ancestral Protein Reconstruction Methods”. en. In : *PLoS Computational Biology* 2.6. Sous la dir. de David Hillis, e69. doi : 10.1371/journal.pcbi.0020069.

Wood, Derrick E. et Steven L. Salzberg (2014). “Kraken : Ultrafast Metagenomic Sequence Classification Using Exact Alignments”. en. In : *Genome Biology* 15.3, R46. doi : 10.1186/gb-2014-15-3-r46.

Wu, Chung-I (nov. 2001). “The genic view of the process of speciation”. en. In : *Journal of Evolutionary Biology* 14.6, p. 851-865. doi : 10.1046/j.1420-9101.2001.00335.x.

Xia, Xichao et al. (avr. 2016). “Molecular cloning, characterization, and the response of Cu/Zn superoxide dismutase and catalase to PBDE-47 and -209 from the freshwater bivalve Anodonta woodiana”. en. In : *Fish & Shellfish Immunology* 51, p. 200-210. doi : 10.1016/j.fsi.2016.02.025.

Yadav, Sushma et Faizan Ahmad (août 2000). “A New Method for the Determination of Stability Parameters of Proteins from Their Heat-Induced Denaturation Curves”. en. In : *Analytical Biochemistry* 283.2, p. 207-213. doi : 10.1006/abio.2000.4641.

Yang, Ziheng (sept. 1994). “Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites : Approximate Methods”. en. In : *Journal of Molecular Evolution* 39.3, p. 306-314. doi : 10.1007/BF00160154.

— (avr. 2007). “PAML 4 : Phylogenetic Analysis by Maximum Likelihood”. en. In : *Molecular Biology and Evolution* 24.8, p. 1586-1591. doi : 10.1093/molbev/msm088.

Zal, Franck et al. (1995). “Reproductive Biology and Population Structure of the Deep-Sea Hydrothermal Vent Worm *Paralvinella grasslei* (Polychaete : Alvinellidae) at 13°N on the East Pacific Rise”. In : *Marine Biology* 122, p. 637-648.

Zdobnov, Evgeny M. et al. (jan. 2021). “OrthoDB in 2020 : Evolutionary and Functional Annotations of Orthologs”. en. In : *Nucleic Acids Research* 49.D1, p. D389-D393. doi : 10.1093/nar/gkaa1009.

Zeinali, Farrokhzad, Ahmad Homaei et Ehsan Kamrani (août 2015). “Sources of marine superoxide dismutases : Characteristics and applications”. en. In : *International Journal of Biological Macromolecules* 79, p. 627-637. doi : 10.1016/j.ijbiomac.2015.05.053.

Zeldovich, Konstantin B., Igor N. Berezovsky et Eugene I. Shakhnovich (jan. 2007). “Protein and DNA Sequence Determinants of Thermophilic Adaptation”. en. In : *PLoS Computational Biology* 3.1. Sous la dir. de Philip E. Bourne, e5. doi : 10.1371/journal.pcbi.0030005.

Zhang, Chao et al. (mai 2018). “ASTRAL-III : Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees”. en. In : *BMC Bioinformatics* 19.S6, p. 153. doi : 10.1186/s12859-018-2129-y.

Zhong, Min et al. (déc. 2011). “Detecting the Symplesiomorphy Trap : a Multigene Phylogenetic Analysis of Terebelliform Annelids”. en. In : *BMC Evolutionary Biology* 11.1, p. 369. doi : 10.1186/1471-2148-11-369.

Zukiene, Rasa et al. (jan. 2010). “Acute Temperature Resistance Threshold in Heart Mitochondria : Febrile Temperature Activates Function but Exceeding it Collapses the Membrane Barrier”. en. In : *International Journal of Hyperthermia* 26.1, p. 56-66. doi : 10.3109/02656730903262140.

Titre : Evolution de la thermophilie au sein de la lignée des annélides polychaetes Alvinellidae.

Mots clefs : thermophilie ; sources hydrothermales ; Alvinellidae ; Evolution moléculaire ; ASR ; Biologie structurelle

Résumé : Les Alvinellidae (annélides polychètes terebellides) constituent une famille d'espèces endémiques des sources hydrothermales profondes, dispersées entre l'océan Pacifique et Indien. Depuis leur découverte avec l'espèce emblématique *Alvinella pompejana*, le ver de Pompéi, ces animaux ont suscité l'intérêt de la communauté scientifique. En effet, si les sources hydrothermales constituent des environnements réputés extrêmes (gradients de température, absence de photosynthèse, anoxie du milieu, présence de divers métaux et sulfides issus de la percolation du fluide hydrothermale dans la croûte basaltique, pH acide), les Alvinellidae sont parvenus à coloniser des niches écologiques variées et montrent une grande diversité morphologique, physiologique et génétique, inter et intra-espèces. Dans le cadre de cette thèse, nous nous sommes plus particulièrement intéressés aux adaptations permettant à ces vers de faire face à des régimes thermiques contrastés. *A. pompejana*, par exemple, est thermophile, survivant à des températures proches de 50°C. D'autres espèces en revanche, comme *Paralvinella grasslei*, sont considérées psychrophiles, vivant à distance des cheminées hydrothermales à des températures entre 10 et 25°C. Plus spécifiquement, nous avons étudié l'acquisition de la thermophilie/psychrophilie au cours de l'évolution de la lignée, en essayant de répondre à la question du phénotype thermique de l'ancêtre des Alvinellidae. Pour cela, nous avons établi la phylogénie moléculaire des Alvinellidae, sur la base des données moléculaires transcriptomiques récupérées pour onze des quatorze espèces de la famille au cours de plusieurs campagnes scientifiques. Ce premier résultat amène à conclure à un ancêtre datant de la fin du Crétacée (entre 60 et 90 millions d'années), déjà présent dans les sources hydrothermales du Pacifique Est. La radiation des Alvinellidae à cette époque a été rapide, en quelques millions d'années, aboutissant à l'apparition de plusieurs espèces présentant de forts taux de tri incomplet de lignée et d'introgression interspécifique. Les résultats de cette phylogénie nous ont permis d'établir le modèle permettant de construire des propositions statistiques de protéines appartenant aux ancêtres de la lignée. Trois protéines ont été choisies, à savoir la malate déshydrogénase cytosolique, la superoxyde dismutase Cu/Zn et une hémoglobine intracellulaire, pour être reconstruites, exprimées et expérimentalement caractérisées. En effet, pour des organismes ectothermes comme les Alvinellidae, il est attendu que les protéines des espèces thermophiles soient en moyenne plus stables thermiquement que les protéines issues des espèces psychrophiles. Ces reconstructions ancestrales nous ont permis de conclure que l'ancêtre de la lignée était un ver déjà adapté aux environnements chauds, et que la psychrophilie de certaines espèces de la lignée est un caractère dérivé acquis plus récemment. Enfin, dans une dernière partie, je me suis intéressé à l'optimisation des modèles de reconstruction des séquences protéiques ancestrales. Ces modèles sont basés sur la diversité des séquences contemporaines et leurs relations phylogénétiques. J'ai essayé d'implémenter ces approches en utilisant deux types d'informations supplémentaires : celles liées aux événements d'insertions/délétions de séquence, et celles concernant l'évolution de la structure secondaire des protéines et la variabilité temporelle des fréquences attendues des résidus aux différentes positions des protéines. Je montre que l'introduction de ces deux derniers types de paramètres dans les méthodes ASR est bénéfique et aboutit à des modèles ayant de meilleures vraisemblances. Toutefois, l'optimisation de ces modèles, nécessairement probabilistes, ne garantit pas un meilleur résultat pour l'expérimentateur, et les limites de ces modèles à estimer l'incertitude des séquences ancestrales inférées sont discutées.