La structure des protéines et son évolution

Mathilde Carpentier

Protein structure and evolution

Habilitation à diriger des recherche

présentée et soutenue publiquement le 25 avril 2023

par

Mathilde Carpentier

Composition du jury

Rapporteurs :	Dr. Juliette Martin Dr Julian Echave Dr. Alexandre de Brevern
Examinateurs :	Pr. Ingrid Lafontaine Pr Guillaume Lecointre Dr. Frederic Cazals

Institut de Systématique, Évolution, Biodiversité (ISYEB) — UMR 7205

Mis en page avec la classe thesul.

Remerciements

La recherche n'est pas une activité facile ou confortable. L'émerveillement est fréquent, devant la richesse et la complexité du monde, devant la simplicité, la clarté et la beauté d'une idée, devant le plaisir et la joie de voir le monde sous un angle nouveau et les possibilités que cela ouvre. Et pourtant, les moments de découragement sont très très réguliers. Trop vaste. Trop difficile. Est-ce pertinent ? Cela n'a t il pas déjà été fait 1000 fois ? Quelle finalité à tout cela ? Heureusement, entre ces moments, il y a les gens, beaucoup de gens bien avec qui échanger, discuter, réfléchir, travailler. L'ABI (Atelier de Bioinformatique) est mon havre, mon refuge, et je ne saurai jamais remercier assez ceux qui m'y ont accueillie : d'abord Joël, mon directeur de thèse avec qui j'appris la recherche, et avec qui j'ai continué la recherche :-), mais aussi tous les membres, présents ou passés de de l'ABI, en particulier Sophie Brouillet, Martine Boccara, Henry Soldano, Eric Duchaud, Guillaume Achaz, Lucie Bittner, Dominique Higuet, Guillaume Sapriel, Sophie Pasek. Je remercie aussi l'ISYEB et tous ses membres de nous avoir si bien accueilli au Muséum d'Histoire Naturelle.

Je remercie aussi toutes les personnes avec qui j'ai collaborées ces dernières années, sans lesquelles rien de tout cela n'aurait été possible. Je remercie en particulier Jacques, mon camarade pour de nombreux articles. Nos rendez-vous hebdomadaires me manqueront, autant pour la science que pour les autres discussions. Je remercie aussi Guillaume L. avec qui cela a été un vrai plaisir, scientifique et humain, de travailler. Et aussi bien sûr Martin qui a stoïquement supporté les tâtonnements de mon premier (co)encadrement de thèse. Je remercie aussi tous les étudiants qui m'ont offert leurs temps lors de ces stages et qui ont accepté de suivre des pistes plus ou moins difficiles. Je remercie aussi les étudiants que j'ai eu en cours, pour les échanges chaleureux et ce que cela m'a apportée. Chercher de nouveaux angles pour expliquer des notions fondamentales est un exercice très enrichissant.

Merci à tous mes re-lecteurs, François-Gilles Carpentier, Joël Pothier, Jacques Chomilier, Dominique Higuet et Sophie Brouillet.

Je remercie les membres du jury qui ont accepté de faire partie de ce jury et de lire ce manuscrit. J'en suis très honorée.

Je remercie mes parents, sans qui rien de tout cela n'aurait été possible.

Et surtout, je remercie, Joël, Alice et Abel, sans qui rien de tout cela n'aurait de sens.

Et puis, cela paraîtra peut-être étrange, mais il y a longtemps que j'ai envie de remercier tous les auteurs de tous les livres que j'ai lus. Je lis beaucoup, vraiment beaucoup. Je ne peux pas ne pas lire, ou alors j'étouffe. J'ai l'impression d'avoir grandie entourée de la sagesse humaine accumulée depuis ces quelques millénaires, et d'y avoir un mieux compris la complexité du monde, et des gens. J'ai aussi l'impression quand sans eux, je ne serai rien, ou presque. Je les remercie donc de m'avoir tant donné.

Je dédie ce manuscrit à Joël, Alice et Abel.

Table des matières

Abréviations et acronymes

ABI : Atelier de Bioinformatique
AFP : *Aligned Fragment Pair*
CATH : *Class Architecture Topology Homology*
EC :*Enzyme Classification*
GO : *Gene Ontologyl*
HMM : *Hidden Markov Model*
InDel : Insertion ou délétion
ML : *Maximum Likelyhood* ou maximum de vraisemblance
MCL : *Markov Cluster Algorithm*
MNHN : Muséum National d'Histoire Naturelle
PDB : *Protein Data Bank*
RMN : Résonance Magnétique Nucléaire
SCOP : *Structural Classification Of Proteins*
SSE : *Secondary Structure Element*
SU : Sorbonne Université
SVM : *Support Vector Machine*
UE : Unité d'enseignement
UFR : Unité de Formation et de Recherche

Première partie

Recherche

CHAPITRE 1

Introduction

« L'esprit scientifique nous interdit d'avoir une opinion sur des questions que nous ne comprenons pas, sur des questions que nous ne savons pas formuler clairement. Avant tout, il faut savoir poser des problèmes. Et quoi qu'on dise dans la vie scientifique, les problèmes ne se posent pas d'eux-mêmes. C'est précisément ce sens du problème qui donne la marge du véritable esprit scientifique. Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de questions il ne peut y avoir de connaissances scientifiques. Rien ne va de soi. Rien n'est donné. Tout est construit. »
Gaston Bachelard, La formation de l'esprit scientifique, Vrin, Paris, 1938, p.14.

« Ce n'est pas le chemin qui est difficile, c'est difficile qui est le chemin. »
Sören Kierkegaard

Apprentissage de la recherche

J'ai découvert la recherche scientifique avec l'analyse et l'alignement des structures protéiques et elles sont devenues mon objet de recherche. Je n'ai depuis jamais changé d'objet et sa complexité ne cesse de m'émerveiller. Je m'intéresse aux relations entre la séquence, la structure, la fonction et l'évolution des protéines.

Lors de ma thèse, j'ai d'abord développé un des premiers programmes permettant de retrouver dans une banque toutes les structures protéiques similaires à une structure requête en un temps linéaire avec la taille de la banque (Carpentier et al., 2005 ; Alland et al., 2005). Ce « BLAST structural » a été nommé YAKUSA. La linéarité est importante pour permettre une recherche suffisamment rapide dans la PDB dont la taille croît très rapidement (bien que plus doucement que celles des banques de séquences). Nous avons utilisé à la fois une représentation simplifiée des structures (les angles α Levitt and Warshel 1975 ; Levitt 1976, qui sont les angles dièdres définis par 4 carbones α successifs) et un algorithme inspiré de celui de BLAST (Altschul et al., 1990, 1997) mais adapté aux structures de protéines. Son efficacité repose sur la construction d'un automate selon l'algorithme d'Aho-Corasick (Aho and Corasick, 1975). Nous avons aussi développé un score probabiliste qui est une approximation par paires des chaînes de Markov (les *Mixture Transition Model*, Berchtold and Raftery (2002) ; Raftery (1985)) permettant de mesurer la rareté des similitudes structurales identifiées. Je poursuis toujours le développement et la mise à jour de ce programme avec des projets tuteurés (en 2016 et en 2020) et des stages (stage en 2018 de O. Foucault et A. Lanrezac). Une nouvelle version sera bientôt disponible (en collaboration avec S. Brouillet). Je me suis ensuite orientée vers deux sujets (toujours pendant ma thèse) : la classification des protéines selon leurs similitudes structurales et l'alignement structural multiples des protéines.

Les deux principales classifications structurales SCOP (Murzin et al., 1995) et CATH (Orengo et al., 1997) existaient alors déjà (en 2003), mais elles ne nous satisfaisaient pas car nous voulions que les critères de similitude soient locaux pour favoriser la définition ultérieure des blocs structuraux conservés et que la procédure soit totalement automatique. Nous avons donc décidé de construire notre propre classification. Nous avons aussi choisi de ne pas découper les structures en domaines car leur définition est toujours problématique (voir page 8) mais nous voulions par contre les structures puissent appartenir à plusieurs familles. La procédure de classification se déroulait en trois étapes : i) comparaison de toutes les structures deux à deux, ii) classification en paires de structures similaires et non similaires, iii) regroupement en familles des structures similaires. J'ai comparé 2 à 2 les structures disponibles alors dans la PDB ayant moins de 90% d'identité de séquence entre elles (15056 structures à cette époque) et j'ai tenté de construire des familles de structures protéiques à partir des similitudes structurales trouvées. Plusieurs critères étaient utilisés (la longueur de l'alignement, le *RMSD*, *etc.*) et deux méthodes de classifications ont été testées pour déterminer si deux structures sont similaires ou non : les Machines à vecteurs de support (SVM pour *Support Vector Machines*) et les lois Normales Multivariées. Les structures similaires ont ensuite été regroupées en familles de structures avec la méthode MCL (*Markov Cluster Algorithm*, Enright et al. 2002).

Cette partie de mon travail n'a finalement pas été publiée car je n'ai finalement pas été totalement convaincue de la pertinence de la classification obtenue. Les résultats étaient globalement cohérents avec les autres classifications : 1700 familles de 2 structures ou plus ce qui est du même ordre de grandeur que le nombre à cette époque de *superfamily* de SCOP (1539) et le nombre de familles de niveau H (Homologous superfamily) de CATH (1467). La famille la plus peuplée contenait 223 structures et seules 5 familles avaient plus de 100 protéines. La taille moyenne des familles étaient de 6,6 structures et 50% des familles contenaient 3 protéines ou moins. J'ai vérifié manuellement beaucoup de familles et en ai déduit que même si la classification était globalement bonne, il y avait néanmoins trop de mauvaises classifications. En effet, les protéines contenant beaucoup d'hélices α étaient fortement attirées par des familles de protéines ayant elles-mêmes beaucoup d'hélices α mais sans autre similitude que ce soit à un niveau local ou à un niveau global. Les hélices α sont des structures secondaires très régulières et leur similitude structurale seule n'est pas l'indice d'une homologie. D'autres similitudes sont absolument nécessaires comme d'autres similitudes structurales locales (boucles etc.) ou globales (topologie globale des structures secondaires). Ensuite, les similitudes entre protéines étant locales, elles ne sont pas transitives : soit une protéine A qui partage le domaine *a* avec la protéine B qui elle même partage le domaine *c* avec la protéine C. La protéine A peut ne partager aucune similitude avec la protéine C. Ainsi, des protéines pouvaient être regroupées dans une même famille alors que leurs similitudes correspondaient à des régions différentes.

Ces travaux ont été poursuivis après mon départ du laboratoire pour résoudre le problème de la non transitivité des similitudes et une nouvelle méthode permettant de mieux prendre en compte la localité des similitudes identifiées (Santini et al., 2012) a été développée. Néanmoins, maintenir à jour ce type de classification est très coûteux en temps, et même si cette dernière classification était de très bonne qualité en 2012, j'utilise pour mes travaux désormais toujours CATH, SCOPe (nouvelle version de SCOP, Chandonia et al. 2019) ou dernièrement ECOD (Cheng et al., 2014) qui sont à jour et de très bonne qualité. La localité des similitudes ne me semble maintenant plus un critère si prégnant pour constituer des familles qui ont un sens. L'homologie des protéines me paraît bien plus importante dans le cadre de mes recherches où l'évolution a pris un rôle central.

En parallèle de cet essai de classification, j'ai développé 3 méthodes d'alignement multiple de structures mais n'ai publié que la dernière des 3. Les deux premières utilisent la représentation en angles α et sont des extensions d'algorithmes déjà connus dans le domaine de l'analyse de séquences que j'ai modifiés pour les adapter aux structures de protéines : la méthode des *m*-diagonales (Schuler et al., 1991) nommée Mafias et une variante du « Gibbs sampling »(Lawrence et al., 1993) nommée Gakusa . La troisième méthode nommée Triades (Pisanti et al., 2006, 2005) utilise une représentation en distances internes et est une extension de l'algorithme KMRC (Soldano et al., 1995) pour la recherche de "motifs relationnels flous". Cette méthode a une portée générale et nous l'avons appliquée à la recherche exhaustive de motifs structuraux locaux. Les résultats de ces différentes approches m'ont amenée à m'interroger sur ce qu'est une structure protéique, comment la représenter et quel est le sens d'un alignement structural.

Les structures protéiques comportent beaucoup de structures secondaires [1] et celles-ci sont structuralement extrêmement similaires d'une protéine à l'autre, sans qu'aucune relation de parenté entre protéines ne soit nécessaire pour cela. La présence de ces motifs "répétés" rend l'alignement structural local difficile et leur proportion importante rend aussi l'alignement global difficile. Dans Yakusa, la représentation en angles α a permis de développer un algorithme très rapide mais il a été nécessaire d'ajouter des critères globaux de similitude (distance quadratique moyenne entre les C_α ou RMS). Dans Mafias et Gakusa, qui font des alignements locaux sans prendre en compte l'agencement global des structures, les résultats sont mitigés : certains blocs structuraux identifiés sont pertinents, d'autres ne le sont pas. La meilleure manière de représenter les structures pour les comparer m'a semblé être une représentation permettant de trouver les similitudes locales et globales *i.e.* la représentation en distances internes, qui est celle utilisée dans Triades. Cette méthode était très prometteuse. Elle est réellement multiple, exhaustive, et générique. Elle permet de trouver tous les motifs à la fois selon des critères structuraux (de distances internes) et de séquences en acides aminés. Néanmoins, le nombre de motifs était rapidement bien trop important pour tenir en mémoire. Il aurait fallu trier, filtrer, regrouper certains motifs, mais la méthode ne serait alors plus exacte. Je poursuis néanmoins le développement de cette méthode par l'encadrement de stagiaires (Huiyuan Li en 2019, Miara Rakotomavo en 2020) et de projet tutorés (2020).

Mes 3 années de thèse se sont alors achevées avec pour ma part bien plus de doutes et de questionnements qu'au début, ce qui est probablement le signe d'un apprentissage réussi... La comparaison et la classification des structures protéiques se sont révélées des tâches complexes et ardues car il existe de multiples manières de représenter les structures et aussi de mesurer leurs similitudes qui ont toutes leurs avantages et leurs inconvénients. Pour choisir parmi ces représentations et mesures, le critère qui m'a semblé le plus pertinent était l'histoire évolutive : les positions à aligner sont les positions homologues et la distance calculée doit être proportionnelle ou au moins cohérente avec les distances évolutives. Je me suis alors intéressée à l'évolution de la structure des protéines et à sa modélisation. Je me suis alors intéressée à l'évolution de la structure des protéines et à sa modélisation. Pour avancer, j'avais besoin de mieux comprendre et si nécessaire de développer les modèles sous-jacents pour représenter et manipuler les structures. Mes questions étaient multiples : Comment représenter les structures pour l'alignement ? L'information structurale permet-elle de mieux identifier les positions homologues que l'information de séquence ? Les similitudes structurales observées résultent-elles d'une homologie ? Quel impact a une substitution sur une structure protéique ? Et une insertion ? Et une délétion ? Les protéines partageant le même repliement sont-elles homologues ?

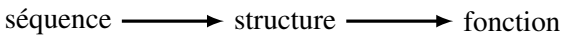
J'ai travaillé ces problématiques sous 3 angles que je présenterai dans ce manuscrit mais avant de poursuivre, un retour sur quelques concepts fondamentaux, qui ne sont finalement pas si simples, m'a été nécessaire.

1. 30% des résidus sont dans des hélices α et 20% dans des feuillets β en moyenne au total dans les protéines globulaires de la PDB

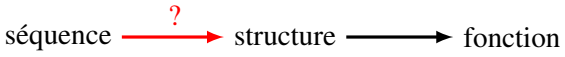
Concepts généraux sur les protéines, leurs structures et leur évolution

Plusieurs hypothèses sont fondamentales dans l'étude des protéines et de leurs structures. Elles sont de fait des paradigmes. Je souhaite ici les poser clairement mais aussi montrer leurs limites car, bien qu'il soit indispensable d'avoir des paradigmes pour avancer dans nos réflexions, il est aussi important de savoir les remettre en cause en particulier en cas de blocage ou d'incohérence.

Mon point de départ est que la séquence protéique (1D) se replie en une structure en 3 dimensions qui elle-même permet d'assurer une fonction biologique, ce qui est classiquement représenté par le schéma ci-dessous.



Relation séquence-structure



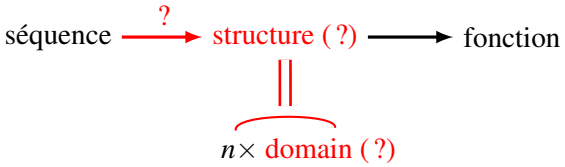
Les protéines se replient en un objet en trois dimensions que l'on nomme structure, qui est plus ou moins stable et dont l'aspect dynamique est difficile à prendre en compte. La première hypothèse est que toute l'information nécessaire au repliement d'une protéine est inhérente à sa séquence d'acides aminés. Cette hypothèse est issue des expériences de synthèse chimique et de repliement de protéines (Haber and Anfinsen, 1962 ; Hirschmann et al., 1969 ; Gutte and Merrifield, 1971) et du prix Nobel de Christian Anfinsen (Anfinsen, 1973). Ce paradigme est renforcé par les mesures de variations d'énergie libre entre les états dépliés et repliés qui sont le plus souvent faiblement négatives (Nikam et al., 2021) et par les méthodes de synthèse de protéines par sur-expression dans des bactéries où les protéines forment des corps d'inclusion où elles sont dénaturées puis retrouvent spontanément leur état replié ensuite (Cabrita and Bottomley, 2004). Néanmoins, ce paradigme a été remis en cause car les expériences de repliement *in vitro* citées précédemment ont été faites sur des protéines bien plus courtes (94 acides aminés en moyenne) que les protéines présentes dans les organismes [2] (Sorokina et al., 2022) et pourrait donc ne pas être généralisable. Les processus de repliement *in vitro* et *in vivo* se déroulent dans des conditions très différentes et beaucoup de protéines *in vitro* sont instables (Chi et al., 2003). *In vivo*, les protéines chaperones sont très souvent présentes lors de l'étape de repliement (Saibil, 2013). Ces chaperones n'affectent habituellement pas l'état final et sont présentes surtout pour éviter l'agrégation mais il existe cependant des cas documentés où les chaperones sont absolument nécessaires car les états non repliés et repliés sont très proches énergiquement (voire l'état déplié peut être plus favorable que l'état replié) et la barrière énergétique pour passer de l'un à l'autre est importante (Pauwels et al., 2007). Une étude du protéome soluble d'E. coli a conclu qu'un tiers de celui-ci n'est pas intrinsèquement repliable sur des échelles de temps physiologiques (To et al., 2021). Enfin, nos modèles physiques sont incapables de simuler le repliement de protéines

2. voir page 8 pour les longueurs des protéines dans les organismes

de taille un peu conséquente malgré l'accroissement des capacités de calculs. Tout cela aboutit à une certaine remise en cause de ce paradigme pour un modèle plus complexe où l'état replié ne serait pas le minimum global d'énergie libre, mais plutôt un minimum local sur un paysage fluctuant d'énergie libre. Le rôle des interactions est probablement crucial et certains chercheurs proposent même un modèle où la variation d'énergie libre serait positive pour la majorité des protéines, qui se replieraient dans leurs conformations natives uniquement par le biais d'interactions avec la machinerie moléculaire énergie-dépendante, en particulier, le système de traduction et les protéines chaperones (Sorokina and Mushegian, 2018 ; Sorokina et al., 2022). Le second paradigme est que les séquences similaires se replient en des structures similaires. Plusieurs exceptions ont été mises en évidence comme l'existence de protéines métamorphes qui peuvent adopter plusieurs conformations (Andreeva and Murzin, 2006 ; Dishman and Volkman, 2022), l'existence des protéines désordonnées qui n'adoptent pas seules une structure stable (Oldfield and Dunker, 2014) ou la construction (artificielle) de protéines très proches en séquence qui pourtant sont très différentes en structure (Alexander et al., 2009). La proportion de protéines métamorphiques a été évaluée à 4% des protéines de la PDB (Andreeva and Murzin, 2006 ; Dishman and Volkman, 2022). La proportion de résidus dans des régions désordonnées est plus difficile à évaluer et varie de 7,4% à 45% (voir la page 23 pour plus de détails à ce sujet).

Malgré leurs limites, ces deux paradigmes restent le cadre dans lequel nous réfléchissons la plupart du temps car ils semblent adéquats pour représenter ce que nous observons.

Les domaines

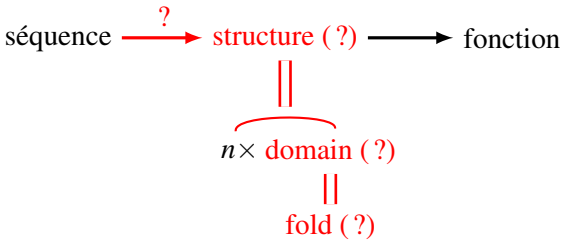


Les protéines sont de tailles très variables si on les prend dans leur forme quaternaire, de quelques nanomètres et quelques centaines d'acides aminés à plusieurs centaines de nanomètres et plusieurs millions d'acides aminés par exemple pour les capsides virales. La taille d'une chaîne polypeptidique est moins variable, en moyenne d'environ 400 à 470 acides aminés chez les eucaryotes et 250 à 350 chez les bactéries et les archées (Brocchieri and Karlin 2005, Tiessen et al. 2012, Milo et al. 2010, BNID 106444) . Une chaîne polypeptidique peut être composée de plusieurs domaines. Un domaine structural est un polypeptide de 50 à 250 acides aminés pouvant se replier indépendamment du reste de la protéine dans une forme compacte et stable (Xu and Nussinov, 1998). Il porte en général une fonction plus ou moins spécifique.

Dans une même chaîne, il y a le plus souvent de 1 à 4 domaines. Le nombre moyen de domaines par chaîne est légèrement plus élevé chez les eucaryotes (Kauko and Lehto, 2018). Le domaine est considéré comme l'unité évolutive des structures 3D des protéines (Doolittle, 1995 ; Dawson et al., 2017). Les domaines constituent en effet des blocs pouvant être recombinés pour créer des protéines qui acquièrent des fonctions différentes (Ye and Godzik, 2004 ; Vogel et al., 2004). Il est néanmoins

possible qu'un niveau inférieur puisse aussi constituer une unité évolutive : des fragments de 20 à 80 résidus communs à différents domaines ont été identifiés et auraient au moins pour certains une origine évolutive commune (Kolodny et al., 2021 ; Kolodny, 2021). Le découpage d'une protéine en domaines peut varier suivant les critères et les méthodes (Holland et al., 2006) mais les domaines présents dans les classifications sont cohérents dans 70% à 80% des cas (Getz et al., 2002 ; Csaba et al., 2009). Il faut néanmoins garder à l'esprit que 20-30% des domaines diffèrent suivant les classifications ce qui explique en partie les différences observées entre les classifications structurales.

Les repliements ou *folds*



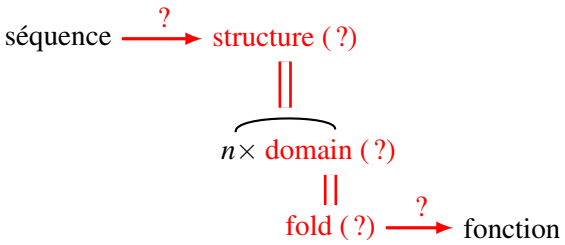
Un domaine est caractérisé structuralement par la succession en 3 dimensions de ses structures secondaires ; cette topologie est appelée le repliement ou *fold* en anglais [3]. Les protéines homologues partagent la plupart du temps le même *fold*. Les domaines sont classés selon leur structure dans des classifications structurales dont les trois principales sont actuellement CATH (Class Architecture Topology Homology) (Dawson et al., 2017), SCOP (Structural Classification Of Protein) (Murzin et al., 1995), devenue SCOPe (Chandonia et al., 2017) et ECOD (Evolutionary Classification of Protein Domains) (Cheng et al., 2014). Ces trois classifications sont hiérarchiques. Les quatre premiers niveaux hiérarchiques de SCOP sont : *class*, *fold*, *superfamily* et *family*, dans CATH ils sont *class*, *architecture*, *topology* et *homology* et dans ECOD ils sont *architecture*, *X-group*, *H-group*, *T-group*. Dans SCOP, les niveaux *class* et *fold* sont entièrement basés sur la structure, tandis que *superfamily* et *family* reposent également sur des similarités de séquence et/ou de fonction ; dans CATH, les trois premiers niveaux (CAT) sont entièrement basés sur la structure et dans ECOD, la classification est basée sur l'évolution et la structure. Même s'il existe certaines divergences entre les classifications, la meilleure correspondance pour le niveau de *fold* de SCOP est le niveau T dans CATH (Csaba et al., 2009) et le niveau X dans ECOD (Cheng et al., 2014). La définition d'un *fold* n'est pas sans ambiguïté (Kolodny et al., 2006). Il n'est pas rare que l'attribution d'un domaine à un *fold* soit difficile (Getz et al., 2002) et les classifications montrent des divergences même si elles sont globalement en accord : dans la comparaison de CATH et SCOP réalisée par Csaba et al. (2009), 80% des paires de domaines classés dans le même *fold* SCOP sont aussi dans la même *topology* CATH et une grande partie des différences est due aux quelques *superfold* (Rossmann folds, immunoglobulin et quelques autres). Il a été montré aussi que ces classifications hiérarchiques, bien qu'étant indispensables pour comprendre et analyser

3. Le terme de repliement a un double sens en français : il peut référer au *fold* mais aussi processus de repliement ou *folding* en anglais. J'utiliserai le terme de repliement pour *fold* et processus de repliement pour *folding*.

les structures et leur évolution biaise notre regard et masquent des relations d'homologie entre certains repliements (Cuff et al., 2009).

L'existence réelle des *fold* a été souvent mise en cause. La question de savoir si l'espace des repliements est discret ou continu a été beaucoup débattue (Shindyalov and Bourne, 2000 ; Dokholyan et al., 2002 ; Taylor, 2002 ; Rost, 2002 ; Kolodny et al., 2006 ; Skolnick et al., 2009 ; Nepomnyachiy et al., 2017). Il est maintenant admis que cet espace est continu mais aussi que les deux points de vue peuvent être fusionnés en un univers physique des protéines qui est théoriquement continu mais dont les populations (le nombre de protéines existant réellement) le rendent discret en pratique dans la plupart des cas (Harrison et al., 2002 ; Sadowski and Taylor, 2009 ; Sadreyev et al., 2009). Ainsi, la classification hiérarchique est justifiée tant que les similitudes sont suffisamment fortes tandis qu'avec des similitudes plus faibles, l'espace de structure continu est mieux représenté par un réseau (Pascual-García et al., 2009 ; Sippl, 2009 ; Mura et al., 2019). Une classification a pris en compte la continuité de l'espace dans son processus de construction (Xu and Zhang, 2016) et seule la classification de 3,4% à 12% des domaines varie, ce qui confirme la pertinence du point de vue discret pour la plupart des analyses. Les *folds* ont donc visiblement une existence physique, mais correspondent-ils aussi à des fonctions spécifiques et ont-ils un sens du point de vue évolutif ?

Relations structure-fonction



La fonction n'est pas une chose simple à définir. Dans GO [4] (Ashburner et al., 2000 ; The Gene Ontology Consortium, 2021), la fonction a trois caractéristiques différentes : le compartiment cellulaire (le noyau, le ribosome, *etc.*), le processus biologique (par exemple le métabolisme, la transduction du signal *etc.*) et la fonction moléculaire (la liaison, la catalyse, le transport, *etc.*). Dans le cadre de l'étude des structures protéiques, l'intérêt est principalement porté sur les mécanismes catalytiques et les interactions potentielles avec d'autres molécules, deux aspects qui sont couverts par la catégorie fonction moléculaire (Dessailly and Orengo, 2009).

Les variations en termes de fonction enzymatique des *folds* sont assez fortes : 25 % des repliements (ici niveau *homology* ou *superfamily* de CATH) comportent des membres avec des fonctions enzymatiques différentes mais au-delà de 40% d'identité de séquence, la variation du numéro EC [5] est rare (Todd et al., 2001). La diversité fonctionnelle est d'abord observée dans les *folds* très répandus dans les génomes (Lee et al., 2005) aussi nommés *superfold* (Orengo et al., 1994) comme par exemple les

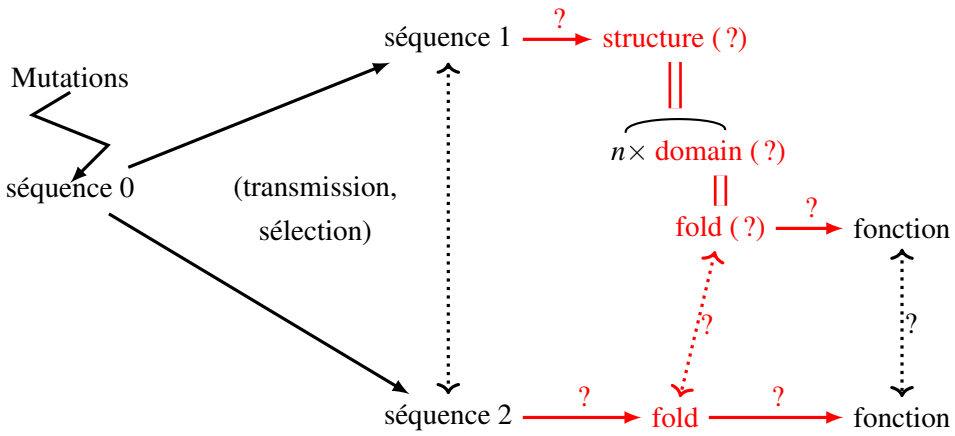
4. *Gene Ontology*
5. *Enzyme Classification*

TIM-like $(\beta/\alpha)8$ *barrel folds* (Nagano et al., 2002) ou les *Rossman folds*. Ces *superfolds* sont subdivisés en de nombreuses sous-familles dont les fonctions varient (Moult and Melamud, 2000 ; Nagano et al., 2002) et il a été montré au moins pour certains d'entre eux qu'ils sont apparus plusieurs fois au cours de l'évolution (évolution convergente) (Anantharaman et al., 2003). La diversité fonctionnelle existe aussi dans des familles moins répandues. Elle peut être due à de la divergence de fonction mais aussi à de la perte comme l'indique la présence assez fréquente de protéines non-enzymatiques dans des familles (*folds)* de protéines ayant des fonctions enzymatiques (Todd et al., 2002). A des niveaux de similitudes encore plus proches, quelques exemples de protéines homologues ayant des fonctions différentes existent (Piatigorsky et al., 1994 ; Jeffery, 2003 ; Whisstock and Lesk, 2003). Il existe aussi des cas où la diversité fonctionnelle proviendrait du changement des domaines adjacents et non du domaine lui-même (Hegyi and Gerstein, 2001 ; Todd et al., 2001). Cette diversité fonctionnelle est principalement [6] due aux substitutions mais aussi insertions/délétions qui jouent un rôle important dans les changements fonctionnels (Pascual-García et al., 2010). La plupart des repliements conservent un certain degré de points communs fonctionnels entre leurs membres, malgré leur divergence (voir Dessailly and Orengo 2009 pour une revue détaillée sur le lien structure - fonction des *folds*). Les mécanismes permettant l'apparition de nouvelles fonctions ne sont pas élucidés. La conservation serait plutôt au niveau du mécanisme catalytique que du substrat (Gerlt and Babbitt 2001 ; Todd et al. 2001). Les protéines enzymatiques peuvent être capables de catalyser d'autres réactions que la réaction qu'elles catalysent principalement (Khersonsky et al., 2006 ; Tawfik, 2010) ce qui faciliterait cette évolution.

L'annotation fonctionnelle des séquences protéiques est classiquement faite par transfert des annotations des séquences suffisamment proches, méthode qui est connue pour avoir des limites et générant des erreurs si elle est mal utilisée en particulier à cause de l'hypothèse sous-jacente selon laquelle les protéines homologues ont des fonctions similaires (Devos and Valencia, 2000, 2001). A plus de 30% d'identité de séquence, la fonction enzymatique peut être prédite avec une précision d'au moins 90%. A moins de 30%, même si la variation fonctionnelle est significative, les données structurales apportent des informations précieuses tant par l'identification des *folds* (Whisstock and Lesk, 2003 ; Adams et al., 2007) que celle des sites actifs ou de liaison.

6. Il existe des cas où le changement de l'environnement dans lequel évolue la protéine (Jeffery, 2003) modifie la fonction

Évolution des structures protéiques



Pour comprendre la diversité des protéines, de leurs structures et de leurs fonctions, il est rapidement indispensable d'ajouter la dimension évolutive. " Rien n'a de sens en biologie, si ce n'est à la lumière de l'évolution ", ce célèbre aphorisme de Theodosius Dobzhansky (1900 – 1975) est pertinent pour la structure des protéines comme pour tous les autres objets biologiques.

La séquence et la structure d'une protéine peuvent être considérées comme son génotype et son phénotype, respectivement, le repliement des protéines, qui fait correspondre la séquence à la structure, jouant le rôle du développement (Marcos and Echave, 2020). Les mutations des protéines, modifications transmissibles d'une génération à l'autre, ont principalement leur origine dans des événements modifiant l'ADN [7]. Ces événements sont très variés : mutations ponctuelles (transition, transversion), insertions, délétions, recombinaisons homologues et non homologues, duplications de gènes ou de chromosomes entiers, incorporation de transposons, de séquences virales ou de plasmides, *etc*. Les conséquences sur les séquences protéiques sont les mutations ponctuelles ou substitutions, les insertions et délétions et permutations circulaires, les combinaisons de domaines et, lors des duplications, la divergence peut être accélérée. Seules les mutations ponctuelles et les insertions/délétions sont explicitement modélisées dans les algorithmes d'alignement de séquences mais les recombinaisons homologues et non homologues (Trudeau et al., 2013 ; Carbone and Arnold, 2007 ; Crameri et al., 1998) et les fusions de domaines (Pandya et al., 2013) jouent aussi un rôle important dans l'évolution des protéines. On peut subdiviser l'évolution de la structure des protéines en deux parties : l'évolution au sein d'une même famille voire même d'un *fold*, nommée micro-transitions dans (Jayaraman et al., 2022), et l'évolution entre *folds*, avec l'apparition de *folds* et le passage d'un *fold* à l'autre, nommée macro-transitions dans (Jayaraman et al., 2022).

Même si tous les phénomènes peuvent participer aux micro-transitions et aux macro-transitions, ils sont plus ou moins importants dans l'un ou l'autre cas. Par exemple les substitutions induisent le plus souvent des changements mineurs dans la structure des protéines et participent surtout aux micro-transitions mais quelques cas ont été documentés où une mutation modifie complètement le *fold*

7. Il existe dans une moindre mesure des modifications de l'ARN comme l'épissage alternatif et des modifications post-traductionnelles

(macro-transitions) (Shortle, 2009 ; Alexander et al., 2009). De manière générale, les micro-transitions sont plutôt dues aux substitutions, insertions, délétions et un peu à la fusion de domaines dans le cas des *superfolds* qui sont particulièrement robustes (Pandya et al., 2013). Côté macro-transitions, plusieurs exemples du passage d'un repliement à un autre par une suite de changements (surtout des ajouts ou des suppression d'éléments de structure secondaire) ont été montrés (Grishin, 2001 ; Andreeva and Murzin, 2006 ; Alva et al., 2008, 2010 ; Farías-Rico et al., 2014 ; Toledo-Patiño et al., 2019 ; Kolodny et al., 2021). Néanmoins, aucun lien évolutif n'a été mis en évidence entre la plupart des *folds*. La fusion de domaines semble jouer un rôle important : il a été mis en évidence une origine pour partie archée et pour partie bactérienne de certaines protéines eucaryotes (Méheust et al., 2018) ce qui pourrait avoir eu un rôle dans l'apparition des nombreux *folds* spécifiques aux eucaryotes. Une source de nouveaux *folds* pourrait être les protéines *de novo* provenant de l'expression d'ADN codant (Carter, 2021) ou non codant (Lange et al., 2021). Les combinaisons de domaines (Yu et al., 2019) et les dynamiques d'apparition et de disparition des repliements au travers de la reconstruction de phylogénies (Winstanley et al. 2005, Yang et al. 2005 et aussi Caetano-Anollés and Caetano-Anollés 2003 et articles postérieurs). Le passage d'un repliement à un autre et l'apparition de nouveaux repliements est néanmoins un territoire assez peu exploré et qui nous apportera très probablement des surprises.

L'évolution des structures au sein des familles (micro-transitions) a fait l'objet de plus de recherches, que ce soit par des analyses des données ou par l'élaboration de modèles mathématiques. Je parlerai par la suite beaucoup des « sites »des protéines. Un site correspond à un acide aminé ou résidu dans une protéine. Comme les acides aminés peuvent changer au cours de l'évolution par substitution d'un acide aminé par un autre, on ne parle plus d'acide aminé mais de site ou position dans une protéine. Entre différentes protéines homologues, les sites peuvent être homologues, c'est-à-dire qu'ils ont évolués par substitution depuis la dernière protéine ancestrale commune. Ces sites homologues correspondent aux colonnes d'acides aminés alignés dans les alignements multiples. L'objectif des méthodes d'alignement est de les trouver, en plus d'identifier ceux qui apparaissent ou disparaissent (par insertion ou délétion). Il a rapidement été remarqué que la structure contraint l'évolution des séquences – par exemple, les sites exposés évoluent deux fois plus vite que les sites enfouis (Goldman et al., 1998) – et qu'elle est est remarquablement conservée au sein des familles protéiques Chothia and Lesk (1986) ; Russell et al. (1997). Ainsi, de nombreuses matrices de substitutions des acides aminés prenant en compte la structure ont été développées (Overington et al. 1990 ; Wako and Blundell 1994a,b ; Koshi and Goldstein 1997 et voir Perron et al. 2019b pour une liste plus complète). Ces matrices donnent de très bons résultats (Le and Gascuel, 2010) mais la nécessité d'avoir au moins une structure résolue explique probablement leur utilisation assez faible. Il est possible que la facilité avec la laquelle les structures sont prédites maintenant change les choses à l'avenir. Récemment, une matrice de substitution prenant en compte l'état conformationel des chaines latérales a été calculée (Perron et al., 2019a) comme quoi ce champs est toujours actif et de nouvelles idées peuvent toujours émerger. Dans les protéines, les taux d'évolution – le nombre de substituions fixées par unité de temps – varient entre les sites. De nombreuses études de la divergence des protéines au sein des

familles de protéines ont cherché à identifier les causes biologiques et physiques sous-jacentes aux variations observées et à contraindre l'évolution des séquences par la prise en compte de la structure (Parisi and Echave, 2001 ; Bastolla et al., 2003 ; Arenas et al., 2013). Même si la structure est ici considérée comme fixe au cours du temps, cela a permis de bien mieux comprendre et modéliser la variabilité des taux d'évolution. De manière générale, cette variabilité est visiblement principalement régis par la sélection pour la stabilité (Echave et al., 2016 ; Echave and Wilke, 2017). Je reviendrai un peu sur ce point dans le chapitre 3 page 41.

Dans tous les travaux précédemment cités la structure ne varie pas au cours de l'évolution, or c'est ce qui m'intéresse principalement ici : lorsque les substitutions s'accumulent, les séquences divergent, mais les structures changent aussi. De nombreux scores géométriques de divergence structurale globale ont été développés comme le RMSD (Chothia and Lesk, 1986), le TM-score (Zhang and Skolnick, 2004) ou un score dépendant de la conservation des contacts (Pascual-García et al., 2010). Ces scores sont plus ou moins corrélés avec la divergence des séquence et les structures évoluent plus lentement que les séquences (Rost, 1997 ; Wood and Pearson, 1999 ; Illergård et al., 2009). L'impact des substitutions a beaucoup été étudié pour tenter prédire l'effet qu'elles ont sur la stabilité de la molécule (cf. chapitre trois, page 40 pour plus de détails sur ce point).

Quelques modèles mathématiques tentent de saisir à la fois les variations en structures et en séquences. Dans les premiers travaux de Gutin and Badretdinov (1994) et de Grishin (1997), la structure est simplement représentée avec les coordonnées cartésiennes de certains de ses atomes dont les coordonnées varient selon des processus de diffusion. La distance entre deux structures est mesurée avec le RMSD. Challis and Schmidler (2012) et Herman et al. (2014) ont également utilisé les coordonnées cartésiennes pour construire des modèles probabilistes bayésiens de l'évolution de la structure des protéines. Dans les travaux de Challis and Schmidler (2012), l'évolution des séquences est modélisée par un processus Markovien à temps continu et à états finis, composé d'un modèle *birth-death* pour simuler l'apparition et la disparition des insertions et délétions (*indels*) d'une une matrice de substitution JTT modifiée pour les substitutions. L'évolution est structures est modélisée en faisant varier les coordonnées des C_α selon un mouvement brownien avec un coefficient de réversion moyen, nommé processus Ornstein-Uhlenbeck (OU). Les simulations pour obtenir la distribution postérieure sont réalisées par MCMC (*Markov chain Monte Carlo*). Pour des raisons de calcul, cette approches trait les coordonnées cartésiennes associées aux atomes comme évoluant indépendamment les unes des autres. Les résultats obtenus montrent que l'ajout de cette information structurale améliore grandement l'estimation de paramètres évolutifs tels que le temps d'évolution entre les espèces, les topologies d'arbres et l'alignement. Herman et al. (2014) ont généralisé la procédure d'inférence de Challis and Schmidler (2012) afin que les ensembles de données composés de plus de deux protéines puissent être analysés conjointement, et pour tenir compte de l'incertitude dans les coordonnées spatiales expérimentales. Le cout en temps de calcul de cette procédure est très important mais elle donne de très bons résultats pour reconstruire simultanément des phylogénies et les alignements. Néanmoins, ces modèles sont encore perfectibles de multiples manières (pour l'indépendance des positions entre elles, l'indépen-

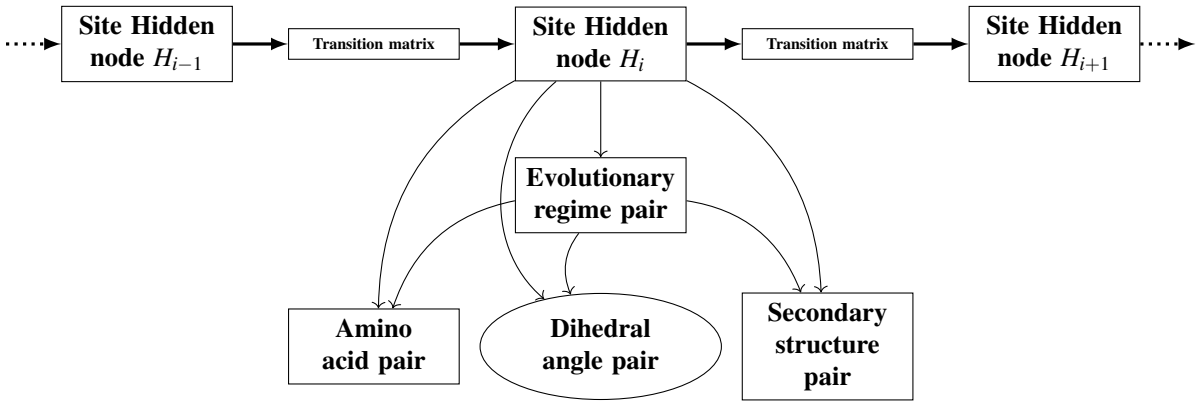


FIGURE 1.1 – Représentation du HMM décrit dans (Golden et al., 2017). Les flèches horizontales entre les nœuds cachés (*Hidden Nodes*) représentent les dépendances de voisinage entre les sites alignés. Ce modèle permet de modéliser le chemin évolutif entre une paire de protéines. Les flèches entre les nœuds cachés et les nœuds de la troisième ligne codent l'indépendance conditionnelle entre les variables de paires de sites d'acides aminés (un de chaque protéine), les paires de sites d'angles dièdres et les paires de sites de classes de structures secondaires. Les cercles représentent des variables continues et les rectangles des variables discrètes. Le régime dévolution correspond aux deux modes d'évolution possibles : l'évolution constante et les événements de saut.

dance de la séquence vis à vis de la structure) et ils ne permettent pas une réorganisation importante de la structure pouvant mener à la création de nouveau repliements.

Deux modèles récents permettent de prendre en compte la dépendance des acides aminés en 3D dans des modèles bayésiens. Le modèle de Larson et al. (2018) est une extension de celui de Challis and Schmidler (2012) pour prendre en compte cette dépendance avec une dérive spatiale corrélée des acides aminés. Golden et al. (2017) ont développé un modèle stochastique génératif qui permet de capturer les dépendances locales en travaillant dans dans l'espace des angles dièdres (ϕ, ψ). La trajectoire évolutive entre les structures d'une paire de protéines est modélisée par une marche aléatoire dans l'espace des angles dièdres à l'aide d'un processus de diffusion angulaire sur le tore bidimensionnel. La dépendance locale entre les sites voisins en séquence est modélisée avec un modèle HMM (*Hidden Markov Model*). L'évolution des séquence et des indels sont modélisés selon les mêmes principes que dans les modèles précédents. Un modèle de Markov en temps continu permet de modéliser aussi l'évolution des structures secondaires. Les différentes dépendances sont elles modélisées avec un modèle HMM contenant ces différentes compostantes (séquence, angle, structures secondaires) pour chaque site, avec des matrices de transitions entre les sites (voir figure 1.1). Deux modes d'évolution sont modélisés : l'évolution constante et les événements de saut pour permettre le couplage entre les 3 types d'états lorsqu'ils sont différents entre les 2 protéines de la paire.

Une autre catégorie de modèles non-bayésiens permettent de rendre compte de la déformation des structures au cours de l'évolution. Ceux si ont été initiés par les travaux de (Leo-Macias et al., 2005b,a) où les auteurs montrent les ensembles de conformations observées résultant de la divergence

des protéines de même famille sont très similaires à ceux produits par les fluctuations thermiques. Ces travaux modélisaient les structures par un réseau élastique et comparaient les analyses en modes normaux avec les analyses en composantes principales des structures des protéines appartenant aux mêmes familles. Les déformations associées aux divergences des séquences des protéines ayant un même repliement se font dans le sens des mouvements de moindre énergie qui sont les plus amples. Ainsi, les modes normaux indiquent la direction des déformations évolutives possibles Ces résultats ont été confirmés par d'autres travaux (Friedland et al., 2009) utilisant le modèle *backrub* (Lauck et al., 2010) (voir page 43 pour une description du modèle) ou utilisant des dynamiques moléculaires (Velázquez-Muriel et al., 2009). D'autres analyses des déformations se basent sur les alphabets structuraux (Agarwal et al., 2011). Les variations de l'alphabet structural au sein des protéines d'un même famille ont été comparées aux ensembles des structures RMN et les correlations sont très fortes sauf dans les boucles (Mahajan et al., 2014).

Par analogie avec les modèles de séquences, certaines études ont proposé que la sélection dirigerait les déformations des structures en ce sens au cours de l'évolution (Leo-Macias et al., 2005b,a ; Haliloglu and Bahar, 2015). Cependant, d'autres études montrent plutôt que la sélection n'ait aucun effet sur les modèles de structure, qui seraient déterminés par la sensibilité mutationnelle (Echave, 2008 ; Echave and Fernández, 2010 ; Marsh and Teichmann, 2014). Dans ces travaux, les structures sont modélisées par des réseaux élastiques (*Linearly Forced Elastic Network Model* ou LFENM) et les mutations sont représentées par un terme perturbateur linéaire (variation de la longueurs des ressorts). Ces perturbations rendent compte de l'effet des mutations sur la conformation d'équilibre (Echave, 2008). Ce modèle prédit que la conformation d'équilibre divergera le long des modes normaux de faible énergie même en cas de mutations aléatoires non sélectionnées, ce qui remet en doute l'interprétation fonctionnelle (Echave and Fernández, 2010). Par contre, ce type de modèle a récemment permis de montrer que si la variabilité structurale locale des structures est bien liée sensibilité mutationnelle (*ie. les régions flexibles, exposées et peu denses mutent beaucoup tandis que les regions rigides enfouis et très denses mutent peu* , la sélection (négative) a un effet apparemment secondaire de mise à l'échelle. L'effet des mutation est mesuré avec un RMSD local calculé pour chaque position entre la structure de départ et les structures mutées, c'est-à-dire dont le réseau élastique aura été perturbé. Lorsque la sélection augmente, les positions ayant un grand RMSD sans sélection ont un RMSD encore plus grand tandis que les positions ayant un RMSD plutôt faible ont un RMSD encore plus petit.

Un modèle utile d'évolution des protéines n'est pas simplement un modèle qui s'adapte bien aux données. Pour qu'un modèle d'évolution des protéines soit utile, il est également bénéfique d'établir un lien solide entre les paramètres du modèle et les caractéristiques biologiques que ces paramètres représentent. Bien que les meilleurs modèles d'évolution des protéines soient aujourd'hui beaucoup plus réalistes que les modèles originaux de Dayhoff (1972), la compréhension de l'évolution des protéines en est encore à ses débuts. Idéalement, un modèle d'évolution des protéines devrait être lié à la fois à la structure tertiaire de la protéine, à l'ADN qui code pour la protéine et à toutes les

interactions qui peuvent jouer dans son évolution. Le développement de modèles a donc encore de beaux jours devant lui. Les logiciels mettant en œuvre des modèles d'évolution des protéines pour l'analyse des données sont très demandés. Il faut souligner que même les modèles existants d'évolution des protéines, aussi irréalistes soient-ils, sont préférables à l'absence de modèles. Les modèles explicites d'évolution fournissent une base sur laquelle l'histoire phylogénétique peut être déduite, les hypothèses d'évolution peuvent être évaluées, la structure des protéines peut être prédite et notre compréhension de l'évolution peut être quantifiée.

Au cours de ces années de recherche j'ai poursuivi le travail abordé en thèse autour de l'alignement des structures protéiques et de l'annotation structurale mais j'ai aussi tenté de mieux comprendre l'évolution des structures tant du côté des micro-transitions et que de celui des macro-transitions, sachant que cette meilleure compréhension devrait me permettre de développer de meilleurs modèles et donc de mieux résoudre les problèmes initiaux d'alignement et d'annotation. Chacun de ces 3 thèmes fait l'objet d'un chapitre dans ce manuscrit. Dans le premier chapitre je présenterai les travaux que j'ai faits autour de la comparaison et l'alignement des protéines. Les questions auxquelles j'ai essayé de répondre sont : l'information structurale permet-elle de calculer de meilleurs alignements ? Comment allier l'information structurale et de séquence pour calculer de meilleurs alignements ? Les contacts sont-ils conservés ? Peut-on les aligner ? Peut-on ajouter de l'information contextuelle pour améliorer l'alignement ? Est-il possible et utile d'intégrer la co-évolution des positions ? Dans le second chapitre je présenterai les travaux que j'ai menés dans le cadre de la micro-évolution. Je me suis principalement intéressée à l'effet des substitutions sur la structure, et dans une moindre mesure à l'effet des insertions/délétions. J'ai tenté de répondre aux questions suivantes : l'effet d'une substitution unique est-il visible (mesurable) dans les structures résolues expérimentalement ? L'effet est-il uniquement au niveau de la mutation ? des résidus en contact avec le résidu muté ? ou ailleurs ? Ce phénomène est-il général ou variable suivant les protéines, les familles, les catégories de protéines, les régions des protéines, les structures secondaires ? Y a-t-il une relation entre la variation d'énergie libre (stabilité) et les variations structurales du squelette ? Et pour les insertions délétions : l'effet sur la structure d'une insertion ou délétion induit-il une déformation prévisible grâce aux modes normaux ? Dans le troisième chapitre, je présenterai les travaux que j'ai menés dans le cadre de la macro-évolution. Je m'y suis demandée si la convergence vers un même *fold* est fréquente et aussi quel est le rythme d'apparition des repliements.

CHAPITRE 2

Comparaison et alignement des protéines : comment allier la structure et la séquence

2.1 Motivations

Les technologies de séquençage toujours plus performantes nous permettent d'accroître exponentiellement la quantité de séquences protéiques disponibles dans les bases de données à un rythme que les méthodes expérimentales d'annotation sont très loin de pouvoir suivre. Pourtant, une séquence brute seule apporte très peu d'information : nous avons alors besoin de transférer les connaissances issues d'expériences menées sur d'autres protéines. Pour cela, nous comparons les nouvelles séquences aux séquences pour lesquelles nous connaissons les activités biochimiques, les fonctions, les structures... Si les ressemblances sont suffisantes, nous supposons que les caractéristiques (fonction, ligands, substrats, localisation...) de ces séquences sont similaires et/ou que les séquences sont homologues, *i.e.* dérivent d'un ancêtre commun (Whisstock and Lesk, 2003). Ainsi, les programmes d'alignement de séquences qui nous permettent d'identifier et de quantifier ces similitudes sont devenus centraux : ils sont à la base de l'annotation mais aussi de beaucoup d'autres domaines de la bioinformatique comme les reconstructions d'arbres phylogénétiques ou les prédictions des structures.

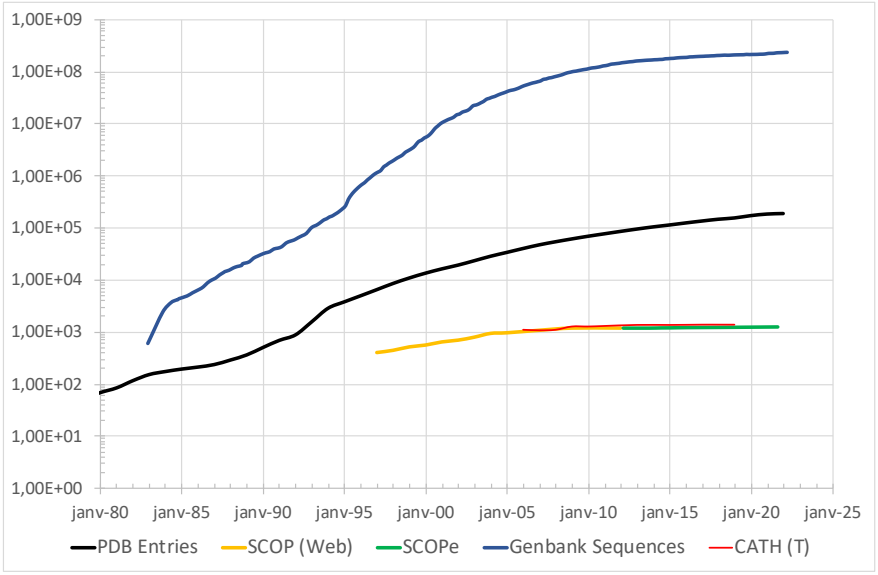
Les premières méthodes d'alignement développées dans les années 70 (Needleman and Wunsch, 1970) permettaient d'aligner 2 séquences. Elles reposent sur un modèle d'évolution prenant en compte les substitutions et les insertions/délétions et sur un algorithme de programmation dynamique. Des heuristiques comme BLAST (Altschul et al., 1990), suffisamment rapides pour scanner les banques de séquences, ont ensuite été développées. Puis vint PSIBLAST (Altschul et al., 1997) version multiple de BLAST (alignant plusieurs séquences). Il permettait d'annoter de 40 à 60% des séquences (Iliopoulos et al., 2001). Des méthodes heuristiques d'alignement multiple comme ClustalW (Thompson et al., 1994) sont apparues et ont aussi été développées pour permettre l'identification des positions homologues, *i.e.* celles qui étaient présentes dans la protéine ancestrale et qui ont évolué par substitution d'acides aminés. Néanmoins, les alignements multiples de séquences partageant moins de 30% d'acides aminés identiques étaient considérés comme non fiables (Levitt, 1997). On remarque toujours qu'une transition s'opère à ce taux de divergence en dessous duquel seule la composition hydrophiles/hydrophobes serait conservée (Kinjo and Nishikawa, 2004). Une petite révolution a alors eu lieu avec le développement de méthodes utilisant les chaînes de Markov cachées (*Hidden Markov Model* ou HMM) qui permettent de mieux modéliser la conservation des acides aminés à chaque position mais aussi surtout la probabilité de survenue des insertions et délétions. Le programme de recherche dans une banque HMMER (Eddy, 1995) a été le premier publié et reste un des plus utilisés. Depuis, de nombreuses améliorations ont été apportées, notamment les méthodes d'alignement HMM-HMM par HHblits (Remmert et al., 2012 ; Steinegger et al., 2019) qui permettent de rechercher des homologues plus lointains mais il n'y a pas eu de nouvelle révolution. Pourtant, si les séquences de génomes complets de Bactéries ou d'Archées sont maintenant annotées à plus de 85%, celles des Eucaryotes et Virus le sont seulement à environ 50% (Perdigão et al., 2015). De plus, nous avons maintenant accès aux séquences issues des organismes non cultivables et la proportion de séquences non annotées augmente alors fortement : 40% à 90% de séquences orphelines dans les métagénomes marins procaryotes de TARA (Villar et al., 2015) et jusqu'à 95% de séquences orphelines dans les

viromes (Reyes et al., 2012).

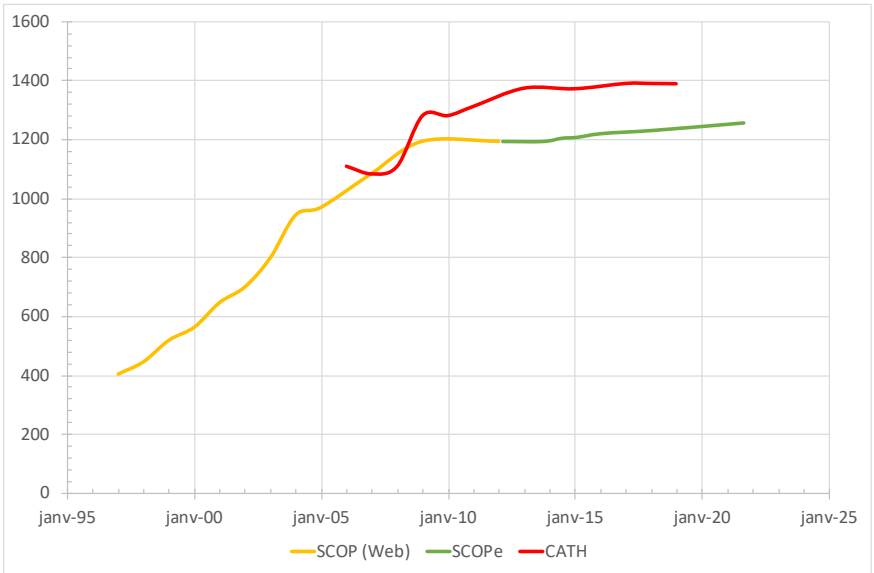
Peut-être abordons-nous un monde de protéines totalement nouvelles et inconnues, mais certains éléments indiquent cependant que nous sommes probablement toujours en présence de protéines connues mais ayant divergé au-delà des seuils acceptables pour les méthodes actuelles d'alignement. Premièrement, la structure des protéines est connue pour être plus conservée que la séquence : il a été montré que 50 à 80% des acides aminés peuvent changer sans que la structure soit significativement altérée (Schaefer and Rost, 2012) et que la structure est de 3 à 10 fois plus conservée que la séquence si l'on compare la conservation des structures secondaires, des contacts ou de l'accessibilité au solvant dans les structures avec le taux d'évolution des séquences (Illergård et al., 2009). Deuxièmement, on constate que le nombre de repliements structuraux différents est très faible comparativement à la diversité des séquences : les 150 000 structures protéiques déterminées expérimentalement déposées dans la PDB (Berman et al., 2000) sont classées en un peu moins de 1400 repliements (topologies) dans la classification structurale CATH (Sillitoe et al., 2019) et de 1300 dans SCOPe [8] (Fox et al., 2014) et ce nombre n'augmente plus ou presque depuis 5 ans (cf. figure 2.1). Troisièmement, la distribution des repliements dans les génomes suit une loi de puissance : certains repliements sont extrêmement fréquents et beaucoup le sont peu (Koonin et al., 2002). A partir de ces trois éléments, il semble raisonnable de penser qu'une grande majorité de ces nouvelles séquences non annotées font partie des repliements déjà connus et très fréquents. Il est probable que de nombreux repliements peu fréquents restent encore à déterminer expérimentalement (Koonin et al., 2002) mais ceux-ci représenteraient une faible part des séquences non annotées. L'information structurale semble donc être une clé pour permettre l'exploitation de ces nouvelles données, d'autant plus que l'utilisation de l'information structurale est deéjà connue pour améliorer l'annotation fonctionnelle (Moult and Melamud, 2000).

On peut néanmoins alors se demander si notre connaissance de la structure des protéines est biaisée notamment à cause des méthodes à notre disposition pour les étudier : les séquences non annotées ne seraient-elles pas désordonnées, c'est-à-dire des protéines pour lesquelles nous ne disposons pas ou presque de structure expérimentale ? Les protéines peuvent être divisées en quatre classes selon leur morphologie : les protéines globulaires qui sont dans un environnement aqueux, les protéines membranaires, les protéines fibreuses qui forment de vastes agrégats et constituent pour la plupart le cytosquelette, et enfin les protéines dites désordonnées qui sont généralement petites et n'ont pas de structure fixe inhérente. Ces dernières sont absentes de la PDB et par conséquent de CATH et SCOP aussi. Les protéines globulaires peuvent aussi avoir des régions désordonnées, le plus souvent aux extrémités N et C terminales, mais aussi dans une moindre mesure dans les boucles. Les informations structurales sur ces régions sont rassemblées dans la banque DisProt (Quaglia et al., 2022) mais nos connaissances sont beaucoup moins étendues que pour les régions structurées. La proportion de résidus qui sont dans des régions désordonnées est prédite par différentes méthodes (actuellement principalement du *deep learning*, *cf.* Necci et al. 2021) et est assez variable suivant les règnes et sui-

8. nouvelle version étendue de SCOP (Murzin et al., 1995)



(a)



(b)

FIGURE 2.1 – (a) Croissance du nombre de *folds* de SCOP, de *Topology* de CATH et d'entrées dans la PDB et Genbank en fonction du temps, entre 1980 et 2022. L'échelle est semi-logarithmique. (b) Croissance du nombre de *folds* de SCOP et du nombre de *Topology* de CATH depuis leurs créations.

vant les publications. Ces proportions vont ainsi de 20.5% à 45% chez les eucaryotes, 8.5% à 28% chez les bactéries, et 7.4% à 38%chez les archées suivant les publications Peng et al. (2015) ; Xue et al. (2012). La proportion de protéines comportant des régions désordonnées longues (>30 résidus) serait de 33% chez les eucaryotes, 4% chez les bactéries et de 2% chez les archées d'après Ward et al. 2004. Par conséquent, il semble improbable que toutes les séquences non annotées, qui représentent de 40% à 95% des séquences, soient désordonnées, mais celles-ci devraient néanmoins en constituer une part non négligeable.

Récemment, la méthode de prédiction de structure par *deep learning Alphafold 2* (Jumper et al., 2021) qui a révolutionné la bioinformatique structurale, a été appliquée pour prédire toutes les structures du protéome humain (Tunyasuvunakool et al., 2021). Pour 98% des protéines une structure a pu être prédite, mais 30% des résidus du protéome ont des indices de fiabilité bas et ces régions chevauchent largement les régions prédites comme désordonnées (Ruff and Pappu, 2021). Nous sommes donc capables, à partir des séquences uniquement, d'inférer la structure, et donc potentiellement la fonction des protéines. Cette information est donc bien présente, et accessible actuellement grâce aux réseaux de neurones. Il est sûrement possible de l'identifier, la comprendre et de l'isoler avec des méthodes moins coûteuses en calculs et en mémoire et surtout plus compréhensibles et analysables que le *deep learning* développé par *Deep Mind*.

Mon intuition est donc qu'il reste beaucoup de similitudes et d'homologies à identifier entre les protéines en particulier en s'aidant de leurs structures et que le faire nous permettra d'accroître notre compréhension du vivant. La méthode pour les identifier qui me semble la plus prometteuse est d'allier les informations des séquences et de structures. Le but des alignements de séquences de protéines est d'aligner (mettre en correspondance) des acides aminés supposés homologues qui dérivent d'une séquence ancestrale par des substitutions. Les modèles d'évolution sous-jacents prennent en compte trois types d'évènements : les insertions, les délétions et les substitutions. L'alignement optimal peut être calculé pour l'alignement de deux séquences (Needleman and Wunsch, 1970 ; Smith and Waterman, 1981) mais même si le calcul de l'alignement optimal de n séquences est possible, il devient rapidement trop coûteux car la complexité est en $O(n^2 2^n l^n)$ avec n le nombre de séquences et l la longueur moyenne des séquences (Konagurthu and Stuckey, 2006). Par conséquent, les méthodes d'alignement multiple utilisées sont des heuristiques qui alignent progressivement les séquences en commençant par les plus similaires.

Dans les alignements structuraux, les modèles évolutifs sous-jacents sont beaucoup moins explicites. La difficulté principale est que si une substitution, insertion ou une délétion est un événement discret et bien défini en séquence, les variations correspondantes dans la structure sont continues et dépendent très fortement du repliement et de la localisation dans la structure. Les méthodes d'alignement de structures sont bien plus diverses que les méthodes d'alignement de séquences. Elles varient quant à la représentation des structures, les mesures de similitudes et les algorithmes utilisés. Néanmoins, la majorité des méthodes d'alignement de structures recherchent de petits blocs structuraux conservés selon des scores géométriques puis les assemblent en un alignement en prenant plus ou

moins en compte la flexibilité de la molécule (Carpentier and Pothier, 2007 ; Berbalk et al., 2009).

Les positions alignées sont donc similaires du point de vue des conformations structurales locales et/ou globales mais l'implication d'une homologie a été questionnée (Godzik, 1996). La convergence structurale a été mise en évidence pour des domaines structuraux au moins pour certains repliements comme le *Rossmann-like fold* (Medvedev et al., 2019), mais c'est un phénomène qui semble néanmoins plutôt rare. Des fragments structuraux similaires (autour de 30 résidus la plupart du temps) ont souvent été identifiés dans des protéines de familles et de repliements différents, avec des fonctions diverses et dans des organismes divers (Lamarine et al., 2001). Il semble qu'au moins une partie de ces peptides puissent avoir des origines évolutives communes car des similitudes en séquences, quoique faibles, sont aussi retrouvées (Alva et al., 2015 ; Nepomnyachiy et al., 2017). Ils pourraient être les briques de base permettant l'émergence de nouveaux repliements grâce à des événements de duplication et de fusion. L'unité évolutive la plus petite dans les structures pourrait alors être ces fragments structuraux ; pourtant les alignements générés sont bien des alignements d'acides aminés et non de blocs structuraux. L'articulation entre les événements de substitution d'acides aminés et la conservation de ces blocs structuraux n'est pas évidente et on peut se demander s'il est possible d'inférer un alignement des acides aminés à partir d'un alignement structural. En d'autres termes, les méthodes d'alignement structural sont-elles vraiment supérieures pour retrouver les résidus homologues ? C'est une question que je suis venue à me poser assez rapidement après ma thèse. J'ai tenté d'y répondre en menant une étude comparative approfondie des performances des méthodes d'alignement basées sur la séquence et/ou sur la structure dont je vais présenter rapidement les principaux résultats et conclusions qui ont été publiés dans (Carpentier and Chomilier, 2019). En parallèle, j'ai aussi poursuivi le développement de méthodes d'alignement de séquences (stages de Damien Monet en 2012 puis de Benjamin Roques en 2015, en collaboration avec Guillaume Achaz), d'alignement de structures *pairwise* (projets étudiants et stage d'Ophélie Foucault et André Lanrézac en 2018 pour l'amélioration de Yakusa) et multiples (stage de Huiyuan Li en 2019 et de Miara Rakotomavo en 2020). Je continue aussi d'explorer l'annotation structurale et ce qu'elle peut apporter (articles (Glasner et al., 2011 ; Rochat et al., 2019), stages de Stéphane Peugeot en 2012, Baptiste Mossoti en 2015, Emile Faure en 2019 et Théo Jamay en 2021 et actuellement Jeremy Rousseau en 2022 en collaboration avec Lucie Bittner).

2.2 Contributions

2.2.1 Comparaison des programmes d'alignement de séquences et de structures

1) Méthodologie

Les alignements de références La méthode classique pour évaluer les performances des méthodes d'alignement est de mesurer leur aptitude à reconstruire des alignements de références. Ces alignements de références sont souvent des alignements vérifiés manuellement (Bahr et al., 2001 ; Thompson et al., 1999a, 2005 ; Andreeva et al., 2007),. Ils sont construits et vérifiés principalement en utilisant

les informations de séquence et de structure mais aussi à partir d'autres informations comme la fonction et les motifs associés (Thompson et al., 2011). Nous avons sélectionné 847 alignements, contenant au moins trois chaînes ou domaines protéiques, dans cinq bases de données d'alignements multiples de référence : BALIBASE 2 (Thompson et al., 1999a), BALIBASE 3, dont les alignements sont construits et vérifiés manuellement (Thompson et al., 2005), HOMSTRAD (Mizuguchi et al., 1998b) dont les alignements (structuraux) sont calculés avec MNYFIT (Sutcliffe et al., 1987), STAMP (Russell and Barton, 1992) et COMPARER (Sali and Blundell, 1990), annotés avec JOY (Mizuguchi et al., 1998a) et vérifiés manuellement, OXBENCH (Raghava et al., 2003) dont les alignements sont calculés avec STAMP (Russell and Barton, 1992) et SISYPHUS (Andreeva et al., 2007) qui regroupent des domaines de SCOP (Murzin et al., 1995) avec des relations structurales non triviales et dont les alignements multiples sont construits manuellement.

Les programmes Nous avons distingué trois catégories de programmes d'alignement multiple : les programmes d'alignement de séquences, les programmes d'alignement de structures et les programmes d'alignement séquence+structure. Pour être inclus dans cette étude, un programme devait : (i) être disponible en téléchargement, (ii) produire un fichier contenant l'alignement de séquences, (iii) fonctionner sans erreur. Chaque alignement multiple devait être calculé en moins de deux heures. Certains programmes n'ont pas produit suffisamment d'alignements pour permettre une analyse significative de leurs performances et ils ont été exclus s'ils ont produit des alignements pour moins de 70% de l'ensemble de données. Comme notre objectif principal est d'étudier les performances des méthodes d'alignement basées sur la structure ou sur la séquence+structure, nous avons essayé d'être aussi exhaustifs que possible pour ces dernières. J'ai recherché tous les programmes d'alignement multiple de structures et de séquence+structure. J'en ai identifié 41 mais seuls 14 d'entre eux étaient utilisables (*cf.* table 2.1 pour la liste des 14 programmes et en annexe table 7.1 page 133pour la liste complète). Nous avons été surpris par le faible nombre de méthodes d'alignement utilisant la séquence et la structure. Nous n'avons pas inclus les méthodes améliorant les alignements après coup, comme STACCATO (Shatsky et al., 2005). Pour le versant de l'alignement de séquences, nous n'avons testé que les programmes plus populaires selon les dernières études (Thompson et al., 2011 ; Le et al., 2017). Nous avons sélectionné neuf programmes basés sur les séquences, cinq programmes basés sur les séquences et les structures (TCOFFEE/3DCOFFEE est exécuté avec SAP ou TM-ALIGN) et onze programmes basés sur les structures.

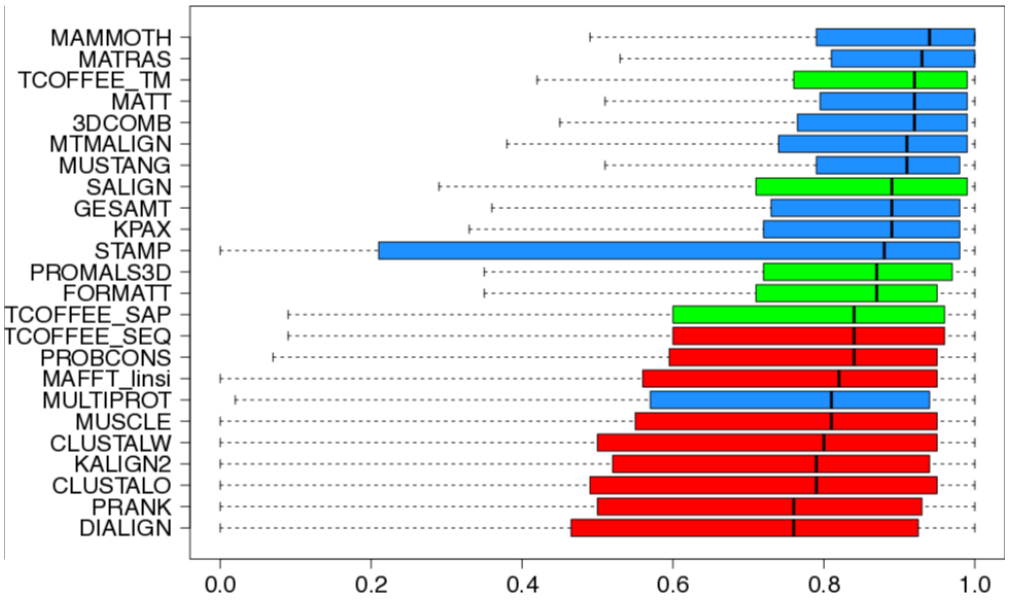
Évaluation de la qualité des alignements Les alignements produits par chaque programme sont évalués par comparaison avec les alignements de référence au moyen de deux scores, comme dans Thompson et al. (1999b) : (i) la fraction de paires de résidus dans l'alignement de référence correctement identifiée par une méthode donnée, connue sous le nom de Sum-of-Pairs (SP) score ; (ii) le Column Score (CS) score qui décrit la fraction de colonnes de référence identifiées.

TABLE 2.1 – Les programmes testés dans cette étude.

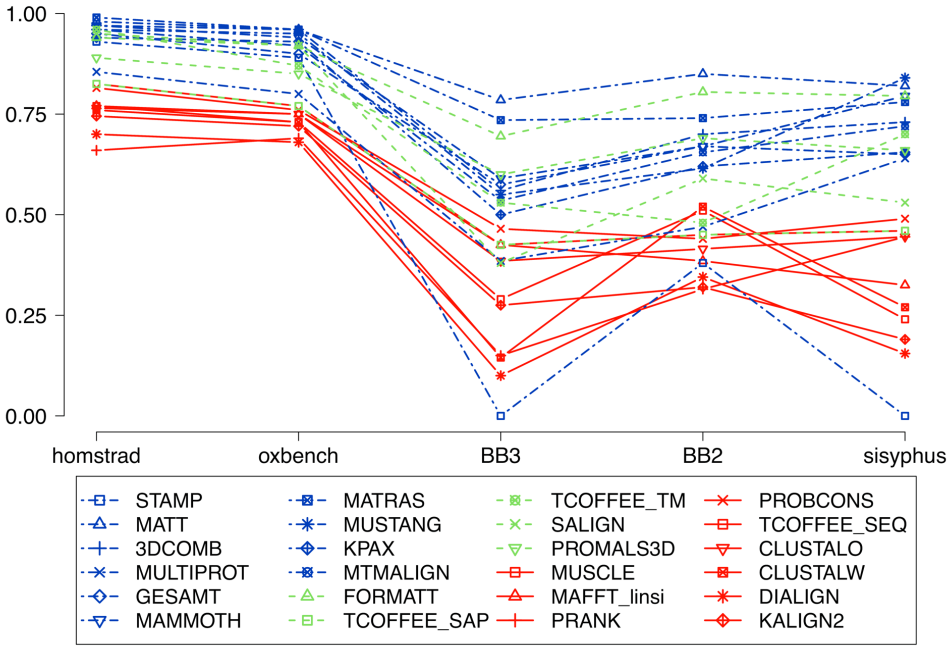
Type	Name	Description	Version	Ref.	Year
SEQ	CLUSTALO	Seeded guide trees and HMM profile-profile	1.2.0	(Sievers et al., 2011 ; Goujon et al., 2010)	2010
SEQ	CLUSTALW	Classical progressive aligner	2.1	(Larkin et al., 2007 ; Thompson et al., 1994)	1994
SEQ	DIALIGN	Greedy and progressive approaches for segment-based multiple alignment	TX, 1.0.2	(Al Ait et al., 2013 ; Morgenstern et al., 1998 ; Morgenstern, 1999)	1998
SEQ	KALIGN2	Wu-Manber string-matching algorithm, improving both accuracy and speed	2.04	(Lassmann and Sonnhammer, 2005 ; Lassmann et al., 2009)	2005
SEQ	MAFFT_linsi	Fast progressive aligner with iteration and refinement using consistency score	7.215	(Katoh and Standley, 2013 ; Katoh et al., 2002)	2002
SEQ	MUSCLE	Fast progressive aligner with iteration and refinement	3.8.31	(Edgar, 2004b,a)	2004
SEQ	PRANK	Phylogeny-aware progressive aligner ; correcting treatment of insertions	v.100701	(Löytynoja and Goldman, 2005)	2005
SEQ	PROBCONS	Probabilistic variant of the consistency algorithm	1.12	(Do et al., 2005)	2005
SEQ	TCOFFEE_SEQ	Consistency-based progressive aligner	11.00. 8cbe486	(Notredame et al., 2000)	2000
SEQ/ STRUCT	PROMALS3D	Derives constraints through structure-based alignments ; combines them with sequence constraints when constructing MSAs	NA	(Pei and Grishin, 2007 ; Pei et al., 2008)	2008
SEQ/ STRUCT	TCOFFEE_SAP	TCOFFEE + pairwise structure alignments by SAP	11.00. 8cbe486	(O'Sullivan et al., 2004 ; Orengo and Taylor, 1996)	2004
SEQ/ STRUCT	TCOFFEE_TM	TCOFFEE + pairwise structure alignments by TMALIGN	11.00. 8cbe486	(O'Sullivan et al., 2004 ; Zhang and Skolnick, 2005)	2004
SEQ/ STRUCT	SALIGN	DP with a score that is a sum of an affine gap penalty and of terms depending on various sequence and structure features	Modeler version : 9.18	(Madhusudhan et al., 2009)	2007
SEQ/ STRUCT	FORMATT	MATT with sequence information	1.02	(Daniels et al., 2012)s	2005
STRUCT	3DCOMB	Identifies structurally similar pairwise fragments and assemblies according to pivot structures. Score : TM-Score (Zhang and Skolnick, 2005)	1.06	(Wang et al., 2011b)	2011
STRUCT	GESAMT	Clustering of small structurally similar pairwise fragments . Score : Q-Score (Krissinel and Henrick, 2004)	7.0	(Krissinel, 2012 ; Winn et al., 2011)	2012
STRUCT	KPAX	DP + alignment optimization. Score : Gaussian structural similarity score	5.0.5	(Ritchie et al., 2012)	2005
STRUCT	MAMMOTH	AFPs alignment by DP. Progressive multiple alignment with a guide tree.Score : probability of residue random match of two different folds (Ortiz et al., 2002).	NA	(Lupyan et al., 2005)	2005
STRUCT	MATRAS	Progressive multiple alignment (guide tree) by DP. Score : PAM like matrices computed on SSE conservation or C_α internal distances.	1.2	(Kawabata, 2003 ; Kawabata and Nishikawa, 2000)	2000
STRUCT	MATT	AFPs chaining by DP. Score : based on RMS for AFP and on a geometrical transformations to allowing flexibility for chaining.	1.0	(Menke et al., 2008)	2008
STRUCT	MISTRAL	Superposition by minimizing interaction energy and residue one-to-one correspondence afterwards. Score : interaction energy and RMS.	3.6	(Micheletti and Orland, 2009)	2009
STRUCT	MTMALIGN	Progressive multiple alignment (guide tree) by DP. Score : TM-Score.	20171124	(Dong et al., 2018)	2017
STRUCT	MULTIPROT	With each structure as a pivot, detection of all AFPs, assembling to build the longest consistent alignment. Score : alignment length, consistency and RMS.	1.93	(Shatsky et al., 2004)	2004
STRUCT	MUSTANG	AFP and progressive multiple alignment with a tree. Score : C_α internal distance (DALI like, (Holm and Sander, 1993))	3.2.3	(Konagurthu et al., 2006)	2005
STRUCT	STAMP	Iterative superposition and alignment of C_α by DP with a guide tree. Score : C_α distances and conformational similarity	4.4	(Russell and Barton, 1992)	1992

2) Résultats et discussion

Je ne détaillerai pas ici tous les résultats de cette étude ; je soulignerai simplement quelques points et les conclusions principales qui sont : i) les programmes d'alignement de structures ont de meilleurs résultats que les programmes alliant séquence et structure et que les programmes d'alignement de séquences ; deux méthodes d'alignement de structures se démarquent des autres mais elles ont des représentations, scores ou méthodes d'alignement des structures très différentes ; ii) il ne semble pas que cette différence soit due à la méthode de construction des alignements de références ; ces différences sont retrouvées pour tous les taux d'identité de séquences et quelles que que soient les structures secondaires, iii) le nombre de gaps ajoutés par les programmes d'alignement structuraux est plus important que le nombre ajouté par les méthodes d'alignement de séquences.



(a)



(b)

FIGURE 2.2 – (a) Boites à moustaches des scores CS pour les 531 alignements calculés par toutes les méthodes. Les programmes sont ordonnés selon leur score médian. Les couleurs représentent les 3 catégories de programmes : rouge pour ceux alignant les séquences, bleu pour ceux alignant les structures et vert pour ceux alignant séquences et structures. Ces couleurs seront conservées dans toutes les figures suivantes représentant les scores CS ou SP. (b) Scores CS médians pour chaque programme et chaque banque. Figures provenant de (Carpentier and Chomilier, 2019)

Les programmes d'alignements de structures sont meilleurs Les meilleurs programmes d'alignement sont ceux utilisant uniquement l'information structurale (cf. figure 2.2a), devant les programmes d'alignement de séquences mais aussi plus étonnamment, devant les programmes utilisant à la fois l'information de structure et de séquence. Ces derniers ont néanmoins de meilleures performances que les programmes d'alignement de séquence. Il faut cependant noter que MAMMOTH a échoué à aligner un grand nombre d'alignements (225 sur 847 alignements au total) tandis que MATRAS a calculé tous les alignements. Les analyses de performance ont été réalisées uniquement sur les alignements que tous les programmes ont calculés (535 sur 847 alignements au départ), ce qui avantage très probablement MAMMOTH car il est le programme échouant à aligner le plus d'alignements, il "choisit" donc ses alignements.

Nous avons analysé un peu plus en détail les résultats en calculant des tests de Friedman pour savoir si les performances de certains programmes étaient significativement différentes. Il y a 6 groupes de programmes ayant des performances similaires entre eux et différentes des autres selon les tests de Friedman (cf. figure 2.3a). Les deux premiers groupes sont MATRAS et MAMMOTH, puis MATT, TCOFFEE_TM, 3DCOMB, MUSTANG et TCOFFEE_SAP. Les trois derniers groupes contiennent tous les programmes basés sur les séquences ainsi que FORMATT, MULTIPROT et MISTRAL. TCOFFEE_SEQ et PROBCONS sont les deux meilleurs programmes d'alignement multiple de séquences.

Nous avons aussi classé par classification ascendante hiérarchique les programmes en fonction de leurs scores CS. Dans cette classification, les programmes sont donc regroupés en fonction des alignements qu'ils arrivent à calculer correctement et de ceux qu'ils n'arrivent pas à retrouver. Tous les programmes basés sur séquence+structure ou sur la structure seule, à l'exception de STAMP, sont dans le même sous-arbre et de même, tous les programmes basés sur la séquence sont également ensemble (cf. figure 2.3b). Ce résultat signifie que globalement, les programmes basés sur les séquences réussissent et échouent sur les mêmes alignements, qui sont différents, au moins pour certains d'entre eux, des alignements pour lesquels les programmes basés sur les structures réussissent ou échouent. Nous avons aussi ordonné les alignements avec la même procédure et 4 zones différentes apparaissent : les alignements "faciles" où les scores de tous les programmes sont bons donc proches de 1 (zone A dans la figure 2.3b), les alignements difficiles pour tous les programmes (zone B), les alignements difficiles pour les programmes alignant les structures, faciles pour les programmes alignant les séquences (zone C) et les alignements difficiles pour les programmes alignant les séquences, faciles pour les programmes alignant les structures (zone D). L'analyse détaillée de cette matrice permet d'identifier précisément quels sont les alignements qui sont difficiles pour certains groupes de programmes, et donc les forces et faiblesses des différentes approches et donc devrait permettre de mieux les combiner ou de les améliorer. Je n'analyserai pas en détail cette matrice mais il est intéressant de noter que les approches alignant avec la séquence et la structure ont bien des profils intermédiaires entre celles alignant la séquence et celles alignant la structure : par exemple, PROMALS3D a des scores assez hauts (jaunes) pour les alignements normalement difficiles pour les programmes alignant les

structures (zone C), mais il a visiblement plus de difficultés que ceux-ci pour les alignements de la catégorie D.

MAMMOTH et MATRAS ressortent comme les meilleurs méthodes d'alignement selon cette étude et pourtant ce sont deux méthodes très différentes en dehors de la procédure générale :

étape 1 : calcul d'un score entre toutes les paires de protéines ($N(N-1)/2$ scores) ;

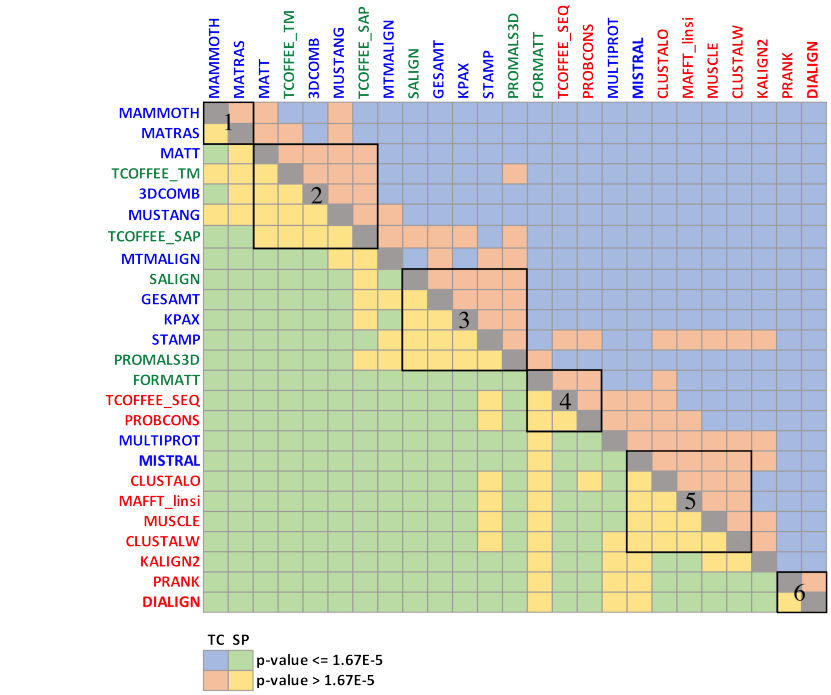
étape 2 : construction d'un dendrogramme à partir des scores calculés ;

étape 3 : alignement de toutes les protéines dans l'ordre indiqué par le dendrogramme (en commençant par la paire de structures les plus similaires).

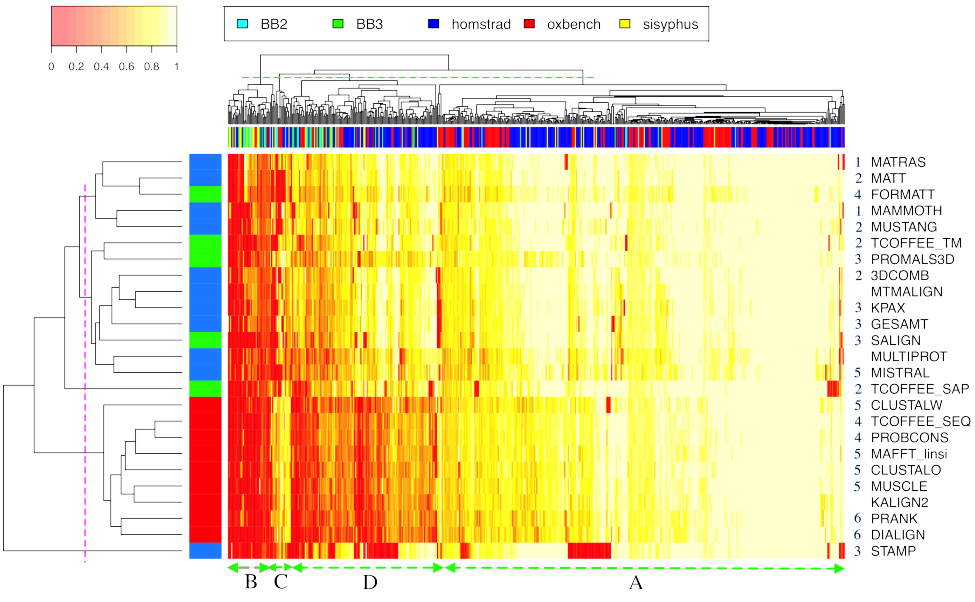
Dans le programme MAMMOTH, à l'étape 1, des fragments structuraux similaires de 7 résidus sont recherchés en calculant une variante du RMS (*URMS*) (Chew et al., 1999 ; Kedem et al., 1999). Un alignement local est calculé ensuite à partir de ces fragments et ces scores permettent de construire le dendrogramme. L'alignement des structures est fait ensuite par programmation dynamique par rapport à une structure moyenne et il est affiné par des étapes de superposition des structures. Dans la méthode MATRAS, trois scores de similarité structurale sont calculés d'une manière analogue aux scores de similarité des acides aminés des matrices PAM (Dayhoff et al., 1983). Les matrices ont été construites à partir de paires de protéines homologues d'une banque non-redondante à 95% d'identité extraite de SCOP (Murzin et al., 1995), regroupées et alignées grâce à leurs séquences. Ces trois scores de similarité structurale nécessitent la définition d'états discrets et il y en a 3 catégories (et donc 3 matrices) : i) des paires de structures secondaires, représentées par des vecteurs, et de leur orientation, ii) des angles (ϕ, ψ) discrétisés et regroupés selon les structures secondaires et l'enfouissement, iii) des distances internes entre C_β discrétisées avec un pas de 1ÅLa similarité entre toutes les paires de structures est calculée et un arbre est construit par la méthode UPGMA. L'alignement entre deux groupes suit celui entre les deux structures appartenant à chacun des groupes qui s'alignent le mieux. MATRAS est un programme relativement ancien ((2003 ; 2000) et qui n'a pas été mis à jour depuis. Il fonctionne pourtant toujours remarquablement bien.

Un deuxième groupe de méthodes d'alignement de structures est proche en termes de performances : MATT, MUSTANG et 3DCOMB (basés sur la structure), TCOFFEE_TM et TCOFFEE_SAP (séquence+structure). Ces sept programmes construisent les alignements à partir de fragments de quelques résidus alignés par paires.

Dans TCOFFEE_TM et TCOFFEE_SAP, l'ajout d'informations sur la structure améliore clairement l'alignement obtenu par TCOFFEE_SEQ, mais ce n'est pas le cas pour MATT et FORMATT. Les programmes basés sur la cohérence des fragments (*consistency based*) (TCOFFEE, PROBCONS) sont les mieux classés des programmes alignant des séquences comme dans (Thompson et al., 2011).



(a)

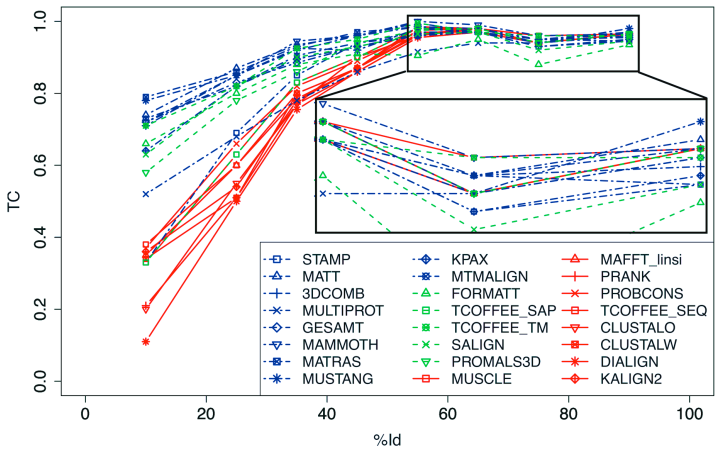


(b)

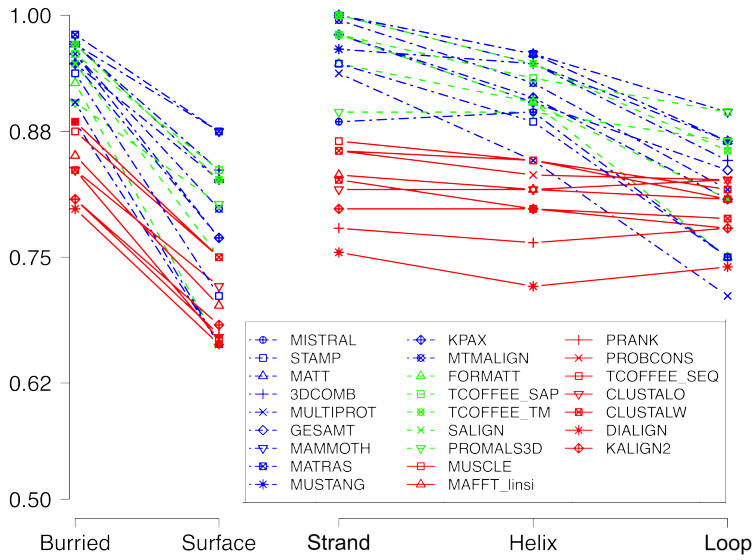
FIGURE 2.3 – (a) Résultats des tests de Friedman. Les cases au-dessus de la diagonale représentent les tests réalisés avec le score CS et celles en dessous avec le score SP. Les programmes sont classés selon leurs scores SP médians. Les cases orange et jaunes représentent des tests non significatifs et les cases bleues et vertes les tests significatifs (seuil de 5% et correction de Bonferroni pour tests multiples). Nous avons délimité 6 groupes de programmes pour lesquels aucun test n'est significatif avec aucun des programmes du groupe (carrés noirs numérotés de 1 à 6). (b) Carte des scores CS (entre 0 et 1) ordonnés selon les lignes et les colonnes avec des classifications ascendantes hiérarchiques. Les numéros devant les noms des programmes sont les numéros des carrés définis dans la figure de gauche. A la base du dendrogramme à gauche, les couleurs représentent les 3 catégories de programmes : rouge pour ceux alignant les séquences, bleu pour ceux alignant les structures et verts pour ceux alignant séquences et structures. Figures provenant de (Carpentier and Chomilier, 2019)

Observe-t-on des biais suivant les alignement de références ou la localisation résidus ? On peut s'interroger sur un biais possible des alignements de référence en faveur des programmes d'alignement de structure. En effet, les similitudes structurales sont utilisées pour construire ou valider les alignements de références. Néanmoins, les similitudes de séquences sont tout autant utilisées. Nous avons comparé les performances des programmes en fonction des banques de références et la variation la plus notable est entre les alignements vérifiés manuellement (BB2, BB3 et SISYPHUS) et les alignements automatiques (HOMSTRAD et OXBENCH). Les premiers sont plus difficiles et produisent des performances plus variables (cf. figure 2.2b). Dans tous les cas, les programmes d'alignement de structures ont pour la plupart les meilleurs résultats, quelles que soient les banques mais aussi quel que soit le pourcentage d'identité des protéines alignées (cf. figure 2.4a). Ce résultat était pour moi plutôt inattendu : je pensais *a priori* que lorsque le pourcentage d'identité est suffisamment haut les alignements purement structuraux seraient moins justes que ceux reposant sur la séquence car on compare alors des séquences proches donc souvent identiques à des structures qui, elles, sont variables même lorsque les séquences sont strictement identiques (Kosloff and Kolodny, 2008). Les programmes d'alignement de structures sont pourtant, pour la plupart aussi bons, voire meilleurs que les programmes d'alignement de séquences. La structures secondaire a par contre l'effet attendu sur la qualité de l'alignement : les résultats sont meilleurs pour les résidus enfouis ou situés dans des structures secondaires régulières, mais plus étonnamment les variations de performance sont pratiquement les mêmes quel que soit le type des méthodes d'alignement (cf. figure 2.4b).

Les *gaps* Le dernier élément que nous avons regardé est la proportion de *gaps* (uniquement les ouvertures) dans les alignements. Celle-ci est clairement différente entre les programmes alignant les séquences et ceux alignant les structures (voir figure 2.5). Les programmes basés sur la structure, à l'exception de MAMMOTH, ont tendance à surestimer le nombre d'indels et les programmes basés sur la séquence ont tendance à sous-estimer le nombre de *gaps*. Une amélioration de la gestion des *gaps* pour les méthodes basées sur la structure et sur la séquence+structure devrait améliorer leurs performances, comme cela a été souligné par ailleurs (Golubchik et al., 2007).

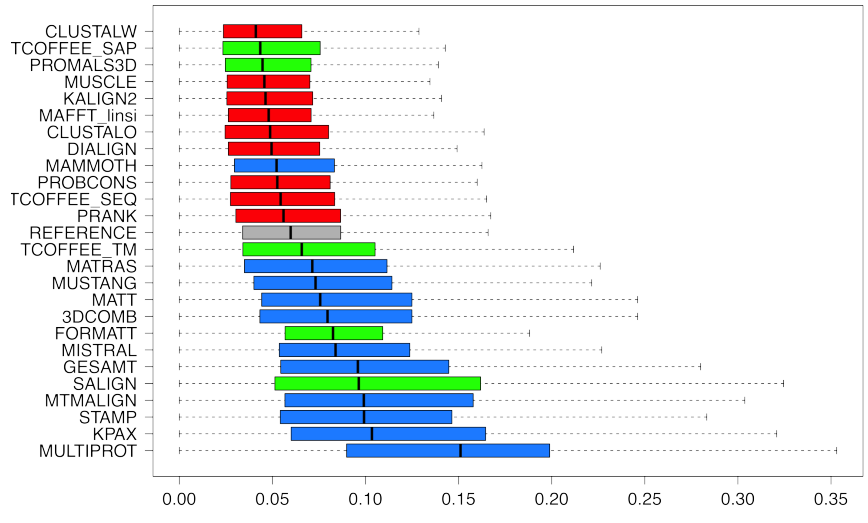


(a)



(b)

FIGURE 2.4 – (a) Scores CS médians de chaque programme en fonction du pourcentage d'identité de l'alignement de référence. (b) Scores CS médians de chaque programme selon l'enfouissement ou la structure secondaire des résidus alignés. Figures provenant de (Carpentier and Chomilier, 2019)



[!ht]

FIGURE 2.5 – Boites à moustaches de la fréquences des *gap* dans les alignements des différents programmes. Figure provenant de (Carpentier and Chomilier, 2019)

3) Conclusion

Cette étude a montré que, si plusieurs structures d'une famille sont connues, l'alignement le plus fiable est celui de la structure. Cependant, même si la proportion de familles protéiques avec au moins une structure disponible s'élève à plus de 50% (8700 sur un total de 17 425 familles ou domaines dans la base de données PFAM (El-Gebali et al., 2019)), nous avons la plupart du temps beaucoup plus de séquences que de structures disponibles et il est dommageable de se priver de ces informations sur la variabilité dans les familles. L'utilisation de méthodes basées sur les séquences et les structures avec toutes les séquences et structures connues permettrait de combiner toutes les informations disponibles. Il est néanmoins décevant que les méthodes combinant les informations provenant des structures et des séquences protéiques ne donnent pas de meilleurs résultats que celles qui se basent uniquement sur la structure. Une piste est d'analyser plus en détail les alignements posant problème aux différentes méthodes, ce que j'ai commencé à faire.

Expliciter le lien entre séquence et structure devrait permettre d'améliorer les alignements. Ce lien existe pour la plupart des protéines globulaires connues, comme nous le montre les succès de la méthode *alphafold* (Jumper and Hassabis, 2022 ; Jumper et al., 2021) pour prédire la structure à partir de la séquence. Il a été montré récemment que même lorsque les modèles structuraux prédits par *alphafold* ont une résolution basse, l'alignement multiple de ces structures est pratiquement aussi bon que l'alignement multiple des structures expérimentales (Baltzis et al., 2022).

Il est intéressant de noter que Kawabata and Nishikawa pour le programme MATRAS ont estimé empiriquement les taux de changement entre les catégories de structures secondaires et les environnements d'accessibilité aux solvants, ce qui est une modélisation empirique de l'évolution des structures. Les résultats de MATRAS sont toujours excellents bien que la méthode ait maintenant 20 ans, ce qui est un très encouragent sur l'intérêt de prendre en compte l'évolution des structures pour l'alignement. Les récents développement de modèles bayésiens combinant à la fois la séquence et la structures sont aussi très prometteurs. J'ai déjà évoqué les travaux de Golden et al. (2017) qui ont développé un modèle stochastique génératif qui permet de capturer les dépendances locales en travaillant dans dans l'espace des angles dièdres et des travaux de Larson et al. (2018) ou la divergence des coordonnées 3D des acides aminés est corrélée pour modéliser la dépendance entre positions. Des travaux encore plus récents présentent des méthodes d'alignement combinant information de séquence et de structure. Ainsi, Fallaize et al. (2020) ont développé une méthode bayésienne pour l'alignement des structures protéiques et un cadre qui devrait permettra d'inclure d'autres informations telles que la séquence d'acides aminés, et pourrait être adapté à d'autres situations telles que les protéines flexibles ou les échanges de domaines. Herman (2019) a développé une implémentation de leur modèle bayésien(Challis and Schmidler, 2012 ; Herman et al., 2014) pour l'alignement de protéines en ajoutant le module *StructAlign* dans leur programme *StatAlign* (Novák et al., 2008). Les meilleurs résultats de ces nouvelles méthodes sont à confirmer avec des *benchmark* plus complets, mais ils devraient *a priori* se confirmer. Si les temps de calculs sont raisonnables, on peut espérer que nous soyons sur la bonne piste pour combiner séquences et structures pour l'alignement, mais aussi pour l'inférence

phylogénétique.

2.3 Conclusion et perspectives

2.3.1 Autres travaux (en cours)

Pendant toutes ces années, j'ai aussi poursuivi le développement de méthodes d'alignement de séquences et de structures. Ces travaux n'ayant pas été (encore) finalisés, je ne les aborderai que rapidement.

En collaboration avec Guillaume Achaz, nous avons développé une méthode d'alignement multiple de séquences adaptée pour les séquences d'ADN génomique répétées. Les motifs répétés (assez stricts) sont d'abord identifiés et il faut ensuite les étendre. Nous avons dans un premier temps tenté de le faire avec les méthodes classiques d'alignement multiple mais les résultats n'étaient pas satisfaisants : il faut que l'alignement soit ancré sur le motif de initial ce qui n'était pas permis par ces méthodes. Nous avons alors développé une méthode d'alignement multiple exacte de séquences en utilisant le fait que nous ne voulions pas trop de *gaps* (l'alignement doit rester près de la diagonale) pour borner l'espace de recherche (*bounding*). Ce programme permet d'aligner jusqu'à 7 séquences simultanément. Il faut maintenant développer la partie permettant d'aligner ces alignements de 7 séquences pour obtenir l'alignement de plus de séquences. Ce travail a été réalisé notamment lors de l'encadrement des stages de Damien Monet (L3) en 2012 puis de Benjamin Roques (AgroParisTech) en 2015.

1) Alignement de structures

J'ai également encadré plusieurs stages et projets dont le but étaient de poursuivre le développement de Yakusa (stages de M1 d'Ophélie Foucault et André Lanrézac en 2018 et projets étudiants en 2020). Comme dans BLAST, Yakusa identifie de nombreuses similitudes locales (*Structural High Scoring Pairs*) qu'il faut ensuite filtrer et combiner. L'objectif de ces projets et stages étaient d'améliorer la complexité de la phase de combinaison pour limiter la phase de filtrage qui était jusqu'alors très drastique. Cela permet d'améliorer les résultats et un article est en cours d'écriture en collaboration avec Sophie Brouillet pour présenter cette nouvelle version.La publication sera accompagnée d'une nouvelle version du site.

J'ai aussi poursuivi le développement du programme Triades pour l'adapter à l'alignement multiple de cartes de contact. L'alignement de paires de protéines par l'alignement de leurs cartes de contacts [9] a été résolu de manière exacte (Andonov et al., 2011) mais il n'existe pas de programme d'alignement multiple pour l'alignement des cartes de contacts. L'algorithme KMR est particulièrement adapté à la résolution de ce type de problème (Crochemore and Rytter, 1991). Huiyuan Li (L2)

9. Les cartes de contacts sont des matrices $n \times n$ n étant le nombre de résidus dans la protéine contenant 1 si les deux résidus sont en contact et 0 sinon.

en 2019 et Miara Rakotomavo (M2) en 2020 ont commencé à étendre le programme pour qu'il puisse traiter ce type de données.

2) Combiner l'information de séquence et de structures

Le premier lien qui a été mis en évidence entre la séquence et la structure est l'alternance de résidus hydrophiles et hydrophobes. L'analyse de cette alternance est très riche d'information quant au repliement des protéines globulaires comme l'a montré la méthode *Hydrophobic Cluster Analysis* (Eudes et al., 2007 ; Woodcock et al., 1992). Ensuite, les alphabets structuraux ont beaucoup été explorés (voir (Joseph et al., 2010) pour une revue). Ces blocs ont été très utilisés pour prédire la structure lorsqu'aucune séquence homologue n'est connue (de Brevern et al., 2000 ; Simons et al., 1999). Avec l'augmentation du nombre de séquences, des méthodes de prédiction des contacts en fonction de la co-évolution des paires de positions ont ensuite été développées (Morcos et al., 2011) et cela a été utilisé pour la prédiction de structure (Kuhlman and Bradley, 2019).

J'ai exploré ces idées pour les adapter à l'alignement de protéines lors de deux collaborations récentes.

En collaboration avec Olivier Gascuel (DR CNRS à l'ISYEB) nous avons commencé à explorer de nouveaux scores de substitution d'un acide aminé par un autre prenant en compte l'environnement de l'acide aminé. Pour l'instant, cet environnement est uniquement représenté par les résidus voisins en séquence mais il est prévu de l'étendre aux résidus voisins en 3D. Nous avons co-encadré le stage de Pauline Turk (AgroParisTech) où un modèle Bayésien naïf a été développé et a permis de déterminer la fenêtre avec laquelle le gain d'information est le plus important. Ce gain est néanmoins assez faible et nous nous proposons à l'avenir de tester des réseaux de neurones pour capturer plus d'informations.

Je me suis aussi intéressée à l'apport des co-variations pour l'alignement en collaboration avec François Coste (CR INRIA à l'IRISA de Rennes) et Hugo Talibart (en post-doc à l'ABI sous ma supervision en 2022). La forte conservation des structures implique qu'il existe pour maintenir les repliements des protéines au cours de l'évolution des contraintes fortes sur les protéines, et pourtant la conservation des séquences peut être très faible. La conservation ne serait alors pas tant au niveau des acides aminés eux-mêmes qu'au niveau de leurs interactions – ce sont les distances et les contacts qui sont conservés. Les travaux de M. Weigt et coll. ont permis de modéliser ces contraintes à l'aide des modèles de Potts, modèle provenant de la physique statistique où ils permettent de modéliser l'interaction des spins dans des réseaux cristallins (Weigt et al., 2009). Inférés à partir d'un ensemble de séquences alignées, ces modèles comportent des paramètres permettant à la fois de modéliser la conservation des acides aminés aux différentes positions mais aussi les corrélations entre les positions : ils permettent de prendre en compte les interactions des acides aminés. François Coste et Hugo Talibart ont développé PPAlign, un moteur performant d'alignement de paires de modèles de Potts (Talibart and Coste, 2021). L'objectif de ce projet était d'adapter cette méthode à l'alignement multiple de protéines et l'annotation. Nous avons adapté la méthode d'inférence des modèles pour qu'elle soit plus rapide, modélise au mieux les insertions et délétions et soit robuste au biais d'échantillon-

nage des séquences alignées. Un article est en cours de rédaction. Il faudra ensuite tester le potentiel de cette nouvelle méthode pour l'alignement multiple, pour la recherche d'homologie lointaine et la caractérisation fonctionnelle des protéines alignées.

CHAPITRE 3

Quel est l'effet d'une mutation dans les structures ?

3.1 Motivations

Dans le cadre de l'évolution des protéines au sein d'une même famille structurale (*microtransition* dans Jayaraman et al. 2022), les mécanismes étudiés sont principalement les substitutions d'acides aminés et dans une moindre mesure les insertions et délétions (indels), bien que tous les autres mécanismes dont nous avons parlé précédemment puissent aussi être importants. Mon objectif est ici d'essayer de comprendre comment la structure se déforme au cours de l'évolution au sein d'une même famille structurale mais il m'a semblé important d'abord de me replacer dans le cadre plus général des connaissances que nous avons accumulées sur les mutations, leur fréquence, leur distribution, leurs effets, et les pressions de sélection qui sont à l'œuvre. Je n'aborderai ici que les substitutions d'acides aminés et la co-évolution. Les insertions et délétions seront plutôt abordées dans mon projet en conclusion de cette partie recherche. Comme le sujet est ici les protéines, je n'évoquerai pas le cas des mutations silencieuses et des mutations dans l'ADN non-codant.

Les mutations se produisent au hasard partout dans l'ADN selon de divers mécanismes, mais ensuite, la sélection joue un rôle majeur et change beaucoup leur distribution. Dans une protéine, la substitution d'un acide aminé par un autre modifie la molécule (certains atomes sont différents), mais les conséquences sur ses propriétés biologiques - concept qui regroupe à la fois l'activité chimique et la stabilité physique mais aussi les autres caractéristiques qui sont soumises à une pression sélective - sont diverses et variables. Il est intéressant dans un premier temps de tenter d'évaluer quelles substitutions sont possibles en fonction de leur localisation dans la molécule. II semble assez évident que toutes les substitutions ne sont pas observées partout car certaines sont purgées extrêmement rapidement. Il est par exemple très peu probable qu'un acide aminé hydrophobe comme le tryptophane soit remplacé par en un acide aminé hydrophile polaire comme l'aspartate. Il est par contre beaucoup plus difficile d'évaluer ce qui est possible : quels acides aminés sont possibles à un site donné ? Tous les sites d'une protéine peuvent-ils muter ? À quel point l'épistasie joue-t-elle ? Existe-t-il des mutations "neutres" ? De très nombreuses études ont été publiées soulignant l'impact fort de certaines mutations sur tout le panel des propriétés biologiques des protéines : sur leur mécanisme de repliement (*folding*) (Choe et al., 2000 ; Fersht et al., 1992 ; Lindberg et al., 2006 ; Matthews and Fersht, 1995 ; Pilipczuk et al., 2017 ; Viguera et al., 1995), sur leur structure (Alexander et al., 2009 ; Religa et al., 2005) et sur tout le réseau de liaisons fortes et faibles qu'entretiennent les atomes entre eux (Zhou et al., 2007), sur la stabilité de la molécule (Gromiha and Sarai, 2010) et sa dynamique (Clarkson and Lee, 2004), sur sa propension à former des agrégats (Ahn et al., 2016 ; Buell et al., 2011), sur ses affinités avec d'autres partenaires (Ray et al., 2022) et plus largement sur sa fonction biologique, le phénotype et la fitness qui en dépendent (Bobay et al., 2010 ; Gong et al., 2009). Les travaux cités ici présentent des cas où les effets sont importants, et pourtant, la majorité des mutations sont considérées comme neutres ou quasi-neutres. Ce concept de neutralité des mutations m'a longtemps posé problème ; il dépend en fait simplement du cadre dans lequel on se place.

3.1.1 Caractérisation des mutations observées

Les expériences de *deep mutational scanning* mesurent de manière systématique l'effet d'un très grand nombre de substitutions sur l'activité d'une protéine (Chen et al., 2020 ; Fowler et al., 2014 ; Newberry et al., 2020) ou ses affinités de liaison avec un partenaire (Heyne et al., 2021). La proportion de substitutions délétères vis-à-vis de l'activité est de 33 à 40% suivant les protéines testées (Guo et al., 2004 ; Bershtein et al., 2006 ; Camps et al., 2007 ; Smith and Raines, 2006). Ce taux est déjà important mais il est encore supérieur *in vivo* où la tolérance aux substitutions observée dans les protéines homologues est bien plus faible (Rockah-Shmuel et al., 2015 ; Boucher et al., 2016 ; Young and Fields, 2015). Les effets délétères principaux identifiés sont une altération du repliement et/ou une diminution de la stabilité. Les mutations qui altèrent la fonction biochimique sont probablement purgées encore plus intensément et sont donc très peu observées (Firnberg et al., 2014 ; Rockah-Shmuel et al., 2015). Le taux d'évolution ou nombre de substitutions (fixées) par unité de temps évolutive varie suivant les protéines (Zhang and Yang, 2015), entre les sites d'une même protéine (Echave et al., 2016) et même au cours des périodes d'évolution (Lopez et al., 2002). Le taux d'évolution global d'une protéine dépend principalement de son niveau d'expression et non l'importance fonctionnelle de la protéine (Zhang and Yang, 2015). Cela serait principalement dû aux contraintes sur l'usage du code et la séquence protéique elle même pour éviter les mauvais repliements de protéines (*misfolding*) qui sont très couteux du point de vue métabolique et toxiques lorsqu'elles s'agrègent Drummond et al. 2005). Ce type de distribution des taux d'évolution peut être reproduit avec des modèles de protéines en lattice (Yang et al., 2010 ; Lobkovsky et al., 2010) utilisés pour étudier les mécanismes de repliements (*folding*) des protéines. Il semble aussi que les protéines de structure plus « robustes »– ayant un coeur hydrophobe important ou ayant une forte densité de contacts entre résidus – évoluent plus rapidement (Bloom et al., 2006 ; Franzosa and Xia, 2009, 2012 ; Scherrer et al., 2012 ; Faure and Koonin, 2015). La variabilité entre les sites d'une même protéine dépend principalement de la structure, avec les résidus enfouis, localisés dans les zones denses et rigides dans mutant peu et les les résidus exposés, dans les zones flexibles et peu denses mutant beaucoup plus. Des contraintes dues à la fonction peuvent aussi être isolées : les résidus des sites actifs ou voisins de ceux ci évoluent plus lentement. Les expériences de *deep mutational scanning* ont par contre montré que toutes les positions peuvent muter : le nombre de positions où toute substitution ou presque est fortement délétère est très faible (McLaughlin et al., 2012 ; Firnberg et al., 2014). Par exemple, dans la TEM-1 $\beta - lactamase$ d'*Escherichia coli*, seules 8 positions sur 264 sont très réfractaires à toute substitution, 4 positions du site actif et 4 positions très conservées (Firnberg et al., 2014). L'épistasie joue un rôle très important. La substitution d'un acide aminé par un autre peut être neutre ou bénéfique dans une séquence et délétère dans une autre (Kondrashov et al., 2002 ; Lehner, 2011 ; Weinreich et al., 2005). Ainsi, 90% des substitutions possibles ne sont neutres ou bénéfiques que dans les séquences où elles sont observées. Des que les séquences diffèrent un peu, et que donc l'environnement en acides aminés change, ces mêmes substitutions deviennent délétèrent (Breen et al., 2012). Les expériences d'accumulation progressives de mutations réalisées montrent aussi que la très grande majorité des sub-

stitutions (67 à 80%) deviennent rapidement délétères (Rockah-Shmuel et al., 2015 ; Boucher et al., 2016). De même, les réversions deviennent aussi très rapidement improbables à cause de l'épistasie (McCandlish et al., 2016). Il est néanmoins aussi possible qu'une mutation délétère puisse être fixée si une ou des mutations précédentes le permettent (mutations permissives ou habilitantes) ou que des mutations ultérieures la compensent (mutations compensatoires). Le modèle de mutation permissives et compensatoires a particulièrement été utilisé pour modéliser la variation de stabilité ($\Delta\Delta G$) lors de l'évolution des protéines (Tokuriki and Tawfik, 2009). Néanmoins, les protéines globulaires présentent en général une épistasie négative quant à l'effet sur la stabilité : les effets délétères de deux mutations différentes sont supérieurs à la somme des effets individuels (Bershtein et al., 2006). Ceci est cohérent avec le faible nombre de chemins mutationels (séries ordonnées de mutations) existant par exemple pour acquérir une meilleure résistance à un antibiotique (Weinreich, 2006). L'espace des séquences est donc visiblement très contraint par la structure

L'identification de corrélations des substitutions entre les sites des protéines a fait l'objet de très nombreuses études (Göbel et al., 1994 ; Neher, 1994 ; Shindyalov et al., 1994 ; Taylor and Hatrick, 1994 ; Pazos et al., 1997 ; Korber et al., 1993 ; Chelvanayagam et al., 1997 ; Pollock et al., 1999 ; Pritchard et al., 2001 ; Tuff and Darlu, 2000 ; Fukami-Kobayashi et al., 2002 ; Govindarajan et al., 2003 ; Dutheil et al., 2005 ; Dutheil and Galtier, 2007 ; Dutheil, 2012 ; Avino and Poon, 2019 ; Colavin et al., 2022 ; Chaurasia and Dutheil, 2022). Ces corrélations ont permis de beaucoup améliorer la prédiction des structures secondaires et tertiaires des protéines (Weigt et al., 2009 ; Taylor et al., 2013 ; Stein et al., 2015) et des partenaires d'interaction des protéines (Pazos and Valencia, 2001 ; Goh et al., 2000 ; Jothi et al., 2006 ; Li et al., 2022). Ceci est probablement dû au fait que 75% des résidus co-évoluant sont à moins de 5Å et seulement 3% à plus de 15Å (Anishchenko et al., 2017). Néanmoins, la co-évolution s'avère plutôt faible entre la majorité des sites (Chaurasia and Dutheil, 2022). Le signal le plus fort provient de sites dans l'empilement des acides aminés dans les hélices α (Pollock et al., 1999) et de co-variations de compensation de charge (Chelvanayagam et al., 1997). D'après McLaughlin et al. (2012), seuls 20% des acides aminés co-évolueraient et seraient organisés en quelques « secteurs » qui seraient surtout le résultat d'une adaptabilité fonctionnelle aux variations environnementales (McLaughlin et al., 2012 ; Halabi et al., 2009 ; Lockless and Ranganathan, 1999 ; Süel et al., 2003 ; Chakrabarti and Panchenko, 2009).

Pour résumer, sur une chaîne protéique, les mutations que nous observons sont loin de représenter toutes les mutations possibles, mais pratiquement tous les sites peuvent muter. Les substitutions acceptables dépendent de la localisation et une contrainte forte est la conservation des propriétés physico-chimiques locales de la zone : encombrement stérique, hydrophobie/hydrophilie, stabilité, *etc.*, ce qui permet de conserver la structure de la protéine. Les mutations sont souvent neutres dans un certain contexte (pour une séquence donnée) et délétères dans d'autres, et ainsi les réversions peu fréquentes. Pourtant, le nombre de résidus pouvant être identifiés comme co-évoluant est finalement assez faible : quelques résidus co-évolueraient et formeraient un petit nombre de réseaux. Ces réseaux seraient surtout importants pour la fonction et l'adaptabilité aux changements environnementaux. Ce

seraient eux que les méthodes d'analyse de couplage et de prédiction de contact identifient, en plus des quelques cas bien répertoriés où des co-évolutions sont avérées comme le cas des compensation de charge.

Mon objectif est maintenant de tenter de comprendre quel est l'effet sur la structure de ces mutations et s'il est possible de le modéliser.

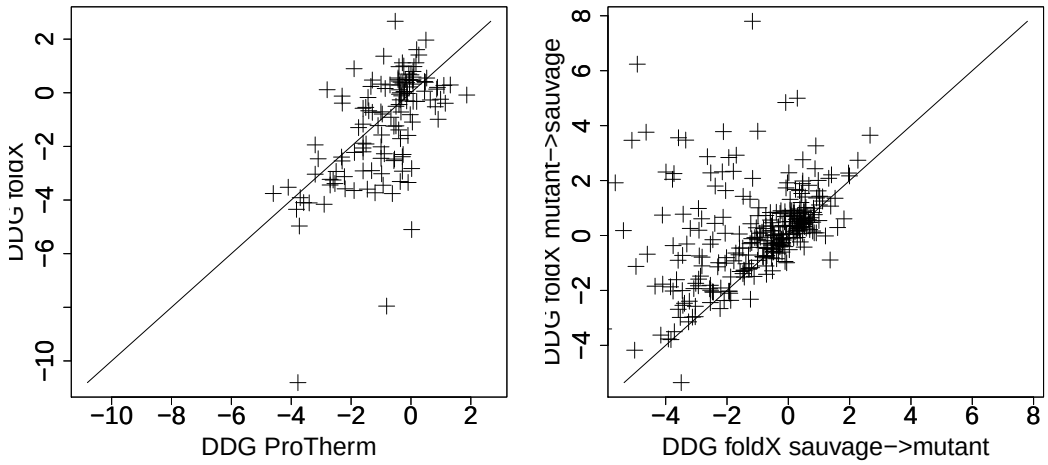
3.1.2 Effet des mutations sur la stabilité

Les protéines sont "marginalement" stables : il y a typiquement une différence d'énergie libre de repliement (ΔG) de 3 à 7 kcal / mol entre les conformations repliées et dépliées (Privalov, 1979). Cette faible stabilité des protéines est supposée être soit le résultat d'un équilibre entre la fonction et la stabilité (DePristo et al., 2005), soit le résultat d'un équilibre entre les mutations déstabilisantes et la faible stabilité des protéines (Taverna and Goldstein, 2002b,a ; Zeldovich et al., 2007 ; Wylie and Shakhnovich, 2011 ; Bloom et al., 2007 ; Liberles et al., 2012). Les substitutions ont alors un effet non négligeable sur la stabilité des protéines : l'effet d'une seule mutation est en moyenne de -1,68 kcal / mol pour les mutations déstabilisantes et de 1,33 kcal / mol pour les mutations stabilisantes selon la base de données Protherm [10] (Gromiha and Sarai, 2010 ; Nikam et al., 2021), qui rassemble les mesures de $\Delta\Delta G$ pour plusieurs milliers de protéines avec leur structure. Le $\Delta\Delta G$ est l'écart entre les variations d'énergie libre de repliement de la protéine mutée et la protéine native dans Protherm.

De nombreux algorithmes et serveurs web ont été développés pour fournir une estimation de la variation de l'énergie libre de Gibbs ($\Delta\Delta G$) sous l'effet d'une mutation ponctuelle : FoldX (Guerois et al., 2002) et Rosetta (Kellogg et al., 2011) en sont parmi les plus connus. SPROUTS (Lonquety et al., 2009) est lui un serveur web combinant les résultats de plusieurs méthodes. Ces méthodes tentent de prédire si une mutation donnée va être déstabilisante, neutre ou stabilisante. La comparaison entre la variation d'énergie prédite et expérimentale donne globalement des résultats satisfaisants (cf figure 3.1 (a)) : la corrélation entre les prédictions de FoldX et les mesures de Protherm est de 0,59. Néanmoins, dans certains cas l'écart peut être assez important (Lonquety et al., 2009). Il est intéressant de noter que les résultats des prédictions donnent des résultats assez différents selon que l'on va du sauvage vers le mutant ou du mutant vers le sauvage au signe près (cf. figure 3.1 (b)) . De plus, dès que les mutations s'accumulent, les résultats sont beaucoup plus mitigés (Barnes et al., 2022).

Dans la plupart des méthodes, les déformations du squelette ne sont pas prises en compte ; seules les chaînes latérales sont autorisées à bouger. L'amplitude des mouvements des chaînes latérales est extraite des bibliothèques de rotamères et leurs degrés de liberté dépendent fortement de la conformation du squelette (Levitt et al., 1997 ; Levitt, 1997 ; Dunbrack, 2002 ; Gautier and Tufféry, 2003). La prise en compte de la déformation du squelette devrait améliorer notre compréhension des conséquences de la substitution. Seuls deux modèles ont été définis pour l'accommodation du squelette sous l'effet d'une substitution d'acides aminés. Le premier est le mouvement *backrub* (Lauck et al., 2010),

10. https://web.iitm.ac.in/bioinfo2/prothermdb/Statistics.html



(a) (b)

FIGURE 3.1 – (a) prédiction de $\Delta\Delta G$ par le programme FoldX pour 130 protéines présentes dans Protherm et appartenant aux 11 familles pour lesquelles au moins 20 structures de mutants ponctuels différents sont connus. (b) prédiction par FoldX du $\Delta\Delta G$ pour les familles dans lesquelles au moins 20 structures de mutants ponctuels différents sont connues. En abscisse la valeur prédite de $\Delta\Delta G$ de la native vers le mutant, et en ordonnée la valeur prédite de $\Delta\Delta G$ du mutant vers la native. Figures provenant de (Carpentier and Chomilier, 2021)

dérivé de l'observation des positions alternatives des atomes de la chaîne latérale dans les structures cristallographiques à ultra-haute résolution. Un mouvement *backrub* peut être représenté par une rotation autour des angles pseudo-dièdres de trois peptides adjacents, sans effet sur le reste de la protéine. Il a été utilisé avec succès pour améliorer le calcul du $\Delta\Delta G$ de Rosetta (Smith and Kortemme, 2008). Les déformations du squelette de la protéine sont donc influencées par les conformations des chaînes latérales de manière prévisible (Davis et al., 2006) mais la conformation de la chaîne latérale dépend du squelette (Dunbrack and Karplus, 1994). Un autre modèle a été conçu par Bordner and Abagyan (2004) à partir d'une base de données de 2141 paires de structures protéiques ne différant que par une seule mutation ponctuelle. Le squelette est d'abord transformé en une géométrie idéalisée, et après l'introduction de la mutation, les angles dièdres χ des chaînes latérales sont minimisés. Ce modèle a également été utilisé avec succès pour améliorer la prédiction de l'énergie libre de Gibbs après une mutation. Il n'existe donc à ma connaissance que deux modèles permettant de modéliser la déformation induite par une mutation ponctuelle, modèles assujettis à la prédiction de l'effet sur la stabilité. La déformation structurale induite par une substitution, une insertion ou une délétion est donc assez peu étudiée en tant que telle, et ce point m'a semblé important pour l'évolution et l'alignement des structures protéiques, pour comprendre comment les structures se déforment.

3.1.3 Effet des mutations sur le squelette

L'effet des mutations sur la structure globale est faible. Il a été montré que 50 à 80% des acides aminés (sites) peuvent être substitués par un autre sans altération significative de la structure de la protéine (Sander and Schneider, 1991 ; Shakhnovich and Gutin, 1991 ; Schaefer and Rost, 2012). De nos jours, la résolution de la structure des protéines déterminée par cristallographie et diffraction des rayons X est de l'ordre de 1Å, ce qui signifie que la précision des structures obtenues est de l'ordre du 1/10 d'Å(Luzzati, 1952). Le déplacement des atomes du squelette causé par une substitution est typiquement aussi du 1/10 d'Å, ce qui nous place à la limite de la résolution. La première question que nous nous sommes posée a alors été si l'effet sur la structure d'une substitution est mesurable avec nos méthodes expérimentales.

J'ai d'abord participé à un travail conjoint avec l'équipe d'Irena Roterman (Cracovie, Pologne, financé par PHC Polonium en 2012 et 2013) qui visait à identifier les acides aminés qui sont au coeur du repliement des protéines et qui sont particulièrement bien conservés (Banach et al., 2015). Une des familles étudiées étaient les lysozymes pour lesquels les structures d'un très grand nombre de mutants sont disponibles. Dans cette famille, les variations structurales les plus importantes entre la structure sauvage et les mutants sont toujours localisées au même endroit, au niveau de deux boucles particulièrement flexibles et ce quelle que soit la localisation de la mutation : ces variations structurales sont-elles réellement dues aux mutations ? Nous avons alors voulu savoir si les variations structurales induites par une simple substitution étaient mesurables dans les structures résolues par cristallographie et rayons X ou si ce que nous observions n'était que des variations normales dues à la flexibilité intrinsèque de la molécule et aux changements de conditions expérimentales. Ce travail a été réalisé avec J. Chomiler (DR CNRS, IMPMC) et S. Shanthirabalan, ingénieure d'étude dont j'ai encadré le travail et qui a été financée par l'ANR TempoMut (ANR-12-JSV7-0007) dont j'étais partenaire. Nous avons montré que les variations dans les structures sont plus fortes lorsque les conditions expérimentales changent que lorsqu'il y a une substitution, mais qu'il y a bien des variations au niveau de ces mutations. De plus, si on prend en compte la flexibilité de la molécule, on se rend alors compte que de très faibles variations au niveau des résidus enfouis ou dans des structures secondaires régulières sont en fait comparativement plus importantes (Shanthirabalan et al., 2018) que dans les regions très flexibles. Nous avons ensuite montré que la méthode que nous avons développée permet aussi d'identifier des variations structurales distantes des mutations (Carpentier and Chomilier, 2020). Ces travaux ont aussi donné lieu à deux chapitres de livre : (Carpentier and Chomilier, 2021) et (Carpentier and Chomilier, ress). Cette méthode permet donc d'identifier quels sont les acides aminés dont la position dans la structure est anormalement perturbée suite à une mutation. Elle nécessite d'avoir les structures du "mutant" et du "sauvage" (structure de référence) car ce n'est pas une méthode prédictive. Son inconvénient est qu'actuellement elle nécessite un nombre suffisant de structures (une vingtaine) pour estimer la variabilité structurale normale. Nous avons encadré un stage de master en 2020 (Nicolas Dibot, Master de bioinformatique de Nantes) pour pallier cet inconvénient en estimant la variabilité avec les modes normaux. Ces travaux sont en cours de finalisation.

3.2 Contributions

3.2.1 Développement d'un méthode pour identifier les mutations perturbantes

Nous avons rassemblé 11 familles protéiques comportant au moins 20 mutants avec une seule substitution par rapport à une structure de référence, pour un total de 580 mutants et 11 structures de références (Shanthirabalan et al., 2018). Ces 11 familles sont présentées table 3.1.

TABLE 3.1 – Les 11 familles étudiées

Protein	#Members	Reference	Length	Class	organism
T4 lysozyme	147	1lw9A	164	Alpha	Phage T4
Human lysozyme	124	2nwdX	130	Alpha	H. sapiens
Transferase	78	2dekA	265	Alpha beta	P. Horikoshii
Anhydrase II	69	2iliA	255-260	Alpha beta	H. sapiens
Staphylococcal nuclease	31	1ey0A	149	Beta	S. aureus
Catalase HPII	29	4bflA	753	Multi-domain	E. coli
Ribonuclease A	25	2e3wA	124	Alpha beta	B. taurus
Hen lysozyme	24	2vb1A	129	Alpha	G. gallus
Transthyretin	22	4fi8A	126-127	Beta	H. sapiens
Reaction centre	22	2j8cM	302-314	Alpha beta	R. sphaeroides
Benzoylformate decarboxylase	20	5deiA	524-536	Alpha beta	P. putida

Pour identifier finement l'effet sur la structure (ici le squelette peptidique) d'une substitution, nous avons calculé des variations structurales locales du squelette entre la structure de référence et chaque structure des protéines mutantes. Nous avons pris en compte uniquement les positions des C_α car les tests que nous avons effectués avec tous les atomes du squelette ne donnaient pas de meilleurs résultats. Les deux protéines (mutante et sauvage) comparées ont des séquences extrêmement proches (substitution d'un seul acide aminé et souvent des acides aminés manquant aux extrémités), il est donc aisé de trouver l'alignement (la correspondance) entre acides aminés des deux protéines. Ensuite, les deux structures (en vert et en rouge figure 3.2) sont superposées pour minimiser les distances entre les paires de C_α en correspondance. Nous calculons alors les distances entre les C_α des deux structures pour de petits fragments (en bleu figure 3.2) en calculant le RMS (Root Mean Square deviation). Une longueur de 3 résidus a été choisie car c'est celle pour laquelle l'effet local mesuré est le plus fort. Nous appelons ces RMS "RMSnG" (RMS3G par exemple), pour rappeler qu'ils résultent d'une superposition globale et qu'ils sont calculés sur n résidus. Il est aussi possible de ne superposer localement que ces n C_α, et de calculer un "RMSnL" après cette superposition locale. La taille optimale des fenêtres est alors de 7 et non de 3 et nous nommerons alors ce RMS RMS7L.

Les RMSn(G/L) ont comme attendu des valeurs très faibles (en moyenne de 0,25Å$\pm$0,25 pour les

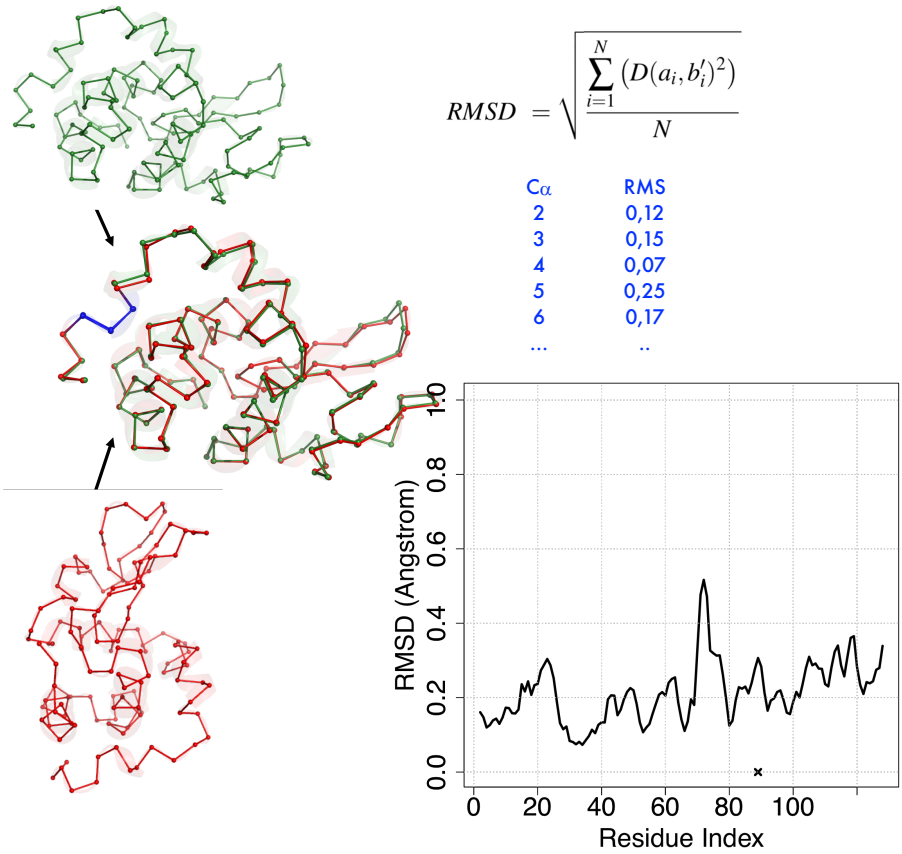


FIGURE 3.2 – Gauche : superposition de deux structures de lysozyme. Leur écart est mesuré par le RMS glissant calculé pour les fragments de trois résidus successifs, par exemple le fragment en bleu. Le RMS (formule en haut) est la racine de la moyenne de la somme des carrés N distances D entre les paires de C_α a_i et b_i', (b_i' représente les $C_\alpha b_i$ après superposition). N vaut dans cet exemple. Ce calcul est effectué pour tous les fragments possibles ; il y a donc autant de RMS que de résidus dans la protéine, sauf pour les 2 $C_\alpha$aux extrémités. Si on trace le graphique des RMS le long de la séquence, obtient un profil (en bas à droite). La mutation est localisée à l'endroit de la croix (protéine 2hef chaîne A, mutation I89A). Figures provenant de (Carpentier and Chomilier, 2021)

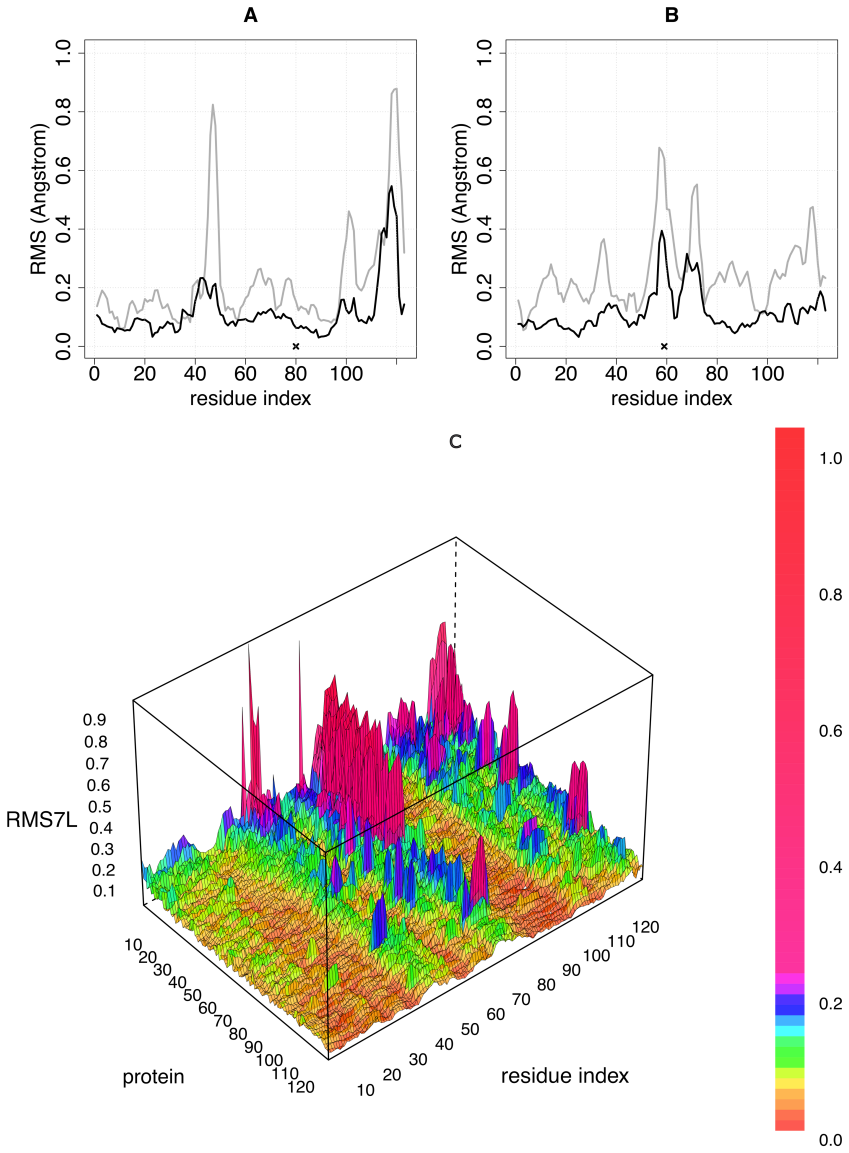


FIGURE 3.3 – RMS3G et RMS7L des deux protéines 1b5x chaîne A (A) et 2hee chaîne A (B). Les superpositions sont soit locales (ligne noire, RMS7L) soit globales (ligne grise, RMS3G). La croix indique l'emplacement de la mutation et correspond à un déplacement maximal au site de la mutation dans la figure (B) mais pas dans la figure (A). (C) Représentation en 3D des RMS7L à chaque position pour toutes les chaînes de la famille du lysozyme humain. Le RMS7L s'étend jusqu'à 2,15Å mais l'axe z a été coupé à 1Å. Figures provenant de (Carpentier and Chomilier, 2020)

RMS3G par exemple). Par contre, les plus grandes valeurs pour un profil sont rarement à l'endroit de la mutation (cf. les deux profils figure 3.3). Si l'on compare les profils des protéines d'une même famille, on observe que les régions avec les plus grands RMS sont souvent les mêmes, quelle que soit la position de la mutation et que certaines protéines « bougent » globalement très peu alors que d'autres « bougent » beaucoup (cf. figure 3.3 en bas). Il a par ailleurs été montré que même pour des séquences identiques, les structures des protéines varient souvent, principalement en raison de différentes interactions protéine-protéine, de différentes interactions ligand ou protéine-ligand ou de différents solvants (Kosloff and Kolodny, 2008 ; Vetrivel et al., 2019). Il est alors difficile de savoir si un grand RMSn(G/L) est dû à la mutation, à la flexibilité intrinsèque de la protéine ou à une variabilité globale résultant des diverses conditions de cristallisation. Pour isoler l'effet dû aux mutations, il faut neutraliser les deux autres sources de variabilité.

Considérer les rangs à la place des valeurs est une transformation robuste utilisée dans de nombreux tests statistiques. Les RMSn(G/L) de chaque profil sont d'abord classés par ordre croissant puis les rangs sont divisés par le nombre de RMSn(G/L) dans le profil (c'est-à-dire la longueur de la chaîne). On obtient alors des p-valeurs empiriques, sans dimension. Si les mutations n'avaient pas d'effet particulier sur le RMSn(G/L), la distribution de ces p-valeurs devrait être uniforme, ce qui n'est pas observé : les p-valeurs des sites mutés sont largement en dessous de la diagonale dans la figure 3.4 (points noirs). Elles sont donc plus petites qu'attendu par hasard, alors que le même nombre de p-valeurs calculées à partir de résidus non mutés tirés au hasard dans notre échantillon est lui bien proche de la diagonale (courbes grises de la première courbe "All"). Par contre, pour les régions flexibles ou exposées, les p-valeurs des sites non mutés ont tendance a être plus grandes (figure 3.4, courbes grise des graphiques "exposed" et "loops"). Néanmoins, si une mutation a lieu dans une région flexible ou exposée, le RMS a toujours tendance a être plus grand qu'attendu par hasard.

Cette première transformation permet de prendre en compte la variabilité expérimentale mais pas la flexibilité intrinsèque de la molécule. En effet, les RMSn(G/L) des régions très flexibles seront toujours plus grands que ceux des régions plus rigides, ce qui se retrouve aussi dans les p-valeurs empiriques. Pour neutraliser cet effet, nous avons effectué alors à nouveau classé les p-valeurs empiriques de chaque position (site) dans chaque famille. Nous obtenons une nouvelle p-valeur empirique qui est alors nommée *p-rank* pour la différencier de la première. Cette nouvelle valeur permet bien de neutraliser l'effet des variabilités expérimentales et aussi celui de la flexibilité intrinsèque aux protéines d'une même famille. Une manière de s'en rendre compte est de prendre les 5% des plus grands RMS3G, et des 5% des plus petites p-valeurs empiriques et *p-rank* et de regarder si les sites extraits sont les lieux des mutations. Parmi les 5% des plus grands RMS3G, on trouve 12% qui sont centrés sur une mutation ; avec les 5% des meilleures p-valeurs empiriques, on en trouve 15% et avec les 5% des meilleurs *p-rank* 25%. Ces deux transformations permettent donc bien de mieux isoler les régions localement perturbées par une mutation et 1/4 de positions les plus perturbées correspondent aux positions mutées. Nous avons aussi montré que même si les RMS3G sont plus grands dans les boucles et les positions exposées, l'effet relatif d'une mutation est plus important dans les brins et les positions

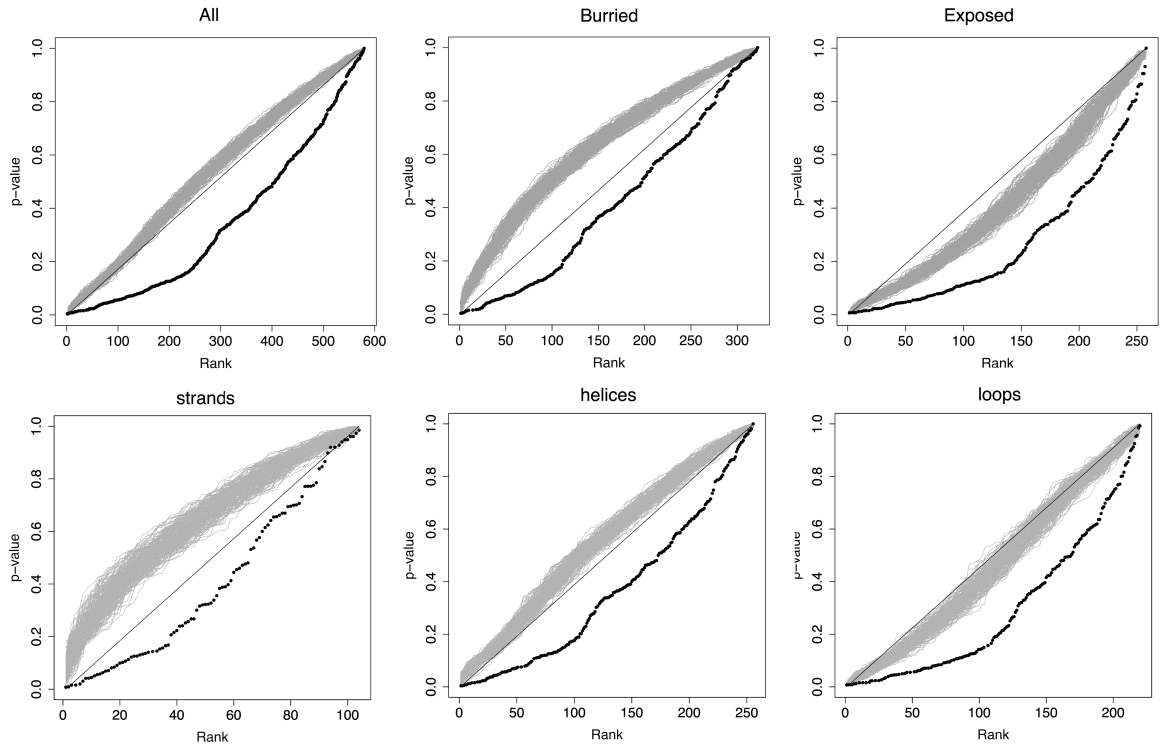


FIGURE 3.4 – Courbes de la p-valeur empirique pour RMS3G en fonction de leur rang, pour les mutations (en noir) et la distribution aléatoire (en gris clair) avec dans l'ordre : toutes les mutations, celles survenant dans des résidus enfouis, exposés, dans des hélices alpha des feuillets beta et des boucles. Les distributions aléatoires ont été générées 200 fois en tirant au hasard des positions dans les mutants parmi les positions non mutées. Figures provenant de (Shanthirabalan et al., 2018).

enfouies si l'on considère plutôt le *p-rank*. Nous avons aussi confirmé l'absence de corrélation entre le RMS et les variations d'énergie libre mais trouvé une petite corrélation entre de grands RMSn(G/L) et les erreurs de prédiction de $\Delta\Delta G$ (Shanthirabalan et al., 2018).

3.2.2 Application de la méthode pour identifier des effets longue portée

Avec cette méthode, nous identifions donc des sites potentiellement perturbés structuralement par une substitution mais ceux-ci sont donc souvent différents du site de mutation lui-même. Nous nous sommes intéressés ensuite à la caractérisation de ces sites perturbés dans la familles des lysozymes humains car c'est la famille ayant le plus de structures mutantes disponibles (122 mutants différents au moment de l'étude) (Carpentier and Chomilier, 2020).

Nous avons montré que toutes les mutations n'ont pas les mêmes effets. Les résidus ont été regroupés dans 3 catégories : hydrophobe (F, I, L, M, V, W, Y), boucle (D, G, N, P, S) et autre (A, R, C, Q, E, H, K, T) (Wojcik et al., 1999). Les mutations ayant le plus d'effets sont celles qui substituent un résidu autre vers hydrophobe ou boucle et hydrophobe vers autre. Nous avons aussi sélectionné pour chaque mutant les 5% de résidus les plus perturbés, c'est-à-dire ceux ayant les valeurs de *p-rank* les plus faibles. Nous avons montré que la distribution des distances entre le résidu muté et les 5% de résidus les plus perturbés est exponentielle décroissante avec la distance. Nous avons ensuite croisé les résidus sectionnés avec toutes les perturbations potentielles évidentes : le résidu muté, ceux en contact avec celui-ci, ceux qui sont en contact avec un ligand absent ou différent dans la structure de référence. Ces résultats sont illustrés et résumés figure 3.5. Un tiers des résidus perturbés sont en contact direct avec la position mutée ($(42+241)/899$) et moins de 0,5% sont en contact avec un ligand ; les deux tiers restants sont des effets longue portée potentiels (cf. Table 3.2). Nous montrons aussi que 34% des résidus mutés sont perturbés et 20% des résidus en contact avec les résidus mutés le sont aussi.

TABLE 3.2 – Évaluation des déplacements significatifs en nombre et en pourcentage pour le RMS3G et le RMS7L, restreints aux 5% les plus importants des 3 méthodes. La première ligne (ALL) indique le nombre de positions dans l'ensemble des données qui sont soit mutées, soit en contact avec le résidu muté ou avec un ligand, puis la somme des trois colonnes précédentes, "autre" est le nombre restant de positions et nous avons enfin la somme totale des sites. Pour le RMS3G et le RMS7L, les mêmes nombres et leurs proportions sont rapportés, limités aux 5% supérieurs des 3 méthodes. Le nombre total de résidus sélectionnés peut différer (dernière colonne) en raison des ex aequo. Table extraite de (Carpentier and Chomilier, 2020)

RMS	Transformation	Mutated		Contact		Ligand		Sum		Other		Total	
ALL	None	122	100%	1222	100%	149	100%	1493	100%	13635	100%	15128	100%
RMS3G	RMS	3	2%	31	3%	0	0%	34	2%	723	5%	757	5%
	p-value	25	20%	107	9%	13	9%	145	10%	516	4%	661	4%
	p-rank	36	30%	205	17%	3	2%	244	16%	624	5%	868	6%
RMS7L	RMS	2	2%	31	3%	1	1%	34	2%	723	5%	757	5%
	p-value	20	16%	113	9%	4	3%	137	9%	595	4%	732	5%
	p-rank	41	34%	241	20%	3	2%	285	19%	614	5%	899	6%

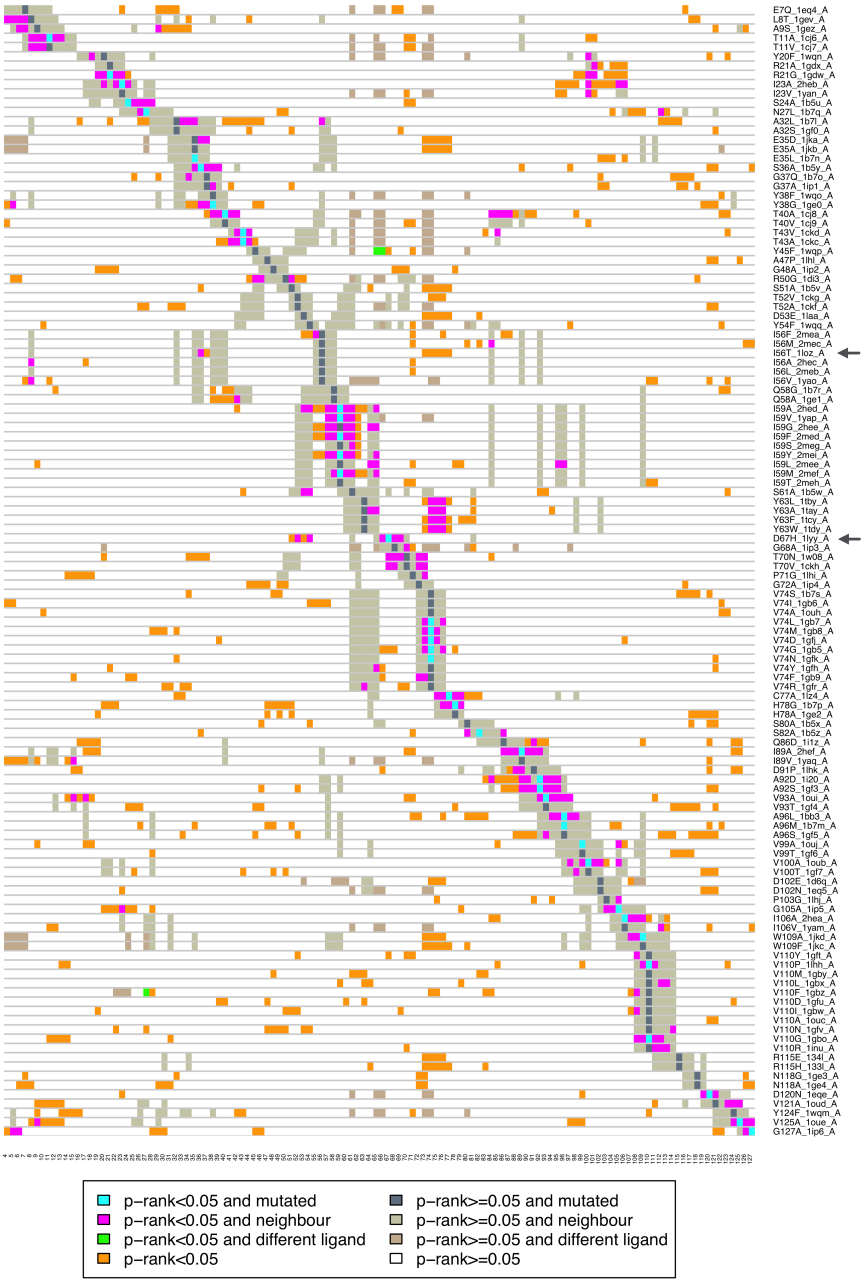


FIGURE 3.5 – RMS7L *p-rank* pour les 122 mutants de la famille des lysozymes humains (en vertical) en fonction de la séquence (en horizontal). Les couleurs les plus vives (orange, cyan, magenta et vert) mettent en évidence les positions parmi les 5% de valeurs *p-rank* les plus faibles. Les autres résidus potentiellement perturbés (mutés, en contact avec la mutation ou en contact avec un ligand absent ou différent dans la structure de référence) sont en gris. Les flèches indiquent les deux protéines amyloïdes, avec les mutations I56T et D67H. On peut remarquer que lorsqu'il y a différentes substitutions à une même position elles produisent des déplacements dans les mêmes régions. Par exemple, les mutations aux positions 21 et 23 entraînent un déplacement important à longue distance, sur le fragment 95 à 106. Autre exemple, les quatre mutations aux positions 21 et 23 (R21A, R21G, I23A et I23V) influencent principalement la structure de l'extrémité N-terminale de l'hélice alpha qui est en face des mutations. Figure provenant de (Carpentier and Chomilier, 2020)

Pour valider la pertinence de ces sélections nous avons croisé nos résultats avec les nombreuses structures résolues et analysées par les chercheurs de l'Institut de recherche sur les protéines de l'Université d'Osaka, en particulier par K. Takano (Takano et al., 2000, 2001a, 1995, 1997, 2001b,c ; Funahashi et al., 2000, 2002). Dans la plupart des cas, ils n'ont pas identifié de variation structurale notable dans la chaîne principale sauf dans le cas d'une dizaine de mutants. Toutes les variations observées par les auteurs sont aussi détectées avec notre méthode, mais nous identifions également d'autres perturbations potentielles. Nous avons aussi confronté nos résultats avec ceux du groupe Dobson (Ahn et al., 2016 ; Booth et al., 1997) qui a travaillé sur les formes amyloïdes des lysozymes. Les deux mutants naturels, D67H et I56T forment des fibrilles amyloïdes dans l'espace extracellulaire de multiples organes et tissus, entraînant une amylose systémique non-neuropathique. Les causes de la formation des fibrilles amyloïdes sont difficiles à déterminer mais selon Ahn et al. (2016), les mutations D67H et I56T diminuent la stabilité de l'état natif. Il existe deux réseaux de liaisons hydrogène des chaînes latérales dans la protéine de type sauvage (Artymiuk and Blake, 1981). Il est intéressant de noter que ces deux réseaux stabilisent le domaine beta et couvrent les diverses régions de la structure primaire qui sont rapprochées lors du repliement 3D. Dans le cas D67H, la structure est perturbée dans deux boucles très flexibles (Fig. 3.6), que nous retrouvons avec le *p-rank* calculé à partir du RMS3G mais pas avec le RMS7L. Ceci est dû à un effet de levier car avec la superposition locale (RMS7L), ces boucles restent inchangées bien que certains résidus à la base de ces boucles soient localement perturbés. Ces résultats sont cohérents avec les travaux de Booth et al. (1997) où ils ont montré que le réseau de liaisons faibles de ces résidus est fortement perturbé. La structure du variant I56T est par contre presque identique à celle de la protéine de type sauvage (Fig. 3.7) ce qui pourrait la rendre plus difficile à analyser mais nous identifions également des résidus déplacés appartenant au réseau de liaisons faibles qui est perturbé selon Booth et al. (1997).

Nous confirmons donc la fiabilité des résidus identifiés comme significativement perturbés par notre méthode en comparant nos résultats à ceux des études expérimentales. Ces résultats montrent aussi que les mutations ont un impact à longue portée sur le squelette des protéines qui peut être détecté, bien que le déplacement des atomes affectés soit faible.

3.2.3 Conclusion et perspectives

Cette méthode est cependant très limitée par la nécessité de posséder de nombreuses structures de la même famille protéique ce qui n'est la plupart du temps pas le cas. Afin de s'affranchir de cette limite, d'autres approches peuvent être envisagées pour évaluer la flexibilité de la chaîne peptidique. La dynamique moléculaire, qui permet de décrire précisément les mouvements d'une protéine quelconque mais qui nécessite des temps de calcul importants n'est pas vraiment envisageable si l'on veut des temps de calcul raisonnables pour ce type d'analyse. Les facteurs de température B, qui sont des informations disponibles dans les fichiers PDB des structures protéiques, et qui correspondent à l'agitation thermique, donc à la flexibilité peuvent nous apporter des informations ainsi que les

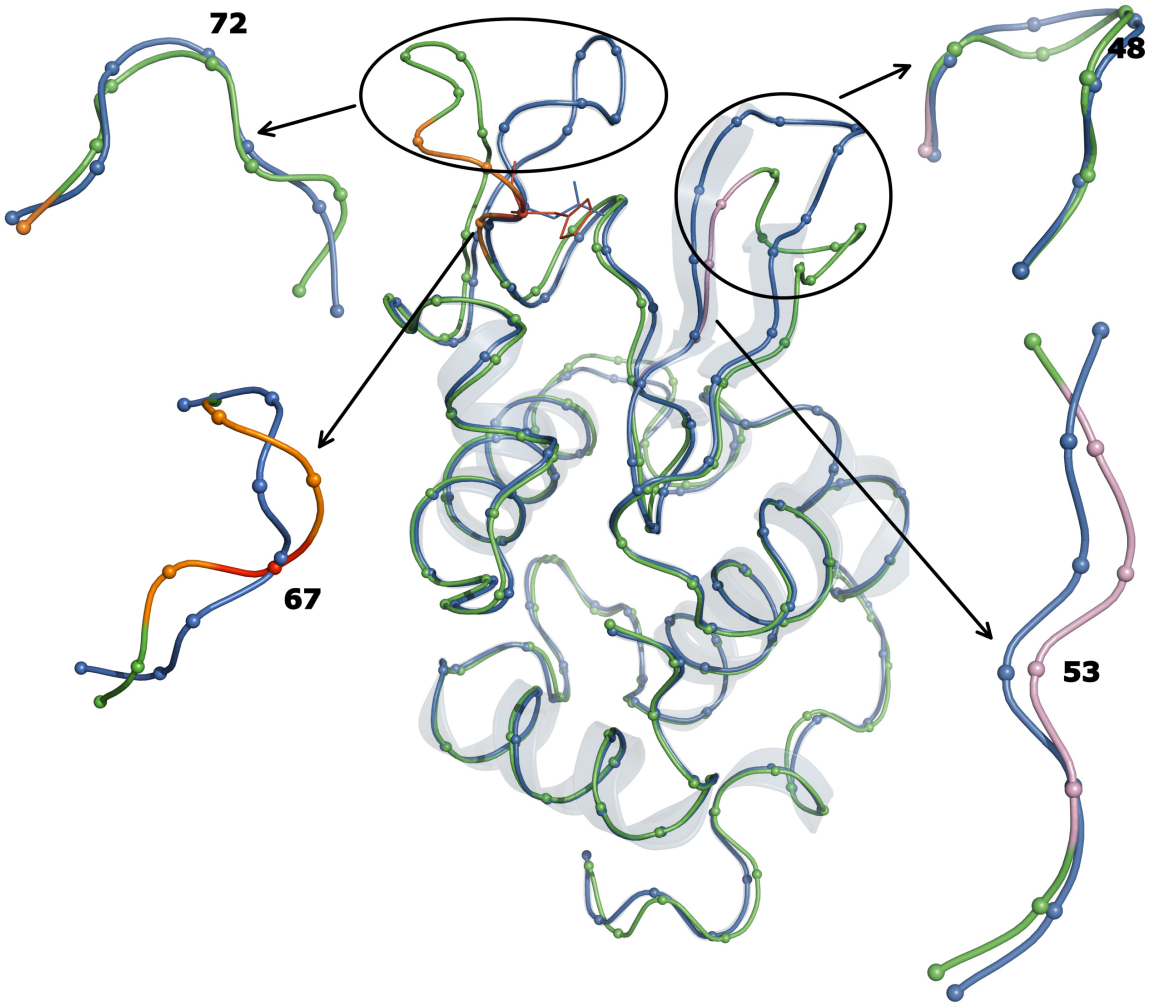


FIGURE 3.6 – Superposition globale de 2nwdX (sauvage, en bleu) et 1lyyA (mutant D67H, en vert). Quatre fragments superposés localement sur une fenêtre de 7 résidus sont représentés sur les côtés, centrés sur les résidus 48, 53, 67 et 72. Le résidu muté 67 est représenté avec sa chaîne latérale par des lignes. Les résidus sélectionnés selon les 5% supérieurs du *p-rank* (RMS7L) sont en orange et en rose. Le résidu muté figure également dans cette sélection. Figure provenant de (Carpentier and Chomilier, 2020).

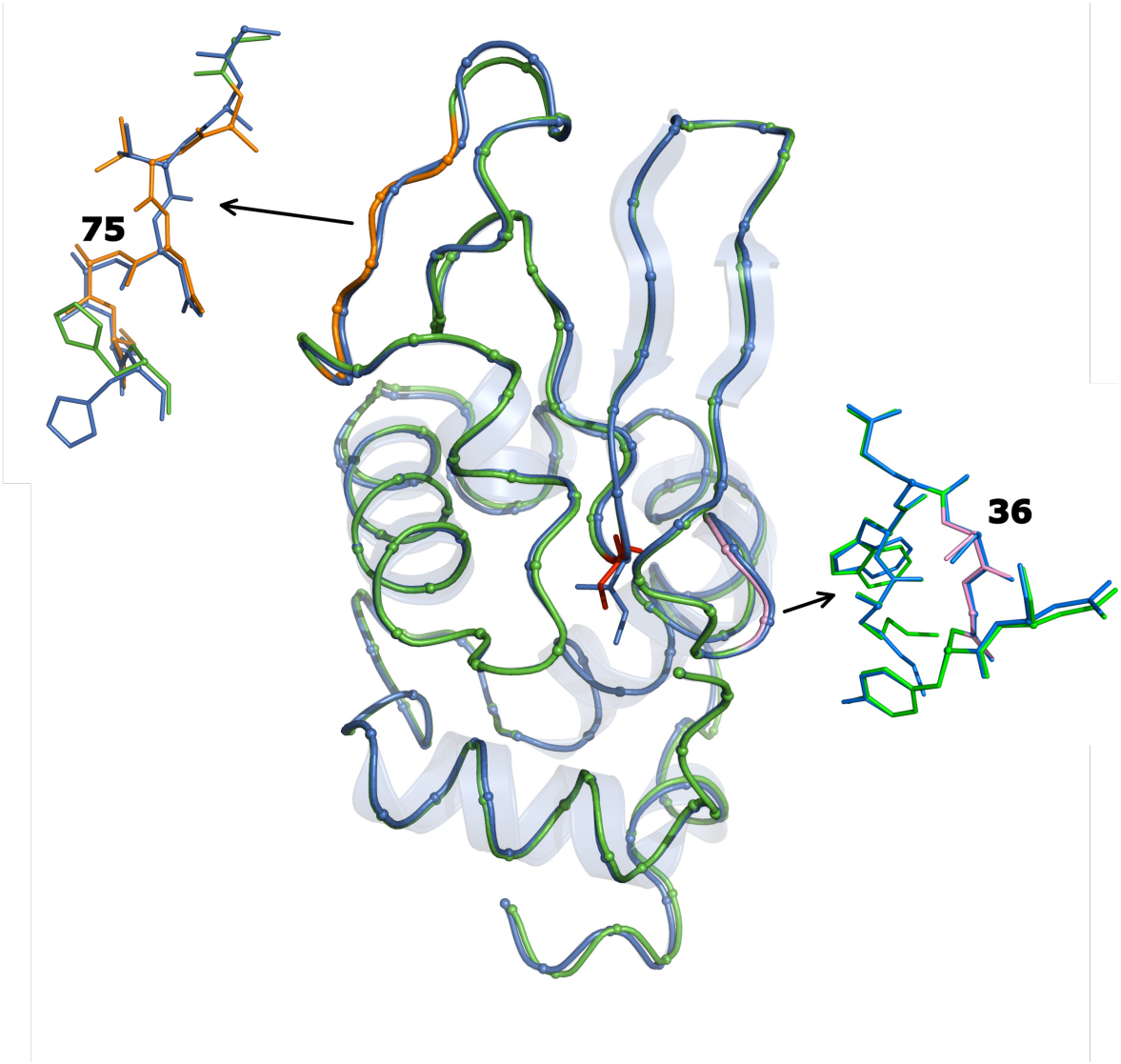


FIGURE 3.7 – Superposition globale de 2nwdX (sauvage, en bleu) et 1lozA (mutant I56T, en vert). Deux fragments, superposés localement sur une fenêtre de 7 résidus, sont montrés en gros plan : centrés sur le résidu 36 à droite et sur le résidu 75 à gauche. Les chaînes latérales des résidus mutés sont représentées par des lignes. Les deux fragments sélectionnés par *p-rank* (RMS7L) sont en orange et en rose. Figure provenant de (Carpentier and Chomilier, 2020).

différents modèles RMN lorsque la structure a été résolue par cette méthode. Toutefois, ces informations ne sont pas toujours disponibles. Des outils développés par d'autres équipes de recherche comme CONCOORD (van Aalten et al., 1997), méthode qui génère un ensemble de conformations possibles d'une protéine, ou PROPHET (Lavery and Sacquin-Mora, 2007 ; Sacquin-Mora and Lavery, 2006), qui donne les régions les plus rigides d'une protéine à partir d'un modèle de réseaux élastiques semblent plus adaptés. Enfin, les modes normaux sont aussi une piste prometteuse car notre méthode ne prend en compte que l'amplitude de la déformation du squelette mais pas la direction de celle-ci. Il serait possible de modéliser les variations attendues des positions de Ca avec des ellipses (Gerstein and Altman, 1995b,a ; Taylor et al., 1983) et non avec des sphères. En effet, même si la déformation à une position donnée est de la même amplitude pour la protéine mutée et pour toutes les autres non mutées, il est possible que la direction de la déformation soit différente. Par conséquent, il pourrait être intéressant de prendre dans notre étude toutes les positions mutées et de comparer leurs directions à celles des positions non mutées. La modélisation par ellipse et l'utilisation des modes normaux ont été abordées lors du stage de Nicolas Dibot (M1) en 2020 [11] et devraient être poursuivie dans les années à venir. A plus long terme, mon objectif est de mettre en place un serveur permettant de répondre à la question "où ma structure est-elle perturbée ?" lorsque l'utilisateur fournit les structures de deux protéines différant d'un acide aminé.

11. Ce stage a malheureusement été réalisé en distanciel à cause du premier confinement dû au Covid19

CHAPITRE 4

Distribution des *folds* dans le vivant

4.1 Motivations

Dans ce troisième chapitre, je présenterai les connaissances que j'ai accumulées et mes contributions sur le thème des macro-transitions, c'est-à-dire autour de la dynamique d'apparition des repliements protéiques. Encore une fois, le champ est vaste. Sont concernés notamment tous les mécanismes participant à la création de nouveaux repliementx (gènes *de novo*, transition d'un *fold* à l'autre, combinaisons de domaines) mais je me suis positionnée jusqu'à présent à un niveau plus général : celui de l'analyse de la répartition des repliements dans le vivant.

Le nombre de repliements connus est limité à quelques milliers (Zhang, 1997 ; Wolf et al., 2000), et est presque constant depuis 2013 selon les classifications structurales CATH et SCOP (*cf.* figure 2.1 page 22). Ce nombre est étonnamment faible, de plusieurs ordres de grandeur inférieur au nombre de séquences dans la biosphère (Landenmark et al., 2015). Il est possible que des contraintes biologiques ou physiques limitent significativement le répertoire de repliements (Mannige, 2014) mais les expériences récentes de *protein design* tendent à) montrer plutôt que le nombre repliements observés est très faible par rapport au nombre de repliements stables possibles (Linsky et al., 2022). Il est possible la convergence structurale soit fréquente. Dans ce cas, les similarités structurales des repliements entre les lignées de l'arbre du vivant seraient des homoplasies, *i.e.* elles ne proviennent pas d'un ancêtre commun. Cependant, il est également possible que les *folds* représentent des caractères intrinsèquement stables même en cas de forte divergence des séquences (Bajaj and Blundell, 1984 ; Illergård et al., 2009) : les repliements protéiques seraient alors monophylétiques à l'exception peut-être de certains des *superfolds* les plus divers (Koonin et al., 2002). Il est même probable que même ces *superfold* aient en fait une origine évolutive commune (Ding and Dokholyan, 2006). Les *superfolds* sont les *folds* les plus fréquents dans les génomes et sont présents dans les trois super-règnes (Orengo et al., 1994). Il en existe une dizaine. Les plus connus sont les *TIM Barrel* (Nagano et al., 2002) et les *Rossmann-like alpha/beta/alpha sandwich fold*(Medvedev et al., 2019) . Il a été démontré que l'évolution convergente est rare en termes de combinaison de domaines structuraux (*superfamily* de SCOP dans Gough (2005)), mais cela n'a pas été fait à ma connaissance pour les repliements eux mêmes.

Comme évoqué page 9, les trois principales classifications structurales sont CATH (Dawson et al., 2017), SCOP (Murzin et al., 1995), devenue SCOPe (Chandonia et al., 2017) et ECOD (Cheng et al., 2014). Le premier niveau dans ces 3 classifications est trop général pour contenir des informations évolutives utiles (Koonin et al., 2002 ; Cheng et al., 2014). Les autres niveaux dans SCOPe sont les *folds*, les *superfamily* et les *family*. Il est établi que les protéines de la même *superfamily* ou *family* de SCOP dérivent d'un ancêtre commun car elles sont construites à partir de similitudes en séquence et/ou en fonction qui permette d'en déduire l'homologie. Par contre l'origine commune des *folds* peut être encore encore discutée au moins pour certains *folds* (Koonin et al., 2002). Comme nous voulions interroger la pertinence des similitudes structurales et non celles des séquences, nous nous sommes principalement intéressés au niveau des *folds* car nous ne souhaitions pas ajouter d'information de séquences. Concernant les deux autres classifications, la meilleure correspondance pour le niveau *fold*

du SCOP est le niveau T dans CATH (Csaba et al., 2009) et le niveau X dans ECOD même s'il existe certaines divergences entre les classifications, (Cheng et al., 2014).

Pour étudier la répartition des repliements dans les génomes, il faut avoir la ou les structures des domaines des protéines des protéomes des génomes complets. Le nombre de structures expérimentales est très réduit mais il est souvent possible de prédire le repliement des domaines des protéines. La méthode classique consiste à chercher des similitudes entre les protéines des protéomes et les protéines des différentes familles structurales des classifications. En effet, chacune des 3 classifications précédentes met à disposition des modèles HMM construits à partir des alignements des séquences des protéines de même famille ou superfamille. Ces HMM permettent d'assigner un ou des repliements à en moyenne 60% des protéines de chaque protéome (voir page 61 pour une analyse un peu plus détaillée) et d'avoir ainsi le "contenu en *folds*" des génomes.

De nombreuses tentatives de reconstruction phylogénétique à partir du contenu en *folds* des génomes ont été faites dès les années 90 même si alors le nombre de génomes disponibles était très limité (Gerstein, 1998 ; Gerstein and Hegyi, 1998 ; Wolf et al., 1999 ; Lin and Gerstein, 2000). Ensuite, le nombre de génomes complets et de structures résolues ayant augmenté de façon exponentielle, des phylogénies plus conséquentes ont été reconstruites à partir des abondances de *folds* (nombre de copies) ou des occurrences (binaires). Caetano-Anollés et son groupe ont reconstruit un très grand nombre phylogénies et ont produit beaucoup d'articles sur le sujet. Ils ont principalement utilisé l'abondance des *folds* et la parcimonie (Caetano-Anollés and Caetano-Anollés, 2003, 2005 ; Caetano-Anollés et al., 2009 ; Kim and Caetano-Anollés, 2011, 2012 ; Wang et al., 2007, 2011a ; Wang and Caetano-Anollés, 2006, 2009). D'autres groupes ont utilisé des méthodes de distance (Yang et al., 2005) ou ont comparé les deux approches (Winstanley et al., 2005). Les travaux de Caetano-Anollés et son groupe sont discutés et assez discutables. En effet, ils ont principalement utilisé l'abondance des *folds*, alors que Yang et al. (2005) et Winstanley et al. (2005) ont trouvé que l'abondance des *folds* n'est pas assez fiable pour calculer des arbres robustes car elle est grandement affectée par la duplication des gènes et des chromosomes, ce qui ne se produit pas selon un processus uniforme (Yang et al., 2005). Ensuite, Caetano-Anollés et coll. ont enraciné leurs phylogénies en choisissant la branche de l'arbre présentant le plus grand nombre d'états de caractères ancestraux (plésiomorphes) (Nasir et al., 2014) ce qui est très discutable. Ils ont déduit de ces phylogénies enracinées que le monde des organismes est tripartite, et que les relations phylogénétiques entre les trois super-règnes sont [archées, [bactéries, eucaryotes]], les archées étant à la racine et paraphylétiques, et les bactéries et les eucaryotes étant monophylétiques (Kim and Caetano-Anollés, 2011). Ces résultats sont en contradiction avec les phylogénies actuelles où les eucaryotes sont très proches des archées (Hug et al., 2016) ou se branchent même en leur sein (Williams et al., 2020). Ces contradictions apparentes interpellent : pourquoi la distribution des repliements protéiques et les mutations de l'ADN ou des protéines ne racontent-elles pas la même histoire ? Cateano-Anollés et coll. soutiennent que l'utilisation des séquences moléculaires est problématique pour la reconstruction des phylogénies pour de nombreuses raisons : saturation mutationnelle, difficultés à identifier les sites homologues dans les

alignements de séquences, échantillonnage des taxons et déséquilibre des arbres, et signatures historiques différentes dans les domaines des protéines multidomaines (Caetano-Anollés and Nasir, 2012 ; Caetano-Anollés et al., 2014). Ils affirment au contraire que les domaines structuraux se diversifient principalement par descente verticale et sont mieux adaptés à la reconstruction de phylogénies (Wang and Caetano-Anollés, 2009 ; Kim and Caetano-Anollés, 2012)). D'autres chercheurs arguent à l'inverse que cette façon d'analyser les données sur les *folds* pour résoudre les relations phylogénétiques profondes n'est pas satisfaisante car les analyses publiées ne tiennent pas compte des histoires évolutives individuelles variées des *folds* dues à l'endosymbiose et aux transferts de gènes. Les simulations qu'ils ont effectuées montrent que l'inférence de la racine dans les arbres construits à partir des *folds* n'est pas fiable et est affectée par la variation de l'abondance et de la distribution des *folds* dans les génomes (Williams et al., 2020).

Ces incohérences posent donc question. Une manière d'aborder ce problème est de confronter la définition, la classification et la distribution des *folds* avec nos connaissances actuelles sur la phylogénie du vivant. C'est ce qui a été réalisé lors de la thèse de Martin Romei que j'ai co-encadrée avec Guillaume Lecointre (Pr MNHN, ISYEB) et Jacques Chomilier (DR CNRS, IMPMC). Plusieurs stagiaires ont aussi participé à ce travail : Léon Faure (M1) qui a annoté les protéines et régions désordonnées dans les protéomes, Pierre Imbert (M1) qui a étudié les combinaisons de domaines et Théo Jamay (M1) qui a réalisé un travail similaire à celui présenté ci-dessous avec la classification CATH. Ce travail a donné lieu à deux publications, une dans le journal Evolution (Romei et al., 2022) et une en cours de publication (Romei et al., soumis).

4.2 Contributions

4.2.1 Méthodologie

Notre démarche a été de projeter la répartition des repliements SCOP ou CATH ou ECOD sur un échantillon de 210 espèces de l'arbre du vivant. Nous avons mesuré la congruence entre l'arbre et la distribution dans les espèces de chaque repliement, fait une analyse en composantes principales pour les branches plus profondes, identifié des *clusters* permettant de définir des blocs de *folds* représentant des synapomorphies caractéristiques de clades et enfin analysé les fonctions associées aux *folds* qui sont des synapomorphies pour les mettre en relation avec d'autres caractéristiques apparues au cours de l'évolution des espèces.

L'arbre du vivant de référence est issu de Lecointre & Le Guyader (Lecointre and Le Guyader, 2017), complété par des bactéries de l'arbre de Hug et al. (2016) et par des archaea Asgard de Adam et al. (2017). Les espèces ont été choisies parmi les protéomes complets disponibles et afin d'échantillonner toutes les lignées principales tout en maximisant l'étendue des temps de divergence des lignées entre elles. Nous avons 70 espèces pour chaque super-règne. L'annotation des protéines pour leur attribuer un ou des repliements est effectuée avec des modèles HMM construits pour caractériser

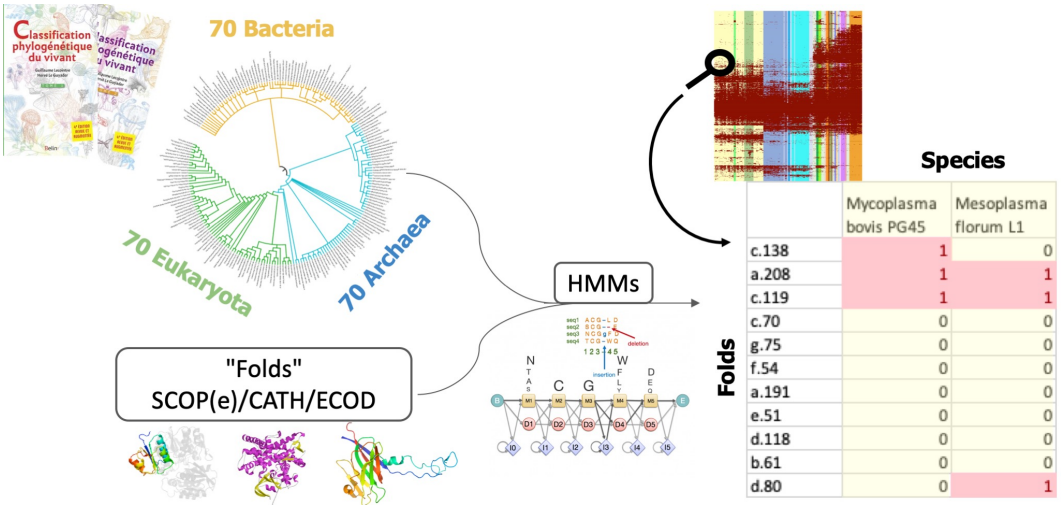


FIGURE 4.1 – Méthodologie développée lors de cette étude. 210 espèces ont été sélectionnées parmi les protéomes complets disponibles. Les HMM provenant des classifications structurales SCOP, CATH et ECOD ont été utilisés pour savoir si les repliements sont présents ou non dans les protéomes ce qui a permis de construire une carte permettant de visualiser la distribution des repliements dans le vivant. Le schéma représentant les HMM provient de `www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/` et est sous licence Creative Commons Attribution 4.0 International (CC BY 4.0) `creativecommons.org/licenses/by/4.0/`

chaque repliement que ce soit dans CATH (Dawson et al., 2017), dans SCOP (Wilson et al., 2009) ou dans ECOD (Cheng et al., 2014) (voir (Romei et al., 2022) pour une description précise de la méthodologie).

Nous savons donc pour chaque espèce si un repliement est présent ou non. Avec les HMM de SCOP nous avons annoté en moyenne dans les protéomes, 64% des protéines, 66% avec CATH et 67% avec ECOD. Les eucaryotes sont légèrement moins annotés que les archées et les bactéries (59 ± 10%, 65 ± 4%, et 67 ± 6%, respectivement avec SCOP) . L'absence d'un repliement est à interpréter avec soin car elle peut aussi signifier qu'il n'a pas été détecté. Afin d'analyser la répartition des repliements nous avons construit des cartes (*heatmap*). La carte générée à partir des données de SCOP est présentée figure 4.2, celles générées à partir des données de CATH et ECOD sont présentées figure 4.3. Les repliements sont en ligne, ordonnés selon une classification hiérarchique et les espèces sont en colonne, ordonnées selon la phylogénie de référence. Les branches sont permutées pour optimiser la visualisation et les *clusters* de *folds* sont délimités en coupant dans l'arbre des *folds*. Le développement de cette procédure de classification, permutation, *clustering*, visualisation a représenté une part importante du travail de thèse de Martin Romei. Il a été effectué en collaboration avec Guillaume Sapriel et il est prévu qu'il soit publié. Nous avons ensuite calculé pour chaque repliement l'indice de rétention (*retention index* ou RI) qui est "la fraction des synapomorphies possibles dans les caractères qui est retenue comme synapomorphie sur l'arbre (Farris, 1989). Sa formule est $RI = \frac{(g-s)}{(g-m)}$ avec g le nombre maximum de pas, s le nombre de pas calculés par parcimonie et m le nombre de pas minimum

(ce minimum est à 1 dans notre cas car le caractère est binaire). Le RI est de 1 si le ou les caractères sont en parfaite adéquation avec l'arbre (dans notre cas, une seule apparition ou disparition dans tout l'arbre). Nous avons aussi calculé les RI pour des distributions aléatoires des *folds* dans les espèces, l'arbre étant conservé.

Nous avons aussi réalisé un *clustering* des repliements selon leur classification ascendante hiérarchique pour identifier des groupes (*clusters)* de repliements. Ceux-ci forment régulièrement des blocs dans la carte, ce qui signifie alors qu'ils sont spécifiques de certains clades. Nous avons extrait les annotations fonctionnelles des repliements prédits avec la classification SCOP (Wilson et al., 2007). Il existe sept catégories générales de fonctions : métabolisme, information, processus intra-cellulaires, régulation, autres, processus généraux et extra-cellulaires. Ces catégories générales sont ensuite divisées en 50 fonctions détaillées. 10,2% des repliements sont associés à plusieurs fonctions. Nous avons analysé les fonctions des repliements partagés entre eucaryotes, archées et bactéries et les fonctions des repliements spécifiques aux eucaryotes. Nous avons reconstruit les états ancestraux, c'est-à-dire les repliements présents à chaque noeud de l'arbre phylogénétique de référence. Nous avons utilisé des méthodes de maximum de vraisemblance et de parcimonie (programme PastML, Ishikawa et al. 2019). Les longueurs de branches sont absentes dans l'arbre original et sont nécessaires pour ces calculs. Nous les avons ajoutées en nous basant sur les informations de *Time Tree of Life* [12] (Kumar et al., 2017).

4.2.2 Cohérence de la distribution des repliements avec la phylogénie

TABLE 4.1 – Indice de rétention moyen calculé pour tous les caractères avec soit tous les organismes, soit uniquement les Bactéries, Eucaryotes ou Archaea (en ligne). Les caractères sont la présence ou l'absence prédite dans les protéomes des repliements SCOP, l'architecture de niveau T de CATH ou l'architecture de niveau X d'ECOD (en colonne).

	SCOP	CATH	ECOD
All	0.56	0.53	0.54
Bacteria	0.29	0.26	0.27
Eukaryotes	0.44	0.43	0.47
Archaea	0.27	0.27	0.27

Nous avons montré que la distribution taxonomique des repliements correspond à l'arbre du vivant de manière significative avec un indice de rétention moyen de tous les repliements de 0,56 avec SCOP, 0,53 avec CATH et 0,54 avec ECOD. Les différents RI calculés avec les 3 classifications sont tous très proches (cf. table 4.1) . 813 (76%) repliements ont un RI significativement élevé, 351 (33%) ont un RI>0,75 et 578 (54%) un RI>0,5. Le nombre de repliements présents est plus important chez

12. http ://www.timetree.org/

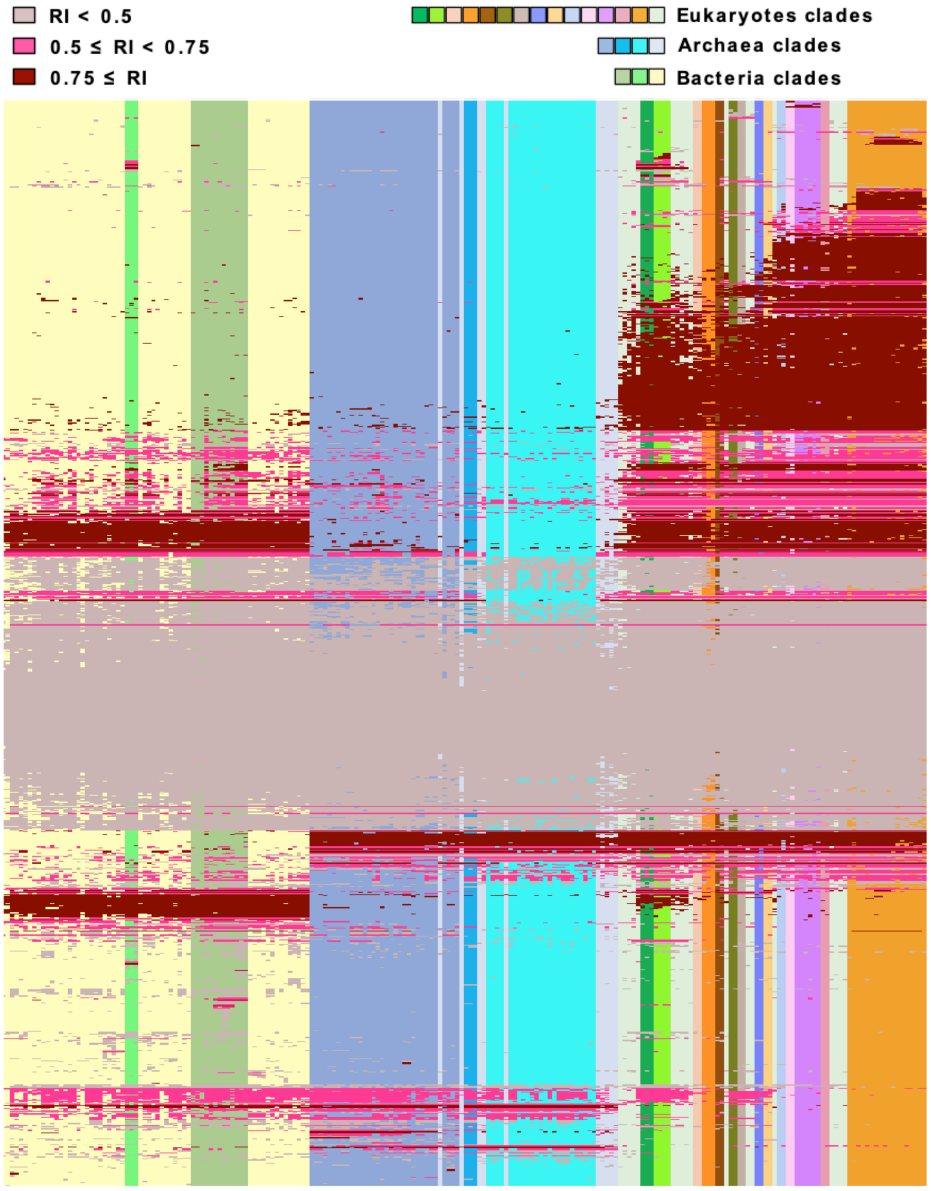
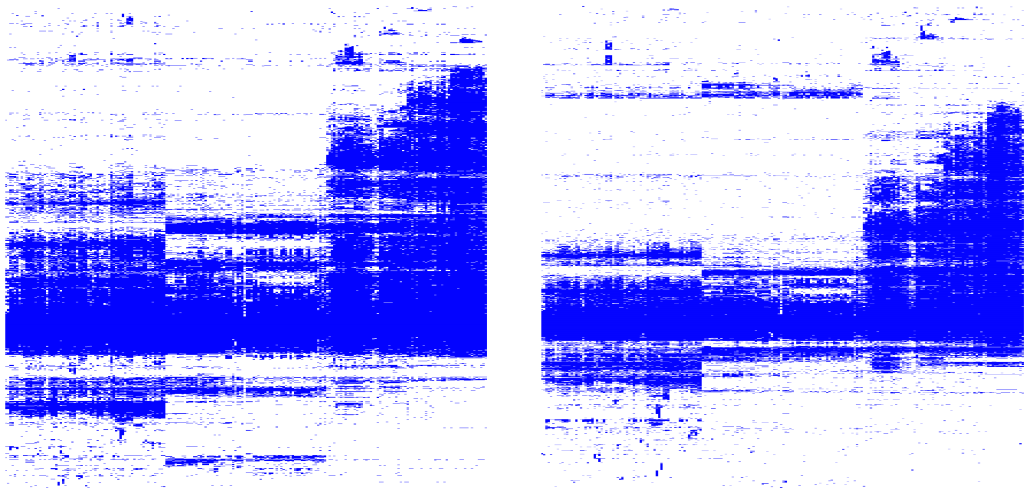


FIGURE 4.2 – Carte colorée montrant la répartition des repliements protéiques à travers le vivant. Les colonnes représentent les espèces ordonnées selon l'arbre phylogénétique de référence. Les 210 espèces, de gauche à droite, sont : 70 bactéries, 70 archées, 70 eucaryotes. Par commodité, les couleurs des colonnes indiquent les groupes taxonomiques selon la nomenclature NCBI. Pour les bactéries, nous avons choisi de ne présenter que deux *phyla* d'intérêt (cyanobacteria et proteobacteria). Les rangées représentent 1 073 repliements de protéines (selon SCOP). Les cellules représentent la présence des repliements dans l'espèce correspondante, colorées selon l'indice de rétention calculé pour la répartition de chaque repliement sur l'arbre phylogénétique de référence. Les cellules rouges foncées correspondent aux repliements qui peuvent être interprétés comme des marqueurs taxonomiques fiables (c'est-à-dire des synapomorphies). Le détail de la légende de couleurs est en annexe pages 134. Figures provenant de (Romei et al., 2022)



(a) (b)

FIGURE 4.3 – Cartes montrant la répartition des topologies CATH des protéines (a) et des niveaux X ECOD (b) à travers la diversité de la vie. Les colonnes sont des espèces ordonnées selon l'arbre phylogénétique de référence. Les 210 espèces, de gauche à droite, sont : 70 bactéries, 70 archées, 70 eucaryotes. Les lignes comportent respectivement 1 295 topologies CATH (a) et 1 863 niveaux X ECOD (b). Les carrés bleus représentent la présence des Topologies/niveaux X dans l'espèce correspondante. Les espèces sont dans le même ordre que dans la carte des plis SCOP afin de faciliter la comparaison des cartes. Figures provenant de (Romei et al., 2022)

les eucaryotes (915) que chez les bactéries (804) ou les archées (627). Par construction, l'indice de rétention moyen calculé sur tous les repliements est différent pour chacun des trois super-règnes et pour l'ensemble des 3 super-règnes : il vaut 0,44 chez les eucaryotes, 0,27 chez les archées et 0,29 chez les bactéries (voir Tableau 4.1). En effet, cette mesure dépend de la distribution des folds mais aussi de l'arbre de référence. Le RI moyen global (0,56) est supérieur aux RI moyens des trois super-règnes car ces derniers sont calculés sur des sous-arbres distincts. La valeur plus élevée du RI moyen global montre que la distribution des repliements est très cohérente avec la séparation en trois super-règnes, mais qu'elle l'est un peu moins dans chacun des super-règnes. La carte est en effet séparée en trois parties distinctes. La "partie eucaryotes" contient 224 repliements spécifiques qui sont regroupés en blocs imbriqués et leur RI est élevé (0,59). Les archées et les bactéries ont moins de repliements spécifiques, 19 et 89 respectivement et ne présentent pas une telle structure emboîtée. Les valeurs RI plus faibles obtenues pour les archées et les bactéries et le peu de repliements spécifiques indiquent que la reconstruction de l'histoire évolutive des organismes à partir des repliements est plutôt risquée au sein des bactéries et des archées tandis que la valeur élevée du RI, la forte proportion de repliements spécifiques et la forme en escalier chez les eucaryotes révèlent de manière surprenante que les repliements peuvent permettre de séparer des clades même récents au sein des eucaryotes. Un nombre assez important de *folds* sont présents dans tous les organismes : 54 plis (4,8%) sont présents dans 100% des organismes, et 181 plis (16,9%) le sont dans 90% des organismes. Ils sont localisés dans la bande centrale noire figure 4.4. Ils ne présentent pas d'intérêt pour l'inférence phylogénétique à partir du contenu en *folds*, mais ils pourraient être intéressants pour identifier des protéines ou gènes présents dans tous les organismes pour reconstruire des phylogénies moléculaires classiques ou peut-être des phylogénies calculées à partir des structures protéiques.

Certains repliements sont partagés exclusivement entre les eucaryotes et les bactéries ou les eucaryotes et les archées et ils forment des blocs dans notre carte (en bleu et rouge figure 4.4). Ces repliements communs reflètent l'origine des eucaryotes et la première endosymbiose. Les archées et les eucaryotes partagent 40 repliements (3,7%) et le RI moyen est de 0,72. 28 repliements (2,6%) sont partagés entre les bactéries et les eucaryotes avec un indice de rétention également de 0,72. La seconde endosymbiose (chloroplaste) est aussi visible dans la carte avec deux blocs de repliements (en orange et en rouge figure 4.4). Le premier groupe (orange) contient 7 repliements avec les eucaryotes photosynthétiques et les cyanobactéries. Le second groupe de 8 repliements (Figure 3, en rouge) est partagé entre les eucaryotes photosynthétiques et toutes les bactéries. Ce second groupe met en évidence le transfert de repliements non spécifiques des cyanobactéries par la seconde endosymbiose.

4.2.3 Analyse des fonctions

Nous avons analysé les fonctions des repliements des deux clusters principaux de repliements partagés entre eucaryotes et archées et entre eucaryotes et bactéries. Il a été trouvé précédemment que chez les eucaryotes, les fonctions héritées des bactéries sont plutôt des fonctions dites opérationnelles

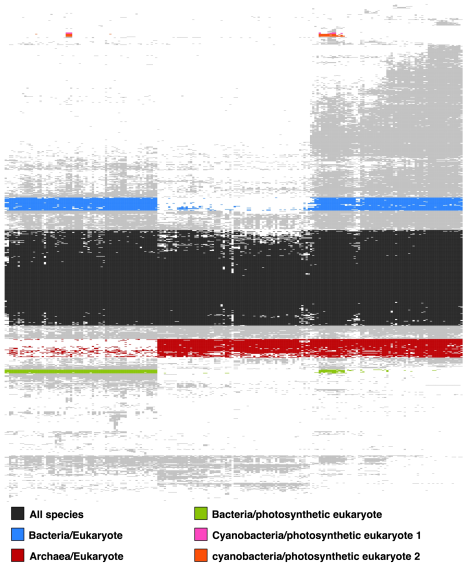


FIGURE 4.4 – Carte des *folds* SCOP et des espèces ordonnées comme dans la Figure 4.2, avec un code couleur montrant les groupes de repliements partagés entre deux super-règnes ou deux clades distants. Les groupes noirs sont des repliements partagés par toutes les espèces. Les groupes bleus et rouges sont des repliements partagés entre les eucaryotes et l'un des deux autres super-royaumes. Les groupes vert et orange sont les repliements partagés entre les bactéries et les groupes d'eucaryotes photosynthétiques. Figure provenant de (Romei et al., 2022)

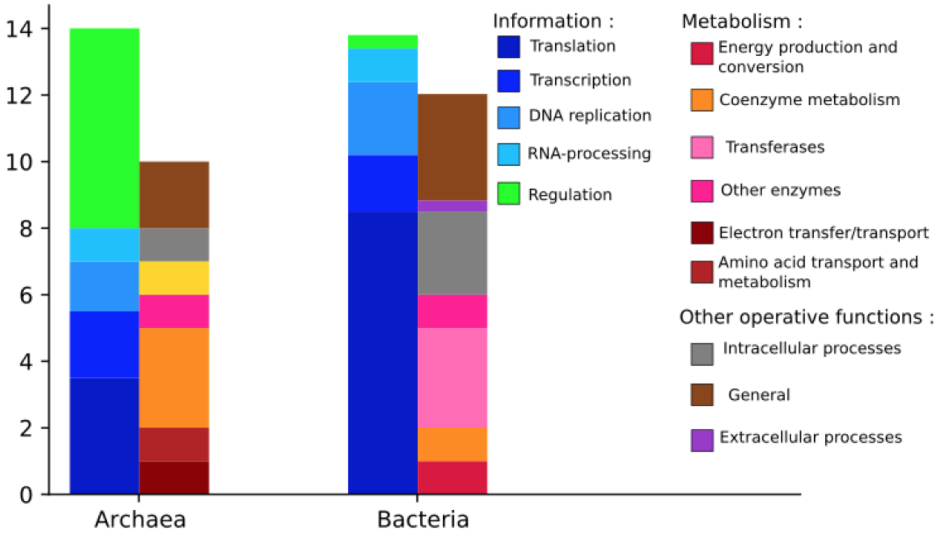


FIGURE 4.5 – Fonctions et nombre de repliements partagés entre eucaryotes et archées (à gauche) et entre eucaryotes et bactéries (à droite). Les fonctions informationnelles sont en bleu et vert et les fonctions opérationnelles en rouge, rose et brun. En ordonnée est indiqué le nombre repliements.

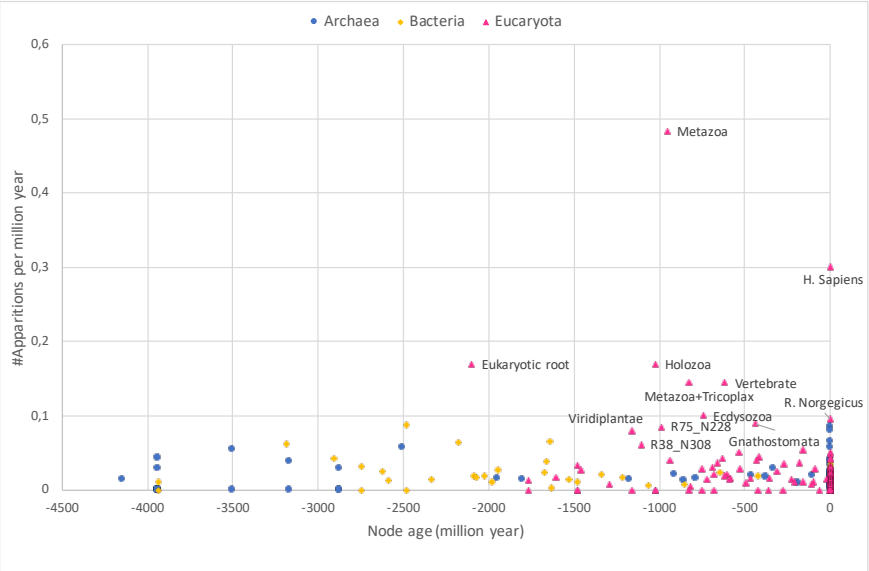


FIGURE 4.6 – Nombre d'apparitions de repliements par million d'années (en ordonnée) et millions d'années en abscisse. Les trois couleurs représentent les nœuds au sein des branches Eukaryota, Archaea et Bacteria.

et que les fonctions héritées des archées sont plutôt des fonctions dites informationnelles (Alvarez-Ponce et al., 2013). Les fonctions informationnelles sont liées à la traduction, à la structure et à la biogénèse des ribosomes, au traitement et à la modification de l'ARN, à la transcription, à la réplication, à la recombinaison et à la réparation de l'ADN, à la structure et à la dynamique de la chromatine. Les fonctions opérationnelles sont toutes les autres fonctions à l'exception des fonctions inconnues ou non annotées. Nos résultats sont légèrement différents : si pour les repliements Archaea/Eukaryota, les fonctions informationnelles sont majoritaires comme cela est attendu, elles le sont aussi pour les repliements Bacteria/Eukaryota. Le nombre de repliements opérationnels est important dans les deux cas, mais il est tout de même plus important dans le cluster Bacteria/Eukaryota (*cf.* figure 4.5).

Selon la reconstruction des états ancestraux calculée par Maximum de Vraisemblance, 268 repliements ont une forte probabilité d'être présents à la racine des 3 super-règnes. Ce nombre est cohérent avec les précédentes études comme celle de Weiss et al. (2016) qui avaient identifié 355 familles ancestrales de protéines. Nous avons calculé le nombre d'apparitions de repliements par million d'années sur chaque branche de l'arbre (*cf.* figure 4.6). Le taux basal d'apparitions est de 0,017 pour les archées, 0,016 pour les bactéries et 0,028 pour les eucaryotes. Le taux le plus élevé est pour les Metazoa (36 apparitions, taux de 0.49), et par ordre décroissant Eukaryota (312, 0.17), Holozoa (14, 0.17), Metazoa stricto sensu (18, 0.15), Vertebrata (10, 0.15), Ecdysozoa (10, 0.1), Gnathostomata (16, 0.09),

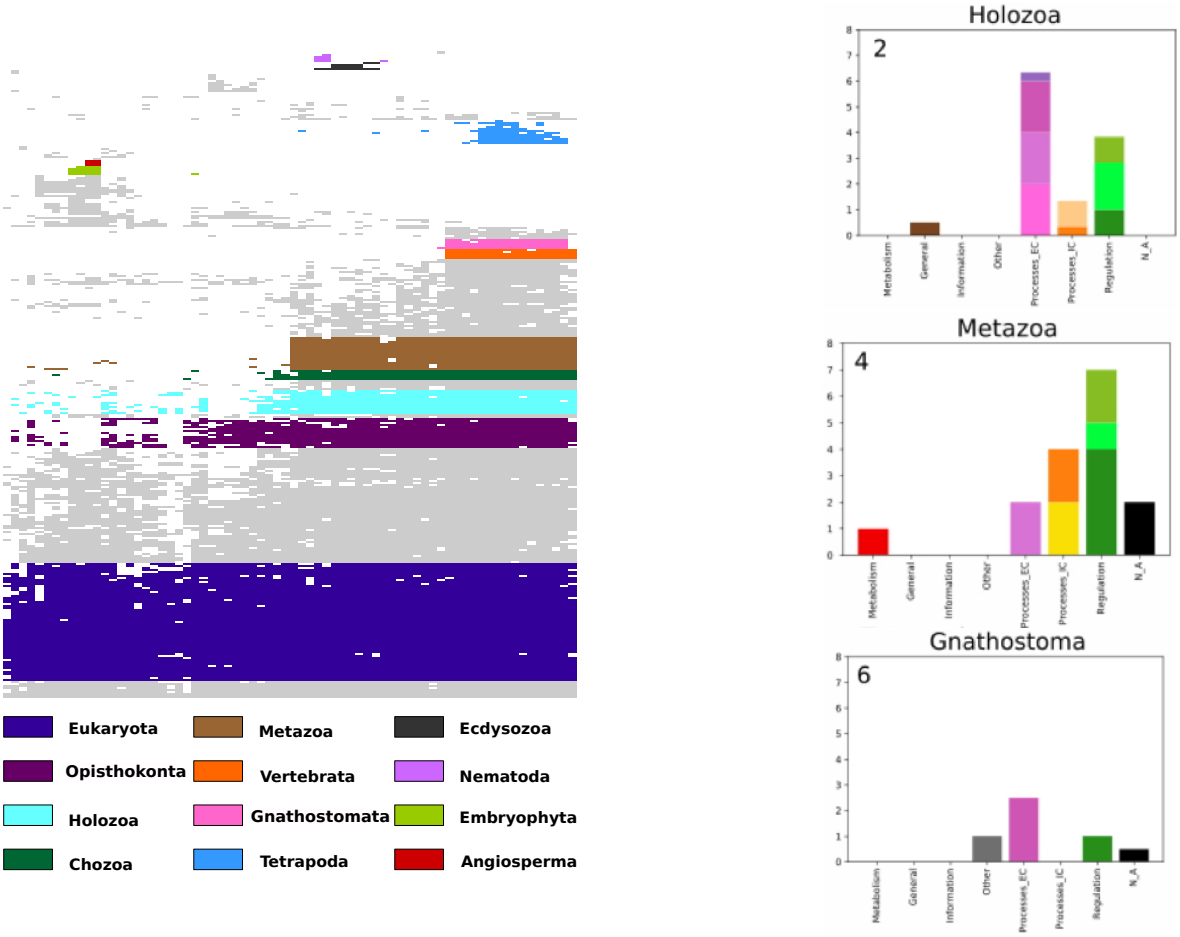


FIGURE 4.7 – (a) Extrait de la partie de la figure 4.2 comportant les *folds* SCOP spécifiques aux eucaryotes. Les blocs de *folds* formant des synapomorphies sont identifiés par des couleurs différentes. (b) Diagrammes des fonctions principales des repliements des clades Holozoan Metazoa et Gnathostomata. Les fonctions principales sont dans l'ordre : Métabolisme, général, information, autre, processus extra-cellulaires, processus intra-cellulaires, régulation et non-attribué. Pour le détail des fonctions, voir en annexe figure 7.2 page 135.

Fungi (10, 0.09), Viridiplantae (49, 0.09), et Opisthokonta (23, 0.06). Le taux d'apparition de plis est également élevé pour Homo sapiens et Rattus norvegicus mais il s'agit probablement d'un artefact : les branches de ces deux espèces sont très courtes (6,5 et 20,9 millions d'années respectivement) et il n'y a que deux apparitions de repliements pour ces deux espèces. Le nombre de repliements présents chez les eucaryotes est plus important que dans les deux autres super-règnes et cela se traduit par un nombre d'apparitions important sur quelques branches alors que les autres branches eucaryotes ont des taux assez proches de ceux des branches archées et bactéries.

Chez les eucaryotes, nous avons identifié 12 clusters qui forment des blocs dans la carte et qui sont spécifiquement associés à des clades (groupes taxonomiques monophylétiques) soutenus de manière fiable par au moins 3 repliements (cf. figure 4.7). Le premier bloc englobe tous les eucaryotes

avec 59 repliements (5,5%). Nous observons une surreprésentation des fonctions liées à la régulation (19 repliements, 26,5%) et aux processus intracellulaires (17 repliements, 24,0 %), ce qui est cohérent avec les études précédentes des repliements spécifiques aux eucaryotes (Kauko and Lehto, 2018). Le regroupement et la valeur élevée de RI permettent l'identification de 11 blocs correspondant à 11 clades : Opisthokonta, Holozoa, Chozoa, Metazoa, Vertebrata, Gnasthostomata, Tetrapoda, Ecdysozoa, Nematoda, Embryophyta et Angiospermae. Au sein de ces 11 blocs de repliements, le métabolisme et les fonctions informationnelles sont clairement sous-représentés avec 6 (7,1%) et 4 (4,7%) repliements, alors que la régulation et les processus extra-cellulaires sont sur-représentés avec 22 (24,5%) et 17 repliements (18,6%) respectivement. Cette surreprésentation peut être liée à l'essor de la multicellularité qui contraint les cellules à faire des jonctions et à communiquer. Les deux derniers blocs correspondent à des clades photosynthétiques dont les repliements ne présentent pas de fonction pouvant être clairement liée aux spécificités de ces clades bien qu'ils constituent des synapomorphies. Par contre, pour les autres, des liens ont pu être identifiés entre les clades et les fonctions associées aux repliements.

Par exemple, le bloc Holozoa est composé de 12 repliements dont 6,5 [13] sont associés aux processus extra-cellulaires (*cf.* figure 4.7 (b)). Les génomes du choanoflagellé (Monosiga brevicollis) et du filastère Capsaspora owczarzaki, le parent unicellulaire connu le plus proche des métazoaires en dehors des choanoflagellés, a permis de découvrir qu'ils avaient un répertoire plus riche de protéines impliquées dans l'adhésion cellulaire et la régulation transcriptionnelle (Suga et al., 2013 ; Lecointre and Le Guyader, 2017) que les autres eucaryotes unicellulaires. Dans le bloc du clade Holozoa (regroupant les métazoaires, les choanoflagellés, les filastéens et les mésomycétozoaires), nous avons bien des repliements associés à l'adhésion cellulaire : les repliements g.64 ("domaine B de la somatomédine") et a.215 ("domaine moyen de la Taline 1"). Chez les métazoaires qui sont des organismes pluricellulaires permanents, on s'attend à ce que l'activité de signalisation entre les cellules augmente et parmi les 16 repliements du bloc, nous avons bien les repliements g.76, a.123 et g.1 qui sont impliqués dans l'activité des récepteurs et la transduction des signaux. La division cellulaire puis le cycle cellulaire doivent être régulés, et les repliements a.77 et g.17 y sont bien impliqués. Le *fold* a.77 est appelé "Death domain", un domaine lié à l'apoptose. Trois repliements des Gnathostomata sont trouvés dans des protéines classées sous la fonction générale "Réponse immunitaire" : a.109, d.9, d.19. Ceci est probablement lié à la montée du système immunitaire adaptatif chez les gnathostomes (Bleyzac et al., 2005). Les repliements sont donc bien des synapomorphies qui peuvent avoir un très grand intérêt pour comprendre l'évolution du vivant et des caractères phénotypiques. Cet aspect n'en est qu'à son début et j'espère pouvoir le poursuivre avec des spécialistes des clades pour apporter des éléments pour la résolution des phylogénies des espèces et la compréhension de l'évolution des caractères.

13. Certains *folds* ont plusieurs fonctions. Par exemple le *fold* g.16 est associé à des fonctions extra-cellulaires et intracellulaires. Il compte donc pour 0,5 dans chacune des catégories.

4.3 Conclusion et perspectives

Ma question initiale était de savoir si les repliements sont de bons marqueurs phylogénétiques. Ils le sont clairement, tant pour les événements anciens que récents : ils sont stables dans le temps et apparaissent de manière assez régulière. En revanche, les transferts horizontaux massifs sont visibles et une étude plus détaillée des repliements dans les bactéries ou les archées nous permettra probablement de mieux les identifier dans ces organismes. Ce travail permet de comprendre pourquoi les phylogénies calculées par parcimonie standard avec les données de repliements sont parfois contradictoires. En effet, un arbre entièrement dichotomique est inadéquat pour représenter la double origine des eucaryotes. Les nombres de repliements partagés soit par les eucaryotes et les bactéries, soit par les eucaryotes et les archées sont assez proches. Selon les espèces et les sélections de repliements effectuées par les auteurs, ce nombre varie légèrement et les eucaryotes peuvent alors basculer côté archées ou côté bactéries. La même inadéquation des représentations dichotomiques se retrouve avec l'endosymbiose du chloroplaste. Une approche plus progressive, par exemple en calculant des phylogénies distinctes pour les archées et les eucaryotes et pour les bactéries et les eucaryotes (Raymann et al., 2015), ou des phylogénies où les différents signaux sont enlevés progressivement (Pisani et al., 2007) pourrait donner de bons résultats.

Je pense également qu'une telle approche, qui s'ancre dans les connaissances actuelles de la phylogénie des espèces permettra également de mieux comprendre les mécanismes d'évolution des repliements et d'apparition de nouveaux repliements. Je reviendrai sur ce point plus en détail dans mon projet. Enfin, nous espérons que cette étude contribuera à donner au lecteur une vision plus claire et plus précise de l'univers des repliements protéiques et de la distribution des repliements dans les organismes vivants.

CHAPITRE 5

Conclusion : projets

Mon projet pour les prochaines années s'articule autour de deux axes. Le premier est la poursuite de l'étude de l'évolution de la structure des protéines, notamment en étudiant la distribution, l'apparition et les modifications des repliements chez les eucaryotes. Le second est l'intégration de contraintes structurales afin d'améliorer les méthodes d'alignement et de détection d'homologie lointaine, c'est-à-dire de mettre de la structure dans la séquence. Ces deux axes se déclinent en projets à court, moyen et long terme. Certains points ont déjà été évoqués dans les conclusions des chapitres précédents.

5.1 Evolution des structures

5.1.1 micro-transitions

Les travaux que j'ai réalisés avec S. Shanthirabalan (I.E. à l'ABI en 2016-2017, sous ma supervision) et J. Chomilier (DR. CNRS de l'IMPMC) sur l'effet des mutations ponctuelles (substitutions) sur les structures sont, en ce qui me concerne, pratiquement achevés (ce qui ne signifie pas que ce champ ne mériterait pas d'être encore investigué). J'aimerais juste pouvoir proposer un service à la question "où ma structure est-elle perturbée ?" où un utilisateur fournit les structures de deux protéines différant d'un acide aminé et le serveur indique les variations structurales inattendues. Comme évoqué page 53, cela nécessite de se libérer de la contrainte d'avoir un assez grand nombre de structures résolues par modélisation et éventuellement simulation. Ces travaux ont été initiés lors d'un stage (N. Dibot, M1) et seront finalisés à court ou moyen terme. Tenter de déterminer quels sont les résidus perturbés par une mutation se rapproche grandement du champ de recherche sur l'allostérie, qui est passionnant. J'ai lu de nombreux articles s'y rapportant et il n'est pas exclu qu'un jour, si j'ai le temps et que l'occasion se présente, je m'y investisse plus activement. Une piste serait de comparer les résidus identifiés avec les réseaux allostériques et de croiser ces résultats avec les données de *deep mutational scanning*.

J'ai assez peu abordé l'effet des insertions et des délétions (InDels) dans ce manuscrit qui sont pourtant très importantes pour l'évolution des protéines. L'analyse de l'évolution des InDels dans les phylogénies de protéines orthologues indique qu'ils sont éliminés (purgés) de 9 à 100 fois plus vite que les substitutions (Tóth-Petróczy and Tawfik, 2013). L'incorporation des InDels est non seulement lente mais aussi non linéaire. En moyenne, 50 substitutions s'accumulent avant l'apparition du premier InDel mais il y a plus de substitutions à proximité des InDels (en séquence et en 3D), ce qui suggère que certaines substitutions sont corrélées aux InDels. Il est possible que ces substitutions soient compensatoires - et arrivent après l'InDel - ou permissives - et arrivent avant l'InDel - (Tóth-Petróczy and Tawfik, 2013). Pourtant, les InDels sont connus pour être indispensables à l'acquisition de certaines nouvelles fonctions (Tawfik, 2006 ; Akiva et al., 2008 ; Neuenfeldt et al., 2008 ; Hashimoto and Panchenko, 2010 ; Britten, 2010 ; Cooley et al., 2010 ; Afriat-Jurnou et al., 2012) et ils peuvent être sous sélection positive (Podlaha et al., 2005 ; Podlaha and Zhang, 2003). Leur rôle important dans l'évolution naturelle des structures (Grishin, 2001) et l'évolution artificielle (Jochens et al., 2009 ; Ochoa-Leyva et al., 2009) est donc pleinement accepté mais loin d'être suffisamment compris.

J'ai encadré deux stages sur ce thème. Le premier, celui de Meriem Aziz (M1) en 2013 portait sur la caractérisation structurale des InDels (localisation, longueur, composition en acides aminés). Le second, celui de Clément Joubert (M2) en 2017, portait sur l'étude de l'effet sur le squelette peptique des InDels. L'objectif était en particulier de tester si les déformations induites par une insertion ou une délétion se font dans le même sens que les modes normaux comme cela a été montré pour des familles de protéines (Leo-Macias et al., 2005b,a). Reprendre les résultats de ces deux stages et poursuivre cette étude pourrait amener des éléments intéressants tant pour la compréhension de l'évolution que pour le développement de fonctions de score des InDels dans les alignements structuraux.

5.1.2 macro-transitions

Les deux points précédents sont des travaux à finaliser et ne représentent pas un nouveau projet de recherche. Mon premier projet principal est me concentrer sur la poursuite de l'étude de la distribution des *folds* dans le vivant pour comprendre les mécanismes évolutifs à l'œuvre pour l'apparition de nouveaux repliements. Je pense que l'ancrage dans la phylogénie des espèces apporte un éclairage très original et intéressant.

A court terme, j'aimerais publier un article quantifiant les convergences des *repliements* protéiques. La question est donc de savoir si la convergence structurale est fréquente ou non. Les résultats présentés dans le chapitre 3 (Distribution des *folds* dans le vivant), montrent que la distribution des repliements est très cohérente avec la phylogénie des espèces mais qu'il y a visiblement beaucoup de transferts de matériels génétiques autres que filiaux. Il faudra donc faire la différence entre les transferts horizontaux et les phénomènes de convergence.

A moyen terme, je souhaite m'intéresser plus spécifiquement aux repliements chez les eucaryotes car les transferts horizontaux y sont moins fréquents chez les bactéries et probablement que chez les archées. L'identification de l'unité évolutive dans les structures protéiques n'est pour moi pas résolue. Je pense simplement (mais c'est peut-être une évidence...) que l'unité évolutive dépend du mécanisme évolutif et donc qu'une protéine peut être découpée en différentes unités suivant le mécanisme qu'on veut étudier. Les résultats présentés dans le chapitre 3 montrent qu'il y a eu beaucoup de nouveaux repliements qui sont apparus chez les eucaryotes. Mon questionnement principal est de savoir si l'évolution des structures protéiques est modulaire chez les eucaryotes. Plusieurs indices m'orientent vers cette hypothèse. Au niveau génomique, les introns et les exons délimitent des modules qui sont remaniés lors de l'épissage alternatif ou du brassage d'exons (*exon shuffling*). Des travaux très intéressants ont aussi montré qu'il existe de nombreux gènes résultant de la fusion de séquences d'origine bactérienne et de séquences d'origine Archaea (Méheust et al., 2018). Au niveau structural, la présence de sous-structures a été décrite à de nombreuses reprises : les super-structures secondaires, les « *reused fragments* » (Nepomnyachiy et al., 2017) ou les TEF (Tightened End Fragments) qui sont des fragments d'environ 30 acides aminés dont les deux extrémités sont spatialement proches dans la protéine (Lamarine et al., 2001). Enfin, certains travaux montrent que certaines sous-structures se chevauchent avec les exons (Gelly et al., 2012) et que les introns anciens (en phase 1) délimitent des exons formant

des modules 3D de 15 à 30 résidus (de Souza et al., 1996).

Dans un premier temps, nous identifierons les introns et nous localiserons entre quels acides aminés ils se situent dans la structure protéique. Nous utiliserons pour cela les bases de données existantes et une collaboration avec Olivier Dameron (INRIA, Rennes) a été initiée pour mieux tirer parti des informations disponibles grâce au Web sémantique et aux données RDF. Nous analyserons ensuite les résultats à la lumière des différentes histoires évolutives déjà identifiées pour les repliements. Nous pourrons ensuite explorer si les introns et la délimitation des domaines sont liés. Enfin, nous essaierons de comprendre à partir de quoi ont été construits les nouveaux repliements, c'est-à-dire les repliements spécifiques aux eucaryotes. Notre hypothèse est que ceux-ci pourraient résulter de la combinaison de sous-structures (qui sont à déterminer), et que les introns pourraient jouer un rôle. Nous mesurerons la cohérence de leur présence dans les espèces en fonction de l'arbre phylogénétique de référence. Ce projet fera l'objet d'un dépôt de sujet de thèse dans les années à venir (il a déjà été proposé une fois mais n'a pas été pourvu) et d'un projet ANR.

5.2 Mettre de la structure dans la séquence

Le second axe est d'intégrer l'information structurale dans les travaux de recherche en bioinformatique et biologie. Les questions qui m'ont le plus souvent été posées sont : la structure pourrait-elle m'aider à mieux comprendre la fonction et les spécificités de ma protéine ? La structure pourrait-elle m'aider à identifier des homologues lointains à ma famille de protéines ? La structure pourrait-elle m'aider à annoter des protéines de fonctions inconnues ? Evidemment, la structure permet bien tout cela et apporte des éléments souvent très intéressants. J'essaie de répondre au cas par cas quand le problème m'est présenté par des biologistes ou bioinformaticiens, mais je tente aussi de développer des outils permettant de mieux y répondre.

J'ai participé à plusieurs projets où la prédiction de la structure a permis de mieux répondre aux questions posées : l'annotation du génome de la bactérie *Dickeya dadantii 3937* (anciennement *Erwinia chrysanthemi*) (Glasner et al., 2011) et la caractérisation fonctionnelle d'une protéine dégradant l'élastine de la bactérie *Flavobacterium psychrophilum* (Rochat et al., 2019). D'autres projets du même type sont en cours, pour l'annotation de lipoprotéines toujours avec E. Duchaud (DR. INRAE Jouys en Josas), la recherche d'homologues lointains aux protéines PSU avec Eduardo Rocha (DR CNRS, Pasteur) [14], et la caractérisation structurale des mutations dans les opsines de papillons Morpho avec Violaine Llaurens (DR CNRS, ISYEB).

Côté développement d'outils, les projets sont un peu plus conséquents : je continue à développer de méthodes d'alignement de séquences et de structures avec l'idée de réussir à combiner l'information structurale et de séquence et j'ai pour projet d'appliquer les méthodes de prédictions de structures aux protéines orphelines (*dark proteome*) de TARA en collaboration avec avec Lucie Bittner (Mcf SU, à l'ABI).

14. *Polarity Suppression Protein*) des satellites de phages (voir Moura de Sousa and Rocha 2022 pour plus de détail sur cette protéine et les satellites de phages.

5.2.1 Alignement de structures

En collaboration avec François Coste (CR INRIA Rennes) et Hugo Talibart (post-doc à l'ABI en 2022 sous ma supervision), nous proposons d'aller vers une nouvelle génération d'outils de comparaison, d'analyse et d'identification d'homologie de séquences alliant à la fois les similitudes des séquences et des structures en utilisant des modèles utilisés en physique statistique : les modèles de Potts. Ces modèles n'ont jamais été utilisés dans ce cadre mais ils ont montré leurs potentiels pour combiner les conservations de séquences et de structures des protéines dans la prédiction des contacts et de structures dans les protéines où ils ont révolutionné le domaine. Nous voulons porter cette révolution à l'alignement de protéines et la recherche d'homologies. Dans ce projet nous proposons de développer une nouvelle méthode permettant l'alignement de 2 modèles de Potts, de construire une banque de modèles pour des protéines virales et d'utiliser ces modèles et cette méthode pour scanner des métagénomes viraux. Cette nouvelle méthode est aussi le premier pas vers une nouvelle méthodologie pour aligner les séquences protéiques et nous explorerons son potentiel aussi dans ce cadre. Ce projet fait l'objet d'un financement Emergence de Sorbonne Université. Si les résultats sont prometteurs, il sera ensuite possible de poursuivre son développement dans le cadre d'autres projets.

5.2.2 Combiner l'information de séquence et de structure

Notre vision de la diversité monde vivant est-elle biaisée ? Cela semble très probable car nos connaissances sont inférées des expériences menées sur un nombre relativement faible d'organismes macroscopiques ou microscopiques cultivables, or la majorité des lignées microbiennes restent à ce jour incultivables. Étudier les données environnementales, et notamment les données omiques (e.g. génomique, transcriptomique, métabolomique), nous permet d'appréhender de manière globale la diversité taxonomique et fonctionnelle de ces lignées. La génération de jeux de données massifs comme ceux de Earth Microbiome Project, Tara Océans ou MetaSub Project ainsi que des métadonnées associées (e.g. site de prélèvement, conditions environnementales) permet à ce jour l'analyse de la structuration de ces communautés à l'échelle de la planète. Au-delà des enjeux de santé publique et de la compréhension voire de la préservation du fonctionnement des écosystèmes et des cycles biogéochimiques associés, ces jeux de données offrent un éclairage inédit pour étudier les mécanismes adaptatifs et évolutifs sous-jacents des populations d'éléments génétiques mobiles (e.g. virus, girus, plasmides), de procaryotes (archées et eubactéries) et d'eucaryotes unicellulaires (« protistes ») voire pluricellulaires (e.g. fungi, larves de Métazoaires). Il est néanmoins difficile d'exploiter cette nouvelle masse d'information car beaucoup des séquences produites sont trop différentes de celles que nous connaissons pour que nous puissions transférer nos annotations fonctionnelles avec les méthodes classiques. Et pourtant, c'est une étape nécessaire pour donner du sens à cette masse de séquences biologiques. Je propose donc ici de tirer partie de la conservation des structures protéiques pour aider à l'annotation.

En collaboration avec Lucie Bittner, nous voulons utiliser les méthodes de prédiction de repliement et d'autres caractéristiques structurales des protéines (régions désordonnées, structures secondaires, segments transmembranaires) pour annoter les séquences issues des bases de données méta-

omiques incluant les données de la « matière noire microbienne» (i.e. des séquences taxinomiquement non attribuées au niveau du genre et/ou fonctionnellement non annotées). Lorsque qu'un repliement aura été prédit, il sera aussi alors possible de proposer une ou plusieurs annotations fonctionnelles. Nous développerons une toute nouvelle méthode d'annotation des données méta-omiques tirant parti de la faible divergence des structures protéiques. Nous établirons ensuite des profils de répartition des repliements pour les différentes communautés étudiées et tenterons de les relier aux conditions environnementales pour identifier les principaux acteurs impliqués dans certaines niches et biomes de l'océan.

Deuxième partie

Autres activités et CV

CHAPITRE 6

Curriculum vitae

Curriculum vitae

MATHILDE CARPENTIER
Née le 4 février 1978
Nationalité Française

Maître de conférences, Sorbonne Université
Co-responsable de l'équipe Atelier de Bioinformatique
ISYEB UMR 7205 - SU MNHN CNRS EPHE UA
45 rue Buffon - CP 50, 75231 Paris Cedex 05 FRANCE

mathilde.carpentier@sorbonne-universite.fr
+33 01 40 79 48 15

Recherche

2016-... **Maître de conférences**, ISYEB UMR 7205, MNHN, CNRS, SU, EPHE (Paris, France).
Thèmes : **Bioinformatique, Biologie Structurale, Evolution des protéines**

2010-16 **Maître de conférences**, IMPMC UMR 7590, Université Pierre and Marie Curie (Paris).
Thèmes : **Bioinformatique, Biologie Structurale, Mutations dans les protéines**

2007 **Chercheur invité**, Stanford University (CA, USA), groupe de Michael Levitt.
Thèmes : **Classification structurale des protéines et évolution de structures protéiques**

2006-10 **Maître de conférences**, groupe Analytical Genomics de l'UMRS 511, UPMC (Paris).
Thèmes : **Métagénomique, bioinformatique et biologie strucurale**

2005-06 **Post-doc** (ATER), Modeling in integrated Biology team, Université D. Diderot (Paris).
Thèmes : **Métagénomique virale**

2002-05 **Thèse**, directeurs : J. Pothier et P. Netter, Atelier de BioInformatique, UPMC.
Titre : **Méthodes de détection des similarités structurales : caractérisation des motifs conservés dans les familles de structures pour l'annotation des génomes.**

Formation académique

2005 **Doctorat** de l'Université Pierre et Marie Curie - Paris 6.

2002 **Master Analyse des génomes et modélisation moléculaire**, Universités Paris 6 et 7.

2001 **Maîtrise de biologie cellulaire et de physiologie, mention génétique moléculaire et cellulaire, option informatique**, Université Paris-Sud 11.

2000 **Licence de biologie cellulaire et de physiologie**, Université de Bretagne occidentale (UBO).

Enseignement

Current **Sorbonne Université , 192h/an**.
Domaines : Bioinformatique, algorithmique et programmation.

2005-06 **ATER à l'Université Denis Diderot - Paris 7, 96h/an**.
Domaines : Bioinformatique et statistiques.

2003-05 **Moniteur à l'Université Pierre and Marie Curie - Paris 6, 50h/an**.

2000-01 **Vacataire à l'Université Paris Sud, 40h**.

1999-00 **Vacataire à Université Bretagne Occidentale, 40h**.

Responsabilités actuelles

Co-responsable de l'équipe Atelier de Bioinformatique (2018-...).

Coordinateur scientifique du pole Analysis de l'UMS 2700 (2021-...).

Membre du comité d'organisation des rencontres annuelles bioinformatiques du MNHN (2019-...)

Membre du conseil d'unité de l'UMR ISYEB 2018-....

Co-organisateur avec M. Cariou des séminaires SAD-ABI-OMICS depuis 2020.

Co-responsable du parcours double majeure biologie-informatique

Co-responsable de l'UE LU3SV687, 30 étudiants

Responsable de l'UE LU1IN002, 700 étudiants

Responsabilités passées

Membre élu du conseil de l'UFR Ingénierie (2020-2021).

Membre du comité d'organisation des rencontres bioinformatiques du MNHN 2019, 2020, 2021, 2022.

Membre du jury de thèse de :
Amélie Héliou, 31/08/2017.
Arnaud Becheler, 31/03/2018.

Membre du comité de thèse de Damien Monet, institut Pasteur (2015-2018)

Membre élu du conseil académique (CAC) de l'UPMC (2016-2017)

Membre du comité d'organisation du Workshop Co-Evolution, Fitness landscapes and Epistasis (Paris, 1-3 mars 2017).

Membre du comité scientifique de JOBIM 2010.

Referee pour divers journaux et conférences (Nucleic Acids Research, BMC Bioinformatics, Bioinformatics, RECOMB, *etc.*).

Financements

Financement ATM indemnité de stage du MNHN 2021-2022
Annotation structurale de métagénomes marins.
Rôle : Porteur

Financement Emergence de Sorbonne Université 2021-2022
Utilisation des modèles de Potts pour l'alignement séquence-structure des protéines.
Rôle : Porteur

Financement ATM du MNHN 2019-2020
Conservation des contacts dans les structures protéiques.
Rôle : Porteur

Financement ATM du MNHN 2018
Détection des transferts horizontaux chez *Penicillium*.
Rôle : Porteur

Financement ANR (ANR-12-JSV7- 0007) 2012-2015
Tempomut : What constraints the order of mutations ?
Rôle : Membre

Financement PHC Polonium Egide under #27748NE 2012-2013
Determination of the core of globular proteins
Rôle : Membre

Financement PEPS BMI du CNRS 2013-2014
A la recherche des séquences virales dans les metagenomes
Rôle : Porteur

Encadrements

Encadrement de 4 étudiants de licence, 11 Masters 1, 7 Masters 2, un ingénieur et un doctorant

Janvier-juin 2007 : Stéphane Peugeot, M2 du Master biologie informatique de Paris 7.
Sujet : Annotation fonctionnelle de gènes orphelins de *Plasmodium falciparum*.

Juin-juillet 2012 : Damien Monet, L3 PIMA (mathématiques et informatique).
Sujet : Développement d'une méthode exacte d'alignement multiple de séquences.

Avril-juin 2013 : Meriem Aziz, M1 ISSD (Paris 7).
Sujet : Etude structurale des insertion et délétions dans les protéines.

Juin-juillet 2015 : Baptiste Mossoti, étudiant en M1 BIM (UPMC).
Sujet : Recherche de motifs structuraux dans les métagénomes viraux.

Avril-septembre 2015 : Benjamin Roques (co-direction avec G. Achaz), M1 AgroParisTech.
Sujet : Développement d'une méthode exacte d'alignement multiple de séquences.

Juin-juillet 2016 : Mojgan Karimi, M1 BIM (UPMC).
Sujet : Test d'une nouvelle méthode de détection de transferts horizontaux.

Février-juillet 2016 : Steven Fletcher, M2 BIM (UPMC)
Sujet : Etude structurale des co-évolutions dans les protéines.

Février 2016-février 2017 : Suvethigaa Shanthirabalan, IE, financé par l'ANR tempomut.
Sujet : Développement d'un potentiel statistique pour la simulation de l'évolution des protéines

Février-juillet 2017 : Clément Joubert, Master M2 BIM (UPMC)
Sujet : Evolution de la dynamique des protéines.

Février-juillet 2018 : Mojgan Karimi, Master M2 BIM (UPMC)
Sujet : Identification de transferts horizontaux chez Penicilium.

Février-juillet 2018 : Martin Romei, Master M2 BIM (UPMC)
Sujet : Evolution structurale des protéines.

2018-2021 : Martin Romei, IPV, Sorbonne Université PhD Grant.
Sujet : Evolution Structurale des protéines

Juillet-août 2018 : Ophélie Foucault et André Lanrezac, Master M1 BIM, Sorbonne Université.
Sujet : Amélioration du programme Yakusa, un BLAST structural

Juin-juillet 2019 : Huiyuan Li, L2 informatique, Sorbonne Université.
Sujet : Adaptation de l'algorithme de recherche de motifs KMR pour les cartes de contacts

Juillet-août 2019 : Emile Faure, M1 Master BIM, Sorbonne Université.
Sujet : Calcul de HMM pour la reconnaissance de repliement.

Janvier-juin 2020 : Miara Rakotomavo, M2 Master Bio-Informatique, Université de Paris
Sujet : Conservation des contacts dans les structures protéiques.

Mars-juin 2020 : Pierre Imbert, M1 Master Bio-Informatique, Université de Paris
Sujet : Combinaison des domaines structuraux dans les génomes.

Mars-juin 2020 : Nicolas Dibot, M1 Master Bio-Informatique, Université de Nantes.
Sujet : Effet des mutations ponctuelles dans les structures protéiques.

Mars-juin 2021 : Théo Jamay, M1 Master Bio-Informatique, Université de Paris
Sujet : Distribution des repliements protéiques dans le vivant.

Janvier-mars 2022 : Sophie Chareyre, L3 CNAM
Sujet : Superposition des structures des protéines

Mars-Août 2022 : Pauline Turk, M2 Paris Saclay et AgroParisTech
Sujet : Deep learning et évolution des protéines

Avril-juin 2022 : Jeremy Rousseau, M1 Master Bio-Informatique, Université de Rennes 1
Sujet : Distribution des repliements protéiques dans les métagénomes.

Avril-juin 2022 : Clara Toussaint, M1 Master Bio-Informatique, Université Paris Saclay
Sujet : Distribution des co-occurences d'acides aminés

Communications orales (selection)

GT MASIM du GDR BIM 25/11/2021, Lyon ;
Protein folds as synapomorphies of the tree of life ;
M. Carpentier.

Rencontres bioinformatique du MNHN 10/12/2020, Paris ;
Protein Multiple Alignments : Sequence-based vs Structure-based Programs ;
M. Carpentier.

Séminaire de l'IRISA 30/01/2020, Rennes ;
Protein Multiple Alignments : Sequence-based vs Structure-based Programs ;
M. Carpentier.

Journée scientifique de l'ISYEB 10/03/2019, Paris ;
Normal Mode Analysis and Protein Structure Evolution ;
M. Carpentier.

CECAM workshop Normal Mode Analysis 13/09/2018, Paris ;
Normal Mode Analysis and Protein Structure Evolution ;
M. Carpentier.

GT MASIM 16/11/2017, Paris ;
Structure et évolution des protéines ;
M. Carpentier.

JOBIM, 2004, Montreal, Canada,
Protein multiple and local structural alignment,
M. Carpentier, S. Brouillet, J. Pothier ; communication and poster.

ECCB (European Conference on Computational Biology), 2003, Paris,
Structural search in databases,
M. Carpentier, J. Pothier ; communication and poster.

Domaines d'intéret

Evolution structurale des protéines

Evolution

Structure des protéines

Dynamique des protéines (en particulier les modes normaux).

Algorithmes et programme d'alignement des structures et des séquences protéiques
Annotation structurale des protéines
Prédiction de la structure des protéines
Métagénomique des virus
Métagénomique environnementale, en particulier TARA
Algorithmes et développement de programmes pour la biologie
Programmer dans des langages variés (C, python, R, shell, awk, scheme...)
Comment enseigner la programmation et l'algorithmique

Publications

YAKUSA : a fast structural databases scanning method
Carpentier, M, Brouillet, S, Pothier, J.
Proteins : Structure, Function, and Bioinformatics (2005), Volume 61 (1), 137-51.

RPBS : a web resource for structural bioinformatics
Alland C, Moreews F, Boens D, **Carpentier M**, Chiusa S, Lonquety M, Renault N, Wong Y, Cantaloube H, Chomilier J, Hochez J, Pothier J, Villoutreix BO, Zagury JF, Tufféry P.
Nucleic Acid Research (2005), Volume 33, *Web Server Issue*, W44-9.

Incremental Inference of Relational Motifs with a Degenerate Alphabet
Pisanti N, Soldano H, **Carpentier,M.**
Lecture Notes in Computer Science, (proceedings CPM, Combinatorial Pattern Matching, 2005), Volume 3537, pages 229-40, editor : Springer.

A Relational Extension of the Notion of Motifs : Application to the Protein Common 3D Substructures Searching Problem
Pisanti N, Soldano H, **Carpentier M**, Pothier J.
Journal of Computational Biology, (2009), Volume 16 (12), 1635-60.

Genome sequence of the plant pathogenic bacterium Dickeya dadantii 3937
Glasner JD, Yang CH, Reverchon S, Hugouvieux-Cotte-Pattat N, Condemine G, Bohin JP, Van Gijsegem F, Yang S, Franza T, Expert D, Plunkett G 3rd, San Francisco MJ, Charkowski AO, Py B, Bell K, Rauscher L, Rodriguez-Palenzuela P, Toussaint A, Holeva MC, He SY, Douet V, Boccara M, Blanco C, Toth I, Anderson BD, Biehl BS, Mau B, Flynn SM, Barras F, Lindeberg M, Birch PR, Tsuyumu S, Shi X, Hibbing M, Yap MN, **Carpentier M**, Dassa E, Umehara M, Kim JF, Rusch M, Soni P, Mayhew GF, Fouts DE, Gill SR, Blattner FR, Keen NT, Perna NT.
Journal of Bacteriology, (2011) Volume 193.(8), 2076-2077 .

Contribution to the Prediction of the Fold Code : Application to Immunoglobulin and Flavodoxin Cases
Banach M, Prudhomme N, **Carpentier M**, Duprat E, Papandreou N, Kalinowska B, Chomilier J, Roterman I.
PLoS ONE, (2015), 10, e0125098.

Structural effects of point mutations in proteins
Shanthirabalan S, Chomilier J, **Carpentier M.**
Proteins, Structure, Function, and Bioinformatics, (2018), Volume 86, 853-867.

Identification of a Novel Elastin-Degrading Enzyme from the Fish Pathogen Flavobacterium psychrophilum
Rochat T, Pérez-Pascual D, Nilsen H, Carpentier M, Bridel S, Bernardet JF, Duchaud E.
Appl. Environ. Microbiol. (2019), Volume 85, e02535-18.

Protein multiple alignments : sequence-based versus structure-based programs
Carpentier M and Chomilier J.
Bioinformatics, (2019), Volume 35, 3970-3980.

Analyses of displacements resulting from a point mutation in proteins
Carpentier M and Chomilier J.
Journal of Structural Biology, (2020),Volume 211 (2), 107543.

Protein folds as synapomorphies of the tree of life
Romei M, Sapriel G, Imbert P, Chomilier J, Lecointre G, **Carpentier M.**
Evolution. 2022 ;76(8) :1706-19.

Origins and functional significance of Eukaryotic protein folds
Romei M, **Carpentier M.**, Sapriel G, Chomilier J, Lecointre G
Soumis (2022).

Book Chapters

Implicit and Explicit Representation of Approximated Motifs ;
N. Pisanti, H. Soldano, **M. Carpentier**, and J. Pothier ;
in *Algorithms for Bioinformatics*, editeur C. Iliopoulos, K. Park et K. Steinhofel, King's College London Press, 2006.

Protein Pairwise Structural Alignment Methods : a Review ;
M. Carpentier, et J. Pothier ;
dans *Recent Advances in Structural Biology*, editeur Alexandre G. de Brevern, ISBN : 978-81-308-0208-4, Research Signpost, Trivandrum, Kerala, India, 2007.

Impact of a Point Mutation in a Protein Structure ;
M. Carpentier, et J. Chomilier ;
editeurs MC Maurel et P. Grandcolas, ISTE edition, 2020.

Analyses of mutation displacements from homology models ;
M. Carpentier, et J. Chomilier ;
in Methods in Molecular Biology, editeur S. Filipek , by Springer Nature, 2022, in press.

CHAPITRE 7

Résumés des activités d'enseignement, d'encadrement, d'animation et de responsabilités diverses

Je me suis attachée ces dernières années à développer de manière équilibrée mon activité d'enseignant-chercheur tant côté charges collectives que recherche et enseignement. J'ai notamment pris des responsabilités dans les équipes : co-responsabilité de l'Atelier de Bioinformatique (ABI), coordination scientifique du pôle analyse du Muséum National d'Histoire Naturelle (MNHN), dans les instances collectives : conseil académique de Sorbonne Université (SU), conseil de l'Unité de Formation et de Recherche (UFR) d'ingénierie, conseil d'unité de l'Institut de Systématique et Biodiversité, participé à l'animation du champ scientifique (organisation de conférences, gestion de séminaires réguliers), pris des responsabilités en enseignement (lancement de la double majeure biologie-informatique, responsabilités d'unités d'enseignement), en plus des activités classiques de communication scientifique, d'encadrement, de recherche de financement et bien sûr la participation aux enseignements et à l'élaboration de leurs contenus.

7.1 Activité d'enseignement

Mon domaine d'expertise est la bioinformatique et j'ai été recrutée à l'UFR d'ingénierie pour enseigner l'informatique et la bioinformatique. C'est une activité que j'apprécie beaucoup, en particulier à cause du plaisir à accompagner les étudiants dans des domaines de connaissances que je trouve passionnants. C'est aussi une activité qui m'a permis (voire obligée à) d'approfondir mes connaissances dans des domaines que je n'aurais pas abordés avec la seule recherche. L'équilibre entre enseignement et recherche est néanmoins parfois difficile à maintenir. J'essaie de développer mon activité d'enseignement d'abord toujours en gardant à l'esprit l'intérêt des étudiants et ensuite selon trois axes : avoir une vue globale et précise de la formation des étudiants de licence en programmation et en algorithmique, contribuer à l'essor de la bioinformatique à SU et intégrer de nouvelles pédagogies dans mes enseignements.

Pour le premier aspect, j'ai participé aux débats autour de la refonte de la licence d'informatique qui a eu lieu pour le contrat quinquennal actuel. Je m'attache depuis à avoir une vision globale de la formation que reçoivent les étudiants en programmation impérative et en algorithmique. Je participe ou ai participé ainsi à la plupart des ue de programmation impératives et d'algorithmique, pour les TD et TP, les cours et l'élaboration des contenus et supports (cf. liste ci-dessous). J'ai aussi pris la responsabilité de l'UE de programmation du second semestre de L1 qui comporte environ 700 étudiants. Côté bioinformatique à SU, je suis responsable de la partie programmation d'une UE du master BIM (Bioinformatique et modélisation), d'une UE de bioinformatique en L3 de Biologie et de la licence disciplinaire intensive biologie-informatique (responsabilité partagée avec d'autres dans ces deux derniers cas). Cette licence permet d'offrir une formation d'excellente qualité à des étudiants qui se destinent souvent à la bioinformatique, la biologie des systèmes ou les neurosciences. Le troisième volet est le développement de nouvelles pédagogies qui est un projet à plus long terme. J'ai commencé à suivre des formations et à utiliser ces connaissances dans mes cours mais je souhaite m'inscrire dans une démarche plus collective qui sera plus efficace. J'ai suivi en 2017 les formations « méthode ABCD

(Activity Based Curriculum Design) » et « apprentissage par problème » (Université de Louvain) qui m'a particulièrement intéressée car je pense que ce type d'approche est particulièrement adapté à l'apprentissage de la programmation. J'ai aussi suivi plusieurs conférences proposées par la Capsule dont j'apprécie beaucoup l'activité. J'ai essayé d'intégrer certaines pratiques dans l'ue du master BIM mais je souhaite développer bien plus fortement cet aspect de mon activité notamment en contactant d'autres personnes de la Capsule pour tenter de constituer un groupe de réflexion et aussi en suivant la formation à l'apprentissage par problème proposé par l'Université de Louvain.

Voici la liste de mes activités depuis 5 ans, je ne suis pas remontée plus tôt.

— Conception (en équipe) du contenu et des supports (TD et TP) de l'ue élements de programmation 1 (LU1IN001).
— Conception (en équipe) du contenu et des supports (CM, TD, TP) de l'ue élements de programmation 2 (LU1IN002).
— Conception du contenu et des supports de la partie programmation de l'ue 4i700 du Master BIM.
— Conception du contenu de l'ue ISUP5G3 d'introduction à python (responsable de l'ue).
— Conception (en équipe) du contenu de l'ue modélisation, algorithmique et programmation pour la biologie (3V687, anciennement 3V686, co-responsable).
— Participation (TD et TP) à l'ue algorithmique appliquée (LU2IN006).
— Participation aux modules d'école doctorale et formation continue OBI.

7.2 Encadrement

Encadrer des stagiaires est aussi un vrai plaisir pour moi, et c'est aussi très enrichissant côté recherche. C'est notamment l'occasion d'explorer des pistes qui resteraient inexplorées sinon. Les travaux effectués lors des stages ne permettent donc pas toujours d'écrire une publication mais ils m'ont permis d'avancer dans mon cheminement et je suis très reconnaissante vis-à-vis des étudiants. J'encadre actuellement chaque année un à deux Master 2, un à deux Master 1 et quelquefois des étudiants de licence. Au total depuis 2007, j'ai donc suivi 4 étudiants de licence, 11 Masters 1, 7 Masters 2, un ingénieur et un doctorant. Beaucoup de ces stages ont été faits en co-direction. J'ai aussi été membre d'un comité de thèse et de deux jurys de thèse.

7.3 Responsabilités collectives

Notre communauté est essentiellement auto-gérée. Il me semble donc indispensable d'y participer : chacun doit faire sa part. J'ai donc été dans plusieurs instances collectives : le conseil académique de Sorbonne Université, le conseil de l'UFR d'ingénierie et le conseil de l'UMR 7205 ISYEB. J'ai aussi pris avec plaisir la co-responsabilité de l'ABI avec Lucie Bittner. L'ABI est une structure ancienne (Danchin, 2000) qui a accompagné l'essor de la bioinformatique en France et déménagé

plusieurs fois. J'espère que j'arriverai à garder son esprit d'accueil, avec la chaleur et la dynamique qui en font un lieu merveilleux pour faire de la recherche. J'ai aussi été coordinateur scientifique du pôle analyse de l'UAR 2AD du MNHN. Le MNHN souhaite développer la bioinformatique au service des thématiques de recherche qui lui sont spécifiques, et je pense qu'il a tout à gagner à cela. J'ai accepté cette responsabilité pour accompagner le MNHN dans la diffusion de la bioinformatique. J'ai cependant mis en suspens cette responsabilité faute de temps suffisant à y consacrer. J'espère pouvoir la faire reconnaître par Sorbonne Université et ainsi demander une décharge d'enseignement me permettant d'y accorder le temps nécessaire.

7.4 Financements

La recherche de financements représente maintenant une part importante de notre travail. J'ai obtenu plusieurs financements en tant que porteur (emergence SU, ATM MNHN, PEPS CNRS, financement de thèse SU) et ai participé à d'autres (ANR, PHC Polonium). J'ai aussi déposé de nombreux projets qui n'ont pas été sélectionnés (ANR, projet Ile de France, FRM, bourses de thèse SU, *etc.*) ce qui est assez chronophage. J'avoue avoir de gros doutes sur l'efficacité de ces pratiques de financement par projets hyper-sélectifs qui sont pourtant de plus en plus répandues. Je m'attache néanmoins à écrire et soumettre des projets régulièrement. Mes prochaines demandes de financement seront une thèse pour le projet d'annotation structurale et fonctionnelle de la matière noire marine, ainsi qu'une ANR aussi sur ce sujet. J'ai participé à la rédaction du projet PEPR *Atlasea* et j'espère donc pouvoir y participer par la suite, toujours sur l'annotation structurale et fonctionnelle. Ensuite, je souhaite aussi (re)soumettre une demande de financement de thèse sur l'évolution des repliements chez les eucaryotes.

7.5 Diffusion de mon activité de recherche

Ma production scientifique, si elle n'atteint pas le niveau de certains, est, j'espère, tout à fait honnête. Je suis très active côté recherche et je publie régulièrement des articles et des chapitres d'ouvrages. Pour être sincère, je ne vois pas comment je pourrais augmenter ma productivité en termes de nombre d'articles, et c'est une question qui pourtant me travaille, ou du moins m'a travaillée. Je ne peux pas travailler plus d'heures car ce serait au détriment de l'équilibre avec ma vie personnelle. Je pourrais passer moins de temps à discuter avec les uns et les autres, mais je crains de perdre alors en créativité et surtout que le métier perde en saveur. Je pourrais essayer de restreindre plus mon domaine de recherche, mais étant curieuse et enthousiaste, c'est assez difficile pour moi (mais j'essaie...). Et puis je reste pratiquement toujours dans mon domaine, et ce que je fais s'intègre bien dans mon questionnement plus global. J'ai beaucoup cherché à améliorer mon organisation et mon efficacité, j'ai beaucoup lu à ce propos. Cela m'a beaucoup aidé et j'ai beaucoup gagné en efficacité, mais j'ai fini par me rendre compte - il n'y a pas si longtemps - que plus j'étais efficace, plus je faisais de choses,

en restant toujours à la limite de ce que je pouvais assumer, et en ayant toujours l'impression d'être débordée. J'ai donc décidé maintenant d'arrêter d'essayer d'optimiser mon temps pour en faire plus (de toute façon je ne suis pas sûre que je pourrais encore faire mieux tout en conservant ma qualité de vie, de travail et de relations avec les autres). J'ai décidé de plutôt me concentrer sur savourer ce que je fais, et tant pis si je ne publie pas plus d'articles par an :-)

Par contre, je fais chaque année de une à trois présentations de mes travaux de recherche mais je pense qu'il faudrait que j'en fasse plus, en particulier à l'international. Côté animation scientifique, je m'occupe chaque année des journées bioinformatique du MNHN et d'un séminaire bioinformatique régulier proposé au MNHN. Enfin, l'aspect diffusion et transfert des connaissances au grand public manque dans mon activité et je m'attacherai à le développer dans les années à venir pour notamment promouvoir la place des femmes en sciences et particulièrement en informatique. Ceci pourra être fait par des présentations dans les lycées et par des articles dans The conversation (theconversation.com), site de diffusion des nouvelles scientifiques très utilisé dans l'UMR.

7.6 Conclusion

Une autre chose que j'adorerais serait de faire des séjours scientifiques à l'étranger. C'est assez difficile pour moi à organiser à cause des enseignements et de la vie de famille, mais je ne désespère pas. J'adore ce métier, j'adore la recherche, et j'espère avoir pu partager mon plaisir au long de ce manuscrit.

Bibliographie et annexes

Bibliographie

Adam, P. S., Borrel, G., Brochier-Armanet, C., and Gribaldo, S. (2017). The growing tree of Archaea : New perspectives on their diversity, evolution and ecology. *The ISME Journal*, 11(11) :2407–2425.

Adams, M. A., Suits, M. D. L., Zheng, J., and Jia, Z. (2007). Piecing together the structure-function puzzle : Experiences in structure-based functional annotation of hypothetical proteins. *Proteomics*, 7(16) :2920–2932.

Afriat-Jurnou, L., Jackson, C. J., and Tawfik, D. S. (2012). Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry*, 51(31) :6047–6055.

Agarwal, G., Mahajan, S., Srinivasan, N., and de Brevern, A. G. (2011). Identification of Local Conformational Similarity in Structurally Variable Regions of Homologous Proteins Using Protein Blocks. *PLOS ONE*, 6(3) :e17826.

Ahn, M., Hagan, C. L., Bernardo-Gancedo, A., De Genst, E., Newby, F. N., Christodoulou, J., Dhulesia, A., Dumoulin, M., Robinson, C. V., Dobson, C. M., and Kumita, J. R. (2016). The Significance of the Location of Mutations for the Native-State Dynamics of Human Lysozyme. *Biophysical Journal*, 111(11) :2358–2367.

Aho, AV. and Corasick, HJ. (1975). Efficient string matching : An aid to bibliographic search. *Comm. ACM*, 18(6) :333–340.

Akiva, E., Itzhaki, Z., and Margalit, H. (2008). Built-in loops allow versatility in domain–domain interactions : Lessons from self-interacting domains. *Proceedings of the National Academy of Sciences*, 105(36) :13292–13297.

Al Ait, L., Yamak, Z., and Morgenstern, B. (2013). DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Research*, 41(W1) :W3–W7.

Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50) :21149–21154.

Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J., Hochez, J., Pothier, J., Villoutreix, B. O., Zagury, J. F., and Tuffery, P. (2005). RPBS : A web resource for structural bioinformatics. *Nucleic Acids Research*, 33(Web Server issue) :W44–9.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–10.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : A new generation of protein database search programs. *Nucleic Acids Research*, 25(17) :3389–3402.

Alva, V., Koretke, K. K., Coles, M., and Lupas, A. N. (2008). Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Current Opinion in Structural Biology*, 18(3) :358–365.

Alva, V., Remmert, M., Biegert, A., Lupas, A. N., and Söding, J. (2010). A galaxy of folds. *Protein Science : A Publication of the Protein Society*, 19(1) :124–130.

Alva, V., Söding, J., and Lupas, A. N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, 4 :e09410.

Alvarez-Ponce, D., Lopez, P., Bapteste, E., and McInerney, J. O. (2013). Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 110(17) :E1594–1603.

Anantharaman, V., Aravind, L., and Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Current Opinion in Chemical Biology*, 7(1) :12–20.

Andonov, R., Malod-Dognin, N., and Yanev, N. (2011). Maximum Contact Map Overlap Revisited. *Journal of Computational Biology*, 18(1) :27–41.

Andreeva, A. and Murzin, A. G. (2006). Evolution of protein fold in the presence of functional constraints. *Current Opinion in Structural Biology*, 16(3) :399–408.

Andreeva, A., Prlić, A., Hubbard, T. J. P., and Murzin, A. G. (2007). SISYPHUS–structural alignments for proteins with non-trivial relationships. *Nucleic Acids Research*, 35(Database issue) :D253–9.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, NY)*, 181(96) :223–230.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 114(34) :9122–9127.

Arenas, M., Dos Santos, H. G., Posada, D., and Bastolla, U. (2013). Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics*, 29(23) :3020–3028.

Artymiuk, P. and Blake, C. (1981). Refinement of human lysozyme at 1.5 Å resolution analysis of non-bonded and hydrogen-bond interactions. *Journal of Molecular Biology*, 152(4) :737–762.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology : Tool for the unification of biology. *Nature Genetics*, 25(1) :25–29.

Avino, M. and Poon, A. F. Y. (2019). Detecting Amino Acid Coevolution with Bayesian Graphical Models. *Methods in Molecular Biology (Clifton, N.J.)*, 1851 :105–122.

Bahr, A., Thompson, J., Thierry, J., and Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE) : Enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research*, 29(1) :323–326.

Bajaj, M. and Blundell, T. (1984). Evolution and the tertiary structure of proteins. *Annual review of biophysics and bioengineering*, 13(1) :453–492.

Baltzis, A., Mansouri, L., Jin, S., Langer, B. E., Erb, I., and Notredame, C. (2022). Highly significant improvement of protein sequence alignments with AlphaFold2. *Bioinformatics*, page btac625.

Banach, M., Prudhomme, N., Carpentier, M., Duprat, E., Papandreou, N., Kalinowska, B., Chomilier, J., and Roterman, I. (2015). Contribution to the Prediction of the Fold Code : Application to Immunoglobulin and Flavodoxin Cases. *PloS one*, 10(4) :e0125098.

Barnes, J. E., Miller, C. R., and Ytreberg, F. M. (2022). Searching for a mechanistic description of pairwise epistasis in protein systems. *Proteins*, 90(7) :1474–1485.

Bastolla, U., Porto, M., Eduardo Roman, M. H., and Vendruscolo, M. H. (2003). Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *Journal of Molecular Evolution*, 56(3) :243–254.

Berbalk, C., Schwaiger, C. S., and Lackner, P. (2009). Accuracy analysis of multiple structure alignments. *Protein Science*, 18(10) :2027–2035.

Berchtold, A. and Raftery, AE. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17 :328–356.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1) :235–242.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D. S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121) :929–932.

Birzele, F., Gewehr, J. E., Csaba, G., and Zimmer, R. (2007). Vorolign–fast structural alignment using Voronoi contacts. *Bioinformatics (Oxford, England)*, 23(2) :e205–11.

Bleyzac, P., Exbrayat, J.-M., and Fellah, J. S. (2005). Émergence du système immunitaire adaptatif : Hypothèses en présence. *médecine/sciences*, 21(2) :210–215.

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006). Structural Determinants of the Rate of Protein Evolution in Yeast. *Molecular Biology and Evolution*, 23(9) :1751–1761.

Bloom, J. D., Raval, A., and Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. *Genetics*, 175(1) :255–266.

Bobay, B. G., Thompson, R. J., Hoch, J. A., and Cavanagh, J. (2010). Long range dynamic effects of point-mutations trap a response regulator in an active conformation. *FEBS Letters*, 584(19) :4203–4207.

Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V., Hutchinson, W. L., Fraser, P. E., Hawkins, P. N., Dobson, C. M., Radford, S. E., Blake, C. C., and Pepys, M. B. (1997). Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature*, 385(6619) :787–793.

Bordner, A. J. and Abagyan, R. A. (2004). Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins*, 57(2) :400–413.

Boucher, J. I., Bolon, D. N. A., and Tawfik, D. S. (2016). Quantifying and understanding the fitness effects of protein mutations : Laboratory versus nature. *Protein Science*, 25(7) :1219–1226.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421) :535–538.

Britten, R. J. (2010). Transposable element insertions have strongly affected human evolution. *Proceedings of the National Academy of Sciences*, 107(46) :19945–19948.

Brocchieri, L. and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10) :3390–3400.

Buell, A. K., Dhulesia, A., Mossuto, M. F., Cremades, N., Kumita, J. R., Dumoulin, M., Welland, M. E., Knowles, T. P. J., Salvatella, X., and Dobson, C. M. (2011). Population of Nonnative States of Lysozyme Variants Drives Amyloid Fibril Formation. *Journal of the American Chemical Society*, 133(20) :7737–7743.

Cabrita, L. D. and Bottomley, S. P. (2004). Protein expression and refolding–a practical guide to getting the most out of inclusion bodies. *Biotechnology Annual Review*, 10 :31–50.

Caetano-Anollés, G. and Caetano-Anollés, D. (2003). An Evolutionarily Structured Universe of Protein Architecture. *Genome Research*, 13(7) :1563–1571.

Caetano-Anollés, G. and Caetano-Anollés, D. (2005). Universal Sharing Patterns in Proteomes and Evolution of Protein Fold Architecture and Life. *Journal of Molecular Evolution*, 60(4) :484–498.

Caetano-Anollés, G. and Nasir, A. (2012). Benefits of using molecular structure and abundance in phylogenomic analysis. *Frontiers in Genetics*, 3 :172.

Caetano-Anollés, G., Nasir, A., Zhou, K., Caetano-Anollés, D., Mittenthal, J. E., Sun, F.-J., and Kim, K. M. (2014). Review Article Archaea : The First Domain of Diversified Life. *Archaea*, pages 1–26.

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., and Mittenthal, J. E. (2009). The origin, evolution and structure of the protein world. *Biochemical Journal*, 417(3) :621–637.

Camps, M., Herman, A., Loh, E., and Loeb, L. A. (2007). Genetic Constraints on Protein Evolution. *Critical reviews in biochemistry and molecular biology*, 42(5) :10.1080/10409230701597642.

Carbone, M. N. and Arnold, F. H. (2007). Engineering by homologous recombination : Exploring sequence and function within a conserved fold. *Current Opinion in Structural Biology*, 17(4) :454–459.

Carpentier, M., Brouillet, S., and Pothier, J. (2005). YAKUSA : A fast structural database scanning method. *Proteins*, 61(1) :137–151.

Carpentier, M. and Chomilier, J. (2019). Protein multiple alignments : Sequence-based versus structure-based programs. *Bioinformatics*, 35(20) :3970–3980.

Carpentier, M. and Chomilier, J. (2020). Analyses of displacements resulting from a point mutation in proteins. *Journal of Structural Biology*, 211(2) :107543.

Carpentier, M. and Chomilier, J. (2021). Impact of a Point Mutation in a Protein Structure. In *Systematics and the Exploration of Life*, chapter 2, pages 17–31. John Wiley & Sons, Ltd, p. grandcolas & m.-c. maurel edition.

Carpentier, M. and Chomilier, J. (in progress). Analyses of mutation displacements from homology models. In *Homology Modeling Methods and Protocols*, Methods in Molecular Biology. Springer, filipek, s. edition.

Carpentier, M. and Pothier, J. (2007). Protein Pairwise Structural Alignment Methods : A Review. *Recent Advances in Structural Biology*, page 14.

Carter, C. W. (2021). Simultaneous codon usage, the origin of the proteome, and the emergence of de-novo proteins. *Current Opinion in Structural Biology*, 68 :142–148.

Chakrabarti, S. and Panchenko, A. R. (2009). Coevolution in defining the functional specificity. *Proteins*, 75(1) :231–240.

Challis, C. J. and Schmidler, S. C. (2012). A Stochastic Evolutionary Model for Protein Structure Alignment and Phylogeny. *Molecular Biology and Evolution*, 29(11) :3575–3587.

Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2017). SCOPe : Manual Curation and Artifact Removal in the Structural Classification of Proteins - extended Database. *Journal of Molecular Biology*, 429(3) :348–355.

Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2019). SCOPe : Classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research*, 47(D1) :D475–D481.

Chaurasia, S. and Dutheil, J. Y. (2022). The Structural Determinants of Intra-Protein Compensatory Substitutions. *Molecular Biology and Evolution*, 39(4) :msac063.

Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H., and Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Engineering*, 10(4) :307–316.

Chen, J. Z., Fowler, D. M., and Tokuriki, N. (2020). Comprehensive exploration of the translocation, stability and substrate recognition requirements in VIM-2 lactamase. *eLife*, 9 :e56707.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H., and Grishin, N. V. (2014). ECOD : An Evolutionary Classification of Protein Domains. *PLoS Computational Biology*, 10(12) :e1003926–18.

Chew, L. P., Huttenlocher, D., Kedem, K., and Kleinberg, J. (1999). Fast detection of common geometric substructure in proteins. *J Comput Biol*, 6(3-4) :313–325.

Chi, E. Y., Krishnan, S., Randolph, T. W., and Carpenter, J. F. (2003). Physical Stability of Proteins in Aqueous Solution : Mechanism and Driving Forces in Nonnative Protein Aggregation. *Pharmaceutical Research*, 20(9) :1325–1336.

Choe, S. E., Li, L., Matsudaira, P. T., Wagner, G., and Shakhnovich, E. I. (2000). Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction. *Journal of Molecular Biology*, 304(1) :99–115.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4) :823–826.

Clarkson, M. W. and Lee, A. L. (2004). Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry*, 43(39) :12448–12458.

Colavin, A., Atolia, E., Bitbol, A.-F., and Huang, K. C. (2022). Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Scientific Reports*, 12(1) :820.

Cooley, R. B., Arp, D. J., and Karplus, P. A. (2010). Evolutionary origin of a secondary structure : π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality. *Journal of molecular biology*, 404(2) :232–246.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W. P. C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, 391(6664) :288–291.

Crochemore, M. and Rytter, W. (1991). Usefulness of the Karp-Miller-Rosenberg algorithm in parallel computations on strings and arrays. *Theoretical Computer Science*, 88(1) :59–82.

Csaba, G., Birzele, F., and Zimmer, R. (2009). Systematic comparison of SCOP and CATH : A new gold standard for protein structure analysis. *BMC Structural Biology*, 9 :23.

Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A., Garratt, R., Thornton, J., and Orengo, C. (2009). The CATH Hierarchy Revisited—Structural Divergence in Domain Superfamilies and the Continuity of Fold Space. *Structure(London, England :1993)*, 17(8-8) :1051–1062.

Danchin, A. (2000). A brief history of genome research and bioinformatics in France. *Bioinformatics (Oxford, England)*, 16(1) :65–75.

Daniels, N. M., Nadimpalli, S., and Cowen, L. J. (2012). Formatt : Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC bioinformatics*, 13(1) :259.

Davis, I. W., Arendall, W. B., Richardson, D. C., and Richardson, J. S. (2006). The backrub motion : How protein backbone shrugs when a sidechain dances. *Structure (London, England : 1993)*, 14(2) :265–274.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017). CATH : An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45(D1) :D289–D295.

Dayhoff, M. (1972). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5 :89–99.

Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983). Establishing homologies in protein sequences. *Methods Enzymol*, 91 :524–545.

de Brevern, A. G., Etchebest, C., and Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3) :271–287.

de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W., and Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 93(25) :14632–14636.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L. (2005). Missense meanderings in sequence space : A biophysical view of protein evolution. *Nature reviews Genetics*, 6(9) :678–687.

Dessailly, B. H. and Orengo, C. A. (2009). Function Diversity Within Folds and Superfamilies. In Rigden, D. J., editor, *From Protein Structure to Function with Bioinformatics*, pages 143–166. Springer Netherlands, Dordrecht.

Devos, D. and Valencia, A. (2000). Practical limits of function prediction. *Proteins*, 41(1) :98–107.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in genetics : TIG*, 17(8) :429–431.

Ding, F. and Dokholyan, N. V. (2006). Emergence of protein fold families through rational design. *PLoS computational biology*, 2(7) :e85.

Dishman, A. F. and Volkman, B. F. (2022). Design and discovery of metamorphic proteins. *Current Opinion in Structural Biology*, 74 :102380.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons : Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2) :330–340.

Dokholyan, N. V., Shakhnovich, B., and Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 99(22) :14132–14136.

Dong, R., Peng, Z., Zhang, Y., and Yang, J. (2018). mTM-align : An algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics (Oxford, England)*, 34(10) :1719–1725.

Doolittle, R. F. (1995). The Multiplicity of Domains in Proteins. *Annual Review of Biochemistry*, 64(1) :287–314.

Dror, O., Benyamini, H., Nussinov, R., and Wolfson, H. (2003). MASS : Multiple structural alignment by secondary structures. *Bioinformatics (Oxford, England)*, 19(90001) :95i–104.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 102(40) :14338–14343.

Dunbrack, R. L. (2002). Rotamer Libraries in the 21st Century. *Current Opinion in Structural Biology*, 12(4) :431–440.

Dunbrack, R. L. and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature structural biology*, 1(5) :334–340.

Dutheil, J. and Galtier, N. (2007). Detecting groups of coevolving positions in a molecule : A clustering approach. *BMC evolutionary biology*, 7 :242.

Dutheil, J., Pupko, T., Jean-Marie, A., and Galtier, N. (2005). A model-based approach for detecting coevolving positions in a molecule. *Molecular Biology and Evolution*, 22(9) :1919–1928.

Dutheil, J. Y. (2012). Detecting coevolving positions in a molecule : Why and how to account for phylogeny. *Briefings in Bioinformatics*, 13(2) :228–243.

Ebert, J. and Brutlag, D. (2006). Development and validation of a consistency based multiple structure alignment algorithm. *Bioinformatics (Oxford, England)*, 22(9) :1080–1087.

Echave, J. (2008). Evolutionary divergence of protein structure : The linearly forced elastic network model. *Chemical Physics Letters*, 457(4) :413–416.

Echave, J. and Fernández, F. M. (2010). A perturbative view of protein structural variation. *Proteins*, 78(1) :173–180.

Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature reviews Genetics*, 17(2) :109–121.

Echave, J. and Wilke, C. O. (2017). Biophysical Models of Protein Evolution : Understanding the Patterns of Evolutionary Sequence Divergence. *Annual review of biophysics*, 46(1) :annurev–biophys–070816–033819.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 3 :114–120.

Edgar, R. C. (2004a). Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Research*, 32(1) :380–385.

Edgar, R. C. (2004b). MUSCLE : A multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1) :113.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1) :D427–D432.

Enright, A. J., van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7) :1575–1584.

Eudes, R., Le Tuan, K., Delettré, J., Mornon, J.-P., and Callebaut, I. (2007). A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Structural Biology*, 7(1) :2.

Fallaize, C. J., Green, P. J., Mardia, K. V., and Barber, S. (2020). Bayesian protein sequence and structure alignment. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 69(2) :301–325.

Farías-Rico, J. A., Schmidt, S., and Höcker, B. (2014). Evolutionary relationship of two ancient protein superfolds. *Nature Chemical Biology*, 10(9) :710–715.

Farris, J. S. (1989). The Retention Index and the Rescaled Consistency Index. *Cladistics*, 5(4) :417–419.

Faure, G. and Koonin, E. V. (2015). Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Physical Biology*, 12(3) :035001.

Fersht, A. R., Matouschek, A., and Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *Journal of Molecular Biology*, 224(3) :771–782.

Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6) :1581–1592.

Fowler, D. M., Stephany, J. J., and Fields, S. (2014). Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*, 9(9) :2267–2284.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014). SCOPe : Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1) :D304–D309.

Franzosa, E. A. and Xia, Y. (2009). Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution*, 26(10) :2387–2395.

Franzosa, E. A. and Xia, Y. (2012). Independent Effects of Protein Core Size and Expression on Residue-Level Structure-Evolution Relationships. *PLOS ONE*, 7(10) :e46602.

Friedland, G. D., Lakomek, N.-A., Griesinger, C., Meiler, J., and Kortemme, T. (2009). A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Computational Biology*, 5(5) :e1000393.

Fukami-Kobayashi, K., Schreiber, D., and Benner, S. (2002). Detecting Compensatory Covariation Signals in Protein Evolution Using Reconstructed Ancestral Sequences. *Journal of Molecular Biology*, 319(3) :729–743.

Funahashi, J., Takano, K., Yamagata, Y., and Yutani, K. (2000). Role of Surface Hydrophobic Residues in the Conformational Stability of Human Lysozyme at Three Different Positions,. *Biochemistry*, 39(47) :14448–14456.

Funahashi, J., Takano, K., Yamagata, Y., and Yutani, K. (2002). Positive Contribution of Hydration Structure on the Surface of Human Lysozyme to the Conformational Stability. *Journal of Biological Chemistry*, 277(24) :21792–21800.

Gautier, R. and Tufféry, P. (2003). Critical assessment of side-chain conformational space sampling procedures designed for quantifying the effect of side-chain environment. *Journal of Computational Chemistry*, 24(15) :1950–1961.

Gelly, J.-C., Lin, H.-Y., de Brevern, A. G., Chuang, T.-J., and Chen, F.-C. (2012). Selective constraint on human pre-mRNA splicing by protein structural properties. *Genome Biology and Evolution*, 4(9) :966–975.

Gerlt, J. A. and Babbitt, P. C. (2001). Divergent evolution of enzymatic function : Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Review of Biochemistry*, 70 :209–246.

Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes : A comprehensive structural census. *Proteins*, 33(4) :518–534.

Gerstein, M. and Altman, R. B. (1995a). Average core structures and variability measures for protein families : Application to the immunoglobulins. *Journal of Molecular Biology*, 251(1) :161–175.

Gerstein, M. and Altman, RB. (1995b). Using a measure of structural variation to define a core for the globins. *Computer applications in the biosciences : CABIOS*, 11(6) :633–644.

Gerstein, M. and Hegyi, H. (1998). Comparing genomes in terms of protein structure : Surveys of a finite parts list. *FEMS Microbiology Reviews*.

Getz, G., Vendruscolo, M., Sachs, D., and Domany, E. (2002). Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins*, 46(4) :405–415.

Glasner, J. D., Yang, C.-H., Reverchon, S., Hugouvieux-Cotte-Pattat, N., Condemine, G., Bohin, J.-P., Van Gijsegem, F., Yang, S., Franza, T., Expert, D., Plunkett, G., San Francisco, M. J., Charkowski, A. O., Py, B., Bell, K., Rauscher, L., Rodriguez-Palenzuela, P., Toussaint, A., Holeva, M. C., He, S. Y., Douet, V., Boccara, M., Blanco, C., Toth, I., Anderson, B. D., Biehl, B. S., Mau, B., Flynn, S. M., Barras, F., Lindeberg, M., Birch, P. R. J., Tsuyumu, S., Shi, X., Hibbing, M., Yap, M.-N.,

Carpentier, M., Dassa, E., Umehara, M., Kim, J. F., Rusch, M., Soni, P., Mayhew, G. F., Fouts, D. E., Gill, S. R., Blattner, F. R., Keen, N. T., and Perna, N. T. (2011). Genome sequence of the plant-pathogenic bacterium Dickeya dadantii 3937. *Journal of bacteriology*, 193(8) :2076–2077.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins : Structure, Function, and Bioinformatics*, 18(4) :309–317.

Godzik, A. (1996). The structural alignment between two proteins : Is there a unique answer ? *Protein science : a publication of the Protein Society*, 5(7) :1325–1338.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2) :283–293.

Golden, M., García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T., and Hein, J. (2017). A Generative Angular Model of Protein Structure Evolution. *Molecular Biology and Evolution*, 34(8) :2085–2100.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149(1) :445–458.

Golubchik, T., Wise, M. J., Easteal, S., and Jermiin, L. S. (2007). Mind the Gaps : Evidence of Bias in Estimates of Multiple Sequence Alignments. *Molecular Biology and Evolution*, 24(11) :2433–2442.

Gong, S., Worth, C. L., Bickerton, G. R. J., Lee, S., Tanramluk, D., and Blundell, T. L. (2009). Structural and functional restraints in the evolution of protein families and superfamilies. *Biochemical Society Transactions*, 37(4) :727–733.

Gough, J. (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics (Oxford, England)*, 21(8) :1464–1471.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(Web Server) :W695–W699.

Govindarajan, S., Ness, J. E., Kim, S., Mundorff, E. C., Minshull, J., and Gustafsson, C. (2003). Systematic Variation of Amino Acid Substitutions for Stringent Assessment of Pairwise Covariation. *Journal of Molecular Biology*, 328(5) :1061–1069.

Grishin, N. V. (1997). Estimation of evolutionary distances from protein spatial structures. *Journal of Molecular Evolution*, 45(4) :359–369.

Grishin, N. V. (2001). Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*, 134(2-3) :167–185.

Gromiha, M. M. and Sarai, A. (2010). Thermodynamic database for proteins : Features and applications. *Methods in Molecular Biology (Clifton, N.J.)*, 609 :97–112.

Guda, C., Lu, S., Scheeff, E. D., Bourne, P. E., and Shindyalov, I. N. (2004). CE-MC : A multiple protein structure alignment server. *Nucleic Acids Research*, 32(Web Server issue) :W100–3.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes : A study of more than 1000 mutations. *Journal of Molecular Biology*, 320(2) :369–387.

Guo, H. H., Choe, J., and Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 101(25) :9205–9210.

Gutin, A. M. and Badretdinov, A. Y. (1994). Evolution of protein 3D structures as diffusion in multidimensional conformational space. *Journal of Molecular Evolution*, 39(2) :206–209.

Gutte, B. and Merrifield, R. (1971). The Synthesis of Ribonuclease A. *Journal of Biological Chemistry*, 246(6) :1922–1941.

Haber, E. and Anfinsen, C. B. (1962). Side-chain Interactions Governing the Pairing of Half-cystine Residues in Ribonuclease. *Journal of Biological Chemistry*, 237(6) :1839–1844.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors : Evolutionary units of three-dimensional structure. *Cell*, 138(4) :774–786.

Haliloglu, T. and Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology*, 35 :17–23.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. (2002). Quantifying the Similarities within Fold Space. *Journal of Molecular Biology*, 323(5) :909–926.

Hashimoto, K. and Panchenko, A. R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proceedings of the National Academy of Sciences*, 107(47) :20352–20357.

Hegyi, H. and Gerstein, M. (2001). Annotation transfer for genomics : Measuring functional divergence in multi-domain proteins. *Genome Research*, 11(10) :1632–1640.

Herman, J. L. (2019). Enhancing Statistical Multiple Sequence Alignment and Tree Inference Using Structural Information. *Methods in Molecular Biology (Clifton, N.J.)*, 1851 :183–214.

Herman, J. L., Challis, C. J., Novák, Á., Hein, J., and Schmidler, S. C. (2014). Simultaneous Bayesian Estimation of Alignment and Phylogeny under a Joint Model of Protein Sequence and Structure. *Molecular Biology and Evolution*, 31(9) :2251–2266.

Heyne, M., Shirian, J., Cohen, I., Peleg, Y., Radisky, E. S., Papo, N., and Shifman, J. M. (2021). Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein–Protein Complexes. *Journal of the American Chemical Society*, 143(41) :17261–17275.

Hirschmann, R., Nutt, R. F., Veber, D. F., Vitali, R. A., Varga, S. L., Jacob, T. A., Holly, F. W., and Denkewalter, R. G. (1969). Studies on the total synthesis of an enzyme. V. The preparation of enzymatically active material. *Journal of the American Chemical Society*, 91(2) :507–508.

Holland, T. A., Veretnik, S., Shindyalov, I. N., and Bourne, P. E. (2006). Partitioning Protein Structures into Domains : Why Is it so Difficult ? *Journal of Molecular Biology*, 361(3) :562–590.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1 :16048.

Iliopoulos, I., Tsoka, S., Andrade, M. A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C., and Ouzounis, C. A. (2001). Genome sequences and great expectations. *Genome Biology*, 2(1) :interactions0001.1–interactions0001.3.

Illergård, K., Ardell, D. H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins*, 77(3) :499–508.

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, 36(9) :2069–2085.

Jakuschev, S. and Hoffmann, D. (2009). A Novel Algorithm for Macromolecular Epitope Matching. *Algorithms*, 2(1) :498–517.

Jayaraman, V., Toledo-Patiño, S., Noda-García, L., and Laurino, P. (2022). Mechanisms of protein evolution. *Protein Science*, 31(7).

Jeffery, C. J. (2003). Moonlighting proteins : Old proteins learning new tricks. *Trends in genetics : TIG*, 19(8) :415–417.

Jochens, H., Stiba, K., Savile, C., Fujii, R., Yu, J.-G., Gerassenkov, T., Kazlauskas, R. J., and Bornscheuer, U. T. (2009). Converting an Esterase into an Epoxide Hydrolase. *Angewandte Chemie International Edition*, 48(19) :3532–3535.

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L. S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadié, H., Schneider, B., Etchebest, C., Srinivasan, N., and De Brevern, A. G. (2010). A short survey on protein blocks. *Biophysical Reviews*, 2(3) :137–147.

Joseph, A. P., Srinivasan, N., and de Brevern, A. G. (2012). Progressive structure-based alignment of homologous proteins : Adopting sequence comparison strategies. *Biochimie*, 94(9) :2025–2034.

Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of Molecular Biology*, 362(4) :861–875.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873) :583–589.

Jumper, J. and Hassabis, D. (2022). Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods*, 19(1) :11–12.

Kaiser, F., Eisold, A., Bittrich, S., and Labudde, D. (2016). Fit3D : A web application for highly accurate screening of spatial residue patterns in protein structure data. *Bioinformatics (Oxford, England)*, 32(5) :792–794.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT : A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14) :3059–3066.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability. *Mol Biol Evol*, 30(4) :772–780.

Kauko, A. and Lehto, K. (2018). Eukaryote specific folds : Part of the whole. *Proteins : Structure, Function, and Bioinformatics*, 86(8) :868–881.

Kawabata, T. (2003). MATRAS : A program for protein 3D structure comparison. *Nucleic Acids Research*, 31(13) :3367–3369.

Kawabata, T. and Nishikawa, K. (2000). Protein structure comparison using the markov transition model of evolution. *Proteins*, 41(1) :108–122.

Kedem, K., Chew, L. P., and Elber, R. (1999). Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, 37(4) :554–564.

Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, 79(3) :830–838.

Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006). Enzyme promiscuity : Evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology*, 10(5) :498–508.

Kim, K. M. and Caetano-Anollés, G. (2011). The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC evolutionary biology*, 11(1) :140.

Kim, K. M. and Caetano-Anollés, G. (2012). The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. *BMC evolutionary biology*, 12(1) :13.

Kinjo, A. R. and Nishikawa, K. (2004). Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in proteins. *Bioinformatics*, 20(16) :2504–2508.

Kolodny, R. (2021). Searching protein space for ancient sub-domain segments. *Current Opinion in Structural Biology*, 68 :105–112.

Kolodny, R., Nepomnyachiy, S., Tawfik, D. S., and Ben-Tal, N. (2021). Bridging Themes : Short Protein Segments Found in Different Architectures. *Molecular Biology and Evolution*, 38(6) :2191–2208.

Kolodny, R., Petrey, D., and Honig, B. (2006). Protein structure comparison : Implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, 16(3) :393–398.

Konagurthu, A., Whisstock, J., Stuckey, P., and Lesk, A. (2006). MUSTANG : A multiple structural alignment algorithm. *Proteins*, 64(3) :559–574.

Konagurthu, A. S. and Stuckey, P. J. (2006). Optimal sum-of-pairs multiple sequence alignment using incremental Carrillo and Lipman bounds. *J Comput Biol*, 13(3) :668–685.

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. (2002). Dobzhansky–Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences*, 99(23) :14878–14883.

Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, 420(6912) :218–223.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein : An information theoretic analysis. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 90(15) :7176–7180.

Koshi, J. M. and Goldstein, R. A. (1997). Mutation matrices and physical-chemical properties : Correlations and implications. *Proteins*, 27(3) :336–344.

Kosloff, M. and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, 71(2) :891–902.

Krissinel, E. (2012). Enhanced fold recognition using efficient short fragment clustering. *Journal of molecular biochemistry*, 1(2) :76–85.

Krissinel, E. and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1) :2256–2268.

Kuhlman, B. and Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11) :681–697.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree : A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7) :1812–1819.

Lamarine, M., Mornon, J. P., Berezovsky, I. N., and Chomilier, J. (2001). Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions : Towards an efficient punctuation of protein folding ? *Cellular and Molecular Life Sciences CMLS*, 58(3) :492–498.

Landenmark, H. K. E., Forgan, D. H., and Cockell, C. S. (2015). An Estimate of the Total DNA in the Biosphere. *PLOS Biology*, 13(6) :e1002168.

Lange, A., Patel, P. H., Heames, B., Damry, A. M., Saenger, T., Jackson, C. J., Findlay, G. D., and Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in Drosophila. *Nature Communications*, 12(1) :1667.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21) :2947–2948.

Larson, G., Thorne, J. L., and Schmidler, S. (2018). Modeling Dependence in Evolutionary Inference for Proteins. In Raphael, B. J., editor, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 122–137, Cham. Springer International Publishing.

Lassmann, T., Frings, O., and Sonnhammer, E. L. (2009). Kalign2 : High-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research*, 37(3) :858–865.

Lassmann, T. and Sonnhammer, E. L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, 6 :298–298.

Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., and Kortemme, T. (2010). RosettaBackrub–a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Research*, 38(Web Server) :W569–W575.

Lavery, R. and Sacquin-Mora, S. (2007). Protein mechanics : A route from structure to function. *Journal of Biosciences*, 32(5) :891–898.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals : A Gibbs sampling strategy for multiple alignment. *Science (New York, NY)*, 262(5131) :208–214.

Le, Q., Sievers, F., and Higgins, D. G. (2017). Protein multiple sequence alignment benchmarking through secondary structure prediction. *Bioinformatics*, 33(9) :1331–1337.

Le, S. Q. and Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Systematic Biology*, 59(3) :277–287.

Lecointre, G. and Le Guyader, H. (2017). *La classification phylogénétique du vivant. 4e ed.* Belin.

Lee, D., Grant, A., Marsden, R. L., and Orengo, C. (2005). Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins*, 59(3) :603–615.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8) :323–331.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. (2005a). An analysis of core deformations in protein superfamilies. *Biophysical Journal*, 88(2) :1291–1299.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. (2005b). Core deformations in protein families : A physical perspective. *Biophysical chemistry*, 115(2-3) :125–128.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1) :59–107.

Levitt, M. (1997). Competitive assessment of protein fold recognition and alignment accuracy. *Proteins : Structure, Function, and Bioinformatics*, 29(S1) :92–104.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. (1997). Protein folding : The endgame. *Annual review of biochemistry*, 66 :549–579.

Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253(5494) :694–698.

Li, S., Wu, S., Wang, L., Li, F., Jiang, H., and Bai, F. (2022). Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. *Current Opinion in Structural Biology*, 73 :102344.

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein,

R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., and Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein science : a publication of the Protein Society*, 21(6) :769–785.

Lin, J. and Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs : Implications for comparing genomes on different levels. *Genome Research*, 10(6) :808–818.

Lindberg, M. O., Haglund, E., Hubner, I. A., Shakhnovich, E. I., and Oliveberg, M. (2006). Identification of the minimal protein-folding nucleus through loop-entropy perturbations. *Proceedings of the National Academy of Sciences*, 103(11) :4083–4088.

Linsky, T. W., Noble, K., Tobin, A. R., Crow, R., Carter, L., Urbauer, J. L., Baker, D., and Strauch, E.-M. (2022). Sampling of structure and sequence space of small protein folds. *Nature Communications*, 13(1) :7151.

Liu, X., Zhao, Y.-P., and Zheng, W.-M. (2008). CLEMAPS : Multiple alignment of protein structures based on conformational letters. *Proteins*, 71(2) :728–736.

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2010). Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7) :2983–2988.

Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, N.Y.)*, 286(5438) :295–299.

Lonquety, M., Lacroix, Z., Papandreou, N., and Chomilier, J. (2009). SPROUTS : A database for the evaluation of protein stability upon point mutation. *Nucleic Acids Research*, 37(Database issue) :D374–9.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an Important Process of Protein Evolution. *Molecular Biology and Evolution*, 19(1) :1–7.

Löytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102(30) :10557–10562.

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics (Oxford, England)*, 21(15) :3255–3263.

Luzzati, V. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica*, 5(6) :802–810.

Madhusudhan, M. S., Webb, B. M., Marti-Renom, M. A., Eswar, N., and Sali, A. (2009). Alignment of multiple protein structures based on sequence and structure features. *Protein Engineering Design and Selection*, 22(9) :569–574.

Mahajan, S., de Brevern, A. G., Offmann, B., and Srinivasan, N. (2014). Correlation between local structural dynamics of proteins inferred from NMR ensembles and evolutionary dynamics of homologues of known structure. *Journal of Biomolecular Structure & Dynamics*, 32(5) :751–758.

Mannige, R. (2014). Origination of the Protein Fold Repertoire from Oily Pluripotent Peptides. *Proteomes*, 2(4) :154–168.

Marcos, M. L. and Echave, J. (2020). The variation among sites of protein structure divergence is shaped by mutation and scaled by selection. *Current Research in Structural Biology*, 2 :156–163.

Marsh, J. A. and Teichmann, S. A. (2014). Parallel dynamics and evolution : Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 36(2) :209–218.

Martin, A. C. R. (1996). ProFit.

Matthews, J. M. and Fersht, A. R. (1995). Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins : Observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry*, 34(20) :6805–6814.

McCandlish, D. M., Shah, P., and Plotkin, J. B. (2016). Epistasis and the Dynamics of Reversion in Molecular Evolution. *Genetics*, 203(3) :1335–1351.

McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature*, 491(7422) :138–142.

Medvedev, K. E., Kinch, L. N., Schaeffer, R. D., and Grishin, N. V. (2019). Functional analysis of Rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLoS computational biology*, 15(12) :e1007569.

Méheust, R., Bhattacharya, D., Pathmanathan, J. S., McInerney, J. O., Lopez, P., and Bapteste, E. (2018). Formation of chimeric genes with essential functions at the origin of eukaryotes. *BMC Biology*, 16 :30.

Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., and Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, 7(1) :339.

Menke, M., Berger, B., and Cowen, L. (2008). Matt : Local Flexibility Aids Protein Multiple Structure Alignment. *PLoS Computational Biology*, 4(1) :e10.

Micheletti, C. and Orland, H. (2009). MISTRAL : A tool for energy-based multiple structural alignment of proteins. *Bioinformatics (Oxford, England)*, 25(20) :2663–2669.

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38(Database issue) :D750–D753.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998a). JOY : Protein sequence-structure representation and analysis. *Bioinformatics (Oxford, England)*, 14(7) :617–623.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998b). HOMSTRAD : A database of protein structure alignments for homologous families. *Protein science : a publication of the Protein Society*, 7(11) :2469–2471.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, 108(49) :E1293–301.

Morgenstern, B. (1999). DIALIGN 2 : Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics (Oxford, England)*, 15(3) :211–218.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). DIALIGN : Finding local similarities by multiple sequence alignment. *Bioinformatics (Oxford, England)*, 14(3) :290–294.

Moult, J. and Melamud, E. (2000). From fold to function. *Current Opinion in Structural Biology*, 10(3) :384–389.

Moura de Sousa, J. A. and Rocha, E. P. C. (2022). To catch a hijacker : Abundance, evolution and genetic diversity of P4-like bacteriophage satellites. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 377(1842) :20200475.

Mura, C., Veretnik, S., and Bourne, P. E. (2019). The Urfold : Structural similarity just above the superfold level ? *Protein Science : A Publication of the Protein Society*, 28(12) :2119–2126.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP : A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4) :536–540.

Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions : The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology*, 321(5) :741–765.

Nasir, A., Kim, K. M., and Caetano-Anollés, G. (2014). Global Patterns of Protein Domain Gain and Loss in Superkingdoms. *PLoS Computational Biology*, 10(1) :e1003452–18.

Necci, M., Piovesan, D., and Tosatto, S. C. E. (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*, 18(5) :472–481.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences ? *Proceedings of the National Academy of Sciences*, 91(1) :98–102.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 114(44) :11703–11708.

Neuenfeldt, A., Just, A., Betat, H., and Mörl, M. (2008). Evolution of tRNA nucleotidyltransferases : A small deletion generated CC-adding enzymes. *Proceedings of the National Academy of Sciences*, 105(23) :7953–7958.

Newberry, R. W., Leong, J. T., Chow, E. D., Kampmann, M., and DeGrado, W. F. (2020). Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nature Chemical Biology*, 16(6) :653–659.

Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D., and Gromiha, M. M. (2021). ProThermDB : Thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research*, 49(D1) :D420–D424.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee : A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1) :205–217.

Novák, Á., Miklós, I., Lyngsø, R., and Hein, J. (2008). StatAlign : An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, 24(20) :2403–2404.

Ochagavía, M. E. and Wodak, S. (2004). Progressive combinatorial algorithm for multiple structural alignments : Application to distantly related proteins. *Proteins : Structure, Function, and Bioinformatics*, 55(2) :436–454.

Ochoa-Leyva, A., Soberón, X., Sánchez, F., Argüello, M., Montero-Morán, G., and Saab-Rincón, G. (2009). Protein design through systematic catalytic loop exchange in the (beta/alpha)8 fold. *Journal of Molecular Biology*, 387(4) :949–964.

O'Hearn, S. D., Kusalik, A. J., and Angel, J. F. (2003). MolCom : A method to compare protein molecules based on 3-D structural and chemical similarity. *Protein engineering*, 16(3) :169–178.

Oldfield, C. J. and Dunker, A. K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual Review of Biochemistry*, 83 :553–584.

Oldfield, T. J. (2007). CAALIGN : A program for pairwise and multiple protein-structure alignment. *Acta Crystallogr D Biol Crystallogr*, 63(4) :514–525.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, 372(6507) :631–634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure (London, England : 1993)*, 5(8) :1093–1108.

Orengo, C. A. and Taylor, W. R. (1996). SSAP : Sequential structure alignment program for protein structure comparison. *Methods Enzymol*, 266 :617–635.

Ortiz, A. R., Strauss, C. E., and Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory) : An automated method for model comparison. *Protein science : a publication of the Protein Society*, 11(11) :2606–2621.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee : Combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2) :385–395.

Overington, J., Johnson, M. S., Šali, A., and Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity : Templates, key residues and structure prediction. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 241(1301) :132–145.

Pandya, C., Brown, S., Pieper, U., Sali, A., Dunaway-Mariano, D., Babbitt, P. C., Xia, Y., and Allen, K. N. (2013). Consequences of domain insertion on sequence-structure divergence in a superfold. *Proceedings of the National Academy of Sciences*, 110(36) :E3381–E3387.

Parisi, G. and Echave, J. (2001). Structural constraints and emergence of sequence patterns in protein evolution. *Molecular Biology and Evolution*, 18(5) :750–756.

Pascual-García, A., Abia, D., Méndez, R., Nido, G. S., and Bastolla, U. (2010). Quantifying the evolutionary divergence of protein structures : The role of function change and function conservation. *Proteins*, 78(1) :181–196.

Pascual-García, A., Abia, D., Ortiz, A. R., and Bastolla, U. (2009). Cross-Over between Discrete and Continuous Protein Structure Space : Insights into Automatic Classification and Networks of Protein Structures. *PLoS Computational Biology*, 5(3) :e1000331–20.

Pauwels, K., Van Molle, I., Tommassen, J., and Van Gelder, P. (2007). Chaperoning Anfinsen : The steric foldases. *Molecular Microbiology*, 64(4) :917–922.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(4) :511–523.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering, Design and Selection*, 14(9) :609–614.

Pei, J. and Grishin, N. V. (2007). PROMALS : Towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics (Oxford, England)*, 23(7) :802–808.

Pei, J., Kim, B. H., and Grishin, N. V. (2008). PROMALS3D : A tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, 36(7) :2295–2300.

Peng, Z., Yan, J., Fan, X., Mizianty, M. J., Xue, B., Wang, K., Hu, G., Uversky, V. N., and Kurgan, L. (2015). Exceptionally abundant exceptions : Comprehensive characterization of intrinsic disorder in all domains of life. *Cellular and Molecular Life Sciences*, 72(1) :137–151.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., Schafferhans, A., and O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52) :15898–15903.

Perron, U., Kozlov, A. M., Stamatakis, A., Goldman, N., and Moal, I. H. (2019a). Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States. *Molecular Biology and Evolution*, 36(9) :2086–2103.

Perron, U., Moal, I. H., Thorne, J. L., and Goldman, N. (2019b). Probabilistic Models for the Study of Protein Evolution. In Balding, D., Moltke, I., and Marioni, J., editors, *Handbook of Statistical Genomics*, pages 347–30. Wiley, first edition.

Piatigorsky, J., Kantorow, M., Gopal-Srivastava, R., and Tomarev, S. I. (1994). Recruitment of enzymes and stress proteins as lens crystallins. *EXS*, 71 :241–250.

Pilipczuk, J., Zalewska-Piątek, B., Bruździak, P., Czub, J., Wieczór, M., Olszewski, M., Wanarska, M., Nowicki, B., Augustin-Nowacka, D., and Piątek, R. (2017). Role of the disulfide bond in stabilizing and folding of the fimbrial protein DraE from uropathogenic Escherichia coli. *The Journal of Biological Chemistry*, 292(39) :16136–16149.

Pisani, D., Cotton, J. A., and McInerney, J. O. (2007). Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution*, 24(8) :1752–1760.

Pisanti, N., Soldano, H., and Carpentier, M. (2005). Incremental inference of relational motifs with a degenerate alphabet. *Combinatorial Pattern Matching, Proceedings*, 3537 :229–240.

Pisanti, N., Soldano, H., Carpentier, M., and Pothier, J. (2006). Implicit and Explicit Representation of Approximated Motifs. In *Algorithms for Bioinformatics*, volume 6 of *Algorithms for Bioinformatics*. King's College London Press, c. iliopoulos, k. park et k. steinhofel edition.

Podlaha, O., Webb, D. M., Tucker, P. K., and Zhang, J. (2005). Positive Selection for Indel Substitutions in the Rodent Sperm Protein Catsper1. *Molecular Biology and Evolution*, 22(9) :1845–1852.

Podlaha, O. and Zhang, J. (2003). Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proceedings of the National Academy of Sciences*, 100(21) :12241–12246.

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues : Maximum likelihood identification and relationship to structure11Edited by G. Von Heijne. *Journal of Molecular Biology*, 287(1) :187–198.

Pritchard, L., Bladon, P., M O Mitchell, J., and J Dufton, M. (2001). Evaluation of a novel method for the identification of coevolving protein residues. *Protein Engineering*, 14(8) :549–555.

Privalov, P. (1979). Stability of Proteins Small Globular Proteins. In *Advances in Protein Chemistry*, volume 33, pages 167–241. Elsevier.

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., Pajkos, M., Lazar, T., Peña-Díaz, S., Santos, J., Ács, V., Farahi, N., Fichó, E., Aspromonte, M. C., Bassot, C., Chasapi, A., Davey, N. E., Davidović, R., Dobson, L., Elofsson, A., Erdős, G., Gaudet, P., Giglio, M., Glavina, J., Iserte, J., Iglesias, V., Kálmán, Z., Lambrughi, M., Leonardi, E., Longhi, S., Macedo-Ribeiro, S., Maiani, E., Marchetti, J., Marino-Buslje, C., Mészáros, A., Monzon, A. M., Minervini, G., Nadendla, S., Nilsson, J. F., Novotný, M., Ouzounis, C. A., Palopoli, N., Papaleo, E., Pereira, P. J. B., Pozzati, G., Promponas, V. J., Pujols, J., Rocha, A. C. S., Salas, M., Sawicki, L. R., Schad, E., Shenoy, A., Szaniszló, T., Tsirigos, K. D., Veljkovic, N., Parisi, G., Ventura, S., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2022). DisProt in 2022 : Improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research*, 50(D1) :D480–D487.

Raftery, AE. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society*, B, 47 (3). :528–539.

Raghava, G. P. S., Searle, S. M. J., Audley, P. C., Barber, J. D., and Barton, G. J. (2003). OXBench : A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC bioinformatics*, 4 :47.

Ray, D., Nicolas Quijano, R., and Andricioaei, I. (2022). Point mutations in SARS-CoV-2 variants induce long-range dynamical perturbations in neutralizing antibodies. *Chemical Science*, 13(24) :7224–7239.

Raymann, K., Brochier-Armanet, C., and Gribaldo, S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences of the United States of America*, 112(21) :6670–6675.

Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. V., and Fersht, A. R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature*, 437(7061) :1053–1056.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits : Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2) :173–175.

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., and Gordon, J. I. (2012). Going viral : Next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology*, 10(9) :607–617.

Ritchie, D. W., Ghoorah, A. W., Mavridis, L., and Venkatraman, V. (2012). Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics (Oxford, England)*, 28(24) :3274–3281.

Roach, J., Sharma, S., Kapustina, M., and Carter, C. W. (2005). Structure alignment via Delaunay tetrahedralization. *Proteins*, 60(1) :66–81.

Rochat, T., Pérez-Pascual, D., Nilsen, H., Carpentier, M., Bridel, S., Bernardet, J.-F., and Duchaud, E. (2019). Identification of a Novel Elastin-Degrading Enzyme from the Fish Pathogen Flavobacterium psychrophilum. *Appl. Environ. Microbiol.*, 85(6) :e02535–18.

Rockah-Shmuel, L., Tóth-Petróczy, Á., and Tawfik, D. S. (2015). Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLOS Computational Biology*, 11(8) :e1004421.

Romei, M., Sapriel, G., Imbert, P., Jamay, T., Chomilier, J., Lecointre, G., and Carpentier, M. (2022). Protein folds as synapomorphies of the tree of life. *Evolution*, 76(8) :1706–1719.

Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, 2 :S19–S24.

Rost, B. (2002). Did evolution leap to create the protein universe ? *Current Opinion in Structural Biology*, 12(3) :409–416.

Ruff, K. M. and Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. *Journal of Molecular Biology*, 433(20) :167208.

Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison : Assignment of global and residue confidence levels. *Proteins*, 14(2) :309–323.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., and Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds : Analysis of sequence and structure conservation. *Journal of Molecular Biology*, 269(3) :423–439.

Sacquin-Mora, S. and Lavery, R. (2006). Investigating the local flexibility of functional residues in hemoproteins. *Biophysical Journal*, 90(8) :2706–2717.

Sadowski, M. I. and Taylor, W. R. (2009). Protein structures, folds and fold spaces. *Journal of Physics : Condensed Matter*, 22(3) :033103.

Sadreyev, R. I., Kim, B.-H., and Grishin, N. V. (2009). Discrete–continuous duality of protein structure space. *Current Opinion in Structural Biology*, 19(3) :321–328.

Saibil, H. (2013). Chaperone machines for protein folding, unfolding and disaggregation. *Nature Reviews. Molecular Cell Biology*, 14(10) :630–642.

Sali, A. and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of Molecular Biology*, 212(2) :403–428.

Sandelin, E. (2005). Extracting multiple structural alignments from pairwise alignments : A comparison of a rigorous and a heuristic approach. *Bioinformatics (Oxford, England)*, 21(7) :1002–1009.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1) :56–68.

Santini, G., Soldano, H., and Pothier, J. (2012). Automatic classification of protein structures relying on similarities between alignments. *BMC Bioinformatics*, 13(1) :233.

Schaefer, C. and Rost, B. (2012). Predict impact of single amino acid change upon protein structure. *BMC genomics*, 13(Suppl 4) :S4.

Scherrer, M. P., Meyer, A. G., and Wilke, C. O. (2012). Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*, 12(1) :179.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991). A workbench for multiple alignment construction and analysis. *Proteins*, 9(3) :180–190.

Shakhnovich, E. I. and Gutin, A. M. (1991). Influence of point mutations on protein structure : Probability of a neutral mutation. *Journal of theoretical biology*, 149(4) :537–546.

Shanthirabalan, S., Chomilier, J., and Carpentier, M. (2018). Structural effects of point mutations in proteins. *Proteins : Structure, Function, and Bioinformatics*, 86(8) :853–867.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1) :143–156.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2005). Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, 62(1) :209–217.

Shealy, P. and Valafar, H. (2012). Multiple structure alignment with msTALI. *BMC bioinformatics*, 13(1) :1–1.

Sheng, W. and Wei-Mou, Z. (2009). Fast Multiple Alignment of Protein Structures Using Conformational Letter Blocks. *The Open Bioinformatics Journal*, 3(1).

Shindyalov, I. N. and Bourne, P. E. (2000). An alternative view of protein fold space. *Proteins : Structure, Function, and Genetics*, 38(3) :247–260.

Shindyalov, I. N., Kolchanov, N. A., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations ? *Protein engineering*, 7(3) :349–358.

Shortle, D. (2009). One sequence plus one mutation equals two folds. *Proceedings of the National Academy of Sciences*, 106(50) :21011–21012.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7 :539.

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., Ceballos Rodriguez-Conde, F., Dowling, B., Thornton, J., and Orengo, C. A. (2019). CATH : Expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, 47(D1) :D280–D284.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, Suppl 3 :171–176.

Sippl, M. J. (2009). Fold space unlimited. *Current Opinion in Structural Biology*, 19(3) :312–320.

Skolnick, J., Arakaki, A. K., Lee, S. Y., and Brylinski, M. (2009). The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci U S A*, 106(37) :15690–15695.

Smith, B. D. and Raines, R. T. (2006). Genetic Selection for Critical Residues in Ribonucleases. *Journal of Molecular Biology*, 362(3) :459–478.

Smith, C. A. and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4) :742–756.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1) :195–197.

Soldano, H., Viari, A., and Champesme, M. (1995). Searching for flexible repeated patterns using a non-transitive similarity relation. *Pattern Recognition Letters*, 16 :243–246.

Sorokina, I. and Mushegian, A. (2018). Modeling protein folding in vivo. *Biology Direct*, 13(1) :13.

Sorokina, I., Mushegian, A. R., and Koonin, E. V. (2022). Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process ? *International Journal of Molecular Sciences*, 23(1) :521.

Stein, R. R., Marks, D. S., and Sander, C. (30 juil. 2015). Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLOS Computational Biology*, 11(7) :e1004182.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20(1) :473.

Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1) :59–69.

Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M. W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N., Torruella, G., Derelle, R., Manning, G., Lang, B. F., Russ, C., Haas, B. J., Roger, A. J., Nusbaum, C., and Ruiz-Trillo, I. (2013). The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nature Communications*, 4(1) :2325.

Sun, H., Sacan, A., Ferhatosmanoglu, H., and Wang, Y. (2011). Smolign : A Spatial Motifs Based Protein Multiple Structural Alignment Method. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 9(1) :249–261.

Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, Part I : Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein engineering*, 1(5) :377–384.

Takano, K., Funahashi, J., Yamagata, Y., Fujii, S., and Yutani, K. (1997). Contribution of water molecules in the interior of a protein to the conformational stability11Edited by A. R. Fersht. *Journal of Molecular Biology*, 274(1) :132–142.

Takano, K., Ogasahara, K., Kaneda, H., Yamagata, Y., Fujii, S., Kanaya, E., Kikuchi, M., Oobatake, M., and Yutani, K. (1995). Contribution of Hydrophobic Residues to the Stability of Human Lysozyme : Calorimetric Studies and X-ray Structural Analysis of the Five Isoleucine to Valine Mutants. *Journal of Molecular Biology*, 254(1) :62–76.

Takano, K., Yamagata, Y., and Yutani, K. (2000). Role of Amino Acid Residues at Turns in the Conformational Stability and Folding of Human Lysozyme,. *Biochemistry*, 39(29) :8655–8665.

Takano, K., Yamagata, Y., and Yutani, K. (2001a). Contribution of Polar Groups in the Interior of a Protein to the Conformational Stability,. *Biochemistry*, 40(15) :4853–4858.

Takano, K., Yamagata, Y., and Yutani, K. (2001b). Role of amino acid residues in left-handed helical conformation for the conformational stability of a protein. *Proteins : Structure, Function, and Bioinformatics*, 45(3) :274–280.

Takano, K., Yamagata, Y., and Yutani, K. (2001c). Role of non-glycine residues in left-handed helical conformation for the conformational stability of human lysozyme. *Proteins : Structure, Function, and Genetics*, 44(3) :233–243.

Talibart, H. and Coste, F. (2021). PPalign : Optimal alignment of Potts models representing proteins with direct coupling information. *BMC bioinformatics*, 22(1) :317.

Taverna, D. M. and Goldstein, R. A. (2002a). Why are proteins marginally stable ? *Proteins*, 46(1) :105–109.

Taverna, D. M. and Goldstein, R. A. (2002b). Why are proteins so robust to site mutations ? *Journal of Molecular Biology*, 315(3) :479–484.

Tawfik, D. S. (2006). Loop Grafting and the Origins of Enzyme Species. *Science*, 311(5760) :475–476.

Tawfik, O. K. a. D. S. (2010). Enzyme Promiscuity : A Mechanistic and Evolutionary Perspective. *Annual Review of Biochemistry*, 79(1) :471–505.

Taylor, R. W., M Thornton, J., and G Turnell, W. (1983). An ellipsoidal approximation of protein shape. *Journal of Molecular Graphics*, 1(2) :30–38.

Taylor, W. R. (2002). A 'periodic table' for protein structures. *Nature*, 416(6881) :657–660.

Taylor, W. R., Flores, T. P., and Orengo, C. A. (1994). Multiple protein structure alignment. *Protein science : a publication of the Protein Society*, 3(10) :1858–1870.

Taylor, W. R., Hamilton, R. S., and Sadowski, M. I. (2013). Prediction of contacts from correlated sequence substitutions. *Current Opinion in Structural Biology*, 23(3) :473–479.

Taylor, W. R. and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering, Design and Selection*, 7(3) :341–348.

The Gene Ontology Consortium (2021). The Gene Ontology resource : Enriching a GOld mine. *Nucleic Acids Research*, 49(D1) :D325–D334.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W : Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22) :4673–4680.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0 : Latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1) :127–136.

Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods : Current Challenges and Future Perspectives. *PloS one*, 6(3) :e18093.

Thompson, J. D., Plewniak, F., and Poch, O. (1999a). BAliBASE : A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics (Oxford, England)*, 15(1) :87–88.

Thompson, J. D., Plewniak, F., and Poch, O. (1999b). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27(13) :2682–2690.

Tiessen, A., Pérez-Rodríguez, P., and Delaye-Arredondo, L. J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Research Notes*, 5(1) :85.

To, P., Whitehead, B., Tarbox, H. E., and Fried, S. D. (2021). Nonrefoldability is Pervasive Across the E. coli Proteome. *Journal of the American Chemical Society*, 143(30) :11435–11448.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307(4) :1113–1143.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. *Structure (London, England : 1993)*, 10(10) :1435–1451.

Tokuriki, N. and Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, 19(5) :596–604.

Toledo-Patiño, S., Chaubey, M., Coles, M., and Höcker, B. (2019). Reconstructing the Remote Origins of a Fold Singleton from a Flavodoxin-Like Ancestor. *Biochemistry*, 58(48) :4790–4793.

Tóth-Petróczy, Á. and Tawfik, D. S. (2013). Protein Insertions and Deletions Enabled by Neutral Roaming in Sequence Space. *Molecular Biology and Evolution*, 30(4) :761–771.

Trudeau, D. L., Smith, M. A., and Arnold, F. H. (2013). Innovation by homologous recombination. *Current Opinion in Chemical Biology*, 17(6) :902–909.

Tuff, P. and Darlu, P. (2000). Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Molecular Biology and Evolution*, 17(11) :1753–1759.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873) :590–596.

van Aalten, D. M., Conn, D. A., de Groot, B. L., Berendsen, H. J., Findlay, J. B., and Amadei, A. (1997). Protein dynamics derived from clusters of crystal structures. *Biophysical Journal*, 73(6) :2891–2896.

Velázquez-Muriel, J. A., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., and Carazo, J.-M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Structural Biology*, 9(1) :6.

Vetrivel, I., de Brevern, A. G., Cadet, F., Srinivasan, N., and Offmann, B. (2019). Structural variations within proteins can be as large as variations observed across their homologues. *Biochimie*, 167 :162–170.

Viguera, A. R., Blanco, F. J., and Serrano, L. (1995). The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *Journal of Molecular Biology*, 247(4) :670–681.

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., Bittner, L., Blanke, B., Brum, J. R., Brunet, C., Casotti, R., Chase, A., Dolan, J. R., d'Ortenzio, F., Gattuso, J.-P., Grima, N., Guidi, L., Hill, C. N., Jahn, O., Jamet, J.-L., Le Goff, H., Lepoivre, C., Malviya, S., Pelletier, E., Romagnan, J.-B., Roux, S., Santini, S., Scalco, E., Schwenck, S. M., Tanaka, A., Testor, P., Vannier, T., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Acinas, S. G., Bork, P., Boss, E., de Vargas, C., Gorsky, G., Ogata, H., Pesant, S., Sullivan, M. B., Sunagawa, S., Wincker, P., Karsenti, E., Bowler, C., Not, F., Hingamp, P., and Iudicone, D. (2015). Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science (New York, NY)*, 348(6237) :1261447.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, 14(2) :208–216.

Wako, H. and Blundell, T. L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *Journal of Molecular Biology*, 238(5) :682–692.

Wako, H. and Blundell, T. L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *Journal of Molecular Biology*, 238(5) :693–708.

Wang, M. and Caetano-Anollés, G. (2006). Global Phylogeny Determined by the Combination of Protein Domains in Proteomes. *Molecular Biology and Evolution*, 23(12) :2444–2454.

Wang, M. and Caetano-Anollés, G. (2009). The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World. *Structure*, 17(1) :66–78.

Wang, M., Jiang, Y.-Y., Kim, K. M., Qu, G., Ji, H.-F., Mittenthal, J. E., Zhang, H.-Y., and Caetano-Anollés, G. (2011a). A Universal Molecular Clock of Protein Folds and Its Power in Tracing the Early History of Aerobic Metabolism and Planet Oxygenation. *Molecular Biology and Evolution*, 28(1) :567–582.

Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E., and Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome research*, 17(11) :1572–1585.

Wang, S., Ma, J., Peng, J., and Xu, J. (2013). Protein structure alignment beyond spatial proximity. *Scientific Reports*, 3.

Wang, S., Peng, J., and Xu, J. (2011b). Alignment of distantly related protein structures : Algorithm, bound and implications to homology modeling. *Bioinformatics*, 27(18) :2537.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337(3) :635–645.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1) :67–72.

Weinreich, D. M. (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science (New York, NY)*, 312(5770) :111–114.

Weinreich, D. M., Watson, R. A., and Chao, L. (2005). Perspective : Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution ; International Journal of Organic Evolution*, 59(6) :1165–1174.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1(9) :16116.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics*, 36(3) :307–340.

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., and Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, 4(1) :138–147.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2007). The SUPERFAMILY database in 2007 : Families and functions. *Nucleic Acids Research*, 35(suppl_1) :D308–D313.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, 37(Database issue) :D380–D386.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S.,

Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica. Section D, Biological Crystallography*, 67(Pt 4) :235–242.

Winstanley, H. F., Abeln, S., and Deane, C. M. (2005). How old is your fold ? *Bioinformatics (Oxford, England)*, 21(Suppl 1) :i449–i458.

Wojcik, J., Mornon, J. P., and Chomilier, J. (1999). New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *Journal of Molecular Biology*, 289(5) :1469–1490.

Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. (1999). Distribution of protein folds in the three superkingdoms of life. *Genome research*, 9(1) :17–26.

Wolf, Y. I., Grishin, N. V., and Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, 299(4) :897–905.

Wood, T. C. and Pearson, W. R. (1999). Evolution of protein sequences and structures. *Journal of Molecular Biology*, 291(4) :977–995.

Woodcock, S., Mornon, J. P., and Henrissat, B. (1992). Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Engineering*, 5(7) :629–635.

Wylie, C. S. and Shakhnovich, E. I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24) :9916–9921.

Xu, D. and Nussinov, R. (1998). Favorable domain size in proteins. *Folding and Design*, 3(1) :11–17.

Xu, J. and Zhang, J. (2016). Impact of structure space continuity on protein fold classification. *Scientific Reports*, 6(1) :23263.

Xue, B., Dunker, A. K., and Uversky, V. N. (2012). Orderly order in protein intrinsic disorder distribution : Disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure & Dynamics*, 30(2) :137–149.

Yang, A. S. and Honig, B. (2000a). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology*, 301(3) :665–678.

Yang, A. S. and Honig, B. (2000b). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *Journal of Molecular Biology*, 301(3) :679–689.

Yang, A. S. and Honig, B. (2000c). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *Journal of Molecular Biology*, 301(3) :691–711.

Yang, J.-R., Zhuang, S.-M., and Zhang, J. (2010). Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Molecular Systems Biology*, 6(1) :421.

Yang, S., Doolittle, R. F., and Bourne, P. E. (2005). Phylogeny determined by protein domain content. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2) :373–378.

Ye, Y. and Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome research*, 14(3) :343–353.

Ye, Y. and Godzik, A. (2005). Multiple flexible structure alignment using partial order graphs. *Bioinformatics (Oxford, England)*, 21(10) :2362–2369.

Young, D. L. and Fields, S. (2015). The role of functional data in interpreting the effects of genetic variation. *Molecular Biology of the Cell*, 26(22) :3904–3908.

Yu, L., Tanwar, D. K., Penha, E. D. S., Wolf, Y. I., Koonin, E. V., and Basu, M. K. (2019). Grammar of protein domain architectures. *Proceedings of the National Academy of Sciences of the United States of America*, 116(9) :3636–3645.

Zeldovich, K. B., Chen, P., and Shakhnovich, E. I. (2007). Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41) :16152–16157.

Zhang, C. T. (1997). Relations of the numbers of protein sequences, families and folds. *Protein Engineering, Design and Selection*, 10(7) :757–761.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7) :409–420.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4) :702–710.

Zhang, Y. and Skolnick, J. (2005). TM-align : A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7) :2302–2309.

Zhi, D., Krishna, S. S., Cao, H., Pevzner, P., and Godzik, A. (2006). Representing and comparing protein structures as paths in three-dimensional space. *BMC bioinformatics*, 7 :460.

Zhou, R., Eleftheriou, M., Royyuru, A. K., and Berne, B. J. (2007). Destruction of long-range interactions by a single mutation in lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14) :5824–5829.

Annexes

TABLE 7.1 – Liste de tous les programmes cherchés dans l'étude comparative (Carpentier and Chomilier, 2019).

Type	Name	Error/Issue	Ref.	Year
SEQ	CLUSTALO	Ok	(Sievers et al., 2011 ; Goujon et al., 2010)	2010
SEQ	CLUSTALW	Ok	(Larkin et al., 2007 ; Thompson et al., 1994)	1994
SEQ	DIALIGN	Ok	(Al Ait et al., 2013 ; Morgenstern et al., 1998 ; Morgenstern, 1999)	1998
SEQ	KALIGN2	Ok	(Lassmann and Sonnhammer, 2005 ; Lassmann et al., 2009)	2005
SEQ	MAFFT_linsi	Ok	(Katoh and Standley, 2013 ; Katoh et al., 2002)	2002
SEQ	MUSCLE	Ok	(Edgar, 2004b,a)	2004
SEQ	PRANK	Ok	(Löytynoja and Goldman, 2005)	2005
SEQ	PROBCONS	Ok	(Do et al., 2005)	2005
SEQ	TCOFFEE_SEQ	Ok	(Notredame et al., 2000)	2000
SEQ/ STRUCT	PROMALS3D	Ok	(Pei and Grishin, 2007 ; Pei et al., 2008)	2008
SEQ/ STRUCT	TCOFFEE_SAP	Ok	(O'Sullivan et al., 2004 ; Orengo and Taylor, 1996)	2004
SEQ/ STRUCT	TCOFFEE_TM	Ok	(O'Sullivan et al., 2004 ; Zhang and Skolnick, 2005)	2004
SEQ/ STRUCT	SALIGN	Ok	(Madhusudhan et al., 2009)	2007
SEQ/ STRUCT	FORMATT	Ok	(Daniels et al., 2012)	2005
STRUCT	3DCOMB	Ok	(Wang et al., 2011b)	2011
STRUCT	GESAMT	Ok	(Krissinel, 2012 ; Winn et al., 2011)	2012
STRUCT	KPAX	Ok	(Ritchie et al., 2012)	2012
STRUCT	MAMMOTH	Ok	(Lupyan et al., 2005)	2005
STRUCT	MATRAS	Ok	(Kawabata, 2003 ; Kawabata and Nishikawa, 2000)	2000
STRUCT	MATT	Ok	(Menke et al., 2008)	2008
STRUCT	MISTRAL	Ok	(Micheletti and Orland, 2009)	2009
STRUCT	MTMALIGN	Ok	(Dong et al., 2018)	2018
STRUCT	MULTIPROT	Ok	(Shatsky et al., 2004)	2004
STRUCT	MUSTANG	Ok	(Konagurthu et al., 2006)	2005
STRUCT	STAMP	Ok	(Russell and Barton, 1992)	1992
STRUCT	C-BOP	Precalculated alignemments	(Sandelin, 2005)	2005
STRUCT	CAALIGN	Not publicly available	(Oldfield, 2007)	2007
STRUCT	CBA	obsolet python depencies	(Ebert and Brutlag, 2006)	2006
STRUCT	CE/CE-MC	Doesn't work anymore	(Guda et al., 2004)	2004
STRUCT	CLEMAPS	Sever unavailable	(Liu et al., 2008)	2007
STRUCT	CURVE	Not found	(Zhi et al., 2006)	2006
STRUCT	DEEPALIGN	Doesn't compile (missing files)	(Wang et al., 2013)	2013
STRUCT	EPITOPEMATCH	Not for whole structures comparison	(Jakuschev and Hoffmann, 2009)	2009
STRUCT	FIT3D	Not for whole structures comparison	(Kaiser et al., 2016)	2016
STRUCT	MALECON	Not found	S. (Ochagavía and Wodak, 2004)	2004
STRUCT	MASS	Too slow	(Dror et al., 2003)	2003
STRUCT	BLOMAPS	Not found	(Sheng and Wei-Mou, 2009)	2008
STRUCT	MOLCOM	Not found	(O'Hearn et al., 2003)	2003
STRUCT	MSTALI	Too many uncalculated alignments	(Shealy and Valafar, 2012)	2012
STRUCT	MULPBA	Doesn't work (hard coded pathes in the code	(Joseph et al., 2012)	2012
STRUCT	MULTAL	Not found	(Taylor et al., 1994)	1994
STRUCT	POSA	Non sequential alignments, No standard output file	(Ye and Godzik, 2005)	2005
STRUCT	PRISM	Doesn't compile (obsolet dependencies)	(Yang and Honig, 2000b,a,c)	2000
STRUCT	PROFIT	Least square fitting only	(Martin, 1996)	1996
STRUCT	SMOLIGN	Not enough computed alignments	(Sun et al., 2011)	2010
STRUCT	SSAP	Not found	(Orengo and Taylor, 1996)	1989
STRUCT	TETRADA	Not found	(Roach et al., 2005)	2005
STRUCT	UCSF CHIMERA	Not scriptable	(Meng et al., 2006)	2006
STRUCT	VOROLIGN	Not found	(Birzele et al., 2007)	2007

Informations complémentaires chapitre 3

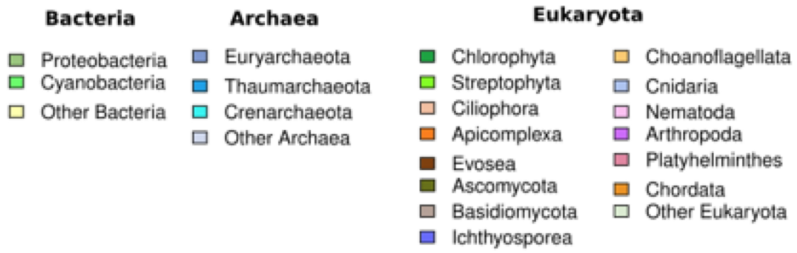


FIGURE 7.1 – Légende supplémentaire détaillée pour la figure 4.2, où chaque couleur représente un groupe taxonomique selon la nomenclature NCBI. Figures provenant de (Romei et al., 2022)

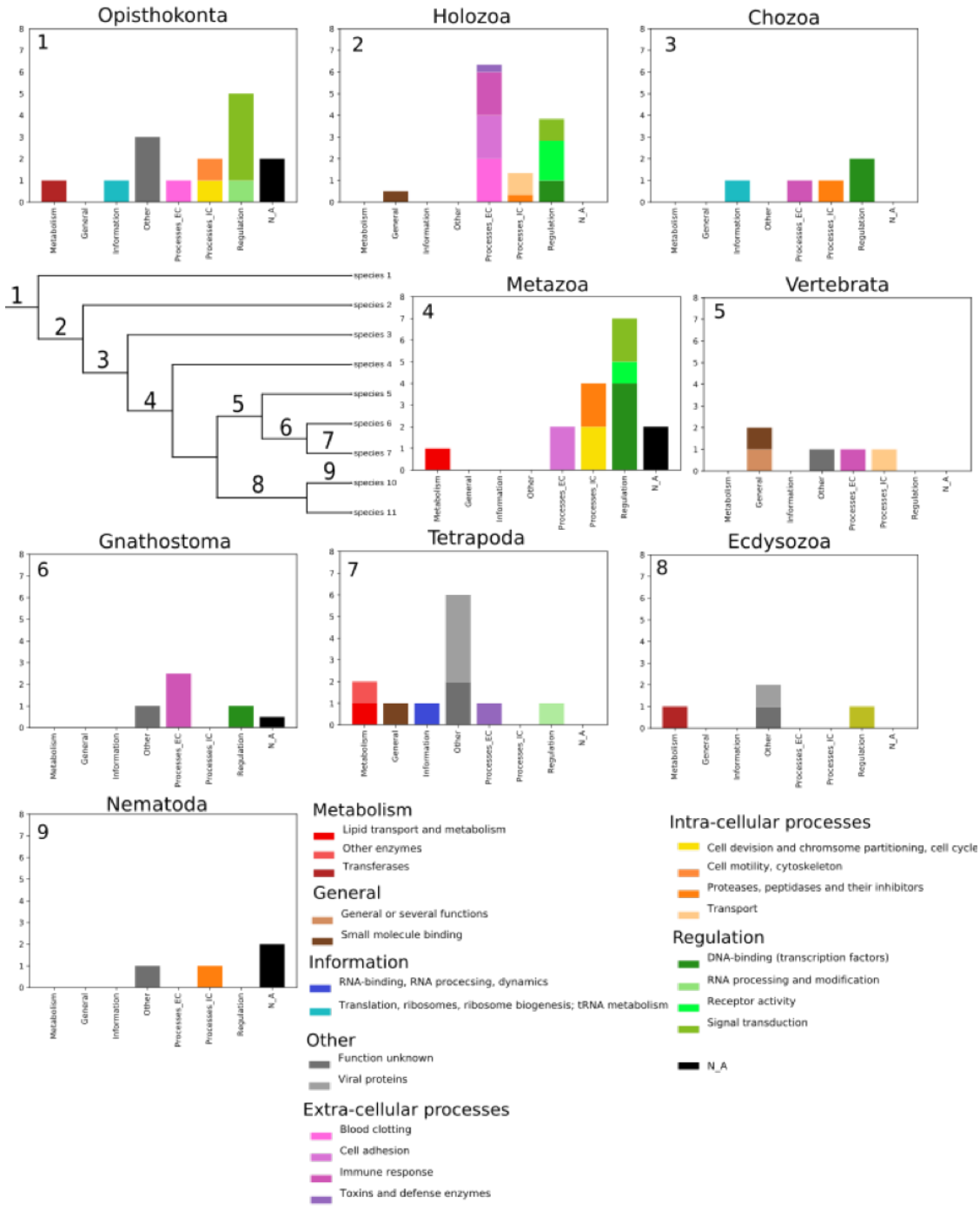


FIGURE 7.2 – Simplification de notre arbre de référence Eucaryote avec seulement des clades avec des repliements spécifiques et histogramme de la répartition des fonctions pour chacun de ces clades. Dans tous les diagrammes, les fonctions principales sont dans le même ordre que dans la légende : Métabolisme, général, information, autre, processus extra-cellulaires, processus intra-cellulaires, régulation et non-attribué.