Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL

Olivier Dennler

Université de Rennes 1

Master 2 Bioinformatique

Parcours Informatique et Biologie Intégrative

Année universitaire 2018 / 2019

Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL

Olivier Dennler

Maître de stage : Nathalie Théret

Co-encadrants : François Coste, Samuel Blanquart, Catherine Belleannée

IRISA

Institut de Recherche en Informatique et Systèmes Aléatoires

Equipe Dyliss

DYnamics, Logics and Inference for biological Systems and Sequences

263 Avenue Général Leclerc, Rennes

Master bioinformatique

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) .....Olivier.....Dennler.....
étudiant(e) en....Master.....2.....Bioinformatique.....
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

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

Date : 14/06/19

Signature :

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Remerciments

Je tiens à remercier Nathalie Théret pour m'avoir donné la possibilité de réaliser ce stage, pour son encadrement, le partage de ses connaissances en biologie et ses relectures de ce mémoire.

Je tiens également à remercier François Coste pour ses divers conseils tout au long de mon stage, pour le partage de ses connaissances sur l'analyse de séquences ainsi que pour ses relectures de ce mémoire.

Je tiens à remercier Samuel Blanquart pour le partage de son expertise et ses précieux conseils en matière de phylogénie, ainsi que pour ses relectures de ce mémoire.

Je tiens également à remercier Catherine Belleannée pour ses divers conseils et ses relectures de ce mémoire.

De plus, je tiens à remercier l'ensemble de l'équipe Dyliss et de Symbiose pour leur accueil et les conditions optimales de recherches mises en places.

Je tiens à remercier le personnel de la cafétéria de l'IRISA pour leurs cafés salvateurs.

Je tiens également à remercier spécifiquement les membres de la DSI (Direction des Systèmes informatiques) pour leur aide et assistance à mes problèmes triviaux.

Dans un dernier temps, je tiens à remercier tout particulièrement les autres stagiaires de Symbiose pour leur présence, sans laquelle ce stage n'aurait pas été le même.

Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL

Les protéines ADAMTS et ADAMTSL sont impliquées dans le remodelage du microenvironnement matriciel et constituent aujourd'hui de nouvelles cibles thérapeutiques dans les pathologies cancéreuses. Les nombreux gènes de la famille et la nature multidomaine des ADAMTS et ADAMTSL ne permettent pas d'utiliser les approches classiques de caractérisation fonctionnelles des protéines. Nous proposons ici une nouvelle méthode de caractérisation en modules fonctionnels adaptée aux protéines multidomaine, se basant uniquement sur les séquences protéiques. Cette méthode allie 2 approches de prédictions fonctionnelles à savoir la conservation des résidus et la phylogénie moléculaire. Après avoir sélectionné 341 séquences d'ADAMTS et ADAMTSL, réparties dans 19 espèces, les modules conservés sont recherchés par une approche *sans a priori* grâce à l'outil `paloma`. L'histoire phylogénétique de ces modules est ensuite élaborée avec l'outil `SEADOG-DM` de réconciliation phylogénétique DGS (Domain-Gene-Species). Enfin les histoires phylogénétiques sont complétées en intégrant des données biologiques issues de bases de données comme les interactions protéiques (PSICQUIC), les motifs (Prosite) et les polymorphismes nucléotidiques (dbSNP). Cette nouvelle stratégie permet de caractériser les différentes protéines et modules et d'étudier des évènements de co-occurence de modules conservés et de fonctions au cours de l'évolution. Cette étude constitue un travail préliminaire et permet d'apporter une preuve de concept vis à vis de la stratégie de caractérisation en modules fonctionnels des ADAMTS et ADAMTSL.

Mots clés : Protéines multidomaines, réconciliation phylogénétique, conservation de séquences, évolution

Functional module characterization of the ADAMTS / ADAMTSL protein family

ADAMTS and ADAMTSL proteins are involved in the remodeling of the matrix microenvironment and are now considered as new therapeutic targets in cancer diseases. The many genes in the family and the multi-domain nature of ADAMTS and ADAMTSL do not allow the use of conventional approaches to functional protein characterization. We propose here a new method of characterization in functional modules adapted to multidomain proteins, based only on protein sequences. This method combines 2 functional prediction approaches, namely residue conservation and molecular phylogeny. After selecting 341 ADAMTS and ADAMTSL sequences, distributed in 19 species, the preserved modules are searched for by an approach without *a priori* thanks to the `paloma` tool. Next, the phylogenetic history of these modules is developed with the `SEADOG-DM` phylogenetic reconciliation tool DGS (Domain-Gene-Species). Finally, phylogenetic histories are supplemented by integrating biological data from public databases such as protein interactions (PSICQUIC), motifs (Prosite) and nucleotide polymorphisms (dbSNP). This new strategy allows to characterize the different proteins and modules and to study co-occurrence events of conserved modules and functions during evolution. This study is a preliminary work and provides proof of concept regarding the functional module characterization strategy of ADAMTS and ADAMTSL.

Keywords : Multidomain proteins, phylogenetic reconciliation, sequence conservation, evolution

Table des matières

1 Introduction

Ce stage porte sur la caractérisation en modules fonctionnels des protéines ADAMTS et ADAMTSL, dans un premier temps je vais introduire ces protéines, leur intérêt, puis différentes approches de caractérisation fonctionnelle, avant de présenter la nouvelle stratégie de caractérisation fonctionnelle mise en place au cours de cette étude.

1.1 La famille des Adamalysines

Les Adamalysines sont des membres de la superfamille des métalloprotéines parmi lesquelles se distinguent les ADAMs (*A Disintegrin And Metalloproteinase*), essentiellement transmembranaires, et les ADAMTS (*ADAM with Thrombospondin Motifs*), qui sont sécrétées[1]. Les ADAMTSL (*ADAMTS-like*) sont des protéines apparentées aux ADAMTS par la présence d'une homologie de séquence au niveau de l'extrémité C-terminale (figure 1).

Les ADAMTS et TSLs sont sécrétées dans l'espace extracellulaire et se lient aux composants de la matrice extracellulaire (MEC). La MEC constitue non seulement un support physique aux cellules dans tous les tissus, mais aussi une structure dynamique qui contrôle l'homéostasie tissulaire. La composition de la MEC varie suivant les organes et cette diversité de composition regroupe plus de 300 protéines (dont les collagènes), glycoprotéines et protéoglycanes, et de très nombreux glycosaminoglycanes qui contribuent à la plasticité de la MEC. A ces composés qui constituent le « *core matrisome* »[3] viennent s'ajouter des composés sécrétés comme les ADAMTS et TSL qui participent au remodelage de la MEC. En effet, de nombreux substrats des ADAMTS sont des composés matriciels, prin-

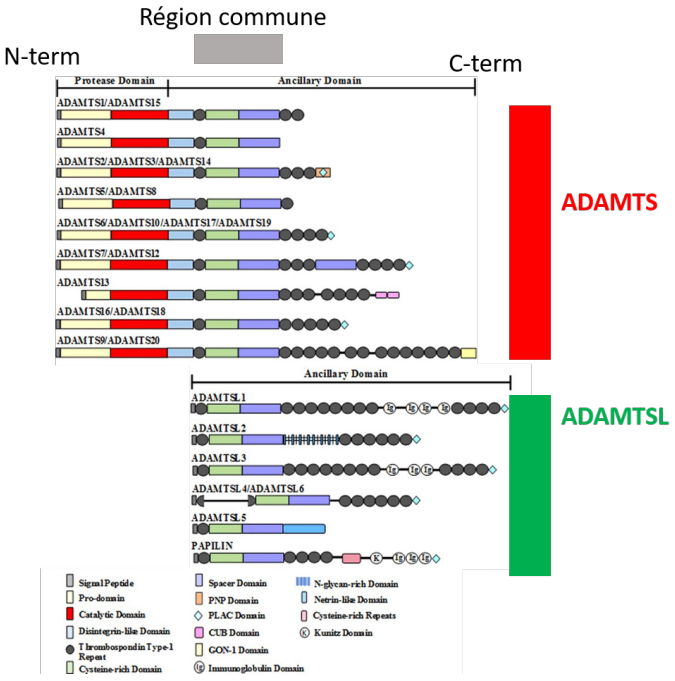


FIGURE 1 – La famille des paralogues humains d'ADAMTS / ADAMTSL et leurs décompositions en domaines connus, figure de Dancevic CM et Al. 2016[2]. Toutes ces protéines partagent une région commune indiquée en gris, les ADAMTSL (vert) contrairement aux ADAMTS (rouge), ne possèdent pas de domaine protéase en extrémité N-terminale. Toutes ces protéines possèdent une extrémité C-terminale variable, composée d'une combinatoire de différents domaines (parmi les 16 types de domaines différents trouvés chez ces protéines).

cipalement des proteoglycans comme les aggrecans et les versicans qui jouent un rôle dans le maintien de l'intégrité de la MEC. Par ailleurs, les ADAMTS et TSL participent à la structuration de la MEC en contribuant notamment à l'organisation des réseaux de fibrilines[4]. Ces activités font de ces protéines des acteurs majeurs dans la régulation de la biodisponibilité des facteurs de croissance et des cytokines qui sont stockés au sein de la MEC. A titre d'exemple, de nombreuses ADAMTS et TSL sont impliquées dans la régulation de l'activité du facteur de croissance TGF-β. Ainsi les protéines ADAMTS2, ADAMTS3 et ADAMTS14 peuvent réguler la réponse au TGF-β par un clivage protéolytique de certains acteurs clés de cette réponse, comme le co-récepteur TGF-βRIII[5]. ADAMTS3 est également capable de cliver la protéine latente du TGF-β(LTBP), tandis qu'ADAMTS2 et ADAMTS14 pourraient agir indirectement, en modulant la rigidité de la MEC via leur activité Pro-Collagène N-Peptidase. Mais ces protéines peuvent aussi agir indépendamment de leur activité protéase et ADAMTS1 a été impliquée dans l'activation non protéolytique de LTBP[6]. Dépourvues de sites catalytiques, les protéines de type ADAMTSL ont été aussi impliquées dans la biodisponibilité du facteur TGF-β. Que ce soit pour ADAMTSL2 qui interagit avec le grand complexe latent du TGF-β[7] ou pour ADAMTSL4 et ADMTSL6 qui sont associées à la fibrillin-1 et dont de nombreuses mutations sont associées à des pathologies en lien avec l'activité du TGF-β[89].

1.2 Les protéines ADAMTS/TSL et la progression tumorale

Les altérations du microenvironnement cellulaire et en premier lieu de la matrice extracellulaire sont directement associées à la progression tumorale[10]. A ce titre le rôle des régulateurs du remodelage de la MEC est essentiel et de très nombreux travaux ont montrés que l'expression et l'activité des protéines ADAMTS / ADAMTSL sont fortement altérées au cours de la progression tumorale[11,12]. L'implication de ces protéines est extrêmement complexe et peut tout aussi bien contribuer à des effets protumoraux qu'antitumoraux comme cela a été démontré dans le contexte de l'angiogenèse[13]. L'exemple le plus documenté est celui de ADAMTS1 qui, par son activité protéolytique sur des composés de la MEC peut générer des facteurs anti-angiogéniques comme pour le nidogen ou des facteurs pro-angiogéniques comme pour les protéoglycans. De la même façon, ADAMTS1 peut piéger le facteur angiogenique VEGF en le liant à son extrémité C-terminale mais peut aussi contribuer à sa libération en activant le remodelage de la MEC (Figure 2).

Cet exemple illustre la complexité des activités des ADAMTS / TSL qui dépend du contexte dans lequel elles sont exprimées et de la nature de leurs interactions avec les composés de leur environnement. Si les adamalysines sont évoquées depuis quelques années comme de nouvelles cibles thérapeutiques d'intérêt, principalement en raison de leur activité enzymatique, les travaux récents démontrent que les ADAMTS ont de nombreux effets indépendants de cette activité comme l'équipe Dymec (équipe en collaboration sur ce projet) l'a montré pour ADAMTS1[6].

Dans ce contexte, l'identification de modules / motifs fonctionnels constitue un verrou dans

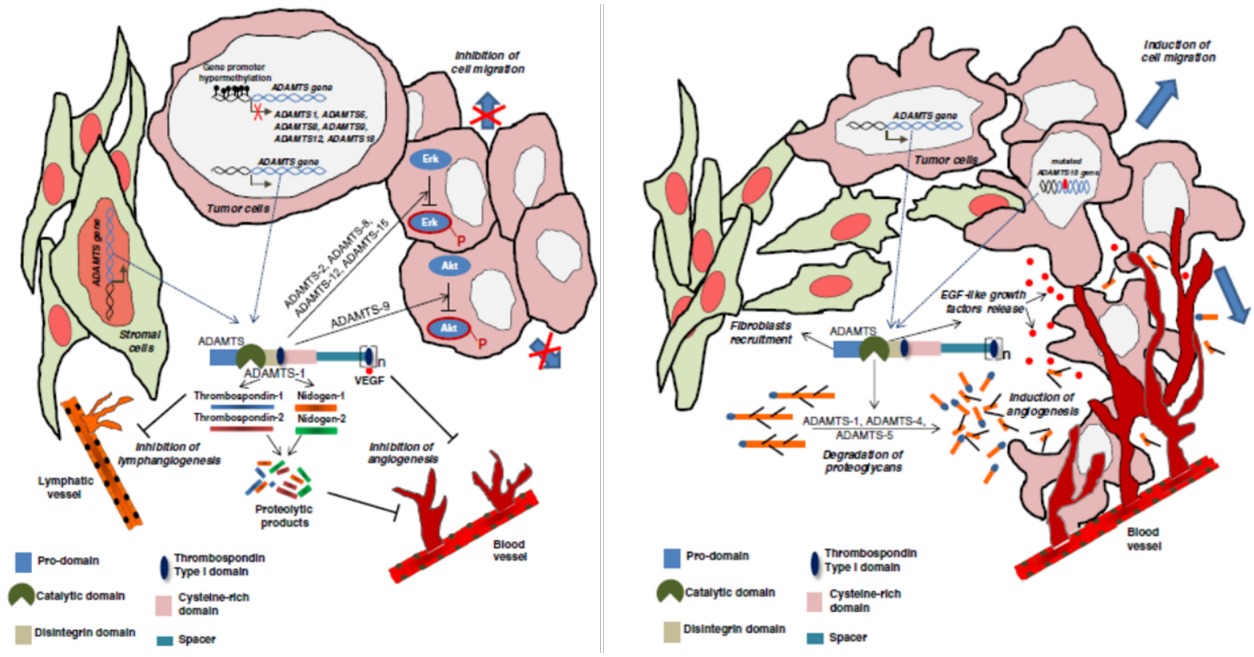


FIGURE 2 – Illustration des effets anti (à gauche) et pro (à droite) tumoraux des ADAMTS, figure de Cal S et al. 2015[11].

la découverte de nouvelles cibles thérapeutiques. En effet la caractérisation des différentes régions responsables des fonctions d'une protéine, permettrait de savoir où cibler précisément cette protéine pour agir sur ces différentes fonctions. C'est pourquoi nous nous sommes intéressés aux méthodes actuelles de prédictions de fonction des protéines afin de pouvoir mettre au point une stratégie adaptée aux ADAMTS / ADAMTSL.

1.3 Méthodes de prédiction de fonctions protéiques

Il existe actuellement 2 principaux types d'approches de prédiction de fonctions protéiques à partir des séquences ; 1) les approches basées sur la conservation de résidus, et 2) les approches basées sur la phylogénie moléculaire.

1.3.1 Conservation de résidus et identification de modules fonctionnels

Résidus conservés

Les différents résidus d'une protéine n'ont pas tous la même importance, certains sont essentiels à la structure ou aux fonctions de la protéine, d'autres sont beaucoup moins importants. Les résidus essentiels subissent une pression de sélection plus importante que les autres résidus de la protéine : ils présentent moins de substitutions et sont plus conservés au cours de l'évolution. La conservation des résidus est l'un des phénomènes les plus étudiés dans le domaine de l'analyse des séquences protéiques[14]. Le principe est d'identifier les résidus conservés qui sont responsables de (ou des) fonction(s) de la protéine[15]. Ces approches sont très utilisées par les biologistes et les bio-

Méthode	Principe	Références
FASTA	Alignement local contre une base de données	Pearso Lipman 1988
BLAST	Recherche de Kmer puis extensions des Kmers	Altschul Gish Miller Lipman 1990, Camacho 2009
ClustalW	Alignement multiple de séquences	Larkin et al. 2007, Thompson Higgins Gibson 1994
MMseqs2	Recherche de Kmer puis alignements successifs (local puis global)	Steinegger Soding 2017
PALOMA	Alignement multiple partiel local	Coste Kerbellec 2006

FIGURE 3 – Exemples de méthodes permettant l'identification de régions conservées ; `FASTA`[16], `BLAST`[17,18], `ClustalW`[19,20], `MMseqs2`[21], `paloma`[22]

informaticiens, et font appel généralement aux méthodes d'alignement multiple des homologues d'une protéine d'intérêt. Le but est de chercher les colonnes les plus conservées de l'alignement multiple afin d'identifier les résidus conservés potentiellement essentiels à la structure et / ou à une fonction de la protéine. Les analyses de conservation sont utilisées pour détecter les résidus impliqués dans des liaisons avec un ligand, pour prédire les interfaces d'interactions protéine-protéine, pour détecter les résidus responsables du maintien de la structure et pour déterminer les spécificités fonctionnelles de la protéine. Les autres méthodes de prédiction de résidus fonctionnels font généralement appel à des informations structurales et sont utilisées quand on dispose d'une structure de la protéine d'intérêt.

Régions conservées

La recherche de régions conservées au sein des séquences protéiques permet la caractérisation de modules ou domaines protéiques[23]. Ces modules ou domaines peuvent être mis en évidence à l'aide de différents algorithmes d'identification de régions conservées (Figure 3). Les modules ou domaines peuvent être représentés sous formes de motifs protéiques, que ce soit par une séquence consensus ou par une matrice des différents résidus possibles à chaque position. Les motifs protéiques décrits dans la littérature sont stockés dans des bases de données de motifs / domaines[24] (Figure 4).

Combinatoire en régions conservées

Les analyses de conservation se basent généralement sur des alignements afin de déterminer les résidus les plus conservés par plusieurs séquences, or les alignements multiples généralement utilisés, vont chercher à obtenir un alignement globalement optimal. Coste (un encadrant du stage) et Kerbellec ont développé un outil d'alignement multiple partiel local (`paloma`)[22] qui permet de détecter l'ensemble des blocs de conservation locale, sans *gaps*, impliquant plusieurs séquences (pas forcément toutes). C'est pourquoi cette méthode est particulièrement adaptée à la recherche

Base de données de motifs / domaines	Description
SWISS-PROT / TrEMBL	Séquences protéiques annotées
PROSITE	Motifs et profils décrivant familles de protéines et domaines
PRINTS	Compendium d'empreintes protéiques
Pfam	Collection d'alignements multiples et de HMMs
SMART	Collection de familles de protéines et de domaines
TIGRFAMS	Familles de protéines basées sur des HMMs
ProDom	Compilation automatique de domaines homologues
InterPro	Documentation intégrée de ressources sur les familles protéiques, les domaines et les sites fonctionnels
IProClass	Classification intégrée de protéines

FIGURE 4 – Exemples de bases de données de motifs / domaines

de régions conservées au sein de protéines possédants une importante combinatoire en domaines, comme c'est le cas pour les protéines ADAMTS / ADAMTSL. Cependant, les approches basées sur la "conservation" des séquences permettent d'observer les résidus qui auraient été sélectionnés par un important nombre de séquences, mais ne permettent pas de corréler la variation des séquences avec des phénotypes. Pour ceci il serait nécessaire de représenter cette variation, par exemple en réalisant une phylogénie des séquences.

1.3.2 Utilisation de la phylogénie moléculaire pour la prédiction de fonctions protéiques

Bien que la majorité des prédictions de fonctions des protéines soit basée directement sur des méthodes de similarité de séquences, il est également possible d'utiliser des méthodes basées sur l'analyse phylogénétiques pour affiner les prédictions de fonction moléculaire d'une protéine[25]. La phylogénomique utilise l'hypothèse que la fonction d'une protéine et sa séquence évoluent en parallèle[26]. Les méthodes phylogénétiques se basent sur la connaissance de l'évolution de la famille de protéines pour décrire comment la fonction moléculaire a pu évoluer. Ces méthodes phylogénétiques utilisent l'histoire de l'évolution des protéines, inférée sous la forme d'un arbre phylogénétique, pour transférer et prédire les fonctions des protéines, à partir des fonctions connues de protéines proches dans l'arbre[27]. Si l'arbre phylogénétique et le transfert de l'information à travers l'arbre sont fait de manière robuste, on évite les biais (e.g. les scores de similarités[28]) que présentent les méthodes de prédiction par similarité de séquences[25]. Les méthodes phylogénétiques ont récemment permis de prédire des fonctions précises de familles de protéines d'intérêts[25].

1.3.3 Phylogénie moléculaire, évolution et cancer

Parce que la progression du cancer est un processus évolutif régit par des contraintes sélectives, les approches de phylogénie moléculaire développées pour modéliser et déduire des relations évolutives parmi les organismes sont aujourd'hui utilisées pour caractériser les étapes de la cancérogenèse à l'échelle des génomes[29]. Croiser l'évolution et le phénotype permet ainsi de caractériser des fonctions. L'application de ces méthodes à l'étude d'un gène pris isolément ou de familles de gènes permet de mettre en évidence des séquences impliquées dans la régulation du processus cancéreux. A titre d'exemple, l'analyse phylogénétique des gènes de la famille des ubiquitines ligase E3, SINA / SIAH a permis d'identifier des motifs fonctionnels conservés au cours de l'évolution et qui constituent des cibles d'intérêt thérapeutique notamment dans le cadre des cancers métastatiques dépendant de l'oncogène K-RAS[30]. Une étude plus récente, basée sur la reconstruction phylogénétique du gène WFDCA, réprimé ou non exprimé dans le cancer du côlon, a permis d'identifier 4 résidus critiques pour l'intégrité de la protéine[31]. Ces approches soulignent l'intérêt de développer des approches de phylogénie moléculaire pour identifier et caractériser de nouveaux motifs / domaines impliqués dans le processus tumoral. L'application de ces méthodes à la famille des ADAMTS / TSL constitue un enjeu particulier en raison du nombre variable de copies paralogues parmi les 19 espèces (i.e. de 26 pour *Homo sapiens* à 7 chez pour *C. elegans*), ainsi que la combinatoire en domaines des paralogues (Figure 1) de cette famille de protéique. Ces différentes caractéristiques rendent la famille de protéine ADAMTS / ADAMTSL non compatible avec les approches de phylogénie classique.

1.4 Réconciliation phylogénétique Domaines-Gènes-Espèces

Réconciliation Gènes-Espèces

Les méthodes classiques de phylogénie moléculaire servent à étudier l'évolution orthologues d'un gène, c'est à dire à regarder l'histoire du gène au cours de l'évolution des espèces. Or les gènes peuvent également évoluer de manière paralogue. Un gène peut évoluer de manière indépendante au sein d'une espèce, il peut être dupliqué, perdu ou même transféré au sein d'une espèce. L'histoire du gène n'est ainsi pas toujours l'histoire des espèces, en particulier chez une famille de protéines multigènes (i.e. comprenant de nombreux paralogues) comme la famille ADAMTS / ADAMTSL. Afin de pouvoir prendre en compte l'évolution orthologue et l'évolution paralogue il existe des outils de réconciliation phylogénétique[32] qui vont permettre de reconstruire la phylogénie de famille multigènes. On parle de réconciliation car l'histoire des espèces (i.e. évolution orthologue) n'est pas l'histoire des gènes (i.e. évolution paralogue).

Protéines multidomaine

De plus, les méthodes classiques de phylogénie moléculaire ne sont pas directement prévues pour prendre en compte le cas particulier des familles de protéines multidomaine. En effet, les gènes des familles de protéines multidomaine sont caractérisés par une mosaïque de segments de séquence, chacun de ces segment de séquence code pour un module fonctionnel ou structural (qui peuvent être aussi nommés domaines) de la protéine[33]. Les familles de protéines multidomaines évoluent par brassage de ces modules, que ce soit par insertion, duplication ou perte de modules[34]. Cette évolution à l'échelle de segments de séquences joue un rôle majeur dans l'évolution de fonctions de protéines multidomaine, comme par exemple cela a été montré pour l'apparition des animaux multicellulaires[35,36] et du système immunitaire chez les vertébrés[37].

Réconciliation Domaines-Espèces

Les méthodes de phylogénie classiques ne sont pas adaptées pour estimer cette évolution à l'échelle de segments de séquence. En effet, les méthodes de phylogénie classique considèrent les protéines comme briques de base de l'évolution. Cela revient à poser l'hypothèse que tous les sites d'un alignement multiple ont la même histoire phylogénétique. Ce qui signifie que chaque protéine possède un ensemble de modules fixes qui évoluent avec elle. Dans le cas de protéines multidomaine ceci n'est pas le cas, des régions différentes d'une protéine multidomaine peuvent résulter d'histoires différentes. Il faut ainsi considérer les modules comme briques de base de l'évolution. Dans ce contexte Stolzer et al. ont développé un outil de réconciliation phylogénétique prenant en compte l'évolution spécifique des modules protéiques[38,39]. On parle de réconciliation car l'histoire d'un module n'est pas l'histoire du gène. Ce modèle réalise 2 réconciliations successives : 1) la réconciliation de l'évolution des espèces avec l'évolution des gènes, puis 2) la réconciliation de l'évolution des gènes avec l'évolution d'un module. Dans la continuité des travaux de Stolzer, Li[40] a récemment développé un nouveau modèle de réconciliation DGS (Domain-Gene-Species), prenant en compte de manière simultanée tous les modules présents et réalisant les deux réconciliations conjointement de manière à optimiser les deux. Ces outils permettent d'obtenir à partir d'un arbre phylogénétique des espèces, d'un arbre phylogénétique des gènes, et d'un arbre phylogénétique pour chaque module, l'histoire phylogénétique la plus parcimonieuse, prenant en compte l'évolution des gènes au sein des espèces et l'évolution des modules au sein des gènes. Pour chaque module, on obtient donc son histoire au cours de l'évolution de la famille de protéines, ainsi que la composition en modules de chacune des protéines ancestrales.

1.5 Prédiction de fonction par phylogénie moléculaire et conservation de séquences

Dans ce contexte, les objectifs du projet ont été d'initier une caractérisation fonctionnelle des protéines de la famille ADAMTS / ADAMTSL en utilisant les nouvelles approches de réconciliation phylogénétique (e.g. réconciliation DGS). Nous proposons une nouvelle stratégie de caractérisation en modules fonctionnels qui va reposer sur : 1) l'identification de modules conservés (impliqués dans les fonctions ou le maintien de la structure de la protéine), 2) l'évolution espèce-gènes-domaines, et 3) les différents types des données disponibles. Nous cherchons à reconstruire l'histoire phylogénétique des différentes protéines et de leurs modules pour pouvoir les mettre en lien avec l'apparition de différentes fonctions et ainsi observer des co-occurences module / fonction dans le but de prédire l'implication du module concerné dans la fonction. Ce travail propose une première phylogénie des ADAMTS et ADAMTSL, une inférence des compositions ancestrales en modules, ainsi qu'un protocole de prédiction d'interaction des protéines modernes et ancestrales.

2 Matériels et méthodes

Dans un premier temps, nous aborderons la stratégie mise en place et son *pipeline* associé, avant de s'intéresser aux étapes principales de ce pipeline. Différents scripts ont étés réalisés en *python3* et en *shell sh* afin de constituer un *pipeline* appliquant notre stratégie, seuls les plus importants seront détaillés dans ce rapport.

2.1 Stratégie et *pipeline* de caractérisation en modules fonctionnels

Une nouvelle stratégie de caractérisation de protéines en modules fonctionnels a été mise en place, afin de prendre en compte les différentes caractéristiques de la famille ADAMTS / ADAMTSL, à savoir une organisation multidomaine, un nombre important de familles de gènes paralogues par espèces (26 chez l'humain) ainsi que des données fonctionnelles de différents types (Figure 5). Cette stratégie consiste dans un premier temps à effectuer une segmentation *sans a priori* en modules des régions conservées dans les séquences protéiques. Cette segmentation permet ainsi d'obtenir un découpage en modules conservés. Dans un second temps, les arbres phylogénétiques des différents niveaux d'évolution (espèces, gènes, modules) sont construits, afin d'effectuer une réconciliation phylogénétique DGS et obtenir une histoire phylogénétique de notre famille de protéine et de ses modules conservés. Dans un dernier temps, les données fonctionnelles des bases de données sont intégrées et transférées le long de cette phylogénie de notre famille de protéine. L'idée est de pouvoir corréler des apparitions de fonctions et de modules au cours de l'évolution, et ainsi de prédire une association module / fonction.

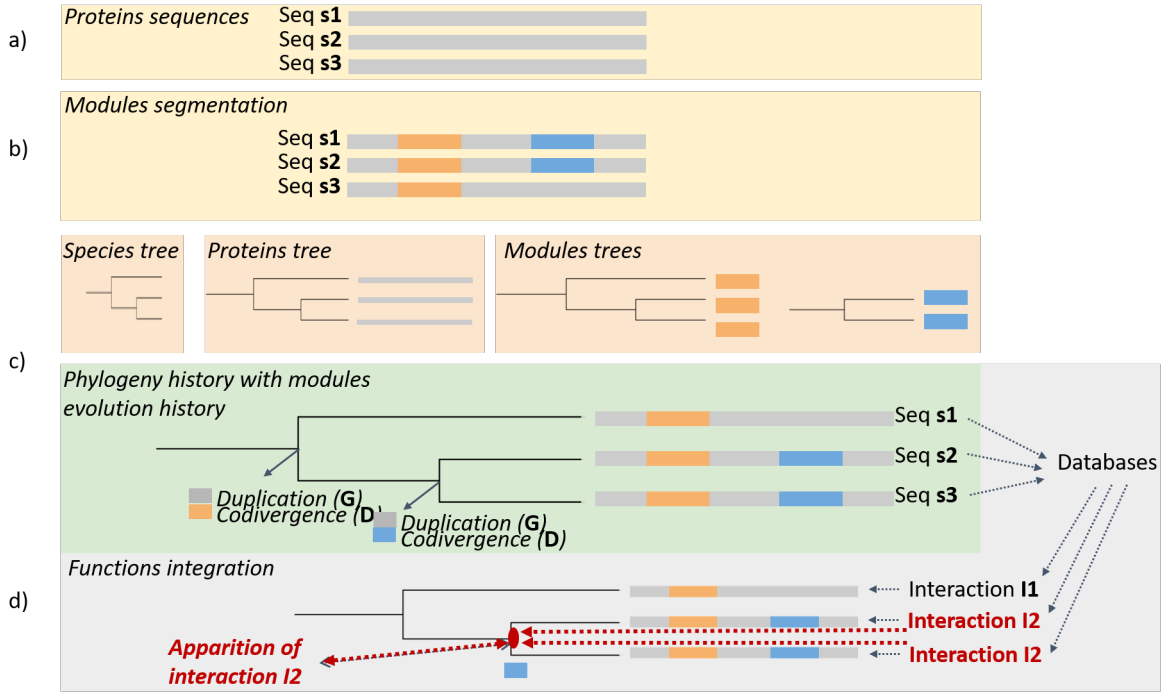


FIGURE 5 – Stratégie de caractérisation en modules fonctionnels mise en place. a) Construction d'un jeu de séquences protéiques (section 2.2). b) Segmentations de ces séquences en modules conservés (section 2.5). Les séquences sont représentées en grises, les modules conservés en orange et en bleu. c) Réconciliation phylogénétique DGS (*Domain-Gene-Species*, section 2.6) pour obtenir l'histoire phylogénétique de la famille de protéines et de ses modules. Différents évènements évolutifs sont attribués à chaque noeud, que ce soit des évènements représentant une évolution à l'échelle des espèces (S), à l'échelle des gènes (G) ou à l'échelle des domaines / modules (D), les différents évènements possibles sont décris en section 2.6. d) Intégration d'informations fonctionnelles (issues de bases de données) à l'arbre obtenu, puis transfert de ces annotations à travers l'arbre (2.7). Ici une information d'interaction protéine-protéine est transférée (en rouge) à l'ancêtre commun des protéines la possédant. Il se trouve que cette interaction protéine-protéine (I2 en rouge), apparaît en même temps que le module bleu, ce qui illustre une association module bleu / interaction I2.

Cette nouvelle stratégie a été implémentée dans un *pipeline* bioinformatique (Figure 6), les étapes principale de ce *pipeline* sont détaillées dans la suite du matériel et méthode.

2.2 Construction du jeu de séquences

Logique du jeu de séquences

La stratégie mise en oeuvre nécessite de prendre en compte : 1) l'évolution orthologue (i.e. les gènes évoluent avec les espèces, deux protéines orthologues sont des protéines homologues chez des espèces différentes, possédant généralement la même fonction et résultantes d'une spéciation), 2) l'évolution paralogue (i.e. les gènes évoluent au sein des espèces, par duplication / perte, deux protéines paralogues sont des protéines homologues d'une même espèces, possédant généralement des fonctions différentes et résultantes d'une duplication du gène). Chaque paralogue est associé à une composition en domaines / modules. Dans le but de pouvoir reconstruire la phylogénie des ADAMTS / ADAMTSL, il est nécessaire de prendre en compte et de réconcilier ces 2 dimensions d'évolution (i.e. évolution orthologue et évolution paralogue). Pour ceci nous avons besoin d'informations de paralogies et d'orthologies, nous allons donc chercher à constituer un jeu de

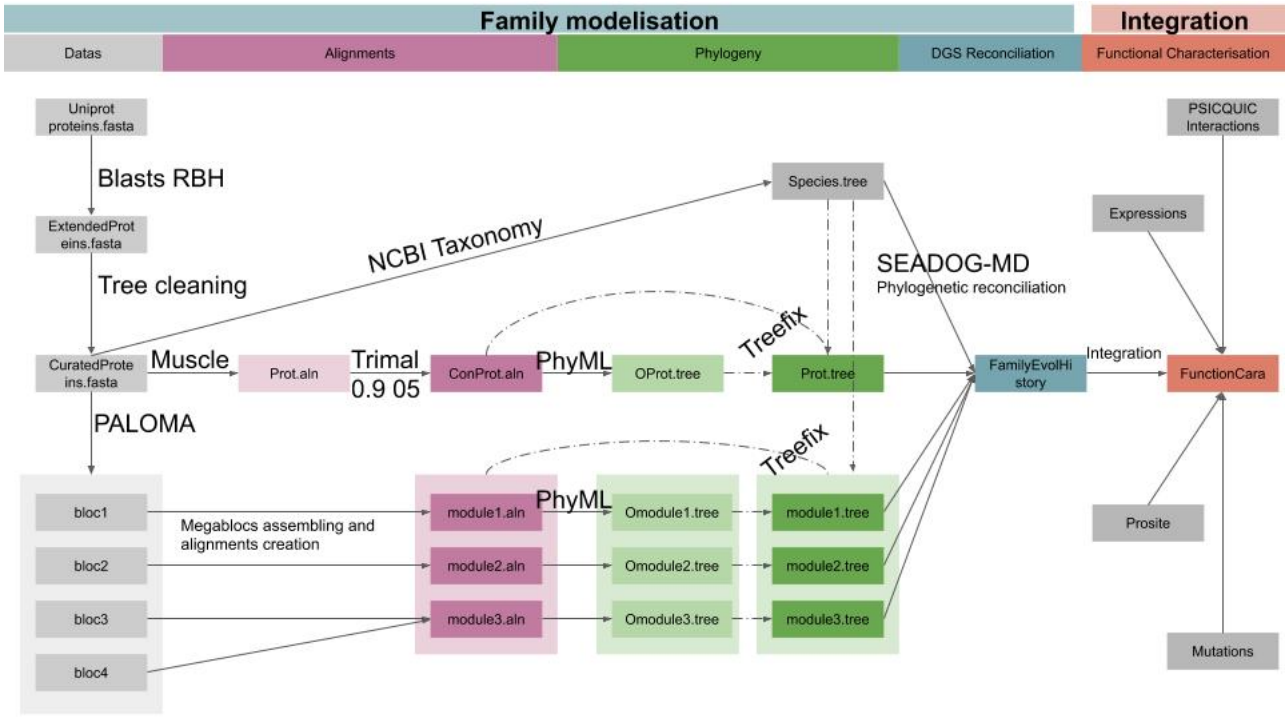


FIGURE 6 – *Pipeline* bioinformatique mise en place pour l'application de la nouvelle stratégie de caractérisation en modules fonctionnels, les différentes étapes de ce *pipeline* sont décrits dans les sections 2.2 à 2.7, les résultats sont visualisés comme décris en section 2.8. Ce *pipeline* consiste à créer un jeu de séquences protéiques (section 2.2), à en chercher les modules conservés (section 2.5), puis à reconstruire les différents arbres phylogénétiques (i.e. arbre des espèces, arbre des gènes, arbres des modules, sections 2.3 à 2.5) nécessaires pour effectuer une réconciliation phylogénétique DGS (Domain-Gene-Species, section 2.6). Différents types de données fonctionnelles sont ensuite intégrées au résultat de cette réconciliation (section 2.7).

séquences protéiques constitué des paralogues d'espèces (i.e. représente l'évolution orthologue) représentatives de l'évolution des métazoaires (Figure 11).

Sélection des espèces

Nous avons identifié 564 espèces possèdant des séquences pour les protéines des familles ADAMTS et ADAMTSL dans la base de données Uniprot[41], avec une surreprésentation des espèces mammifères. Afin de constituer un jeu de données représentatif de l'évolution, nous avons sélectionné 19 espèces possédant des séquences pour les protéines des familles des ADAMTS et ADAMTSL dans la base de données Uniprot. L'utilisation d'espèces différentes nous permettra également d'utiliser les informations fournies chez les différentes espèces dans le but de caractériser les protéines humaines. Ces 19 espèces ont été sélectionnées de manière à représenter la diversité et l'évolution des métazoaires, protostomiens, tuniciers et vertébrés (i.e. différents embranchements de l'arbre de l'évolution des métazoaires).

Récupération des séquences

Les séquences protéiques (au format fasta) ont ensuite été extraites à partir de la base de données

Uniprot. En raison du nombre très important d'isoformes (65 isoformes connus chez l'humain, pour 26 gènes), seules les séquences dites « canoniques » pour chaque protéine et chaque espèce ont été retenues dans cette première étude. Les séquences considérées comme canoniques par Uniprot doivent respecter au minimum l'un des critères suivants : représenter la protéine la plus prévalente, être la plus similaire à ses séquences orthologues, être la plus complète en domaines et annotations, ou à défaut d'aucun de ces critère, être la plus longue. A plus long terme toutes les séquences isoformes devront être intégrées.

`Blasts` RBH

Afin de compléter ce premier jeu de données pour les 19 espèces, des recherches par alignement de séquences de type *Reciprocal Best Hits*[42] (RBH) ont été effectuées pour toutes les séquences manquantes. Pour cela les séquences identifiées chez l'humain (qui est l'espèce possédant le plus de paralogues connus et dont les séquences sont de meilleure qualité) ont été recherchées par alignement de séquences `BLAST`[17](*Basic Local Alignment Search Tool*), dans les génomes des espèces où les annotations sont manquantes. Les meilleures séquences obtenues (*best hit*) ont ensuite été recontrôlées par `BLAST` contre le génome humain dans le but de retrouver l'orthologue humain à l'origine de la recherche. Si l'orthologue humain d'origine est bien identifié, l'orthologue de l'espèce recherchée est validé. Dans le cas contraire, la séquence « *best hit* » est considérée comme trouvée par hasard et n'est pas considérée dans les données. Cette recherche permet de compléter la liste des séquence protéique d'ADAMTS / ADAMTSL qui ne sont pas référencées comme telles dans les bases de données.

Nettoyage du jeu de séquences

Dans un second temps, un arbre phylogénétique des gènes codants pour les séquences protéiques identifiées a été reconstruit par la méthode présentée en section 2.4. Les séquences ne se regroupant pas avec leurs orthologues, ont été retirées du jeu de données. A l'issu de cette étape, le jeu de données nettoyé regroupe 341 séquences protéiques d'ADAMTS / ADAMTSL de 26 familles paralogues chez 19 espèces différentes (section 3.1). Afin de faciliter l'identification des séquences, chaque protéine est référencée par le nom de son orthologue humain précédée du nom de l'espèce d'origine. Par exemple l'orthologue d'ADAMTS1 chez Mus musculus sera noté musmusculusadamts1. Tous les en-têtes des séquences protéiques au format FASTA sont normalisées de la façon suivante :

proteinName Specie (ex : musmusculusadamts1 musmusculus)

2.3 Construction de l'arbre phylogénétique des espèces

Dans le but de pouvoir réaliser une réconciliation phylogénétique, il est nécessaire d'avoir un arbre phylogénétique des espèces. L'arbre phylogénétique des espèces est construit en utilisant la Taxonomie NCBI[43]. Notre arbre a été reconstruit à partir de la version de la base de données disponible le 8 février 2019. La taxonomie NCBI est un regroupement de connaissances taxonomiques issues d'une grande variété de sources (articles publiés, bases de données en lignes, avis d'experts en taxonomies). L'arbre phylogénétique obtenu va également servir à réconcilier les différents arbres qui seront construits par la suite (i.e. arbre des gènes, arbres de modules).

2.4 Construction de l'arbre phylogénétique des gènes

Dans le but de pouvoir réaliser une réconciliation phylogénétique DGS (*Domain-Gene-Species*), il est nécessaire d'avoir un arbre phylogénétique des gènes.

L'arbre phylogénétique des gènes a été réalisé (Figure 6 et 7) en effectuant dans une première étape, un alignement multiple des séquences protéiques complètes en utilisant `MUSCLE`[44] (*multiple sequence comparison by log-expectation*, utilisé avec les paramètres par défauts). L'outil `trimAl`[45] (*a tool for automated alignment trimming in large-scale phylogenetic analyses*) a ensuite été utilisé pour sélectionner les sites communs à toutes les séquences ADAMTS / ADAMTSL (Figure 1), et ainsi se focaliser sur l'histoire évolutive communes des séquences. Les paramètres retenus sont -gt 0.9 -cons 05. `TrimAl` sélectionne les colonnes de l'alignement où sont présentes au minimum 90% des séquences, si le nombre de colonnes sélectionnées est inférieur a 5% de l'alignement, l'outil sélectionne les 5% des colonnes les plus conservées (c'est à dire les colonnes partagées par le plus grand nombre de séquences). A partir des régions sélectionnées par l'outil `trimAl`, un premier arbre phylogénétique a été réalisé avec le logiciel `PhyML`[46] (*maximum-likelihood phylogenetic program*). Cet arbre a ensuite été corrigé par le logiciel `TREEFIX`[47], qui utilise l'arbre issu de `PhyML`, l'alignement de sortis de `trimAl` ainsi que l'arbre phylogénétique des espèces. `TREEFIX` utilise la topologie de l'arbre des espèces pour orienter la reconstruction de l'arbre phylogénétique des gènes en équilibrant le poids des informations issus des séquences (alignement des gènes) et de l'arbre de espèces (e.g. divergences avec les orthologues attendus), grâce à un *framework* de test d'hypothèses statistiques. Toutes les informations disponibles sont ainsi utilisées pour construire l'arbre phylogénétique des gènes le plus probable. Cette correction de l'arbre est recommandée dans le cadre d'une réconciliation phylogénétique, elle constitue une première réconciliation avec l'arbre des espèces en amont (i.e. réconciliation de l'histoire des espèces avec l'histoire des orthologues attendue), tout en utilisant l'alignement multiple.

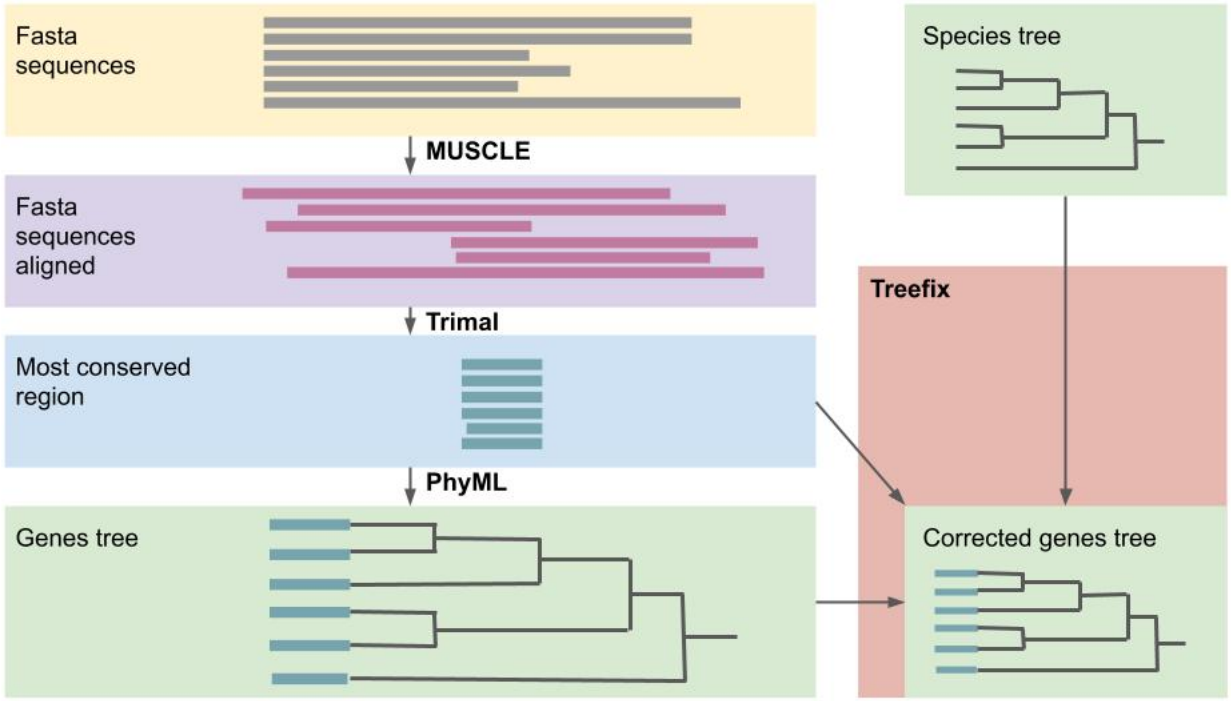


FIGURE 7 – Protocole de création de l'arbre phylogénétique des gènes (section 2.4), les séquences sont dans un premier temps alignées, puis la région la plus conservée est sélectionnée avant d'en construire un arbre phylogénétique, l'arbre est ensuite corrigé en utilisant l'alignement multiple et l'arbre phylogénétique des espèces.

2.5 Recherche de modules conservés

Nous cherchons les différents modules conservés au sein de nos séquences ADAMTS / ADAMTSL, puis pour chacun de ces modules, nous reconstruisons un arbre phylogénétique pour représenter son histoire évolutive. L'obtention d'un arbre phylogénétique par module est nécessaire pour la réconciliation phylogénétique DGS.

2.5.1 Segmentation de séquences en utilisant `paloma`

Le logiciel d'alignement Partiel Local multiple `paloma`[22] a été utilisé afin de rechercher *sans a priori* les régions conservées au sein de nos séquences. `Paloma` va rechercher des blocs de régions localement conservées par plusieurs séquences. Ces "blocs" obtenus sont caractéristiques d'au minimum 2 séquences (paramètre Q), et d'au maximum toutes les séquences. La totalité des blocs obtenus forme un alignement multiple partiel local, ou PLMA (*Partial Local Multiple Alignment*).

Suite à plusieurs tests, `paloma` a été utilisé avec les paramètres suivants :

-Q 2 : Il faut au minimum 2 séquences par blocs (default).

-m 25 : Les segments d'un alignement local doivent être au minimum d'une taille de 25 acides aminés.

-M 40 : Les segments d'alignements locaux utilisés pour rechercher les blocs de régions conservées, sont au maximum d'une taille de 40 acides aminés (maximum).

-t 20 : Poids minimum pour considérer un alignement local (seuil).

-c : Recherche les blocs de faibles consensus (composants connectés).

`Paloma` permet d'obtenir un découpage de nos séquences en petit blocs conservés. C'est néanmoins encore un prototype de recherche peu optimisé ne permettant pas de traiter simultanément l'ensemble de nos séquences d'intérêts en temps raisonnable.

2.5.2 Méthode d'accélération de `paloma` basée sur la redondance

Dans le but de réduire le temps de création des blocs tout en minimisant la quantité d'information perdue, une stratégie basée sur l'élimination de la redondance a été mise en place (Figure 8). Cette méthode consiste à construire avec `MMseq`[48] un jeu de séquences non redondantes. Pour chaque groupe de séquences partageant au minimum 90% d'identité, une seule séquence est conservée dans le jeu non redondant. Cette séquence représentera toutes les autres. En utilisant `MMseq` sur les 341 séquences, il n'en reste que 42 représentatives de toutes les autres. Le but de l'étude est la caractérisation des protéines humaines, c'est pourquoi les 26 paralogues humains ont été ajoutés aux 42 séquences représentatives, ce qui permet de constituer un jeu de 68 séquences représentatif, en terme de diversité, de structures, de fonctions, de compositions en domaines, de l'ensemble du jeu de données (Figure 8A).

Un alignement partiel local multiple (PLMA) de ce jeu non redondant de séquences (68 séquences) a ensuite été réalisé, ce qui permet d'obtenir un découpage en blocs basé sur les séquences les plus représentatives du jeu de données (Figure 8B). La présence de ces blocs a ensuite été détecté par un algorithme naïf de recherche de sous-chaîne, chez les autres protéines (i.e. protéines absentes du PLMA). Si la séquence consensus du bloc est retrouvée chez une protéine (à 90% d'identité), cette protéine est considérée comme ayant le bloc (Figure 8C).

2.5.3 Regroupement des blocs en modules conservés

Plusieurs blocs trouvés par `paloma` peuvent être adjacents, nous cherchons à les regrouper afin de caractériser des régions conservées de taille plus conséquente que celle des blocs. Un algorithme *d'Union-Find*[49] a été utilisé pour regrouper les blocs adjacents. Les blocs sont considérés adjacents si, et seulement si, ils partagent un nombre minimum de séquences où les blocs sont directement contigus. Sur le jeu de données utilisé, il a été choisi de considérer adjacents 2 blocs si ils possèdent 6 séquences contiguës d'un bloc à l'autre. Les blocs adjacents sont ensuite regroupés jusqu'à que cela ne soit plus possible (i.e. jusqu'à qu'on ne trouve plus de blocs adjacents non regroupés). Le regroupement des blocs va représenter un module.

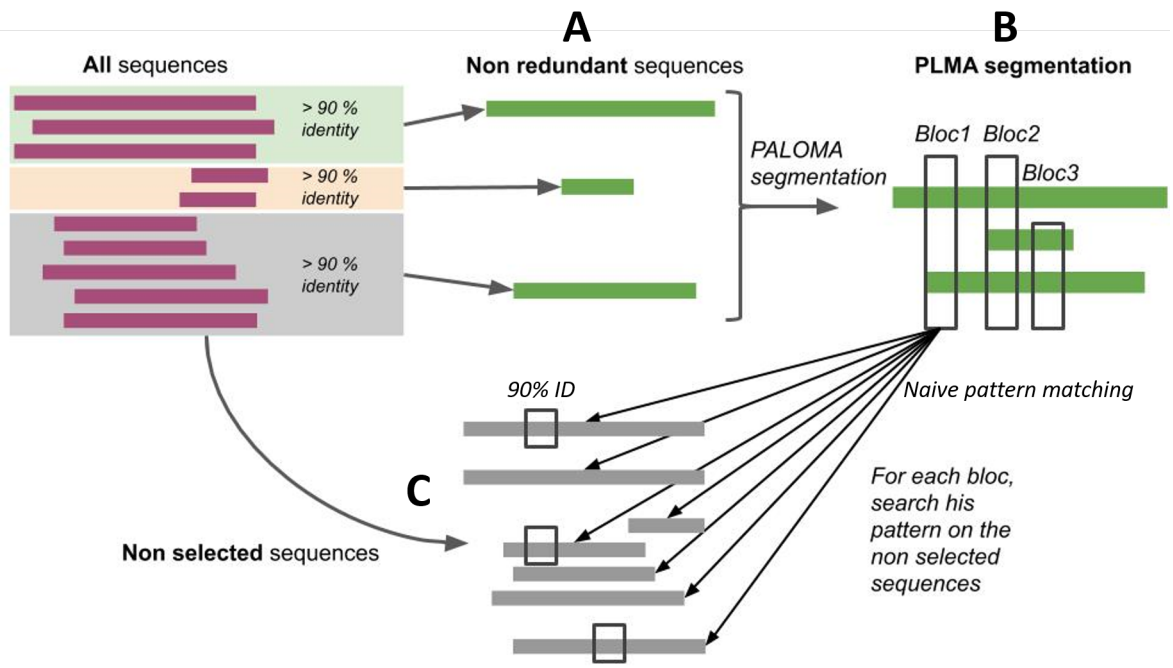


FIGURE 8 – Stratégie d'accélération de la segmentation (section 2.5.2), `paloma` est utilisé pour segmenter un jeu de séquences (ici de 68 séquences) non redondants A, puis les blocs B trouvés sont recherchés dans les séquences non sélectionnées C (ici 273 séquences).

2.5.4 Formatage des modules en alignements multiple

Dans le but de pouvoir réaliser un arbre phylogénétique par module, il est nécessaire de représenter chaque module sous la forme d'un alignement multiple. Les séquences au sein d'un bloc sont déjà alignées. Chaque bloc du module peut être présent chez des séquences différentes et chaque séquence présente dans le module n'est pas nécessairement présente chez tous les blocs du module. Pour chaque bloc du module, si la séquence est présente dans le bloc, la séquence est notée dans l'alignement multiple du module. Si une séquence est présente dans le module mais absente dans le bloc, des *gaps* de la taille du bloc sont notés dans l'alignement multiple. Un alignement multiple (.fasta) de chaque module est ainsi construit en se basant sur sa composition en blocs.

De plus tous les *headers* des fasta de nos modules suivent le format suivant :
moduleID|start|stop|_proteinName (ex : MB66|54|85|_musmusculusadamts1).
Ce formatage est nécessaire pour l'approche de réconciliation phylogénétique utilisée par la suite.

2.5.5 Création des arbres phylogénétiques des modules

Pour pouvoir réaliser une réconciliation phylogénétique DGS, il est nécessaire d'avoir un arbre phylogénétique pour chaque module (Figure 6). Pour chaque alignement de module, un arbre phylogénétique a été réalisé en utilisant `PhyML`[46]. Comme pour la construction de l'arbre des gènes, chacun de ces arbres est ensuite corrigé en utilisant `TREEFIX`[47]. Le logiciel TREFFIX utilise

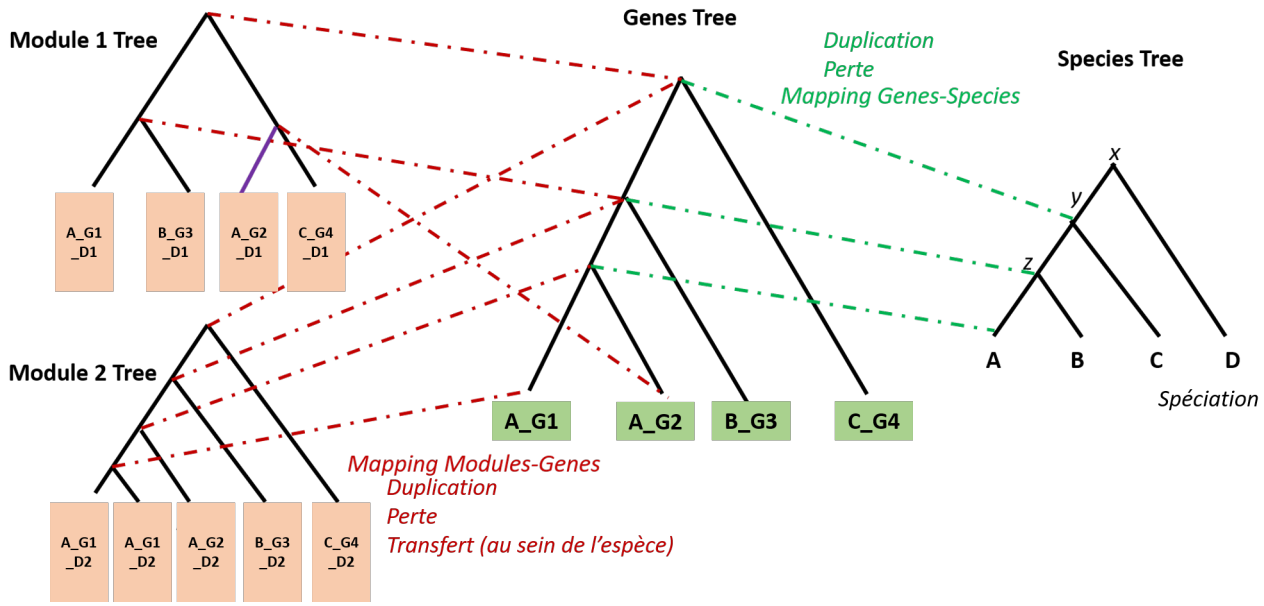


FIGURE 9 – Les différents *mappings* effectués par la réconciliation phylogénétique DGS (Domain-Gene-Species, section 2.6) sont représentés en lignes pointillées, figure inspirée de Li L[40]. Les branches des arbres suivent une normalisation : espèce (pour l'arbre des espèces, *ex : A*), espèce_gène (pour l'arbre des gènes, *ex : A_G1*), espèce_gène_domaine (pour les arbres des domaines, *ex : A_G1_D2*). Deux noeuds sont *mappés* si leurs sous arbres partagent les mêmes feuilles. On associe aux noeuds des évènements évolutifs (spéciation, duplication, perte etc ...) décris dans la section 2.6. Par exemple, le gène G est dupliqué chez l'espèce A (évolution paralogue) en A_G1 et A_G2, et le module A_G1_D2 est dupliqué chez le gène A_G1.

l'arbre phylogénétique du module issu de `PhyML`, l'alignement multiple du module ainsi que l'arbre phylogénétique des espèces, dans le but de corriger l'arbre phylogénétique initial (voir section 2.4).

2.6 Réconciliation phylogénétique DGS (Domain-Gene-Species)

Afin de reconstruire l'histoire phylogénétique de nos protéines ainsi que de leurs modules, une réconciliation phylogénétique a été effectuée en utilisant l'implémentation `SEADOG-MD` du modèle de réconciliation phylogénétique DGS[40].

Nous rappelons que la réconciliation phylogénétique **DGS** utilise l'arbre phylogénétique des espèces, l'arbre phylogénétique des gènes, et les arbres phylogénétique des modules (1 arbre par module) pour estimer l'histoire de la famille de protéines et de ses modules. Pour cela, le logiciel effectue un *mapping* des noeuds équivalents (i.e. les sous arbres avec les mêmes feuilles) des différents arbres et explique les différents *mapping* par des événements évolutifs (Figure 9).

Afin de pouvoir effectuer ce *mapping* les noms des différentes entités (espèces / gènes / modules) des différents arbres doivent respecter des formats spécifiques. Il est nécessaire de pouvoir connaître les gènes d'origine des différents modules, ainsi que les espèces d'origine des différents gènes. Les formatages utilisés sont les suivants :

Format des noms d'espèces : species

Ex : musmusculus

Format des noms de gènes : proteinName_Species

Ex : musmusculusadamts1_musmusculus

Format des noms de modules : moduleID|start|stop|_proteinName

Ex : MB66|54|85|_musmusculusadamts1

Deux *mappings* sont effectués de manières synchrones, 1) un *mapping* est effectué entre l'arbre des espèces et celui des gènes ainsi 2) qu'un *mapping* entre chaque arbre de modules et l'arbre des gènes. Les 2 réconciliations sont ainsi faites en simultané de manière à optimiser conjointement les 2 réconciliations (i.e. réconciliation gène / espèce et réconciliation modules / gène). Chaque *mapping* est expliqué par un événement évolutif défini.

L'implémentation utilisée considère les évènements évolutifs suivants associés à chaque noeud des différents arbres phylogénétiques :

A l'échelle des espèces : spéciation.

A l'échelle des gènes : duplication, perte.

A l'échelle des modules : duplication, perte, transfert (entre gènes de la même espèce).

Chaque évènement possède un coût fixe (qui est paramétrable), ce qui permet d'attribuer un score global à chaque modèle évolutif et de sélectionner le modèle évolutif le plus parcimonieux en minimisant le coût des évènements. Le résultat de la réconciliation est présenté dans un fichier texte, comprenant les arbres phylogénétiques de sorties (espèces, gènes, modules), avec les événements évolutifs attribués à chacun des noeuds, ce qui permet d'obtenir la composition en modules des différentes protéines ancestrales.

L'histoire phylogénétique obtenue va ensuite être corrélée avec des données fonctionnelles dans le but de pouvoir caractériser les fonctions associées aux différentes combinaisons de modules apparus lors de l'évolution des familles de gènes ADAMTS / ADAMTSL.

2.7 Intégration de données fonctionnelles à l'histoire phylogénétique

Afin d'annoter nos modules et de pouvoir observer des co-occurences de fonctions et de modules, les données fonctionnelles issues de différentes bases de données sont intégrées à l'histoire phylogénétique obtenue. Ces données vont servir à annoter les différentes protéines et leurs différents modules. D'une manière générale, des bases de données vont être interrogées avec les protéines actuelles, puis les annotations trouvées vont être transférées aux feuilles de l'arbre, avant d'être propagées à travers l'arbre. Trois types d'annotation fonctionnelle ont été utilisées et intégrées à l'histoire des gènes qui est représentée sous la forme d'un arbre phylogénétique (méthode présentée en section 2.7), où chaque noeud correspond à une protéine caractérisée par sa composition en modules.

1) Données à l'échelle des protéines : interaction protéine-protéine.

2) Données à l'échelle des modules : motif Prosite.

3) Données à l'échelle des acides aminés : mutation référencée.

Pour chacun de ces types de données, l'intégration a été faite d'une manière différente (Figure 10). On va essentiellement interpréter l'évolution des interactions protéine-protéine durant l'évolution des ADAMTS / ADAMTSL.

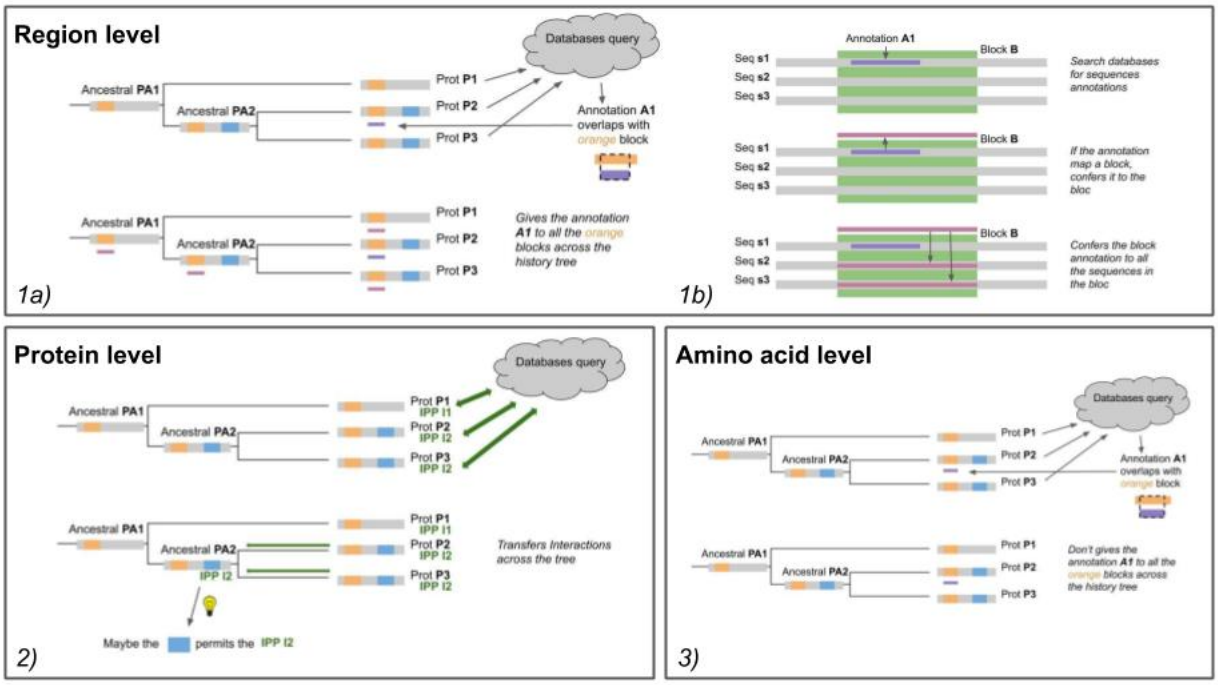


FIGURE 10 – Les différents protocoles d'intégration fonctionnelle des 3 types de données différentes (section 2.7). 2) Les annotations de protéines (e.g. interaction protéine-protéine) sont transférées à l'ancêtre commun des protéines les possédants (section 2.7.1). 1a) Les annotations de régions (e.g. Prosite) sont transférées à toutes les occurrences du module dans lequel elles se trouvent (section 2.7.1) 1b) sur la base de la conservation des séquences au sein du module. 3) Les annotations de résidus (e.g. SNP) sont simplement attribuées au module dans lequel elles se trouvent (section 2.7.3).

2.7.1 Intégration des données de protéines

Pour chaque protéine actuelle (i.e. feuilles de l'arbre), l'API d'interactions protéine-protéine PSICQUIC[50] a été interrogée. Cette API permet d'interroger les bases de données suivantes : Intact, Imex, Mentha, Irefindex, Reactome-fis et Matrixdb (bases de données d'interactions validées de composés de la MEC).

Dans un premier temps les interactions trouvées sont attribuées aux protéines existantes correspondantes. Dans un second temps les interactions sont également attribuées à l'ancêtre commun des protéines possédant cette interaction. C'est le moment d'apparition de l'interaction au cours de l'histoire phylogénétique qui nous intéresse (Figure 10 2)).

2.7.2 Intégration des données de régions de séquences

Pour chacunes des protéines existantes actuellement (i.e. feuilles de l'arbre), l'API Expasy[51] a été interrogée à la recherche de signatures Prosites. Pour chaque signature trouvée, ses positions sur la séquence protéique ont été comparées avec les positions des différents modules inférés (section 2.1). Si une signature Prosite est chevauchante avec un des modules, la signature Prosite est attribuée au module conservé. Les modules représentent des séquences hyper-conservées et on considère que la fonction également est conservée. L'annotation trouvée est ainsi attribuée à toutes les occurrences de ce module (Figure 10 1)), et cela à travers la totalité de l'arbre phylogénétique (protéines actuelles comme protéines ancestrales). Ces annotations sont transférées à titre informatif.

2.7.3 Intégration des données d'acides aminés

Pour chaque protéine actuelle (i.e. feuilles de l'arbre), l'API Uniprot a été interrogée afin de récupérer les informations de SNP (*Single Nucleotide Polymorphism*) issus de de la base de données dbSNP[52]. Pour chaque mutation, sa position a été comparée avec les positions des modules. Si une mutation se trouve dans un des modules, elle est attribuée au module conservé en question, cependant elle n'est pas attribuée aux autres occurrences de ce module dans l'arbre (Figure 10 3)). Ces annotations sont transférées à titre informatif.

2.8 Représentation de l'histoire phylogénétique et des données fonctionnelles

Un script *python3* a été développé, afin de d'intégrer et de visualiser les informations issues de la réconciliation phylogénétique. Ceci afin de créer une modélisation de l'histoire phylogénétique de la famille de protéines, sous forme d'arbre phylogénétique et en utilisant le package python `ETE3`[53]. Ce script prend en entrée (i.e. *input*) le fichier texte .output du programme de réconciliation `SEADOG-DM`, ainsi qu'un fichier texte des correspondances entre les noms normalisés des protéines et leurs identifiants Uniprot. Chaque noeud de cet arbre correspond à une protéine, les feuilles correspondent aux protéines actuelles et les noeuds internes correspondent aux protéines ancestrales. Les différentes informations d'événement évolutif issues de la réconciliation sont utilisées pour récupérer les compositions en modules des différentes protéines. L'intégration des données fonctionnelles à cet arbre permet l'annotation des différents éléments de cet arbre.

Chaque noeud possède un objet python stockant diverses informations sur la protéine :

Une liste des différents modules de la protéine : Pour chaque module on a ; sa position de début, sa position de fin, le nom du module, les annotations du module (signature Prosite, mutations (i.e. SNP) présentes dans le module)

Une liste des évènements évolutifs qui se sont produits à ce noeud. *Ex : Spéciation, co-divergence, duplication, perte*

Une liste des interactions de cette protéine.

L'arbre peut également être coloré en fonction des caractérisations existantes dans la littérature. La sortie (i.e. *output*) de ce script comporte : l'arbre phylogénétique de tous les paralogues et orthologues (format pdf), le sous arbre de l'histoire phylogénétique des paralogues humain (format pdf), un fichier texte pour chacune des protéines (comportant le nom, l'identifiant Uniprot, la compositions en modules et les différentes annotations associées), un répertoire contentant les résultats des requêtes PSICQUIC (i.e. interactions protéine-protéine), un répertoire contenant les résultats des requêtes Expasy Prosite ainsi qu'un répertoire contenant les résultats des requêtes Uniprot.

3 Résultats

3.1 Jeu de données utilisé

8636 séquences d'ADAMTS / ADAMTSL sont répertoriées dans la base de données Uniprot, 330 sont issues des 19 espèces sélectionnées. Les `Blast` RBH permettent de trouver 40 séquences protéiques supplémentaires, ce qui fait un total de 370 séquences protéiques. Après nettoyage des séquences fortement divergentes, il reste 341 séquences protéiques (Figure 11). Un sous jeu de données de 76 ADAMTS de 4 espèces (*Homo sapiens, Bos taurus, Rattus norvegicus, Mus musculus*) a également été utilisé. Notre *pipeline* de caractérisation en modules fonctionnels a été appliquée aux 2 jeux de données (i.e. 341 et 76 séquences). Par souci de taille des figures (Annexe 2), les résultats présentés dans ce mémoire portent uniquement sur le jeu de 76 séquences.

3.2 Arbres phylogénétiques

Ce *pipeline* a permi d'obtenir les arbres phylogénétiques les plus probables pour chacun des éléments de la famille de protéine. Nous obtenons un arbre phylogénétique des espèces, un arbre phylogénétique des gènes ADAMTS / ADAMTSL ainsi qu'un arbre phylogénétique pour chacun des 105 modules trouvés.

3.3 Modularité des paralogues humains

La segmentation en régions conservées des protéines a permis d'obtenir la composition en modules des différents paralogues et orthologues. Chaque module est annoté en fonction des mutations (i.e. SNP) et des signatures Prosite qui lui ont été associés (Figure 12).

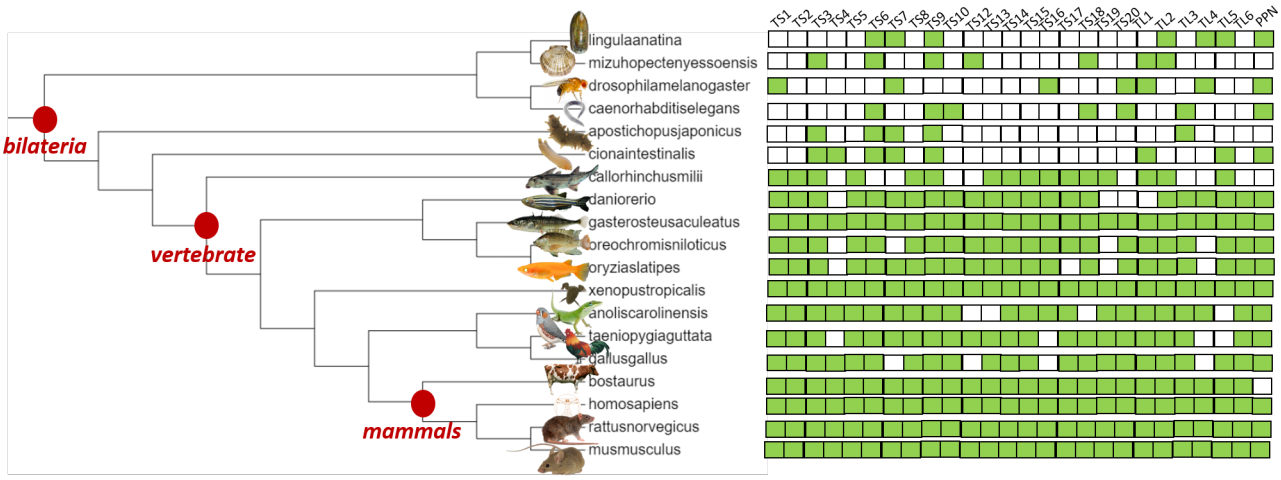


FIGURE 11 – Arbre phylogénétique des 19 espèces sélectionnées et tableau des différents paralogues (les protéines homologues présentes chez une même espèce, souvent de fonctions différentes, correspondant ici à une ligne) et orthologues (les protéines homologues présentes chez des espèces différentes, souvent de même fonctions, correspondant ici à une colonne) du jeu de 341 séquences. Chaque carré représente une séquence protéique potentielle qui a été recherchée, les carrés verts représentent une séquences présentes dans le jeu de données finale de 341 séquences.

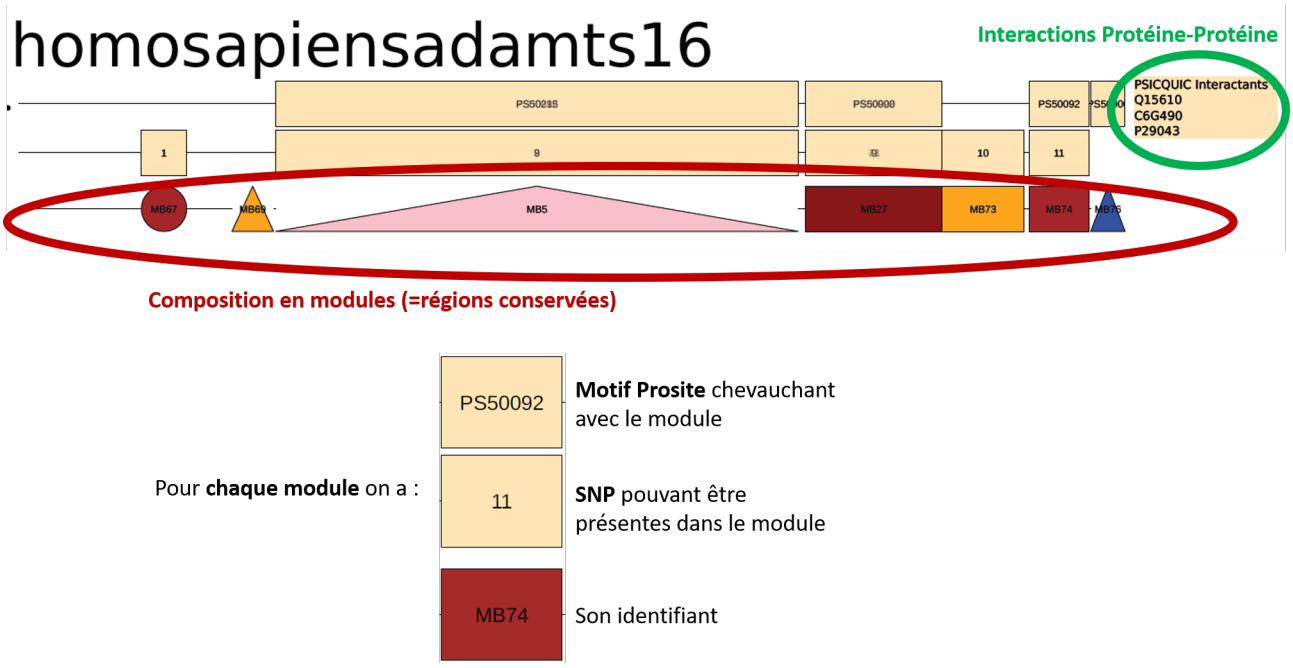


FIGURE 12 – Exemple de caractérisation en modules fonctionnels obtenus par notre *pipeline*, ici ADAMTS16 humaine.

3.4 Histoire des paralogues et orthologues

La stratégie mise au point permet de construire l'histoire phylogénétique des orthologues, des paralogues et de leurs modules conservés, sous la forme d'un arbre phylogénétique (Annexe 1). Chaque noeud de cet arbre phylogénétique correspond à une protéine, les feuilles représentent les protéines actuelles alors que les noeuds antérieurs représentent les protéines ancestrales. Chaque

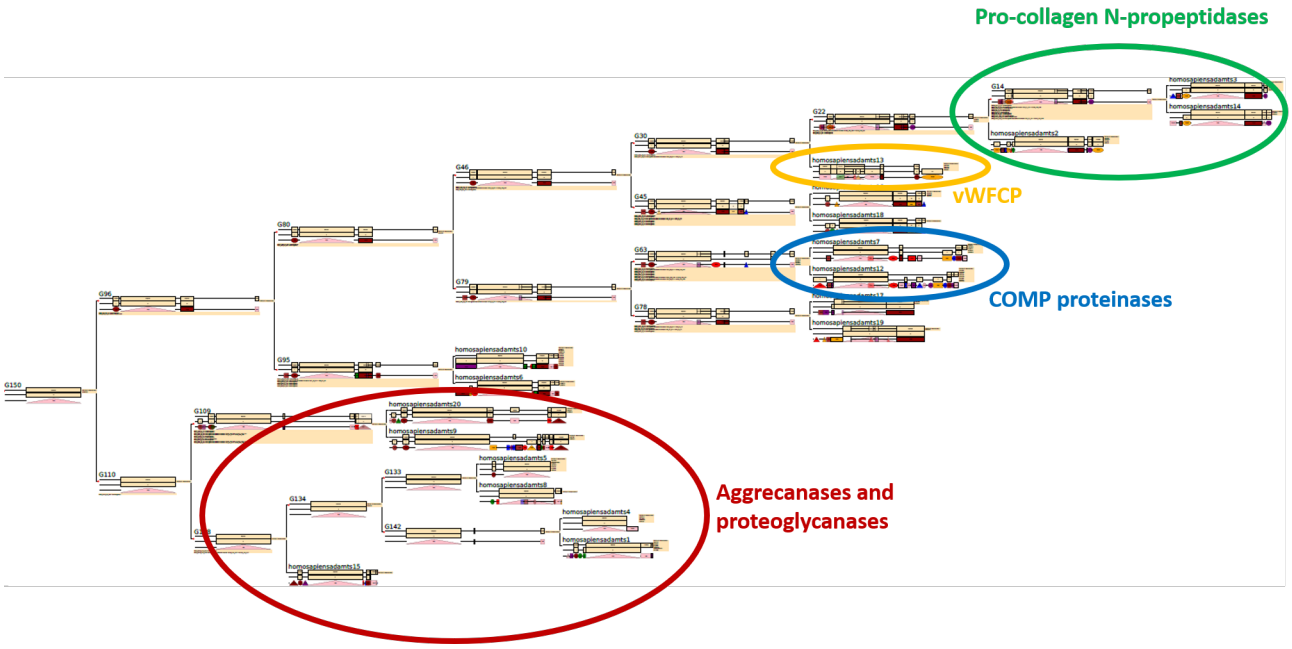


FIGURE 13 – Histoire phylogénétique des paralogues humains et de leurs modules issus de notre *pipeline*. Les annotations fonctionnelles, ici de la littérature, sont représentées par des cercles de couleur entourant les protéines présentant la même annotation. En rouge les protéines de types aggrecanases et proteoglycanases, en bleu les protéinases clivant la protéine COMP, en jaune les protéines clivant le facteur de Von Willebrand, en vert les pro-collagen N-propeptidases. Chaque protéine est caractérisée comme présentée sur la figure 12.

protéine, qu'elle soit ancestrale ou actuelle, est annotée par les évènements évolutifs qui lui sont associés, sa composition en modules conservés (avec leurs annotations) et les protéines avec lesquelles elle est capable d'interagir.

3.5 Histoire des paralogues humains

L'histoire phylogénétique des paralogues et des orthologues permet d'obtenir l'histoire phylogénétique des paralogues humains (Figure 13). Comme le montre la figure 13, l'histoire des paralogues humains que nous avons caractérisée est en accord avec les observations de la littérature. Les protéines ayant des substrats identiques se regroupent dans l'arbre. Nos inférences permettent une première proposition d'enracinement de cet arbre, permise par l'examen des gènes d'autres espèces, ainsi que la première étude de l'histoire conjointe des ADAMTS et ADAMTSL.

3.6 Co-occurence de modules et d'interactions

L'histoire phylogénétique permet de corréler l'apparition des différents modules et des différentes interactions. Il est possible d'observer la co-occurence de modules conservés et d'interactions au cours de l'évolution de la famille de protéines. C'est par exemple ici le cas du module MB64 (50 résidus) qui apparaît en même temps que les interactions avec les protéines COMP et A2MG chez l'ancêtre commun d'ADAMTS 7 et d'ADAMTS 12 (Figure 14). Nos inférences produisent les

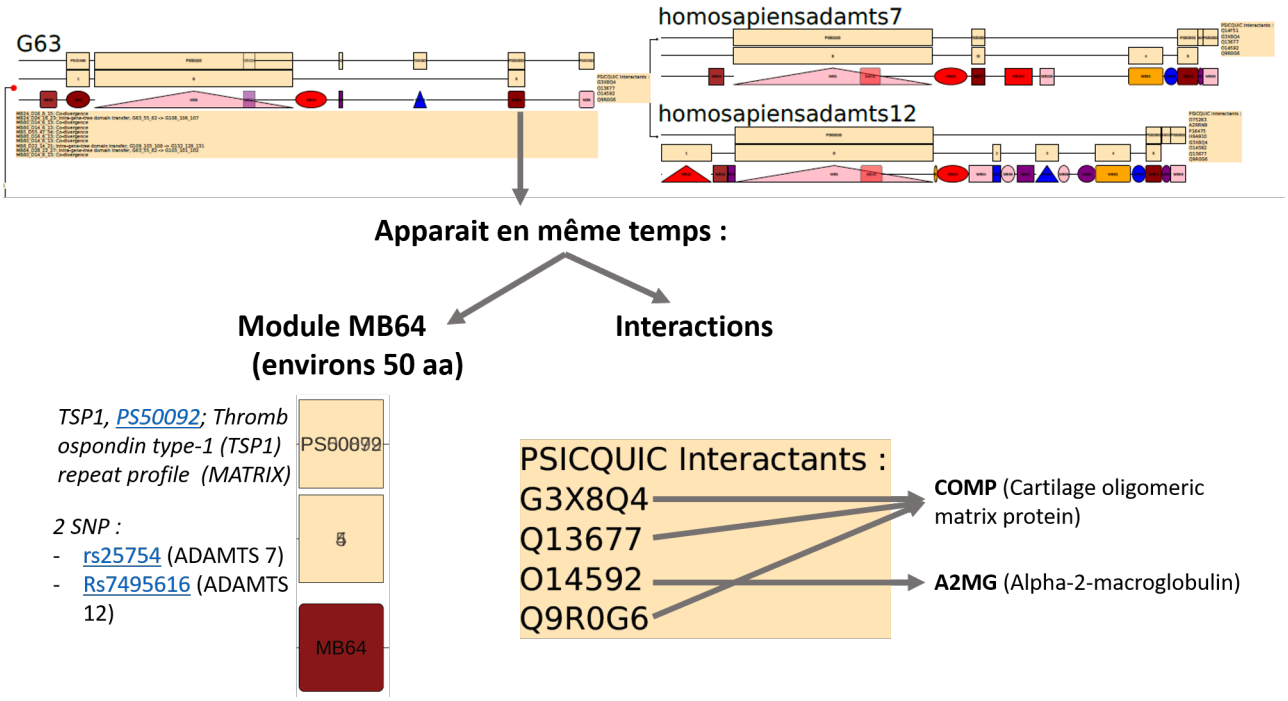


FIGURE 14 – Exemple de co-occurence observable sur l'histoire des paralogues humains, ici du module MB64 et des interactions avec COMP et A2MG chez l'ancêtre commun G63 de ADAMTS7 et ADAMTS12

premières prédictions de modules impliqués dans une interaction protéine-protéine basée sur un principe d'analyse phylogénétique.

4 Discussion

L'objectif de ce stage était d'effectuer une caractérisation fonctionnelle et une phylogénie des protéines de la famille ADAMTS / ADAMTSL. Ces protéines possédant des caractéristiques compliquant les analyses classiques, il a été nécessaire de mettre au point une stratégie d'analyse adaptée et de développer le *pipeline* associé.

Le but était que cette stratégie interprète au mieux les différentes caractéristiques (multidomaines, paralogues, orthologues, multiple niveaux de connaissances) de cette famille de protéines ainsi que les différentes facettes de la problématique (i.e. caractérisation en modules fonctionnels, phylogénie).

Mise en place d'une nouvelle stratégie de caractérisation fonctionnelle

Pour répondre à ce besoin, une nouvelle stratégie a été développée, celle-ci se base à la fois sur 1) le principe d'évolution des gènes au sein des espèces et des domaines au sein des gènes, et sur 2) le principe de modularité fonctionnelle des protéines.

Un *pipeline* permettant d'appliquer cette stratégie a été mis au point. Pour cela il a été nécessaire

de faire fonctionner de manière conjointe des outils de nature très différentes ; 1) un programme de réconciliation phylogénétique Domaines-Gènes-Espèces, 2) un outil de segmentation de séquences et 3) un programme de propagation d'annotations fonctionnelles aux noeuds non annoté de la phylogénie. Le *pipeline* qui a été développé est le premier outil utilisant de manière conjointe ces trois types de méthodes.

Dans le but de faire fonctionner de manière cohérente ce *pipeline*, il a été crucial de régler une grande variété de paramètres, que ce soit sur les alignements multiples, les outils de phylogénie, le programme de segmentation, la valeur des évènements de la réconciliation phylogénétique ou la manière de représenter les informations fonctionnelles.

Découpage en modules conservés

La première étape a consisté à segmenter les séquences protéiques en modules conservés. Pour ceci, le logiciel `paloma` a été utilisé dans le but de prédire des petits blocs extrêmement conservés, ensuite un algorithme a été développé afin de regrouper les blocs adjacents. Il est possible d'obtenir une segmentation en modules plus ou moins précise. Suivant les choix de paramètres de segmentations (i.e. prédiction des blocs et regroupement en modules) il est possible d'obtenir des modules très petits (i.e. la taille d'un bloc issu de `paloma`) ou très grands (i.e. le module recouvre éventuellement la totalité de la protéine). La taille des modules va influencer la précision de l'analyse fonctionnelle. Faire varier la taille des modules revient à choisir le "grain" de l'analyse. Considérer des modules très petits revient à considérer des régions très précises, au risque de ne pas pouvoir les caractériser. Alors que considérer des modules très grands assure une caractérisation de la plupart des modules, mais une moins bonne précision sur la région impliquée. Il est donc nécessaire de faire varier le protocole de construction des modules dans le but d'identifier de nouveaux modules qui répondent au but de l'étude.

Utilisation de la redondance

En raison de la taille de nos données, il est impossible d'utiliser `paloma` sur l'ensemble de nos séquences d'ADAMTS / ADAMTSL. C'est pourquoi nous avons été contraints de mettre au point une méthode pour accélérer la segmentation (i.e. découpage en modules) des séquences. L'élimination de la redondance au sein de notre jeu de données a permis de faciliter la segmentation, il faut noter que la caractérisation des modules se base uniquement sur les informations présentes dans un tout petit jeu de séquences, conçu pour être représentatif de toutes les autres et posséder tous les modules caractéristiques. Cependant, tout module n'existant pas chez au minimum 2 séquences du jeu non redondant, ne sera pas considéré dans notre analyse. L'étude s'intéresse en particulier aux protéines humaines, c'est pourquoi elles ont été ajoutées au jeu non redondant, afin d'avoir la certitude de détecter chez d'autres espèces tous les modules présents chez les protéines humaines. Cette méthode d'accélération permet aussi de réduire le bruit résultant de modules artéfactuels

d'espèces ayant peu d'intérêt dans notre étude. Il est ainsi possible de se focaliser sur les modules présents chez les protéines humaines tout en utilisant les informations que peuvent fournir les protéines d'autres espèces possédant également ces modules.

Réconciliation phylogénétique

Afin de propager l'information d'une protéine ou d'un module aux autres protéines et modules, nous avons reconstruit leur histoire phylogénétique. Il faut noter que les approches de caractérisation fonctionnelle basées sur des modèles phylogénétiques sont généralement plus lentes mais plus précises que les approches classiques par similarités (e.g. les approches de phylogénomiques). Pour cela, il est nécessaire de pouvoir établir une histoire phylogénétique très robuste. Cependant, la quantité d'orthologues, de paralogues et la nature multidomaines de cette famille de protéines ne permettent pas d'obtenir aisément des résultats stables. Un simple changement d'outil d'alignement multiple changera l'arbre phylogénétique obtenu. Pour que le modèle de réconciliation phylogénétique DGS permette d'obtenir la meilleure histoire évolutive, il est nécessaire de construire au préalable les arbres phylogénétiques d'espèces, de gènes et de modules les plus robustes possibles. C'est pourquoi l'utilisation de logiciels tel que `trimAl` et `TREEFIX` est absolument nécessaire pour réduire au maximum les différents biais dans la construction des arbres.

Évolution de fragments de séquences

Pour la cohérence de notre modèle évolutif, les modules sont considérés comme des entités ayant pu évoluer de manière indépendante. Il est souvent question d'évolution distincte des domaines protéiques, or ce ne sont pas exactement les domaines (entité structurale) qui évoluent de manière indépendante. Cette évolution a lieu au niveau des séquences, ce sont des fragments de séquences (e.g. exons / rétrotransposons / transposons, comportant éventuellement un / plusieurs ou aucuns domaines ...) qui évoluent de manières indépendantes. Bien que ces fragments de séquences représentent souvent des domaines, nous ne pouvons espérer obtenir des résultats précis et une cohérence dans notre modèle évolutif en utilisant une définition structurale (domaine) pour représenter des événements au niveau des séquences (séquences conservées). Ce principe d'évolution renforce l'idée d'utiliser des modules prédits avec une approche *sans a priori* sur les séquences, plutôt que des caractérisations de domaines caractérisés dans les bases de données. De plus, la même stratégie a été appliquée en utilisant différents types de domaines et profils HMMs des bases de données (Prosite et TIGRFAMS) (résultats non présentés ici), et dans le cas des ADAMTS / ADAMTSL, la réconciliation phylogénétique était bien plus parcimonieuse avec des modules prédits, qu'avec des domaines issus des bases de données, suggérant des annotations incohérentes. Il est ainsi possible de s'affranchir des caractérisations actuelles et de proposer une caractérisation plus spécifique à notre famille de protéines.

Intégration et propagation de données fonctionnelles

Une fois le modèle évolutif de la famille construit, différents types d'informations de données fonctionnelles y ont été intégrés. Pour chacun des types de données, il a été choisi d'utiliser une technique d'intégration différente. Les données à l'échelle des séquences ont été intégrées d'une manière naïve ; si 2 protéines partagent une caractéristique, il est considéré que leur ancêtre commun la partage aussi. Cependant cette caractéristique n'est pas attribuée à toutes les protéines du sous arbre. Il est en effet impossible de savoir à quel moment cette caractéristique est perdue ou si au contraire elle n'a pas encore été observée. A l'échelle des régions de séquences (e.g. signature Prosite), la très forte conservation de séquences au sein d'un module a été utilisée pour attribuer une annotation présente dans un module d'une séquence, à toutes les occurrences de ce module. Ce transfert d'annotation est cependant questionnable et ne pas être adapté à toutes les signatures Prosite. Dans cette étude, nous avons autorisé ce type de transfert, car ces annotations sont uniquement présentes à titre indicatif et non prédictifs (au contraire des annotations à l'échelle de la protéine). A l'échelle des acides aminés (e.g. SNP), les informations sont attribuées au module de la protéine où elles sont présentes. Les informations de mutation sont très précises et peuvent être responsables d'un phénotype (perte / gain). Il est donc difficile de les caractériser et les transférer peut-être source d'erreur. C'est pourquoi ces informations n'ont pas été transférées à toutes les occurrences des modules, elles sont simplement présentes à titre d'information. Une importante perspective consiste en l'interprétation fiable de ces deux dernières sources d'informations (i.e. informations de régions et informations d'acides aminés).

Caractérisation en modules fonctionnels des ADAMTS / ADAMTSL

L'utilisation de ce *pipeline* sur un jeu de donnée de 341 séquences protéiques d'ADAMTS / ADAMTSL de 19 espèces différentes a permis d'obtenir une première caractérisation en modules fonctionnels ainsi que l'histoire évolutive la plus probable de cette famille de protéines. Ce *pipeline* est extrêmement long à faire tourner, il sert à obtenir des informations sur une famille de protéines complexe que l'on veut étudier avec détails, dans sa version actuelle cette stratégie est bien trop longue pour être utilisé à grande échelle (e.g. cette stratégie est adaptée à l'étude d'une famille de protéines, mais analyser tout un protéome risque d'être très long). Il faudrait soit réduire le nombre de séquences, et donc perdre des informations évolutives (i.e. perte d'espèces), soit utiliser des algorithmes plus rapides mais moins précis au risque de se baser sur une modélisation moins robuste. Ici le sujet porte spécifiquement sur une famille de protéines, c'est pourquoi le choix a été fait d'utiliser les algorithmes les plus longs et robustes, ainsi qu'une grande quantité de séquences, malgré des temps d'exécution conséquents.

Dans un dernier temps, les résultats de la caractérisation en modules fonctionnelles des 341 séquences protéiques d'ADAMTS / ADAMTSL ont été analysés. Cette première caractérisation fonctionnelle permet de mettre en lumière des corrélations intéressantes entre les différents types

d'information. L'apparition synchrone d'un module avec une interaction protéique permet de poser une hypothèse sur le rôle de ce module dans l'interaction en question. Par exemple, dans notre modèle, le module MB64 apparaît en même temps que les interactions avec les protéines extracellulaires COMP et A2GM. Ce module est chevauchant avec un motif thrombospondin, qui est connu pour être impliqué dans de nombreuses interactions. Il serait donc intéressant d'approfondir l'analyse et de vérifier expérimentalement si la présence de ce module est nécessaire pour interagir avec COMP ou A2GM.

Une preuve de concept

Ce type de corrélation est peut être sujet à interprétation, en raison la nature de la stratégie mise en place. Notre modèle étant construit sur une phylogénie robuste, il nous permet de savoir si 2 évènements ont eu lieu en même temps. La précision à l'échelle d'un module est uniquement possible grâce à l'alliance entre segmentation de séquences et réconciliation phylogénétique DGS. Il faut garder à l'esprit que notre stratégie se limite à une intégration de données de différents types. Cette approche permet d'observer des corrélations intéressantes, contribuant aux prédictions sur la fonction des différents modules. Cette étude constitue un travail préliminaire établissant une preuve de concept sur l'utilité d'un protocole pour la prédiction fonctionnelle des ADAMTS / ADAMTSL.

5 Conclusion et perspectives

Au cours de cette étude, j'ai développé une nouvelle stratégie de caractérisation fonctionnelle qui allie deux types d'approches de caractérisation fonctionnelle ; 1) les approches basées sur la conservation de séquences et 2) les approches basées sur la phylogénie des protéines. Cette stratégie a permis de réaliser une première caractérisation en modules fonctionnels des protéines ADAMTS / ADAMTSL humaines, en utilisant les informations de 341 séquences de paralogues et d'orthologues de 19 espèces. De plus, cette méthode permet de proposer de nouvelles prédictions des fonctions de certains modules (i.e. régions conservées) par observation de co-occurences de modules et de fonctions au cours de l'évolution. Cette étude consiste en un travail préliminaire, qui permet d'apporter une preuve de concept vis-à-vis de l'intérêt d'observer l'évolution de régions conservées en parallèle de l'évolution des différentes fonctions des protéines étudiées. Au delà de l'amélioration et de l'optimisation des paramètres du *pipeline* de prédiction, je propose trois types de perspectives : 1) la prise en compte de plus d'annotations fonctionnelles (e.g. données d'expressions protéiques, les différents isoformes et leurs phénotypes), 2) l'utilisation des séquences nucléiques (i.e. informations sur les pressions de sélections), et 3) l'utilisation d'autres méthodes de prédictions fonctionnelles (e.g. *profiling* phylogénétique de blocs protéiques afin de prédire des coévolutions de modules[54]).

Références

1. Van GOOR, H., MELENHORST, W. B. W. H., TURNER, A. J. & HOLGATE, S. T. Adamalysins in biology and disease. *The Journal of Pathology* **219,** 277-286. ISSN : 1096-9896 (nov. 2009).
2. *The ADAMTS hyalectanase family : biological insights from diverse species — Biochemical Journal* `http://www.biochemj.org/content/473/14/2011.figures-only` (2019).
3. NABA, A. *et al.* The extracellular matrix : Tools and insights for the "omics" era. *Matrix Biology : Journal of the International Society for Matrix Biology* **49,** 10-24. ISSN : 1569-1802 (jan. 2016).
4. HUBMACHER, D. & APTE, S. S. ADAMTS proteins as modulators of microfibril formation and function. *Matrix Biology : Journal of the International Society for Matrix Biology* **47,** 34-43. ISSN : 1569-1802 (sept. 2015).
5. BEKHOUCHE, M. *et al.* Determination of the substrate repertoire of ADAMTS2, 3, and 14 significantly broadens their functions and identifies extracellular matrix organization and TGF-β signaling as primary targets. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **30,** 1741-1756. ISSN : 1530-6860 (2016).
6. BOURD-BOITTIN, K. *et al.* Protease profiling of liver fibrosis reveals the ADAM metallopeptidase with thrombospondin type 1 motif, 1 as a central activator of transforming growth factor beta. *Hepatology (Baltimore, Md.)* **54,** 2173-2184. ISSN : 1527-3350 (déc. 2011).
7. LE GOFF, C. *et al.* ADAMTSL2 mutations in geleophysic dysplasia demonstrate a role for ADAMTS-like proteins in TGF-beta bioavailability regulation. *Nature Genetics* **40,** 1119-1123. ISSN : 1061-4036 (sept. 2008).
8. AHRAM, D. *et al.* A homozygous mutation in ADAMTSL4 causes autosomal-recessive isolated ectopia lentis. *American Journal of Human Genetics* **84,** 274-278. ISSN : 1537-6605 (fév. 2009).
9. TSUTSUI, K. *et al.* ADAMTSL-6 Is a Novel Extracellular Matrix Protein That Binds to Fibrillin-1 and Promotes Fibrillin-1 Fibril Formation. *The Journal of Biological Chemistry* **285,** 4870-4882. ISSN : 0021-9258. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836092/` (2019) (12 fév. 2010).
10. PICKUP, M. W., MOUW, J. K. & WEAVER, V. M. The extracellular matrix modulates the hallmarks of cancer. *EMBO reports* **15,** 1243-1253. ISSN : 1469-3178 (déc. 2014).
11. CAL, S. & LÓPEZ-OTÍN, C. ADAMTS proteases and cancer. *Matrix Biology : Journal of the International Society for Matrix Biology* **44-46,** 77-85. ISSN : 1569-1802 (juil. 2015).

12. WAGSTAFF, L., KELWICK, R., DECOCK, J. & EDWARDS, D. R. The roles of ADAMTS metalloproteinases in tumorigenesis and metastasis. *Frontiers in Bioscience (Landmark Edition)* **16,** 1861-1872. ISSN : 1093-4715 (1er jan. 2011).

13. SUN, Y., HUANG, J. & YANG, Z. The roles of ADAMTS in angiogenesis and cancer. *Tumour Biology : The Journal of the International Society for Oncodevelopmental Biology and Medicine* **36,** 4039-4051. ISSN : 1423-0380 (juin 2015).

14. VALDAR, W. S. J. Scoring residue conservation. *Proteins : Structure, Function, and Bioinformatics* **48,** 227-241. ISSN : 1097-0134. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10146 (2019) (2002).

15. CAPRA, J. A. & SINGH, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23,** 1875-1882. ISSN : 1367-4803. https://academic.oup.com/bioinformatics/article/23/15/1875/203579 (2019) (1er août 2007).

16. PEARSON, W. R. & LIPMAN, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85,** 2444-2448. ISSN : 0027-8424. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/ (2019) (avr. 1988).

17. ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403-410. ISSN : 0022-2836 (5 oct. 1990).

18. CAMACHO, C. *et al.* BLAST+ : architecture and applications. *BMC bioinformatics* **10,** 421. ISSN : 1471-2105 (15 déc. 2009).

19. LARKIN, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **23,** 2947-2948. ISSN : 1367-4811 (1er nov. 2007).

20. THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 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,** 4673-4680. ISSN : 0305-1048 (11 nov. 1994).

21. STEINEGGER, M. & SÖDING, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35,** 1026-1028. ISSN : 1546-1696. https://www.nature.com/articles/nbt.3988 (2019) (16 oct. 2017).

22. COSTE, F. & KERBELLEC, G. Learning Automata on Protein Sequences. https://hal.inria.fr/inria-00180429 (2019) (juil. 2006).

23. CASTRESANA, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17,** 540-552. ISSN : 0737-4038. https://academic.oup.com/mbe/article/17/4/540/1127654 (2019) (1er avr. 2000).

24. MULDER, N. J. & APWEILER, R. Tools and resources for identifying protein families, domains and motifs. *Genome Biology* **3,** reviews2001.1-reviews2001.8. ISSN : 1465-6906. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC150457/ (2019) (2002).

25. DENG, S., LU, J., FU, W. & YU, P. *Prediction of protein function by combining phylogenetic tree and mathematical inference* in *2014 7th International Conference on Biomedical Engineering and Informatics* 2014 7th International Conference on Biomedical Engineering and Informatics (oct. 2014), 896-901.

26. BROWN, D. & SJÖLANDER, K. Functional Classification Using Phylogenomic Inference. *PLoS Computational Biology* **2.** ISSN : 1553-734X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1484587/ (2019) (juin 2006).

27. ENGELHARDT, B. E., JORDAN, M. I., REPO, S. T. & BRENNER, S. E. Phylogenetic molecular function annotation. *Journal of physics. Conference series* **180,** 012024. ISSN : 1742-6588. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909777/ (2019) (2009).

28. ROST, B. Enzyme function less conserved than anticipated. *Journal of Molecular Biology* **318,** 595-608. ISSN : 0022-2836 (26 avr. 2002).

29. SOMARELLI, J. A. *et al.* PhyloOncology : Understanding cancer through phylogenetic analysis. *Biochimica Et Biophysica Acta. Reviews on Cancer* **1867,** 101-108. ISSN : 0304-419X (avr. 2017).

30. PEPPER, I. J., VAN SCIVER, R. E. & TANG, A. H. Phylogenetic analysis of the SINA/SIAH ubiquitin E3 ligase family in Metazoa. *BMC evolutionary biology* **17,** 182. ISSN : 1471-2148 (2017).

31. SOLÍS-CALERO, C. & CARVALHO, H. F. Phylogenetic, molecular evolution and structural analyses of the WFDC1/prostate stromal protein 20 (ps20). *Gene* **686,** 125-140. ISSN : 1879-0038 (20 fév. 2019).

32. RUSIN, L. Y., LYUBETSKAYA, E. V., GORBUNOV, K. Y. & LYUBETSKY, V. A. *Reconciliation of Gene and Species Trees* BioMed Research International. https://www.hindawi.com/journals/bmri/2014/642089/ (2019).

33. VOGEL, C., BASHTON, M., KERRISON, N. D., CHOTHIA, C. & TEICHMANN, S. A. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology* **14,** 208-216. ISSN : 0959-440X. http://www.sciencedirect.com/science/article/pii/S0959440X04000454 (2019) (1er avr. 2004).

34. TORDAI, H., NAGY, A., FARKAS, K., BÁNYAI, L. & PATTHY, L. Modules, multidomain proteins and organismic complexity. *The FEBS Journal* **272,** 5064-5078. ISSN : 1742-4658. https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-4658.2005.04917.x (2019) (2005).

35. MIYATA, T. & SUGA, H. Divergence pattern of animal gene families and relationship with the Cambrian explosion. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* **23,** 1018-1027. ISSN : 0265-9247 (nov. 2001).

36. BEN-SHLOMO, I., YU HSU, S., RAUCH, R., KOWALSKI, H. W. & HSUEH, A. J. W. Signaling receptome : a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Science's STKE : signal transduction knowledge environment* **2003,** RE9. ISSN : 1525-8882 (17 juin 2003).

37. PATTHY, L. Modular Assembly of Genes and the Evolution of New Functions. *Genetica* **118,** 217-231. ISSN : 1573-6857. https://doi.org/10.1023/A:1024182432483 (2019) (1er juil. 2003).

38. STOLZER, M., SIEWERT, K., LAI, H., XU, M. & DURAND, D. Event inference in multidomain families with phylogenetic reconciliation. *BMC Bioinformatics* **16,** S8. ISSN : 1471-2105. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4610023/ (2019) (Suppl 14 2 oct. 2015).

39. STOLZER, M. *Phylogenetic Inference for Multidomain Proteins* Thesis (1er juil. 2018). https://kilthub.figshare.com/articles/Phylogenetic_Inference_for_Multidomain_Proteins/6721055 (2019).

40. LI, L. & BANSAL, M. S. *Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model* in *ISBRA* (2019).

41. UniProt : a worldwide hub of protein knowledge. *Nucleic Acids Research* **47,** D506-D515. ISSN : 0305-1048. https://academic.oup.com/nar/article/47/D1/D506/5160987 (2019) (D1 8 jan. 2019).

42. WARD, N. & MORENO-HAGELSIEB, G. Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST : How Much Do We Miss ? *PLOS ONE* **9,** e101850. ISSN : 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0101850 (2019) (11 juil. 2014).

43. SAYERS, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37,** D5-15. ISSN : 1362-4962 (Database issue jan. 2009).

44. EDGAR, R. C. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792-1797. ISSN : 0305-1048. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337/ (2019) (2004).

45. CAPELLA-GUTIÉRREZ, S., SILLA-MARTÍNEZ, J. M. & GABALDÓN, T. trimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972-1973. ISSN : 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712344/ (2019) (1er août 2009).

46. GUINDON, S. & GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52,** 696-704. ISSN : 1063-5157 (oct. 2003).

47. BANSAL, M. S., WU, Y.-C., ALM, E. J. & KELLIS, M. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* **31,** 1211-1218. ISSN : 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393519/ (2019) (15 avr. 2015).

48. TURRO, E., ASTLE, W. J. & TAVARÉ, S. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics (Oxford, England)* **30,** 180-188. ISSN : 1367-4811 (15 jan. 2014).

49. GALIL, Z. & ITALIANO, G. F. Data Structures and Algorithms for Disjoint Set Union Problems. *ACM Comput. Surv.* **23,** 319-344. ISSN : 0360-0300. http://doi.acm.org/10.1145/116873.116878 (2019) (sept. 1991).

50. ARANDA, B. *et al.* PSICQUIC and PSISCORE : accessing and scoring molecular interactions. *Nature Methods* **8,** 528-529. ISSN : 1548-7105 (29 juin 2011).

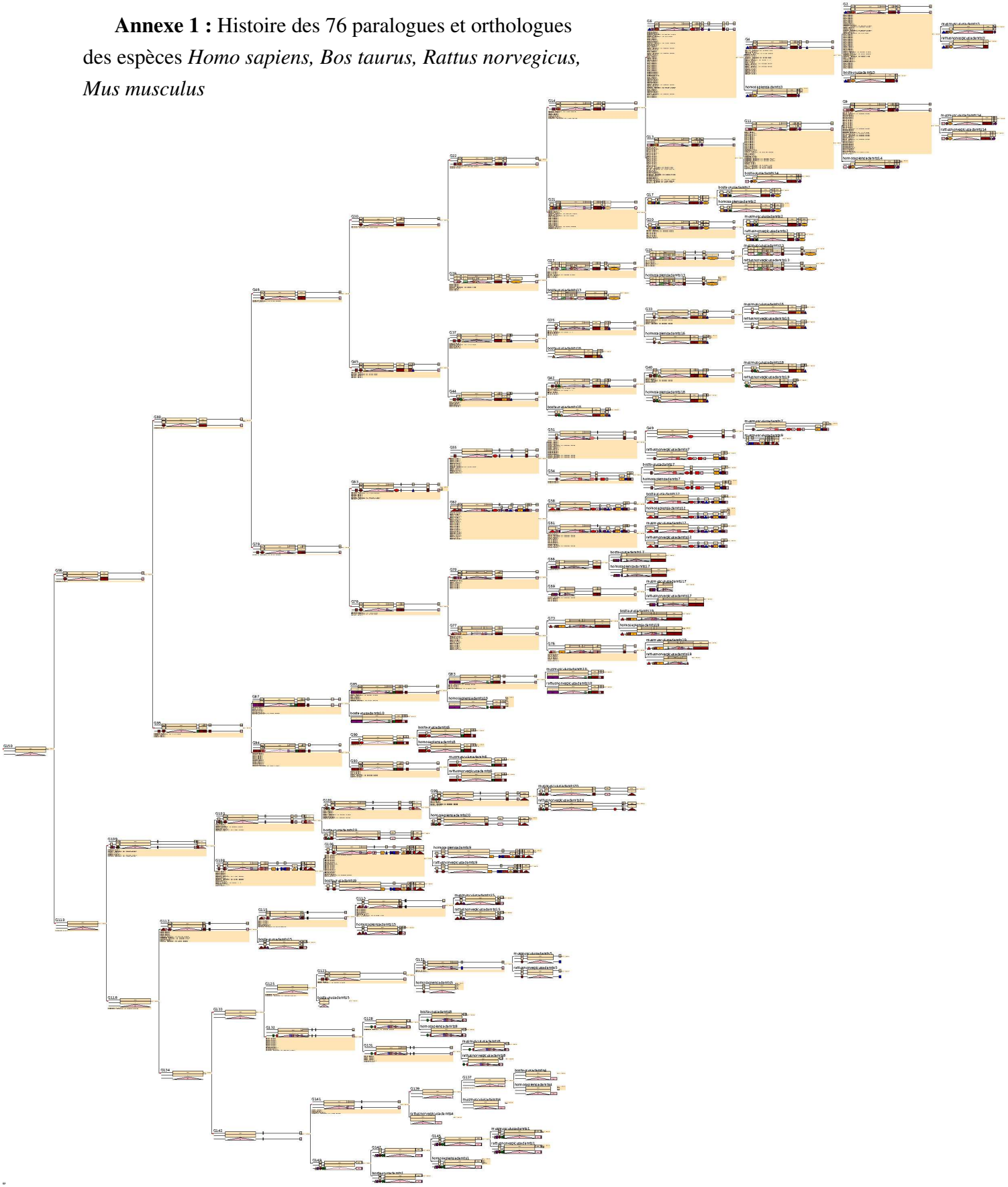
51. De CASTRO, E. *et al.* ScanProsite : detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* **34,** W362-365. ISSN : 1362-4962 (Web Server issue 1er juil. 2006).

52. SHERRY, S. T. *et al.* dbSNP : the NCBI database of genetic variation. *Nucleic Acids Research* **29,** 308-311. ISSN : 1362-4962 (1er jan. 2001).

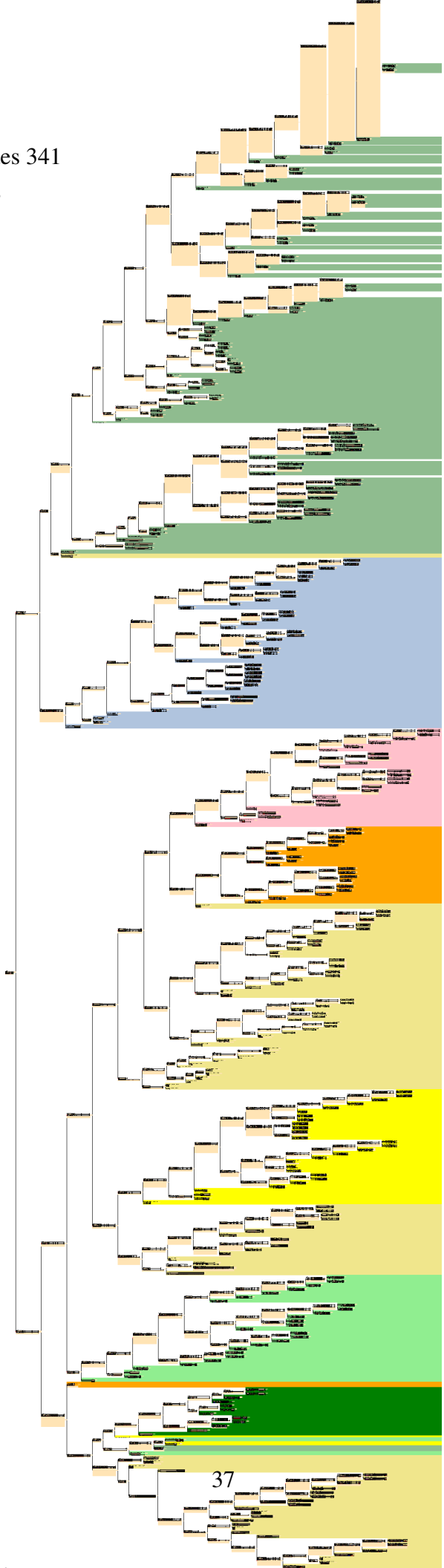
53. HUERTA-CEPAS, J., SERRA, F. & BORK, P. ETE 3 : Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution* **33,** 1635-1638. ISSN : 0737-4038. https://academic.oup.com/mbe/article/33/6/1635/2579822 (2019) (1er juin 2016).

54. KRESS, A., LECOMPTE, O., POCH, O. & THOMPSON, J. D. PROBE : analysis and visualization of protein block-level evolution. *Bioinformatics (Oxford, England)* **34,** 3390-3392. ISSN : 1367-4811 (2018).

Annexe 1 : Histoire des 76 paralogues et orthologues des espèces *Homo sapiens, Bos taurus, Rattus norvegicus, Mus musculus*



Annexe 2 : Histoire des 341 paralogues et orthologues des 19 espèces



Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL

Les protéines ADAMTS et ADAMTSL sont impliquées dans le remodelage du microenvironnement matriciel et constituent aujourd'hui de nouvelles cibles thérapeutiques dans les pathologies cancéreuses. Les nombreux gènes de la famille et la nature multidomaine des ADAMTS et ADAMTSL ne permettent pas d'utiliser les approches classiques de caractérisation fonctionnelles des protéines. Nous proposons ici une nouvelle méthode de caractérisation en modules fonctionnels adaptée aux protéines multidomaine, se basant uniquement sur les séquences protéiques. Cette méthode allie 2 approches de prédictions fonctionnelles à savoir la conservation des résidus et la phylogénie moléculaire. Après avoir sélectionné 341 séquences d'ADAMTS et ADAMTSL, réparties dans 19 espèces, les modules conservés sont recherchés par une approche *sans a priori* grâce à l'outil `paloma`. L'histoire phylogénétique de ces modules est ensuite élaborée avec l'outil `SEADOG-DM` de réconciliation phylogénétique DGS (Domain-Gene-Species). Enfin les histoires phylogénétiques sont complétées en intégrant des données biologiques issues de bases de données comme les interactions protéiques (PSICQUIC), les motifs (Prosite) et les polymorphismes nucléotidiques (dbSNP). Cette nouvelle stratégie permet de caractériser les différentes protéines et modules et d'étudier des évènements de co-occurence de modules conservés et de fonctions au cours de l'évolution. Cette étude constitue un travail préliminaire et permet d'apporter une preuve de concept vis à vis de la stratégie de caractérisation en modules fonctionnels des ADAMTS et ADAMTSL.

Mots clés : Protéines multidomaines, réconciliation phylogénétique, conservation de séquences, évolution

Functional module characterization of the ADAMTS / ADAMTSL protein family

ADAMTS and ADAMTSL proteins are involved in the remodeling of the matrix microenvironment and are now considered as new therapeutic targets in cancer diseases. The many genes in the family and the multi-domain nature of ADAMTS and ADAMTSL do not allow the use of conventional approaches to functional protein characterization. We propose here a new method of characterization in functional modules adapted to multidomain proteins, based only on protein sequences. This method combines 2 functional prediction approaches, namely residue conservation and molecular phylogeny. After selecting 341 ADAMTS and ADAMTSL sequences, distributed in 19 species, the preserved modules are searched for by an approach without *a priori* thanks to the `paloma` tool. Next, the phylogenetic history of these modules is developed with the `SEADOG-DM` phylogenetic reconciliation tool DGS (Domain-Gene-Species). Finally, phylogenetic histories are supplemented by integrating biological data from public databases such as protein interactions (PSICQUIC), motifs (Prosite) and nucleotide polymorphisms (dbSNP). This new strategy allows to characterize the different proteins and modules and to study co-occurrence events of conserved modules and functions during evolution. This study is a preliminary work and provides proof of concept regarding the functional module characterization strategy of ADAMTS and ADAMTSL.

Keywords : Multidomain proteins, phylogenetic reconciliation, sequence conservation, evolution